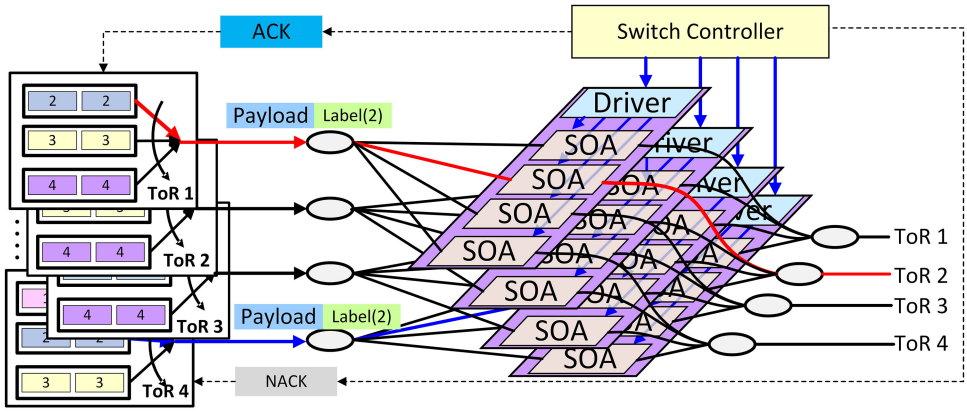
IEEE Photonics Journal
An IEEE Photonics Society Publication

Open Access

Assessment of Adaptive Polling Contention Resolution for Optical Switching in Edge Data Center Networking

Volume 11, Number 5, October 2019

Fu Wang
Bo Liu
Xuwei Xue
Lijia Zhang
Qi Zhang
Qinghua Tian
Feng Tian
Dong Guo
Xiangjun Xin



Packet contention resolution for optical switches

DOI: 10.1109/JPHOT.2019.2939216

Assessment of Adaptive Polling Contention Resolution for Optical Switching in Edge Data Center Networking

Fu Wang,[1,2,4] **Bo Liu**,[3] **Xuwei Xue**,[2] **Lijia Zhang**,[1,4] **Qi Zhang**,[1,4] **Qinghua Tian**,[1,4] **Feng Tian**,[1,4] **Dong Guo**,[1,4] **and Xiangjun Xin**[1,4]

[1]School of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]IPI-ECO Research Institute, Eindhoven University of Technology, Eindhoven 5600 MB, The Netherlands
[3]Institute of Optoelectronics, Nanjing University of Information Science & Technology, Nanjing 210044, China
[4]Beijing Key Laboratory of Space-Ground Interconnection and Convergence, Beijing University of Posts and Telecommunications, Beijing 100876, China

DOI:10.1109/JPHOT.2019.2939216

Manuscript received July 19, 2019; revised August 23, 2019; accepted August 31, 2019. Date of publication September 3, 2019; date of current version September 30, 2019. This work was supported in part by the National NSFC of China under Grants 61425022/61522501/61675004/61307086/61475024/61672290/61475094/61605013/61378061/61675030/61575026; in part by the National High Technology 863 Program of China under Grants 2015AA015501, 2015AA015502, 2015AA015504, 2015AA016904, and 2015AA016901; in part by Beijing Nova Program under Grant Z141101001814048; in part by the Fund of State Key Laboratory of IPOC, Beijing University of Posts and Telecommunications (BUPT); and in part by BUPT Excellent Ph.D. Students Foundation (CX2018305).
Corresponding author: Bo Liu (e-mail: bo@nuist.edu.cn).

Abstract: Edge Data Center (EDC) provides delay-sensitive services for end-users in edge computing. In Edge Data Center Network (EDCN), fast optical switching is the most promising technology due to low latency, high bandwidth, and transparency of rate/modulation format. However, packets contention of optical switching causes optical packets loss leading to low throughput and high latency. For buffer-less optical switches, it is challenging to solve packets contention because of lack of backup resources for contended packets. In this article, we propose a contention resolution, Adaptive Polling Contention Resolution (APCR), and assess the throughput, latency, and packet loss performance in DCN. An experimental demonstration based on flow control is implemented to test different contention resolution for the admissible load. The APCR algorithm can be adaptive to different load by buffer occupation status. APCR employs hybrid polling scheme to improve throughput and decrease average latency by global desynchronizing in high load. We assess the proposed algorithm performance in intra-cluster and inter-cluster EDCN. The results show that the proposed APCR can increase throughput by more than 15% and reduce the contention count by 43.72% at 0.4 loads.

Index Terms: Edge data center, data center networks, optical packets switching, contention resolution.

1. Introduction

With the emergence of delay-sensitive applications, like cloud gaming, Internet of things and autopilot, EDC replaces traditional cloud data center as one of the most promising technology for

future optical network architecture [1]. According to the Cisco global cloud index [2], Global Internet Protocol (IP) traffic will grow 3-fold from 2016 to 2021, reaching 20.6 ZB per year. In front of massive traffic, cloud DCN with centralized nature might be limited for grantee the latency-sensitive applications. Compared with cloud DCN, EDC is closer to end-user, which can provide processing and storage capacity at low latency without moving data to cloud DC. The latency performance of EDCN has a significant impact on the Quality of Service (QoS). In EDCN, traditional electrical switches cannot afford the stringent requirement on bandwidth and latency because of the high energy consumption, cost, and limited bandwidth [3]. Many researchers focus on Fast Optical Switching (FOS) to harness tremendous benefits on flexible granularity, unlimited bandwidth, and low cost [4], [5]. The buffer-less optical switches can provide low-latency transmission with transparent rate and modulation format, insurance delay-sensitive applications in EDCN. However, packet contention occurs when two packets attempt to the same output port at the same time slots. Packets contentions cause packet loss, leading to high latency and low throughput, which is one of the main factors for performance degradation of FOS [6]. Due to lack of optical buffer, it is difficult for FOS to handle optical packets contention in the switch fabric. Therefore, a switch controller is adopted in FOS to realize the packet contention and fabric reconfiguration. An Input-Queue Scheduling (IQS) schemes are needed to manage the packets contention for both FOS and edge electrical switches in switch controller. Lots of researches propose some scheduling algorithms and contention resolutions [6]–[8]. However, efficient contention resolution for FOS still faces a harsh problem.

In electrical input-queue switches, the Virtual Output Queue (VOQ) is proposed to prevent head-of-line blocking, and IQS algorithms select the buffer of VOQ for output port [9]. However, for optical switching, on the one hand, optical switches can prevent the packets contention as electrical switches by IQS algorithms. On the other hand, flow control or optical buffer can be used to solve contention in the switch fabric. IQS algorithm in FOS will introduce a Round-Trip Time (RTT) delay because of the remote interaction between ToRs and FOS. The switch controller carries out the IQS algorithm, and grant the output port to the input port. Lots of works on IQS algorithms, like LQF, MWM, and iSLIP, can be proved to reach 100% throughput under admissible traffic [9]. Input-port scheduling algorithms usually includes three steps (request, grant, and accepted). For FOS, IQS algorithm will introduce RTT delay between request and grant. In existing resolution for optical packets contention, some schemes for optical packets contention often employ Fiber Delay Lines (FDL) and other additional resources [7], [10], [11]. Optical buffers can handle contention while never prevent contention. Some schemes have been proposed to reduce the complexity of optical switches with a simple structure for buffer-less optical switching [12]–[14]. Besides the contention resolution, some researches try to find a way to prevent packets contentions with complicated architectures [15], [16]. However, all schemes are suffered from high complexity and complex switching architecture. The maturity of Software-Defined Optical Networks (SDON) is widely considered to be critical technologies for flexible, reconfigurable optical EDCN, since the virtualization of DCN resources by SDN controller provides coordinated managements for the operator to realize more centralized functions, such as traffic balancing, routing, bandwidth allocation, and service migration [17]–[19]. Preventing contention by centralized scheduling can benefit the FOS performance in EDCN.

In this paper, an SDN-enabled optical EDCN architecture is investigated to enable the contention resolution scheme. Adaptive polling contention resolution is proposed to reduce ToR-to-ToR latency and packets contention in SDN-enabled optical EDCN. The proposed algorithm includes a buffer selection scheme in ToR switches and contention resolution in optical switches. The SDN-enabled optical EDCN architecture is described in Section 2, including the control plane and data plane. In Section 3, we introduce the APCR algorithm. The simulation results are discussed in Section 4. An experimental demonstration is tested in Section 5. The results show that APCR can reduce the ToR-to-ToR by more than 15% and improve the throughput by more than 4.26% compared with the traditional scheme in high traffic load. The contention count is reduced by 43.72% at 0.4 loads. Section 6 is the conclusion.

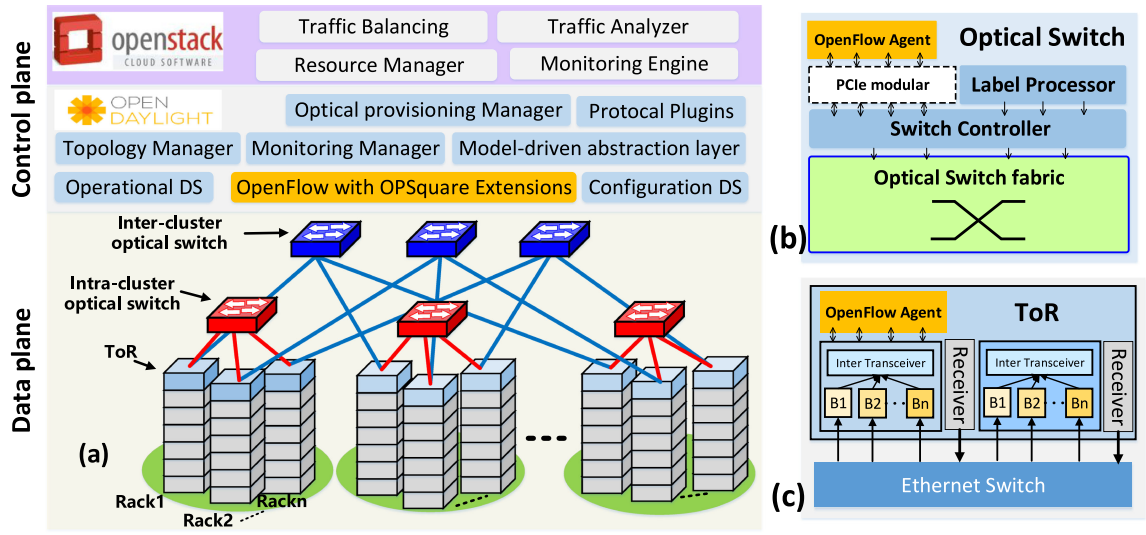


Fig. 1. (a) SDN-based optical switch architecture. (b) Optical switch node architecture. (c) ToR switch architecture.

2. SDN-Enabled Optical EDCN With Flow Control

Fig. 1 shows the SDN-enabled optical EDCN architecture. The architecture includes data plane, control plane, and orchestrator. In the data plane, SOA-based optical switches have been implemented [20]. The nanosecond optical switches can realize the fast packets processing in EDCN. Optical switches employ switch controller for reconfiguration of SOA, Optical Label (OL) processing and traffic monitor shown in Fig. 1(b). ToRs are optoelectronic switches, which aggregate the Ethernet frames into optical packets. The label processor can extract OLs and report them to switch controller for packet processing. An OpenFlow agent is employed for optical switches to implement SDN-enabled functions. In optical EDCN, the Inter-cluster Switches (ESs) and Intra-cluster switches (ISs) are used for inter-cluster traffic and intra-cluster traffic, respectively. Since optical switches only support single-hop transmission without O/E/O conversion, ToRs need to aggregate the Ethernet frames to the different queues of VOQ by the destination. As shown in Fig. 1(c), the Internet frames are cached in different VOQ buffer (B1, B2 . . . , Bn). Bn means the traffic in this buffer would go to output port 'n' The frames with same destination port will be cached in same VOQ buffers.

The control plane includes the SDN controller and OpenFlow agent, as well as the extended OpenFlow protocol. SDN controller is responsible for the management of optical switches and ToR switches. With extended OpenFlow, we can also use the SDN controller to manage the packet contention resolution for optical switches. SDN controller can distribute the contention resolution to switch controller, and then the switch controller can solve the contention by received resolution. The switch controller, shown in Fig. 2(b), carries out the contention resolution in this architecture. OpenDayLight is employed as an SDN controller in our architecture. An OpenFlow agent is used to translate the OpenFlow message into a PCIe signal, since the switch controller, implemented by high-speed FPGA, is simplified to improve the efficient of OL processing and reconfiguration. Except for signal translation, OpenFlow agent can also realize some functions, such as statistics collection, data analysis, and pre-configuration of the flow table. PCIe x8 interface is employed as a control channel in this architecture. The OpenFlow agent for ToR switches can perform more functions like contention resolution. OpenFlow commands can be extended to adapt the management of contention resolution for optical switches in EDCN, and the extended OpenFlow will discuss in Section 3. In this architecture, all functions of ToR and optical switches are under the control of Opendaylight, including flow entry manage, contention resolution, traffic monitor, and topology maintenance.

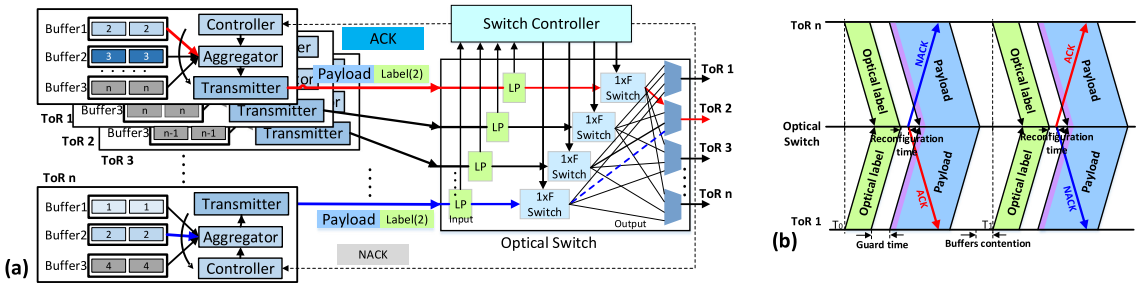


Fig. 2. (a) Packet processing for optical DCN. (b) Schedule process of flow control. (LP: label processor).

Fig. 2(a) shows the packets processing in optical switches. Before transmission of the optical packet, the ToR selects a buffer and makes packet aggregation. Buffer selection means the destination of this packet is specified. Then, ToR would make an OL inform switch controller about the destination, priority, queue size of all VOQ buffer, and payload quantity. ToR switches send the label to switch controller at first, and then send the payload after a guard time. There is a 50 ns guard time between label and payload. When the switch controller receives all OLs, it would solve the packets contention and reconfigure the switches. The cost time for reconfiguration of SOA-based switches array is about 20 ns. The gap time between optical packets is 150 ns. After the reconfiguration of optical switches, the switch controller feedbacks a flow control signal (ACK/NACK). If the optical packet is accepted, an ACK signal will be sent back to ToR. A NACK signal is feedback back to ask for retransmission if a packet is blocked in this time slot. Fig. 2(b) shows the process of interaction between ToR and the optical switch.

3. Adaptive Polling Contention Resolution for SDN-Enabled Optical Switches

At first, we introduce the contention resolution process based on flow control in optical switches. Then we illustrate the local longest queue first scheme. Finally, we introduce the improved APCR buffer scheduling scheme and show an example to describe the algorithm process.

3.1 Contention Resolution Based on Flow Control

In optical switches based on flow control, shown in Section 2, the OL and payload are both sent in every time slots no matter if the switch controller accepts the packet. Flow control ensures packet transmission without the loss of optical packets. The process of FOS includes 4 steps, 'Selection', 'Report', 'Decision', and 'Feedback'.

Selection: each TOR switch selects a buffer before the making an OL. This process is implemented in ToR switch independently without any interaction from other ToRs or optical switch controller. The buffer selection already determines the requested output port of the payload in this time slot.

Request: after buffer selection, the ToR makes an OL for output request and sends it to switch controller. The process would introduce transmission delay because of the Single-Mode Fiber (SMF) between ToR and switch controller. The payload will follow the OL after a guard time.

Decision: after switch controller receives all OLs, the contention resolution is carried out to decide if the optical packet is accepted in this time slot. After contention resolution, optical switch reconfigures the switch fabric for payload.

Feedback: the ACK/NACK messages are sent to ToR to inform if the payload is transmitted successfully. If the payload was accepted, an ACK message informs the ToR to release buffer. If the payload was blocked, a NACK message informs the ToR to retransmit the last optical packet.

In the above process, 'selection' is the most crucial step to impact on the throughput of optical switches. If input packets all attempt to the same output, only one packet can be transmitted

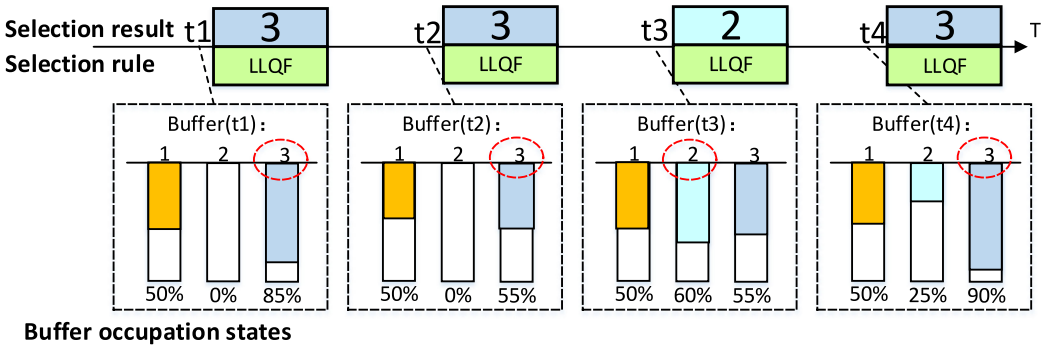


Fig. 3. The buffer selection of LLQF.

successfully. If ToRs randomly select buffer, it would introduce lots of packet contentions. Therefore, a buffer scheduling algorithm dedicates to making scheduling to improve the throughput and reduce the latency.

3.2 Local Longest Queue First

In existing optical switching architecture in DCN, a buffer scheduling scheme is used to improve payload of optical packets, called Local Longest Queue First (LLQF) [21]. The scheme is different from LQF in input-queue scheduling algorithm because only the longest queue of the ToR is selected and reported to optical switches. This scheme tries to maximize the payload of optical packets. In each time slots, the ToR switch only choose the buffer with the longest queue. If the optical contain n outputs. This scheme is of low complexity with O(n). For the switch controller, packets contention would occur. If a contention occurs, one of the packets will be blocked randomly. Fig. 3 shows an example of the buffer selection of LLQF for different time slots in the same ToR switch. Assuming that an optical switch has four ports, there are three buffers in each ToR switch. For time slot t1, queue lengths of Buffer 1, Buffer 2, and Buffer 3 is 50%, 0%, and 85%, respectively. Buffer 3 is selected in time t1 because of its longest queue in the buffer. LLQF scheme tries to ensure that ToR switches transmit as more as payload in a packet. However, it is possible for optical packets from other ToR switches to attempt to occupy the same output 3 in time t1. This scheme would introduce the potential contentions leading to more retransmission.

3.3 Adaptive Polling Contention Resolution

In this section, we introduce the basic polling scheme for bipartite matching. Then we discuss the proposed APCR for SDN-enabled optical switches. To reduce retransmission, we can prevent packets contention by desynchrony of ToRs. The similar schemes have been proposed for input-queued switches [22], [23]. In optical switches, desynchrony is challenging to be implemented because of the remote location of ToRs. However, in SDN-enabled optical DCN, it is possible for ToR switches to realize the global scheduling of buffer selection.

In optical switches, the matching of input and output is the matching problem of a bipartite graph, which has been wildly studied in input-buffer switching algorithms of electrical switches [23]. In traditional electrical switches, the optimal solution can be obtained by iterative algorithms, like MWM, MSM, iSLIP. However, in flow-control-based optical switching, ToR cannot transmit payload after receiving feedback, so how to achieve the maximum match between input and output has always been a difficult problem. Polling scheme is one of the possible schemes to achieve maximum throughput.

Polling scheme periodically selects every buffer in turn. In the polling scheme, the buffer selection depends on a polling sequence from the SDN controller, as shown in Fig. 4. The operator can design the polling sequences for all ToRs to make desynchronizing and distribute sequences to ToRs. Then

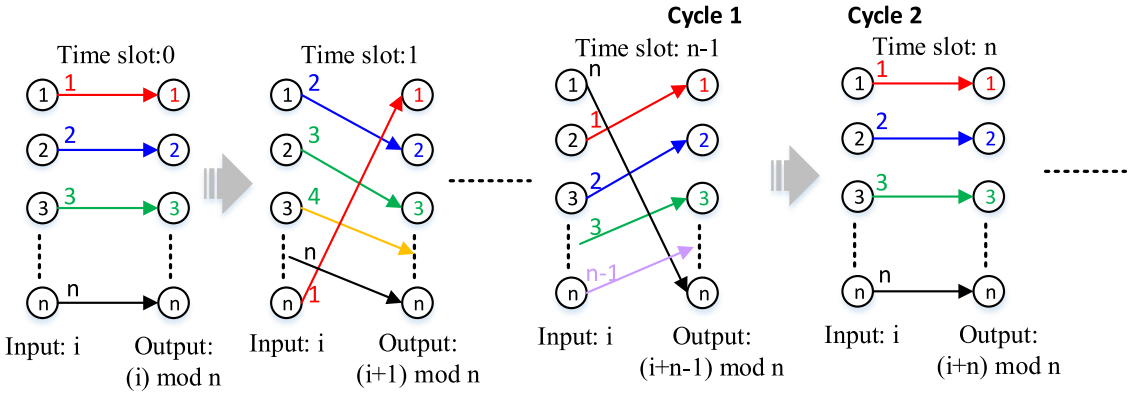


Fig. 4. Input-output matching of polling process.

ToRs will select buffer as the polling sequences. Fig. 4 is an example of a polling process of n-port switches. For any time slot (m), the input (i) would match an output as Equation (1).

$$f(i) = (i + m) \mod n \tag{1}$$

For an n-output switch, the polling period contains (n-1) time slots. Every input will be selected once in a cycle. The f(i) is polling number, indicting which buffer should be selected according to polling sequence. The left points are input ports, while the right points are output ports. The arrows mean the input-output matching. In EDCN, each input matches a ToR. In time t1, the ToR 1 chooses the buffer of output 2, while ToR 2 selects the buffer of output port 3, ToR 3 for output port 4 and ToR n for output port 1, respectively. There is no contention in the polling period. In this scenario, the ToRs select the buffer in a fixed order. If ToRs all select buffer strictly according to polling sequences, the FOS can guarantee contention-free switching. The polling sequences in ToR switches can prevent contentions. In this example, the fixed order for ToR 1 is "1, 2, 3, . . . , n". In optical switching, a ToR need not allocate a time slot for itself. In DCN, the fix order of ToR 1 is "2, 3, . . . , n−1, n". And fix order of ToR 2 is "3, 4, . . . ,n, 1". And so on, the order of n-th ToR is "1, 2, 3, . . . , n−2, n−1".

However, if the traffic is not uniform (like hot-spot load), in which the traffic to every output is not equal, the polling scheme will waste bandwidth if the queue as polling sequence is empty. Polling scheme can significantly reduce throughput if almost all the traffic in a ToR has the same destination. The APCR algorithm improves the polling scheme to arrange the burst traffic and introduces a balancing period to keep load balancing.

The APCR includes two parts: (1) The buffer selection in ToR switches; (2) The contention resolution in optical switch controller. In the load balancing period, the ToR switches choose the buffer with the longest queue, as introduced in Section 3.1. However, in hybrid polling period, the ToRs will check if the buffer as polling sequence is empty before the selection. If the buffer as polling number is empty, the ToR switching will choose the buffer with the longest queue and report the selection to the optical switch controller. If the buffer as polling number is not empty, the ToR will check the Buffer Occupation Ratio (BOR) of total VOQ size. Due to the burstiness of traffic, there is a large amount of empty buffer under low load. The polling process is not suitable for operation under low load conditions. We define a threshold (BOR = 0.4). If the buffer occupation ratio is more than the threshold, APCR will select the buffer as the polling number. If the BOR of all VOQ buffers is less than the threshold, the ToR will select the buffer with the longest queue, which means LLQF will be adopted. Fig. 5 shows an example of APCR. The cycle sequence of ToR 4 is "1, 2, 3, L". Number "2" means that the buffer for output 2 (Buffer 2) would be selected if the Buffer 2 is not empty as well as BOR is more than the threshold. If Buffer 2 is empty or BOR is less than the threshold, the polling number would be ignored. Instead, the buffer with the longest queue is

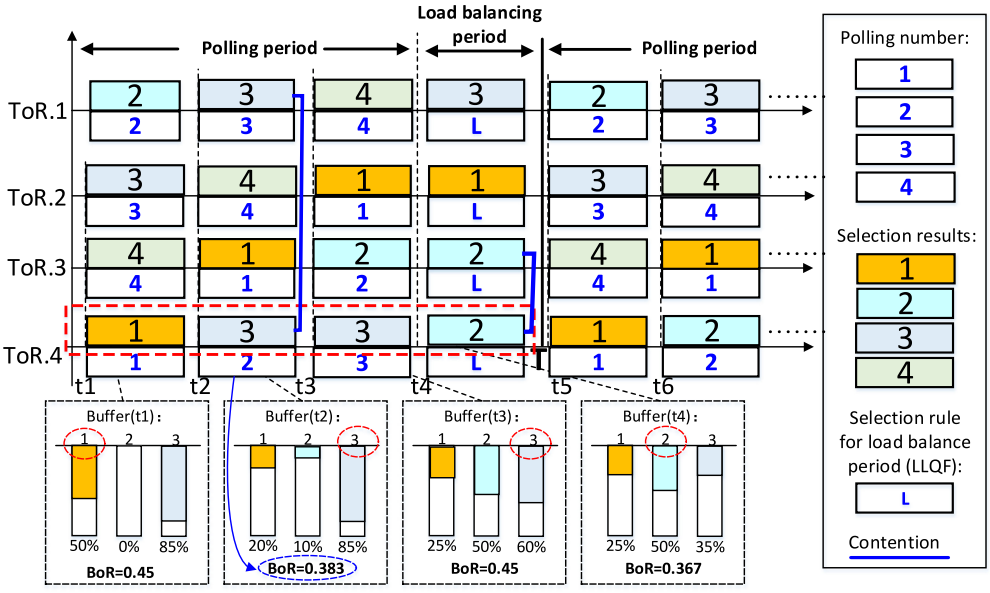


Fig. 5. Schematic diagram of APCR. (M: LLQF).

selected. 'L' is a unique number indicating the load balancing period, which means that the buffer with the longest queue would be selected. In any case of balanced or unbalanced traffic, APCR can change the buffer selection by hybrid polling scheme. As shown in time(t2), the polling number of ToR.4 is '2', but the BOR of ToR.4 is 0.383($<$0.4). The polling number is ignored, and Buffer 3 is selected.

In the SDN-enabled DCN, the SDN controller can dynamically adjust the polling order and length of cycle sequence. The ToR switches report the monitor statistics to the controller. The controller can change the polling sequences according to the real-time traffic case. In a uniform case, the load balancing period should be decreased. For example, in the hot-spot load, load balancing can be implemented by adjusting the number of hot-spot ports in the sequence. In the proposed architecture, the polling sequences can be customized and distributed from the SDN controller by operators. The process of APCR in ToR switches is shown in Algorithm 1. Polling number CS(T) means which values in the cycle sequence should be selected. The current time slot is T. Length of cycle sequence is P. N means the number of output ports, while the BOR(i) means the buffer occupation ratio in *i*th VOQ. L is a unique number, which means the time slots are in load balancing period. The total number of buffers in a ToR is B. The maximum number of iterations is N for searching buffer with the longest queue. The complexity of APCR in ToR is O(N) in the worst case.

When the ToRs complete the buffer selection, they would make labels and send labels to the optical switch controller. The payload will be transmitted after the label and guard time. When the switch controller receives all labels, it will carry out the contention resolution algorithm and then reconfigure the switch fabric as the accepted labels. If the switch controller accepts the transmission request of a ToR switch, the switch controller will feedback an ACK signal to release buffer in the ToR or a NACK signal for retransmission. Algorithm 2 shows the process of APCR contention resolution algorithm for optical switch controller. At first, the packets with higher priority will be granted for transmission. Then, if the priorities of contented packets are the same, the packet with more payload will be granted. After the process of contention resolution, the switch controller will send back the ACK\NACK signal immediately. For an optical switch with N inputs/outputs, the worst case is that N inputs attempt different outputs. For each output, the maximum number of iterations is N. The APCR has a running-time complexity of O(N$\hat{}$2).

Algorithm 1: APCR in ToR Switches.

Given: *CS(i), i* ∈ [1, *P*]; $ *k* ∈ [1, *N*], for ∀ *j* ∈ [1, *N*], BOR(*k*) ≥ ∀ BOR(*j*)
Get CS(T) and BOR(*j*), *j* ∈ [1, *N*]
if CS(T) = L **or** BOR(CS(T)) = 0 **then**
    Find *k* with *BOR(k)*
    Choose Buffer *k*.
else if Σ(BOR(i) < threshold *N **then**
    Find *k* with *BOR(k)*
    Choose Buffer *k*.
else
    Choose Buffer *CS(T)*.
end if
T = *T* + 1
if *T* = *N* + *1* **then**
    T returns to 1
end if

Algorithm 2: Contention Resolution in Optical Switches.

Given: number of ToRs K; available optical label list *label(i), i* ∈ *[1, k]* {Destination *d(i)*, priority *pri(i)*, payload *payd(i)*};
For i **from** 1 **to** *k* **do**
    repeat
        If any *d(j)*, j ∈ {[1, k], j ≠ i} is equal to *d(i)*
            If *pri(i)* > *pri(j)*
                remove the *label(j)*
            elseif *pri(i)* < *pri(j)*
                remove *label(i)*
            else
                If *payd(i)* > *payd(j)*
                    remove the *label(j)*
                else
                    remove the *label(i)*
                endif
            endif
        endif
    until no other destination is equal to *d(i)*
Endfor
Reconfigure the switches as the label

4. Results and Analysis

In order to assess the APCR performance, simulations are built for the numerical results in intra-cluster and inter-cluster. An 8-cluster DCN, shown in Fig. 6(a), is simulated. DCN architecture has been investigated in [24]. An ON-OFF model is adopted to generate traffic. The Internet frames are generated by a bimodal distribution with peaks around 64 B and 1500 B [25]. The Cumulative Distribution Function (CDF) of packets length generated in the simulations shows in Fig. 6(b). Each rack contains 20 servers, and the bandwidth of server is 10 Gb/s. The intra-ToR traffic occupy 50% traffic (intra-ToR: inter-ToR = 1:1). A load of servers can vary from 0.1 to 1 based on an ON-OFF model. The bandwidth of the optical transmitter in ToR switches is 100 Gb/s. The "load = 1" means no OFF period in ON/OFF model. Each ToR includes both inter/intra optical transmitter. The guard time between label and payload is 50 ns, payload time is 980 ns, head of a payload is 20 ns, and

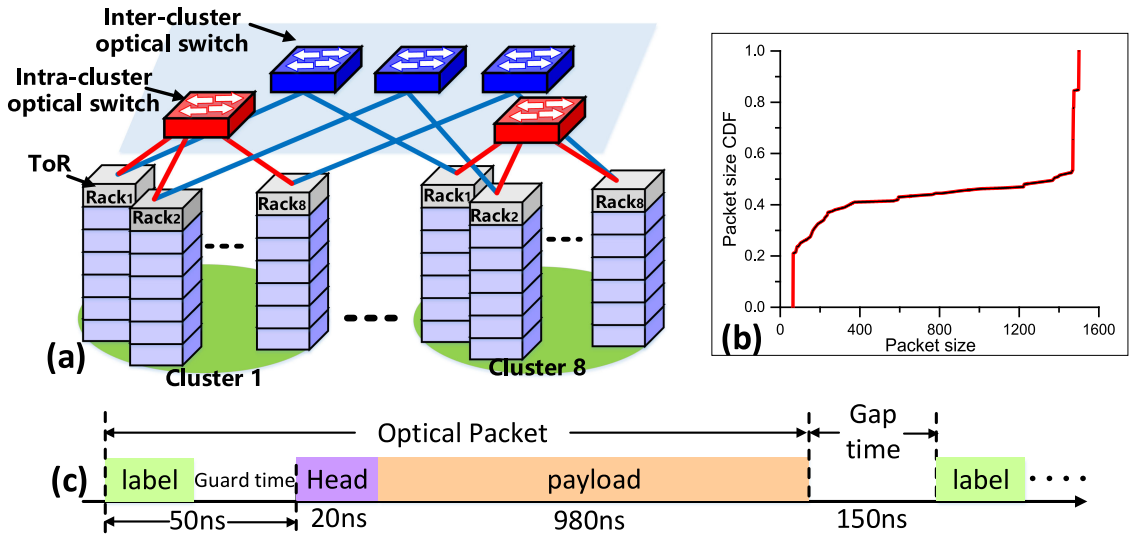


Fig. 6. (a) 8-cluster DCN architecture. (b) CDF of Ethernet frames size. (c) Optical packet format.

gap time is 150 ns. Fig. 6(c) shows the optical packet format. Buffer size of each buffer in ToR switches is range from 12.5 kB to 62.5 kB. The buffer overflow will cause packet loss. The priority of optical packets includes ten levels from 1 to 10 (10 is the highest). The priority scheme used in the simulation is as shown in Equation 2, which means the smallest integer greater than the ten-times BOR.

$$\text{Priority} = \lceil 10 * \text{BOR} \rceil \tag{2}$$

We evaluated the performance of ToR-to-ToR latency, throughput, and packet loss under different load. If two optical packets with payload intending to the same output port, a contention will be counted. We compared the APCR with MWM and LLQF. The simulation time for each result is 20 ms. Only the inter-ToRs traffic was counted in the results since the intra-ToR traffic has no impact on optical switches. MWM algorithm solves the packets contention as electrical switches and introduces an RTT delay (292 ns) between optical label and payload.

4.1 Performance in Optical Switches

At first, we test the contention resolution performance in a 4-port optical switch under different load. In this case, the optical switch consists of 4 input/output ports connected into different ToRs, and each ToR consists of 20 servers. Only the inter-ToR traffic was counted in the statistics. When the Ethernet frames arrive at the ToR switch, the frames are divided into inter-ToR frames and intra-ToR frames. Inter-ToR frames will be buffered into the inter-cluster buffer, while Ethernet switches process the intra-ToR frames. The ToR-to-ToR latency of Ethernet frames countes the time from entering source ToR to entering destination ToR. The traffic case studied in the simulation is that: 50% intra-ToR traffic and 50% inter-ToR traffic (1:1 case). An optical transceiver with the same wavelengths is considered for ToR switches.

Fig. 7(a) shows the average ToR-to-ToR latency. In uniform load, the amount of traffic to each output port is the same. Due to the traffic burst and different contention resolutions, the latency performance is quite different. The ToR-to-ToR latency means the queuing delay, and transmission delay was counted in Fig. 7(a). The transmission delay between ToRs is 200 ns. The APCR keeps the lower delay than LLQF and MWM in high traffic load. At low load, three algorithms perform almost the same for different buffer sizes, because queue delay is slight at low load case. When the buffer size of VOQ is more than 25 kB, the ToR-to-ToR latency increases rapidly. Therefore, in the

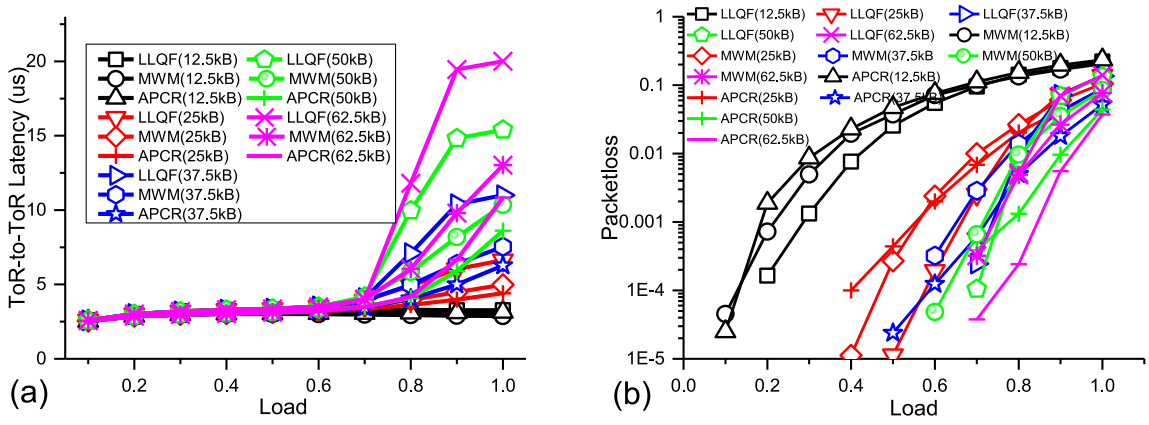


Fig. 7. (a) Load vs. ToR-to-ToR latency under different buffer size. (b) Load vs. Packetloss under different buffer size.

design of optical switches, we need to choose the appropriate buffer size to reduce average delay as well as reducing cost.

Fig. 7(b) shows the packets loss with different load. The buffer size ranges from 12.5 kB to 62.5 kB. When the buffer size is more than 25 kB, the packet loss between different buffer size is not apparent. The too-large buffer size cannot improve packet loss performance effectively. Under high load case, APCR can also reduce the packet loss, especially for large buffer size. When buffer size is 62.5 kB, APCR keeps the best performance for any load. Under a balanced load case, the APCR keep the best performance. The figure also shows that there is no noticeable difference when buffer size is more than 25 kB. However, the necessary buffer size is essential to ensure packet loss performance.

4.2 Performance in DCN

Based on the optical switch simulation performance, we choose 37.5 kB as the buffer size of VOQ to balance the latency and packet loss. A simulation is built to analyze the 8-cluster DCN performance. In this scenario, each cluster contains 8 ToRs, and each ToR includes 20 servers. Total 1280 server are simulated in the results. Only the inter-ToR traffic and inter-cluster traffic is counted in the statistics. When the Ethernet frames arrive at the ToR switch, the frames are divided into inter-cluster frames and intra-cluster frames. Inter-cluster frames will be buffered into the inter-cluster buffer, while the intra-cluster frames will be buffered in the intra-cluster buffer. The Ethernet frames arrive at the destination ToR no more than two hops. The first hop is from the source cluster to destination cluster, and the second hop is from other ToR to destination ToR. The traffic case studied in the simulation is that: the 50% intra-ToR traffic and 50% inter-ToR traffic (1:1), while 75% intra-cluster traffic and 25% inter-cluster traffic (3:1 case). Two optical transceivers have been assigned to the ToR switches, as one for inter-cluster traffic and one for intra-cluster traffic.

4.2.1 Uniform Traffic Load: the performance of different contention resolution is shown in Fig. 8 under uniform traffic load. The ON/OFF model generates steady traffic flows following the Pareto distribution. Fig. 8(a) shows the average ToR-to-ToR latency. The proposed APCR keep the best performance in any load case. When the load is less than 0.6, the MWM algorithm keeps the same performance as APCR. However, when the load is more than 0.6, the latency increases rapidly due to the additional RTT delay introduced by MWM. At high load (load $=$ 1), APCR decreases average latency by 16.84%/36.71% compared with MWM/LLQF. Polling scheme in APCR can prevent packet contention and reduce retransmission to improve latency performance. Fig. 8(b) shows the ToR-to-ToR latency CDF to analysis the delay variation at 0.5 loads. More than 700000 captured frames are used to analysis the CDF for each algorithm. As can be seen between 0 and 20 μs, the latency

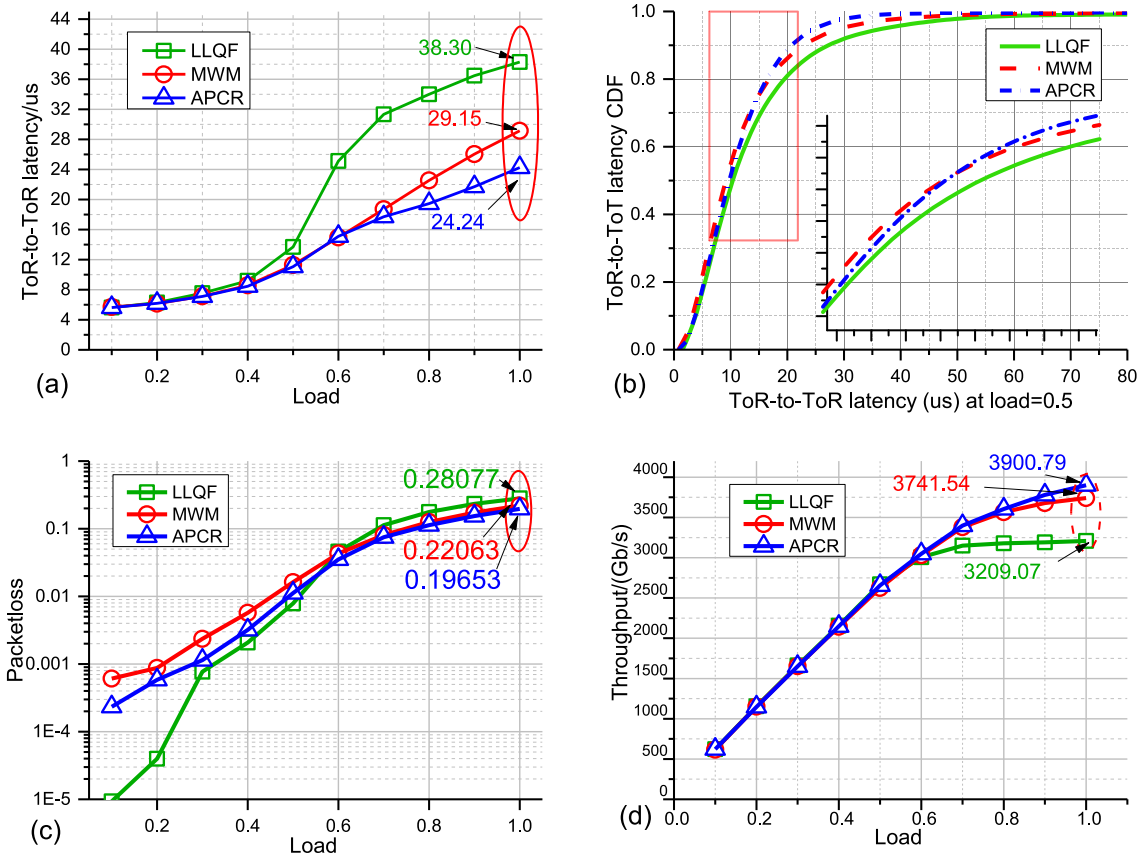


Fig. 8. Under uniform traffic load (a) Load vs. ToR-to-ToR latency. (b) CDF of ToR-to-ToR latency. (c) Load vs. Packets loss. (d)Load vs. Throughput.

distribution of the proposed algorithm is more concentrated. The steeper curve indicates that the APCR algorithm can provide more reliable QoS. Compared with Fig. 8(a) at load = 0.5, APCR and MWM have the same latency performance, but the latency variation of APCR is smaller than MWM. This excellent performance can provide FOS for more latency-sensitive applications in EDCN.

Fig. 8(c) shows the packet loss performance. Under low load, LLQF performs better than MWM and APCR. The LLQF can guarantee a lower packet loss rate, but at the cost of a higher delay, shown in Fig. 8(a). At high load (>0.5), the LLQF's performance is deteriorated rapidly, while the APCR becomes the best of three algorithms because the polling scheme can prevent packets contention leading to more throughput and low latency. In the development of EDCN, the small scale of the DCN results in transient load variability. The low packet loss performance under high load can adapt to the EDCN scenario. Compared with LLQF/MWM, APCR reduce packet loss by 30%/16.84%. Fig. 8(d) shows the throughput for three algorithms. When the load is more significant than 0.6, the difference is noticeable. APCR can improve the throughput by 21.55%/4.26% compared with LLQF/MWM at high load (load = 1). The high throughput of APCR will significantly improve the QoS of the edge data center network.

4.2.2 Hot-Spot Traffic Load: In DCN, especially in EDCN, the burst traffic has a high impact on QoS. Therefore, the performance of optical switching under unbalanced load is essential for EDCN. Hot-spot traffic load is a typical unbalanced load model. In the simulation, we employ a hot-spot traffic model to generate hot-spot traffic upon steady traffic. Each cluster has a hot-spot ToR, and the traffic to the hot-spot ToRs is twice the traffic to the normal ToRs. In this simulation case, the intra-cluster traffic accounts for 75% and inter accounts for 25% (intra-traffic: inter-traffic = 3:1). Intra-Tor traffic accounts for 50%. Only intra/inter-cluster traffic is counted in the results because the

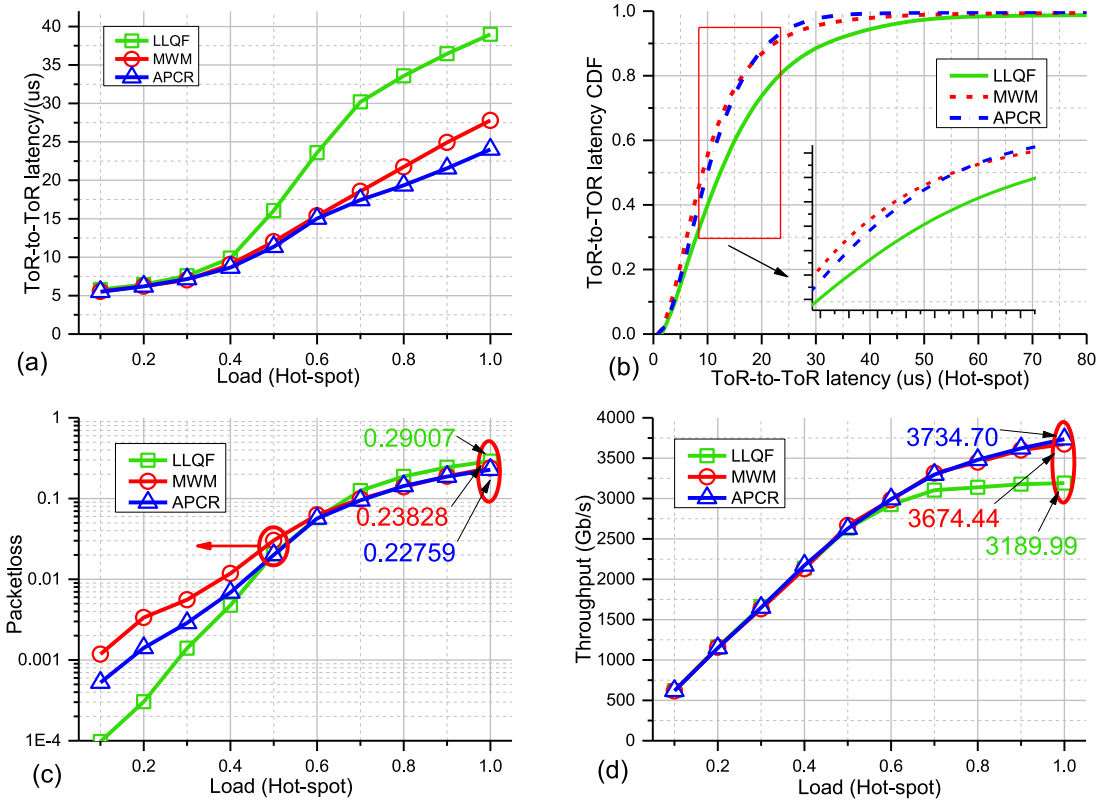


Fig. 9. Under hot-spot traffic load (a) Load vs. ToR-to-ToR latency. (b) CDF of ToR-to-ToR latency. (c) Load vs. Packets loss. (d)Load vs. Throughput.

intra-ToR traffic does not affect optical switches. The simulation results count more than 2 million traffic packets.

Fig. 9(a) shows the performance of ToR-to-ToR latency. The latency performance is similar to the performance in uniform traffic. The proposed algorithm performs better than other algorithms under high load. In low traffic load, APCR and MWM have the almost same performance, better than LLQF. The CDF of ToR-to-ToR latency shows in Fig. 9(b). Hot-Spot traffic has an evident impact on LLQF. However, for APCR and LLQF, there is no obvious difference between uniform traffic and hot-spot traffic. Additionally, the latency distribution of APCR is also the most concentrated, indicting better QoS for low latency applications. Fig. 9(c) shows the packet loss for different algorithms. When the load is 0.5, the packet loss of APCR is same as LLQF. The proposed algorithm performs better at load $>$0.6. APCR reduce packet loss by 21.43%/4.49% compared with LLQF/MWM at high load (load $=$ 1). At non-uniform loads case, especially in low-load scenarios, a large number of buffers are empty, resulting in poor performance of polling schemes. Fig. 9(d) shows the performance of throughput. In a non-uniform traffic load, the throughput of different algorithms also decreased. APCR still keeps the best performance at load $>$0.6. The proposed algorithm improves throughput by 17.07%/1.64% compared with LLQF/MWM. Therefore, the APCR is more stable than the other two algorithms in both uniform and hot-spot traffic. The actual traffic characteristics of the EDCN also affect the results of algorithms., but the hybrid polling scheme will provide a novel idea for the design of optical switching architecture.

4.3 Experimental Demonstration for SDN-Enabled FOS

An experimental demonstration was set up to validate the feasibility of APCR, as shown in Fig. 10(a). It consists of 4 TOR switches implemented by FPGAs. A switch fabric based on SOA arrays is

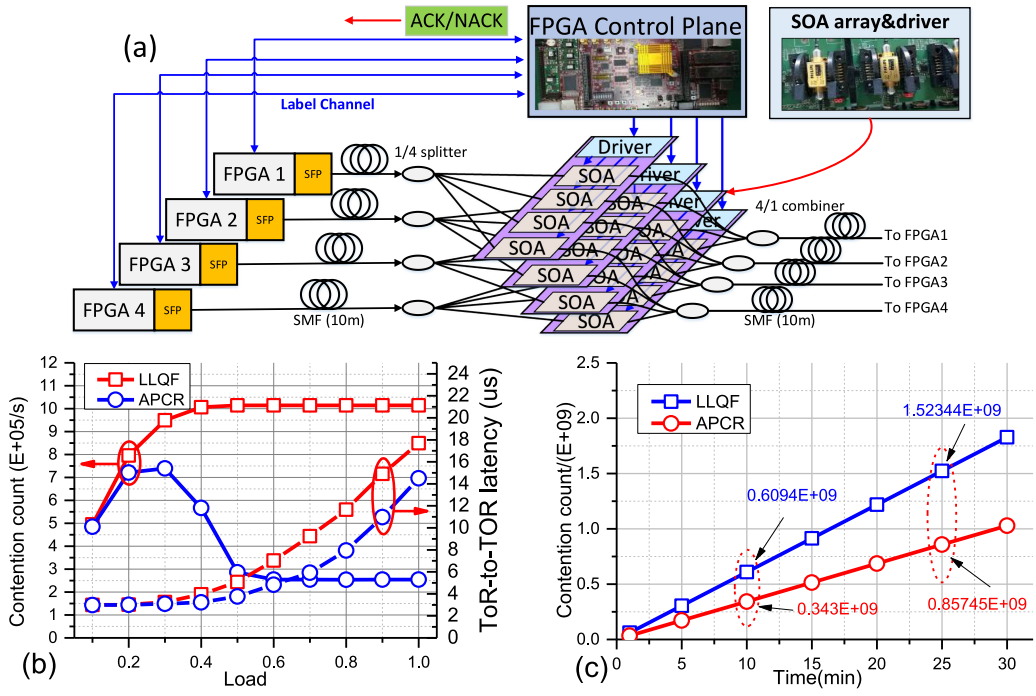


Fig. 10. (a) Experimental setup for 4-ToR cluster (b) Contention counts vs. load for LLQF and APCR. (c) Contention counts vs time for LLQF and APCR.

interconnected with an FPGA based switch controller. Optical flow control is implemented between TORs and the optical switch. Each FPGA based TOR generates Ethernet frames by a random traffic generator emulating the aggregated server's traffic. The Ethernet frame size is following bimodal distribution around 54 B and 1500 B. The traffics will be buffered by the electrical buffers. The electrical buffer consists of four VOQs for each output. Packets from the electrical buffer will be aggregated to optical packet format and transmitted to the optical switch with an attached label to provide the switch port destination. The size of the VOQ is 50 kB. The time length of optical packets is 1 μs. 10 G SFP is employed as an optical transceiver in the setup. We adopt the 1550 nm as data channel. The transmitted power is -2 dBm in our setup. The modulation format is OOK. The guard time between label and payload is 50 ns. The packets gap time is 150 ns. In our demonstration, all TOR transceivers adopt the same wavelength, which means that the optical switch can forward to the destination only one packet per time slot in case of contention. The label is processed by the switch controller, which reconfigures the SOA-based switches according to the received labels. The reconfiguration time of SOA arrays is no more than 20 ns. In the ToR switches, two buffer selection algorithms are achieved (LLQF and APCR). In an optical switch controller, the algorithm, shown in Algorithm 2, solve the contention in the optical switch. The experiments for the two algorithms are based on the same flow control and parameters. Since it is difficult to analyze Ethernet frames in real-time, we have calculated the optical packet contention performance in optical switches. For the APCR algorithm, we define the polling sequences to fix in FPGA-based ToRs. For FPGA 1, the polling sequences is "2,3,4, 99". "3,4,1,99" for FPGA 2, "4,1,2,99" for FPGA 3, and "1,2,3,99" for FPGA 4, respectively. '1', '2', '3', '4' means the buffer for output port 1/2/3/4. '99' is the 'L' in algorithm 1, which means this time slot is in load balancing period, and the buffer with the longest queue will be selected.

Fig. 10(b) shows the contention counts and ToR-to-ToR latency under different load. With the increase of load, the contention count for LLQF increases at load $<$0.5. At high load ($>$0.5), the contention count remains stable. For APCR, the count increase when the load is less than 0.3. However, for load between 0.3 and 0.6, the contention count of APCR decreases rapidly. The

load remains stable for high load ($>$0.6). With the increase of traffic load, the number of optical packets will increase, leading to more contentions. However, the number of empty buffers will decrease with the increase of load. If the queue according to polling number is empty, APCR would select the other buffer and break the contention-less rule. Therefore, the number of the empty queue and load are mutually restrictive to the increase of contention count. When the load increase ($>$0.3 and $<$0.6 shown in Fig. 10(b)), number of empty buffers decreases rapidly. Therefore, the contention count decrease between load of 0.3 and 0.6. When the load is more than 0.6, almost all the buffer is not empty. The ToR switch will select buffer following the polling number strictly, leading to less contention count. The polling scheme can show its superiority under a large load when the traffic in each buffer is relatively balanced. For ToR-to-ToR latency, APCR shows better performance compared with LLQF. Fewer contentions mean that optical switches can achieve more throughput and reduce the probability of packet retransmission. Therefore, the latency of APCR is lower than LLQF. Fig. 10(c) shows the contention count along with time at load $=$ 0.4. The number of contentions increases steadily over time. APCR algorithm shows better performance than LLQF since it reduces 43.72% contention counts at load $=$ 0.4. Experimental results show that APCR algorithm has great potential in preventing packet contentions, improving throughput, and reducing average latency, which provides reliable service for the deployment of EDCN. Compared with LLQF, APCR increase the running-time complexity of buffer selection and require more time for buffer selection, leading to more gap time between optical packets. In our experiment, the gap time between labels and payload is 150 ns, which can keep both high throughput and stable performance.

5. Conclusions

The fast optical switching can provide a solution for the strict requirement from optical EDCN, low latency, stable QoS, and high throughput. It is challenging to solve the contention by the traditional input-queue scheduling algorithm for fast optical switching. In this article, an APCR algorithm is presented that exploits the adaptive polling scheme to prevent packets contention based on flow control. An SDN-enabled EDCN architecture is shown to implement the contention resolution for optical packet switching. Simulations are made for the proposed algorithm and other schemes for Intra/inter-cluster EDCN. The proposed algorithm shows excellent performance in terms of ToR-to-ToR delay, throughput, and latency stability. The APCR also maintains excellent packet loss performance in heavy load scenarios. The proposed algorithms can reduce ToR-to-ToR latency by more than 15% under both uniform and non-uniform traffic. An experimental demonstration validates the effectiveness of the proposed algorithm on preventing contention and reducing ToR-to-ToR latency. APCR can reduce contention count by 43.72% at load $=$ 0.4. The results show the potential advantages of APCR in low latency, high reliability, and high throughput for fast optical switching.

References

[1] D. Puthal, M. S. Obaidat, P. Nanda, M. Prasad, S. P. Mohanty, and A. Y. Zomaya, "Secure and sustainable load balancing of edge data centers in fog computing," *IEEE Commun. Mag.*, vol. 56, no. 5, pp. 60–65, May 2018.
[2] Cisco Systems Inc., *Cisco Global Cloud Index: 2016–2021*, San Jose, CA, USA, White paper, 2018.
[3] W. Shi and S. Dustdar, "The promise of edge computing," *IEEE Comput.*, vol. 49, no. 5, pp. 78–81, May 2016.
[4] J. Chen, Y. Gong, M. Fiorani, and S. Aleksic, "Optical interconnects at the top of the rack for energy-efficient data centers," *IEEE Commun. Mag.*, vol. 53, no. 8, pp. 140–148, Aug. 2015.
[5] T. Segawa, S. Ibrahim, T. Nakahara, Y. Muranaka, and R. Takahashi, "Low-power optical packet switching for 100 Gb/s burst optical packets with a label processor and 8 $\times$ 8 optical switch," *J. Lightw. Technol.*, vol. 34, no. 8, pp. 1844–1850, Apr. 2016.
[6] X. Ye, R. Proietti, Y. Yin, S. Yoo, and V. Akella, "Buffering and flow control in optical switches for high performance computing," *J. Opt. Commun. Netw.*, vol. 3, no. 83, pp. A59–A72, 2011.
[7] H. Liu, Y. Li, H. Peng, J. Huang, and D. Kong, "Multicast contention resolution based on time-frequency joint scheduling in elastic optical switching networks," *Opt. Commun.*, vol. 383, pp. 441–445, 2017.

[8] N. Terzenidis, M. Moralis-Pegios, G. Mourgias-Alexandris, K. Vyrsokinos, and N. Pleros, "High-port low-latency optical switch architecture with optical feed-forward buffering for 256-node disaggregated data centers," *Opt. Exp.*, vol. 26, no. 7, pp. 8756–8766, 2018.
[9] I. Elhanany and M. Hamdi, *High-Performance Packet Switching Architectures*. Berlin, Germany: Sci. Bus. Media, Springer, 2007.
[10] S. Rangarajan, Z. Hu, L. Rau, and D. J. Blumenthal, "All-optical contention resolution with wavelength conversion for asynchronous variable-length 40 Gb/s optical packets," *IEEE Photon. Technol. Lett.*, vol. 16, no. 2, pp. 689–691, Feb. 2004.
[11] R. Proietti, C. J. Nitta, Y. Yin, R. Yu, S. J. B. Yoo, and V. Akella, "Scalable and distributed contention resolution in AWGR-based data center switches using RSOA-based optical mutual exclusion," *IEEE J. Sel. Topics Quantum Electron.*, vol. 19, no. 2, Mar./Apr. 2013, Art. no. 3600111.
[12] M. Asghari and A. G. Rahbar, "Contention avoidance in bufferless slotted optical packet switched networks with egress switch coordination," *Opt. Switching Netw.*, vol. 18, no. pt. 1, pp. 104–119, 2015.
[13] A. G. P. Rahbar and O. W. W. Yang, "Contention avoidance and resolution schemes in bufferless all-optical packet-switched networks: A survey," *IEEE Commun. Surveys Tut.*, vol. 10, no. 4, pp. 94–107, Oct.–Dec. 2008.
[14] S. Fu, P. Shum, N. Q. Ngo, C. Wu, Y. Li, and C. C. Chan, "An enhanced SOA-based double-loop optical buffer for storage of variable-length packet," *J. Lightw. Technol.*, vol. 26, no. 4, pp. 425–431, Feb. 2008.
[15] F. Xue et al.*et al.*, "End-to-end contention resolution schemes for an optical packet switching network with enhanced edge routers," *J. Lightw. Technol.*, vol. 21, no. 11, pp. 2595–2604, Nov. 2003.
[16] S. Yao, F. Xue, B. Mukherjee, S. J. B. Yoo, and S. Dixit, "Electrical ingress buffering and traffic aggregation for optical packet switching and their effect on TCP-level performance in optical mesh networks," *IEEE Commun. Mag.*, vol. 40, no. 9, pp. 66–72, Sep. 2002.
[17] G. M. Saridis et al.*et al.*, "Lightness: A function-virtualizable software defined data center network with all-optical circuit/packet switching," *J. Lightw. Technol.*, vol. 34, no. 7, pp. 1618–1627, Apr. 2016.
[18] J. Zhang et al.*et al.*, "First demonstration of enhanced software defined networking (eSDN) over elastic grid (eGrid) optical networks for data center service migration," in *Proc. Opt. Fiber Commun. Conf.*, 2013, Paper PDP5B.1.
[19] S. Peng et al.*et al.*, "Multi-tenant software-defined hybrid optical switched data centre," *J. Lightw. Technol.*, vol. 33, no. 15, pp. 3224–3233, Aug. 2015.
[20] W. Miao, F. Yan, O. Raz, and N. Calabretta, "OPSquare: Assessment of a novel flat optical data center network architecture under realistic data center traffic," in *Proc. Opt. Fiber Commun. Conf.*, 2016, Art. no. W1J.3.
[21] 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.*, 2018, pp. 1–3.
[22] Y. Jiang and M. Hamdi, "A fully desynchronized round-robin matching scheduler for a VOQ packet switch architecture," in *Proc. IEEE Workshop High Perform. Switching Routing*, 2001, pp. 407–412.
[23] N. McKeown, "The iSLIP scheduling algorithm for input-queued switches," *IEEE Trans. Netw.*, vol. 7, no. 2, pp. 188–201, Apr. 1999.
[24] F. Yan, W. Miao, H. Dorren, and N. Calabretta, "Novel flat data center network architecture based on optical switches with fast flow control," *IEEE Photon. J.*, vol. 8, no. 2, pp. 1–10, Apr. 2016.
[25] T. Benson, A. Anand, A. Akella, and M. Zhang, "Understanding data center traffic characteristics", in *Proc. 1st ACM Workshop Res. Enterprise Netw.*, 2009, pp. 65–72.