Synergistic Switch Control Enabled Optical Data Center Networks

Synergistic Switch Control Enabled Optical Data Center Networks

Xuwei Xue and Nicola Calabretta

In this article, a synergistic optical switch control mechanism is proposed to simultaneously solve the aforementioned challenges. The developed label control technique allows for controlling and configuring packet switching within 43.4 ns.

Abstract

Optical switching, which benefits from high bandwidth and low latency, promises to revolutionize data center networks that suffer from a bandwidth bottleneck and hierarchical structure. However, the lack of adaptable and fast optical switch control schemes, the lack of optical buffers for contention resolution, and the high cost of clock and data recovery (CDR) impede the practical deployment of optical switches in data centers. Networking communities have tried to address these challenges from different perspectives. However, the proposed solutions may be interdependent and thus introduce unexpected issues. In this article, a synergistic optical switch control mechanism is proposed to simultaneously solve the aforementioned challenges. The developed label control technique allows for controlling and configuring packet switching within 43.4 ns. The implementation of clock distribution achieves a fast CDR of 3.1 ns at low cost, and the design of an optical flow control protocol prevents packet loss caused by packet contention.

Introduction

The rapid escalation of high-traffic applications and services, such as big data and cloud computing, has significantly increased the traffic in data centers (DCs). The traffic volume even exceeds that of wide-area telecom networks [1]. Furthermore, due to the inability of the Ball Grid Array (BGA) packaging technique to increase the pin density, current electrical switches suffer from an I/O bandwidth bottleneck and cannot afford the explosive growth in DC traffic. Optical switches that are transparent to the data format and bit rate can theoretically provide unlimited bandwidth. Thus, switching the traffic in the optical domain has been considerably investigated as a future-proof solution to overcome the bandwidth bottleneck of electrical switches [2]. Moreover, all-optical switching could remove the dedicated electrical circuits and devices for various data format modulations and eliminate the power-hungry optical/electrical/optical (O/E/O) conversions, thus significantly decreasing costs, power, and processing delays.

Because of the inability to do information processing at optical switches, a suitable switch control mechanism is required for switch management and traffic forwarding control to effectively utilize high-bandwidth optical switches. Typically, an optical label or header is associated with the data packet, carrying the destination information of the data packet to be processed by the switch controller [3]. Based on the received destination information matrix, the controller will reconfigure the optical switch and then accordingly forward the data packets. Shortening the switch reconfiguration time, compromising both hardware switching time and controlling overhead, is essential to improve the network throughput and latency performance, especially for optical packet switching. Moreover, the overall controlling overhead needs to be independent of the DC network (DCN) scale. Given the scale of modern DCNs, which typically comprise hundreds of thousands of servers organized in few thousand racks, the switch control needs to be performed in parallel and distributed for every optical switch, avoiding a centralized network-wide scale schedule. Furthermore, for optical switch control, the edge nodes connected to the optical switch and controller must be precisely time-synchronized (ideally with a jitter of a few nanoseconds) to align the destination information with the corresponding data packets [4]. Inaccuracy in the synchronization step at any time can be compensated for by a larger customized interpacket gap (IPG). However, this will severely reduce the overall bandwidth utilization. Therefore, the implementation of fast and scalable switch control requires custom hardware support and a fair amount of ingenuity.

The lack of an optical buffer is one of the main technical differences between the electrical switch and optical switch. The electrical switch typically employs random access memory (RAM) to buffer and retransmit packets that lose contention. Due to the lack of effective RAM in the optical domain, packets that lose contention at the optical switches are dropped, thereby introducing a high packet loss rate. Thus, another unsolved critical challenge in building optically switched DCNs is packet contention resolution, which needs to be addressed under the premise of no optical buffer. Despite many approaches based on optical fiber delay lines (FDLs) [5], wavelength conversion [6], or deflection routing [7] that have been proposed to overcome the issue, none are 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 switch-

Digital Object Identifier:
10.1109/MCOM.001.2100683

The authors are with Eindhoven University of Technology, The Netherlands.

0163-6804/22/$25.00 © 2022 IEEE

es, the optical switch creates only momentary physical links between the source and destination nodes [8]. Thus, 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 generated every time the optical switch is reconfigured. Therefore, the receivers have to continuously adjust the local clock (frequency and phase) to properly sample the incoming packets and then correctly recover the data. As no valid data can be received before the clock and data recovery (CDR) is completed, the longer this process takes (hundreds of nanoseconds for off-the-shelf transceivers), the lower the network throughput will be, particularly for intra-data-center scenarios where many applications produce short traffic packets [9]. Burst-mode CDR receivers with nanoseconds data recovery time based on gated oscillators or over-sampling have been extensively investigated. However, these techniques increase the design complexity and cost of the transceivers, which need to be redeveloped and re-evaluated for higher link data rates, making them unsuitable for large-scale DCNs.

The challenges of fast switch control, packet contention resolution, and fast CDR locking have been roadblocks to the deployment of fast optical switches in DCNs. To overcome these issues, networking communities have been investigating various independent solutions, but most of them focus on a single issue. Moreover, each community addresses the challenges and problems from its own perspective. Some techniques are promising solutions for one certain challenge, but also introduce other issues. For instance, clock phase caching was used to implement the fast CDR in [10]. However, aligning the clock phase jitter on the Tx-Rx path results in extra time overhead of the phase shifting at each transmitter, offsetting the fast CDR and yielding certain disadvantages. Any proposed solutions need to be synergistically planned, and offer less network resource use and implementation complexity.

In [11], an optical switch and control system based on software-defined networking (SDN) and wavelength-selective switches is proposed with microseconds reconfiguration time to allocate the optical bandwidth. The SDN control plane enables an optical polling flow control protocol in [12] to solve the packet contention by reducing the head-of-line blocking and packet retransmission times. In this article, we propose and experimentally demonstrate a synergistic control mechanism for the optical switch to comprehensively solve all the aforementioned issues. The nanosecond optical switch and control system are based on a combination of a new label control and synchronization mechanism to achieve a fast system switch control of 43.4 ns; a clock distribution mechanism implementing a 3.1-ns CDR, avoiding the use of high-cost burst mode CDR receivers; and a novel optical flow control (OFC) protocol to solve the packet contention with no need for optical buffers. All these developed techniques are synergistically implemented on the optical label channels and switch controller. This practical and low-complexity scheme significantly decreases network resource occupation and improves network stability.

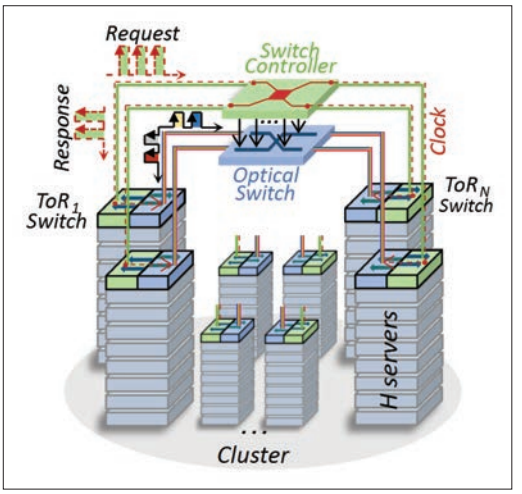

FIGURE 1. Concept of the synergistic switch control system.

The proposed synergistic switch control system is organized at the cluster scale in a parallel fashion, preventing a centralized control for the overall network. This allows the control system to be fully distributed and scalable, independent of the network scale, and it can support DCNs with a large number of racks and servers.

Principle of the Synergistic Switch Control

The proposed synergistic switch control system is organized at the cluster scale in a parallel fashion, preventing centralized control for the overall network. This allows the control system to be fully distributed and scalable, independent of the network scale, and it can support DCNs with a large number of racks and servers. Without a loss of generality and to simplify the explanation, a cluster unit deploying this synergistic control system is used to show its principles, as schematically illustrated in Fig. 1. Each rack consists of H servers and each cluster groups N racks interconnected via the top of rack (ToR) switch. An optical switch and the corresponding controller are used to interconnect all the racks via the data channels and label channels, respectively.

Label Control Mechanism

The optical switch controller and ToR switches in the proposed switch control system are physically implemented by the field-programmable gate array (FPGA). The functional blocks of the FPGA-based ToR switches and switch controller, as well as the optical connectivity, are schematically illustrated in Fig. 2. The Ethernet frames generated by the servers are first processed in the Ethernet switch at each ToR switch. Based on the destination MAC addresses, frames destined for servers located in the same rack (intra-rack traffic) are directly transferred to the destination servers. In contrast, frames destined to servers in other racks (inter-rack traffic) are stored in the electrical buffer (RAM), where each buffer block stores the inter-rack traffic with the same rack destination. The copy of several Ethernet frames in the most occupied buffer block are selected and grouped to generate an optical data packet in every time slot. The aggregated optical data packets transported by a single data channel are then sent to the optical switch, as shown in Fig. 2. Meanwhile, the serial number of the selected buffer block is set as the label request packet, which indicates the destination information of the corresponding aggregated data packet. The label request packets are sent to the switch controller via the label channels. Based on the label request matrix from all the racks, the switch controller arbitrates

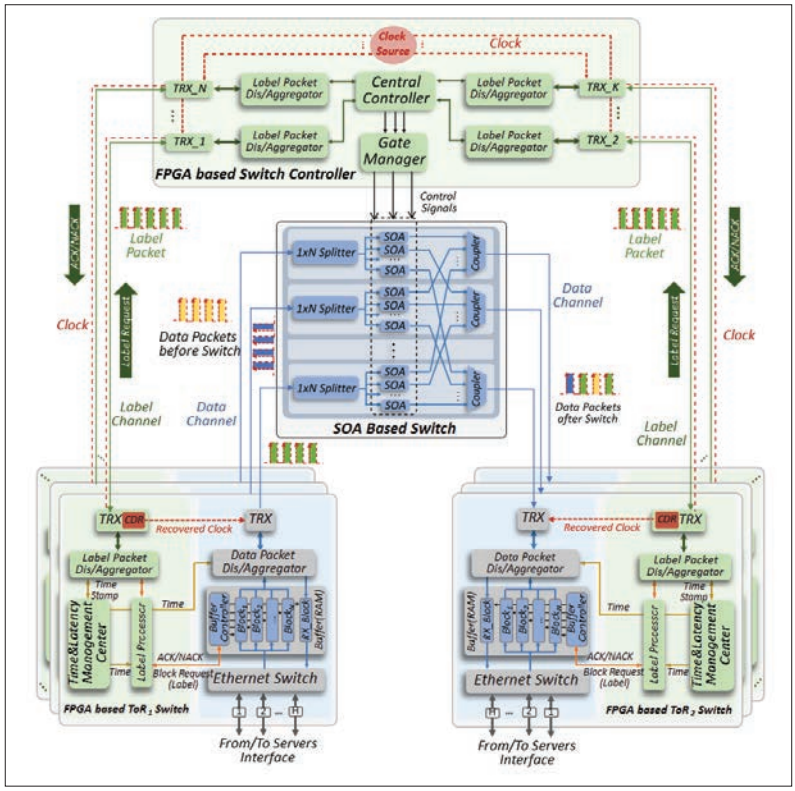

FIGURE 2. Functional blocks of the synergistic switch control system.

the packet contention and then accordingly configures the optical switch to forward the optical data packets to the correct destinations. Note that, exploiting the parallel processing capability of the FPGA-based switch controller, the label matrix can be processed within few clock cycles to implement the nanosecond switch control.

To align the data packets with the corresponding label requests in the proposed control system, the time is accurately synchronized from the switch controller to the edge nodes (ToR switch). At initialization, the time and latency management center in each ToR switch, as illustrated in Fig. 2, sends the timestamp carrying the initial ToR switch time (T_{TX}) to the switch controller via the label channels.The timestamps are processed at the controller and then sent back to the source ToR switches, where the time (T_{RX}) of the received timestamp is recorded. Based on the known signal processing delay at the FPGA-based controller and ToR switch, and the time offset (T_{RX}–T_{TX}), the transmission delay of the label channels can be automatically calculated even if the fiber length of the different label channels varies. Afterward, the switch controller distributes its time to all the connected ToR switches via the label channels. Compensated with the measured fiber delay and the FPGA processing delay, the time of all the ToR switches is synchronized and coherent with the controller. The synchronized time is then used to align the transmission of the data packets and label request packets in each time slot.

Optical Flow Control Protocol

Packet contentions occur at the switch node whenever two or more packets from different source racks have the same output port in the same time slot. An OFC technique is designed and implemented between the ToR switches and optical switch nodes to prevent the packet loss caused by packet contention at the switch node. Based on the received label matrix, the switch controller arbitrates the connection. The optical packet with higher priority is forwarded to its original destination racks, while the packet that lost contention and hence has lower priority is sent to the racks with no destination request. This packet forwarding mechanism guarantees that the receivers of ToR switches are continuously fed with data, receiving the nonstop traffic flow in every time slot. After the label processing, the switch controller generates flow control signals (ACK/NACK) that indicate successful/unsuccessful packet forwarding. As shown in Fig. 2, the flow control signals are sent back to the corresponding ToR switches via the bidirectional label channels. If a ToR switch receives an ACK signal, which indicates the successful forwarding of the optical packet, the corresponding frames are released from the buffer block. In contrast, if a ToR switch receives a NACK signal, which indicates the optical packet is forwarded to the undestined ToR switches, the stored frames are retransmitted in the following time slot until the ToR switch receives the ACK feedback.

Clock Distribution Technique

The locking time of the CDR mostly depends on the time to adjust the variation in the clock frequency to sample the incoming data. Thus, even if the optical switch system takes few nanoseconds to reconfigure (including both hardware switching time and control overhead), the network throughput will still be low because of the long CDR locking time (hundreds of nanoseconds). To accelerate the CDR operation, the bidirectional label channels in our new approach are continuous links that are not only used to send the ACK/NACK signals but also distribute the clock from the switch controller to the ToR switches to synchronize the system clock frequency. As shown in the schematic of the FPGA-based switch controller in Fig. 2, an onboard clock source is employed as the master clock to be distributed to the connected ToR switches by the ACK/NACK signals. At each ToR switch, 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 ToR switches, not only to transmit the data packets and the label signals but also to implement the nanosecond recovery of the data packets. Indeed, once all the network ToR switches 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 a few tens of bits (a few nanoseconds), preventing the need for 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 full vision of the traffic from the racks, exploits the multicast capability of the optical switch to forward packets that have lost

contention to the undestined ToR switches to fill the empty slots. Additionally, the inter-packet gap (IPG) and idle parts of the packets due to the lower traffic load are inserted with pulse transitions ("1010...1010"), as shown in Fig. 3, to maintain the continuous stream of data, similar to 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 receivers' clock phase and providing bit-level synchronization. The start packet delimiter is the 8-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 the source/destination rack is embedded for packet identification. The 32-bit cyclic redundancy check (CRC) is a checksum that is calculated to provide error detection in the case of link errors or packet transmission collisions, which can corrupt the data packet.

Demonstration and Discussion

The proof of concept setup to demonstrate and experimentally investigate the synergistic optical control system is illustrated in Fig. 4. This demonstration consists of four FPGA-based ToR switches (Xilinx UltraScale+ XCVU9P). Each ToR switch is equipped with data channels of 10 Gb/s to deliver the optical data packets. One 4 × 4 SOA-based optical switch with a corresponding FPGA-based switch controller (Xilinx UltraScale+ VU9P) is deployed to connect all the ToR switches. Three buffer blocks, where the buffer size is controllable and variable due to the reprogramming capability of FPGA, are deployed inside each ToR switch. The occupation ratio (occupied bytes/overall buffer size) of each buffer block is monitored by the FPGA-based ToR switch in real time. In each time slot, the traffic stored in the most occupied block will be selected for transmission. Meanwhile, the corresponding label request packet is also delivered to the switch controller on the label channel. The 10 Gb/s label channels efficiently implement the label-controlling mechanism, OFC protocol, and clock frequency distribution. An Anritsu MT1040A Ethernet testing center emulating four servers is connected in this demonstration, generating Ethernet frames at 10 Gb/s with a controllable and variable load. Ethernet frames are randomly generated, with a length between 64 and 1518 bytes. The Ethernet frames generated by servers with the same rack destination are stored in the identical buffer block and are aggregated into a data packet with a length of 2600 bytes. The data packet consists of a 3-byte preamble, a 1-byte start packet delimiter, a 4-byte source rack and destination rack address, and a 4-byte CRC sequence; the remaining parts are the aggregated Ethernet frames.

Investigation Results and Analysis

The fast operations of the label-controlling mechanism and the OFC protocol are first investigated. In each time slot, every ToR switch transmits the LabelRequest signals (as illustrated at the top of Fig. 5) to the switch controller, consisting of the destination information and forwarding priority of the associated data packets. In this demonstration case, the priority order of data packets is set as "1>2>3>4," where packets with the order of

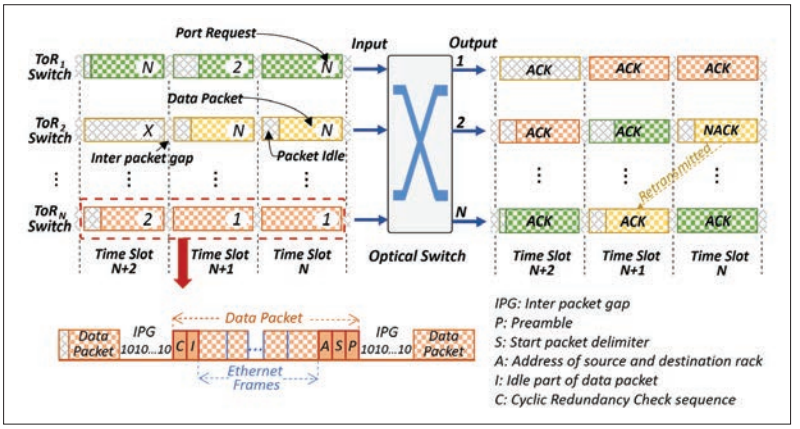

FIGURE 3. Optical data packets flow and pattern.

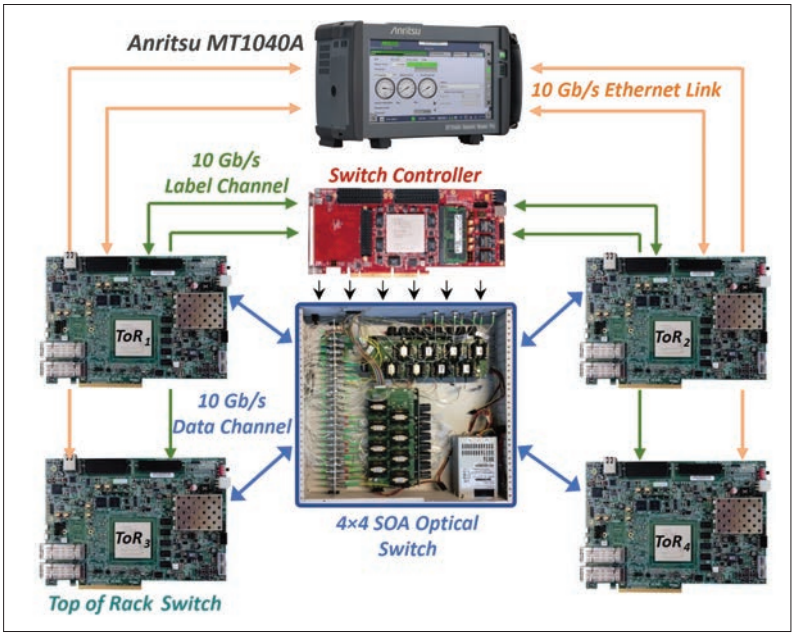

FIGURE 4. Proof of concept of the synergistic switch control system.

"1" have the highest priority. According to the label signals, the switch controller arbitrates the packet contention and then sends the LabelResponse signals to the corresponding ToR switches, as shown at the top of Fig. 5. A response label identical to the request label indicates an ACK acknowledgment, whereas two differing labels indicate a negative acknowledgment, "NACK." The monitored LabelRequest and LabelResponse signals are precisely synchronized at the switch controller. This validates the accurate implementation of the synchronous slotted mechanism. In time slot *N*, the LabelRequests for the ToR_1 switch and ToR_2 switch is 3, which indicates that the data packets from the ToR_1 switch and ToR_2 switch are destined to the ToR_3 switch. Given higher priority, the ToR_1 switch packet is forwarded to the destination ToR_3 switch, while the ToR_2 switch packet with lower priority is sent to the ToR_2 switch to maintain the continuous traffic stream at the receiver (the packet will be dropped at ToR_2 once it is verified that its destination is ToR_3). ToR_1 receives an ACK signal (see label response in Fig. 5) to release the stored packet in the buffer block, while ToR_2 receives a NACK signal to trigger the packet retransmission. In the next time slot, *N* +

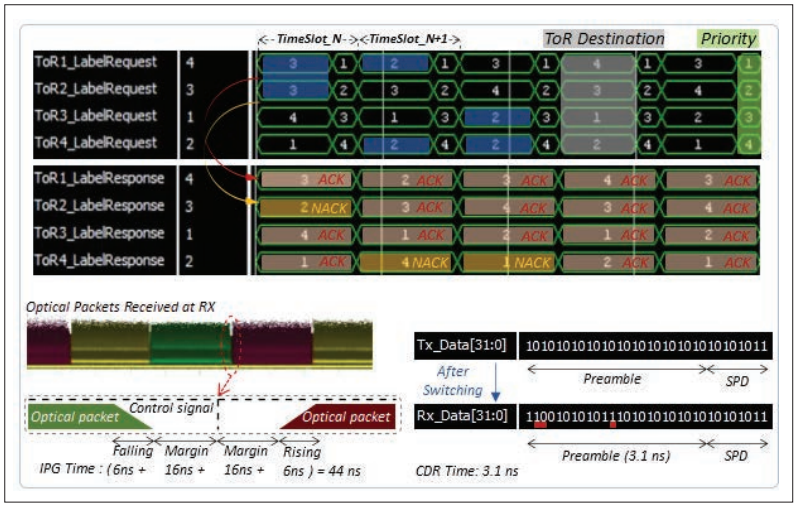

FIGURE 5. Demonstrated results for the synergistic switch control system.

1, ToR_1 sends out a new label request 2, while ToR_2 resends label request 3 until an ACK signal is received.

The overall switch and control time is 43.4 ns, as illustrated in the bottom left of Fig. 5, which consists of the 12.4 ns processing time of label signals, a 3 ns delay at the switch driver, a 6 ns switch rising/falling time, and a 16 ns margin time between the switch control signal and the optical data packet. Compared to typical micro electro-mechanical system (MEMS)-based optical switches exhibiting tens of milliseconds of configuration time [13], the proposed scheme with the implementation of a 43.4 ns switch and control time adapts better to the DCN scenario, where the vast majority of traffic flows are shorter than 10 kB, requiring the fast (nanoseconds) switching configuration time. Moreover, the fast switch and control mechanism enables a short (43.4 ns) IPG, guaranteeing high bandwidth utilization. To further improve this utilization, the CDR needs to be completed as fast as possible. With this aim, the proposed switch control system distributes the clock frequency from the optical switch controller to all the connected ToR switches, and the empty slot is inserted with pulse transitions. Thus, the data can be correctly recovered in 3.1 ns, as can be observed from the receiver data of RX_Data in the bottom right of Fig. 5. The receiver extracts the preamble (3 bytes) and the start packet delimiter (SPD: 1 byte) within one clock cycle (3.1 ns). Compared to the burst-mode digital signal processing technique with a 71.68 ns CDR time [14], the proposed clock distribution technique speeds up the CDR operation by 20 times. For this proposed control system, where the data packet length is 2017 ns (2600 bytes), and the CDR time and the IPG are 3.1 ns and 43.4 ns, respectively, the data bandwidth utilization ratio can achieve 97.7 percent (2017-3.1)/(2017+43.4).

Scalability Discussion

The proposed synergistic switch control system, shown in Fig. 1, targets single-cluster operation. As structurally shown in Fig. 6, based on this control system, large-scale optical DCNs with multiple clusters can also be built, where the inter-cluster switch connects ToR switches located in different clusters. The *i*th ToR switch located in the *i*th cluster are interconnected by the *i*th inter-cluster switch. Single-hop direct interconnection is provided by the intra-cluster switches for intra-cluster communication. For inter-cluster communication, the optical data packet is first switched by the inter-cluster switch to the destined cluster and terminated to the corresponding interconnected ToR switch. The Ethernet switch of the ToR switch processes the received packet and checks the destination address. If the final destination is another ToR switch in the destined cluster, the packets will be forwarded to the final ToR switch destination via the intra-cluster switch. Thus, at most, two hops are sufficient to forward the inter-cluster traffic.

Note that the intra-cluster interconnect network and the inter-cluster interconnect network are two independent subnetworks, as shown in Fig. 6. Each subnetwork has an independent optical switch control system with its own clock frequency distribution mechanism. This is important, as the scalability of this synergistic control system is on a per-cluster scale and not for the whole DCN. This makes the proposed technique fully distributed and scalable even for DCN with a large number of racks. For the intra-cluster network that has independent clock distribution, each ToR switch has its own label channel connected to the switch controller. The clock is distributed via the independent label channels to synchronize the clock of all the ToR switches, and these have no effect on each other. Moreover, the switch controller is implemented by FPGA, which equips sufficient and independent transceivers to distribute the clock via the label channels so that the number of ToR switches does not affect the CDR of 3.1 ns. The proposed synergistic switch and control system is also adaptable to the centralized solutions in the DCN deploying SDN. The SDN control plane can be connected with the FPGA-based ToR switches and switch controller via SDN agents [15]. Effectively integrated with the synergistic optical switch and control system, the SDN controller then monitors the network statistics and accordingly reconfigures the network to provide dynamic quality of service.

Conclusion

To simultaneously overcome the challenges that impede the practical deployments of optical switches in DCNs, we have proposed and experimentally demonstrated a synergistic optical control system based on a novel label control mechanism, OFC protocol, and clock distribution mechanism. Optical label channels transport the label signals for nanosecond packet forwarding, implement the OFC protocol for optical packet contention resolution, and distribute the clock to all ToR switches for nanosecond CDR locking. Experimental demonstrations confirm an overall 43.4 ns optical switching and network control time, 3.1 ns data recovery, and a 97.7 percent bandwidth utilization ratio. The distributed feature of this synergistic control system can be scaled up to build large optical DCNs with independent cluster units. The proposed control system can serve as a starting point for solving the technical challenges from the control perspective, thereby promoting the practical deployments of optical switches in DCs.

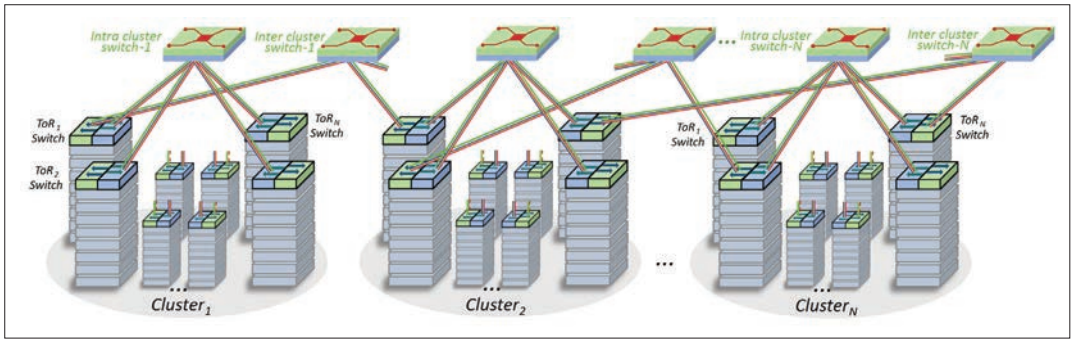

FIGURE 6. Synergistic switch control system enabled a large-scale optical data center network.

References

[1] Cisco, "Forecast and Methodology 2016-2021," White Paper. Cisco Systems," 2018.

[2] N. Parsons and N. Calabretta, "Optical Switching for Data Center Networks," *Springer Handbook of Optical Networks*, Springer, 2020, pp. 795–825.

[3] E. N. Lallas, "A Survey on All Optical Label Swapping Techniques: Comparison and Trends," *Optical Switching Networking*, vol. 31, 2019, pp. 22–38.

[4] H. Ballani *et al.*, "Bridging the Last Mile for Optical Switching in Data Centers," *2018 Optical Fiber Commun. Conf. and Expo.*, 2018.

[5] M. Moralis-Pegios *et al.*, "A Low-Latency High-Port Count Optical Switch with Optical Delay Line Buffering for Disaggregated Data Centers," *Optical Interconnects XVIII*, Int'l. Society for Optics and Photonics, 2018, p. 1,053,805.

[6] R. Farhat, A. Farhat, and M. Menif, "All-Optical Variable-Length Packet Router with Contention Resolution Based on Wavelength Conversion," *Nonlinear Optics and Applications X,* Int'l. Society for Optics and Photonics, 2017, p. 1,022,80X.

[7] J. Kang *et al.*, 2018, July, "Load Balance Based Deflection Routing for Optical Burst Switching," *2018 23rd IEEE Opto-Electronics and Commun. Conf.,* pp. 1–2.

[8] P. J. Argibay-Losada, D. Chiaroni, and C. Qiao, "Optical Packet Switching and Optical Burst Switching," *Springer Handbook of Optical Networks* (Springer, 2020), pp. 665–701.

[9] K. Clark *et al.*, "Sub-Nanosecond Clock and Data Recovery in an Optically-Switched Data Centre Network," *2018 European Conf. Optical Commun.*, 2018.

[10] K. A. Clark *et al.*, "Synchronous Subnanosecond Clock and Data Recovery for Optically Switched Data Centres Using Clock Phase Caching," *Nature Electronics*, vol. 3, no. 7, 2020, pp. 426–33.

[11] X. Xue *et al.*, "Automatically Reconfigurable Optical Data Center Network with Dynamic Bandwidth Allocation," *J. Optics*, vol. 23, no. 11, 2020, p. 114,003.

[12] X. Xue *et al.*, "Experimental Assessments of SDN-Enabled Optical Polling Flow Control for Contention Resolution in Optical DCNs," *J. Lightwave Technology*, vol. 39, no. 9, 2020, pp. 2652–60.

[13] M. Ghobadi *et al.*, "Projector: Agile Reconfigurable Data Center Interconnect," *Proc. 2016 ACM SIGCOMM Conf.*, 2016, pp. 216–29.

[14] J. Zhang *et al.*, "Efficient Preamble Design and Digital Signal Processing in Upstream Burst-Mode Detection of 100G TDM Coherent-PON," *J. Optical Commun. Networking*, vol. 13, no. 2, 2021, pp. A135–A143.

[15] X. Xue *et al.*, "SDN-Controlled and Orchestrated OPSquare DCN Enabling Automatic Network Slicing with Differentiated QoS Provisioning," *J. Lightwave Technology*, vol. 38, no. 6, 2020, pp. 1103–12.

Biographies

XUWEI XUE (x.xue.1@tue.nl) is a postdoctoral researcher in electro-optical communication systems at Eindhoven University of Technology (TU/e). His research directions now are FPGA implementation of scheduling algorithms for low-latency control of fast and large port optical switches, development of optical interconnecting architectures, and development of software defined networking control techniques for optical networks.

NICOLA CALABRETTA is an associate professor in electro-optical communication systems at TU/e. 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 selectors with flexible bandwidth, photonic integrated optical switches, and optical nodes for metro networks.