Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 2903157, 11 pages
https://doi.org/10.1155/2020/2903157

Research Article

PCOI: Packet Classification-Based Optical Interconnect for Data Centre Networks

Rab Nawaz Jadoon,[1,2] **Mohsin Fayyaz**,[3] **WuYang Zhou**,[1] **Muhammad Amir Khan**,[3] **and Ghulam Mujtaba**[3]

[1]*School of Information Science and Technology, University of Science and Technology of China, Hefei 230000, China*
[2]*Department of Computer Science, COMSATS University, Islamabad-Abbottabad Campus, Islamabad 22060, Pakistan*
[3]*Department of Electrical and Computer Engineering, COMSATS University Islamabad, Abbottabad Campus, Islamabad, Pakistan*

Correspondence should be addressed to Rab Nawaz Jadoon; rabnawaz@mail.ustc.edu.cn and WuYang Zhou; wyzhou@ustc.edu.cn

Received 1 January 2020; Revised 21 June 2020; Accepted 30 June 2020; Published 17 July 2020

Guest Editor: Van Huy Pham

To support cloud services, Data Centre Networks (DCNs) are constructed to have many servers and network devices, thus increasing the routing complexity and energy consumption of the DCN. The introduction of optical technology in DCNs gives several benefits related to routing control and energy efficiency. This paper presents a novel Packet Classification based Optical interconnect (PCOI) architecture for DCN which simplifies the routing process by classifying the packet at the sender rack and reduces energy consumption by utilizing the passive optical components. This architecture brings some key benefits to optical interconnects in DCNs which include (i) routing simplicity, (ii) reduced energy consumption, (iii) scalability to large port count, (iv) packet loss avoidance, and (v) all-to-one communication support. The packets are classified based on destination rack and are arranged in the input queues. This paper presents the input and output queuing analysis of the PCOI architecture in terms of mathematical analysis, the TCP simulation in NS2, and the physical layer analysis by conducting simulation in OptiSystem. The packet loss in the PCOI has been avoided by adopting the input and output queuing model. The output queue of PCOI architecture represents an M/D/32 queue. The simulation results show that PCOI achieved a significant improvement in terms of throughput and low end-to-end delay. The eye-diagram results show that a good quality optical signal is received at the output, showing a very low Bit Error Rate (BER).

1. Introduction

DCNs are indispensable entities that enable many of today's services like social networking, search engines, e-mail, and so on. DCNs should be able to satisfy the Quality of Service (QoS) requirements of a huge number of customers in companies like Microsoft, Google, Yahoo, eBay, IBM, and so on. These companies have data centres which have at least 50,000 nodes in a single data centre. Such a massive scale infrastructure needs energy to work. In 2010, the energy consumed by data centres was 1.5 percent of global power consumption. In 2012, the energy consumed by data centres was 120 billion kilowatts [1]. Apart from energy consumption, the DCNs are also growing in the annual network traffic. According to Cisco, the DCNs' traffic is growing at a compound rate of 25 percent up to 2019, reaching to 10 zettabytes (ZB) per year [2]. The data centre traffic is now measured in ZB, and by 2021 more than 95 percent of the total traffic will be terminated and originated in the DCNs. The major concern in designing a DCN is reduction in its power consumption, which impacts the operational expenditure (OPEX). Optical switch technologies are the choice of future [3].

Such challenges of growing traffic demand and energy consumption can be addressed by the introduction of optical interconnects. The optical interconnects in DCNS are able to

provide ultrahigh transmission bandwidth in an energy- and cost-efficient way. There are some obvious benefits of using optical interconnects in DCNs such as large port count, long reach communication, ability to reconfigure, large extinction ratio, and dealing with traffic heterogeneity. Optical DCNs have many communicating nodes, so they have large arrays of similar optical components. Optical components used in optical data centres follow fixed physical laws related to light. Interaction between optical components depends upon the changes in the properties of light brought by the optical components. Optical switching is a key component of today's high-performance communication networks. Optical interconnect is a box with N inputs and N outputs. At any time, the internal interconnection of optical components establishes paths from the inputs to the outputs. Congestion can occur when the number of source nodes trying to access a destination node exceeds the capacity of the destination node, which makes queuing unavoidable. Queuing can be done on the input side, the output side, or both [4].

The signal degradation due to optical components can introduce Bit Error Rate (BER) in optical interconnects. The signal degradation can be measured by analysing the eye diagram of the signal. The eye diagram is created by superimposition of zeros and ones of the optical digital data stream. The eye diagram of the optical signal can give information about jitter, noise, signal amplitude, duty cycle distortion, fall time, and bit period. Different types of signal degradations include reduced amplitude, changed shape, introduction of noise, and change in the fall time or rise time of pulse. Different types of optical components are responsible for such signal degradations. Optical components used in optical data centres include couplers, Directly Modulated Masers (DML), Vertical Cavity Surface Emitting Laser (VCSEL), Fibber Delay Lines (FDL), Fiber Bragg Grating (FBG), and SOA. Couplers cause reduction in amplitude, DMLs and VCSELs cause change in shape of signal, FDLs cause delay in bits, FBGs cause addition of noise, and SOAs introduce spikes in the optical signal.

The signal degradation compensation can be achieved using various methods. Reduction in amplitude is compensated by using amplifiers or increasing the amplitude of transmitted signal. Optical amplifiers can reduce the BER by boosting the amplitude of weak optical signal [5]. Use of DML reduces the cost of communication system as it eliminates the need for separate modulator. However, they introduce nonlinear changes in shape of optical signal which need to be compensated. These changes can be compensated by using predistortion circuits [6]. Changes in shape of optical signal are also introduced by VCSEL. Feedforward equalization is used to compensate these changes [7].

When SOAs are used in optical DCNs, they cause spikes at the start of bit. These spikes gradually reduce with time, depending upon the transition time of SOA. Such distortion can be compensated by increasing the length of bit period, such that it is greater than the transition time of SOA. This results in the reduction of supported bit rate of the system. The equalizer adaptation algorithm can mitigate the timing jitter caused by FDL, which works bitwise to detect the amount of time shift [8]. Uniform noise introduced by FBG can be mitigated by either amplifying the optical signal or increasing the transmitter power.

The objective of this paper is the proposal of a novel architecture which presents various benefits over existing architectures and investigate its performance on the basis of the following techniques:

(1) Mathematical analysis of the input and output queuing is performed

(2) Simulation of TCP protocol

(3) Eye-diagram analysis is also conducted to measure the signal degradation and BER at physical layer

In the rest of paper, Section 2 describes related work of various optical architectures, Section 3 describes the PCOI architecture, Section 4 describes the system model for input and output queue analysis, Section 5 describes the input and output queuing analysis of PCOI architecture, Section 6 describes the TCP simulation of PCOI architecture, Section 7 presents the physical layer analysis in terms of eye diagram and Bit Error Rate (BER), and Section 8 presents the conclusions and future work.

2. Related Work

Various architectures have been proposed in the literature which encounter various problems of optical networks and exploit benefits of optical components. These architectures can be classified into the categories which include (a) reconfigurable architectures, (b) low latency architectures, (c) low blocking probability architectures, (d) low power consumption architectures, (e) scaling link bandwidth architectures, and (f) high radix architectures.

2.1. Reconfigurable Architectures. These architectures have the ability to change the network topology based on change in traffic patterns. Reconfigurable architecture mentioned in the literature includes [9–11] as explained subsequently.

In [9], the architecture presented is based on Wavelength Selective Switch (WSS) and MEMS switch. Every server has a unique wavelength, which is multiplexed together according to destination rack. WSS groups the wavelengths based on destination racks and sends groups of wavelengths to MEMS ports. The topology manager is responsible for configuring WSS and MEMS switch for proper function.

In [10], a reconfigurable architecture based on Arrayed Waveguide Grating Router (AWGR) and Tunable Wavelength Converter (TWC) is presented. Header is extracted from each packet and sent to the control plane. The payload of packet keeps on waiting in FDL until the control unit makes decision about the wavelength to be set for each packet through TWC. Based on change in wavelength, the AWGR routes the packet to proper destination port.

The architecture in [11] is made up of MEMS switches. The control unit is responsible for configuration of MEMS switches. As MEMS switches are slow to configure, so this architecture is more suitable to circuit-switched applications as compared to packet-switched application.

The architecture in [12] uses Software Defined Network (SDN) with optical switching. The building block is a pod, which hosts several racks. The ToRs are connected using a star topology. The switching within the same pod is performed passively using optical filtering. The network can be scaled by using pods in a ring topology. Each ring has WDM traffic and add/drop multiplexing to and from ring is performed on a per-wavelength basis. The data plane exists in TDMA, where time slots are accessed for rack to rack communication.

2.2. Low Latency Architecture. Low latency architectures are the ones which have distributed control which include [10, 13].

In [10], a broadcast-select Spanke-type architecture is presented which minimizes the control decisions in the network to reduce the latency. The packets from source nodes are broadcast to all destination nodes. At the destination node, Wavelength Selector (WS) selects the appropriate destination node. This architecture is scalable without affecting the latency.

In [13], the network is flattened into three stages. The first stage is called input module (IM), the second stage is called central module (CM), and the third stage is called output module (OM). Each module is made up of AWGR and a scheduler. The scheduler at each stage configures the wavelength of flows to be directed to appropriate output port of AWGR. This is also scalable without affecting the latency of the system.

2.3. Low Blocking Probability Architecture. These are the architecture which show a high number of successfully transmitting nodes out of total transmitting nodes at the same time. These architectures include [14–16].

The architecture in [14] is a three-stage network. The first and third stages are based on AWGR, whereas the second stage is a time buffer. Collisions are avoided by using the time buffers. The packets wait in the second stage for contention to be resolved in the third stage.

In [15], the contention is avoided by using space-wavelength multiplexing and broadcast-select scheme. The select unit has two functions: first to select the correct spatial group and second to select correct wavelength. The select units are made up of SOAs. The central scheduler controls the SOA gates. There are two receivers on each destination node, which further reduces the blocking probability.

In [16], the Reflective SOA (RSOA) is behaving as the mutex element. In case of contention, when multiple input ports are trying to access the same destination port, the RSOA grants access to only one input port to transmit. The sender node only starts to transmit the data if a positive acknowledgment is given by RSOA. The acknowledgment of RSOA is made up of reflected optical power. In this case, multiple sender nodes are trying to send to the same output port the reflected power if RSOA drops.

2.4. Low Power Consumption Architectures. The architectures with low energy consumption are [17, 18]. The architecture in [17] has highly distributed control, which makes it scalable to large port counts. Electronic buffer is implemented on each node, which has multiple queues, and each queue has a distinct wavelength. Multiple flows can arrive at a single output port, which can be resolved by Wavelength Selector (WS) partially. When packets from different queues of the same node go to a single destination node, the contention cannot be resolved and only one flow is forwarded. However, when multiple flows from different nodes arrive at the destination, such contention is resolved by WS by changing its wavelength.

In [19], the architecture is made up of clusters, boards, and nodes. Each board has multiple nodes and multiple boards are connected using optical wavelength multiplexing. Vertical Cavity Surface Emitting Lasers (VCSELs) are used to eliminate the need for external modulator, which conserves energy.

In [18], the two main optical communicating devices are AWGR and Microring Resonators (MRR). AWGRs are passive optical devices, which do not need any external energy input. MRRs are also very low power and high bandwidth devices. This architecture is entirely made up of low energy consumption devices, which save energy.

In [20], an architecture is presented which minimizes energy consumption by using a combination of optical cross-connects and WDM rings. Optical circuit switches provide dedicated nonblocking circuits. WDM rings are interrack switching elements.

The architecture in [21] uses space and wavelength multiplexing. The architecture is divided into cards and each card has multiple nodes. Each node has a distinct wavelength. It uses multiple level addressing. The card is selected using couplers in space domain and port is selected in wavelength domain by the wavelength.

2.5. Scaling Link Bandwidth Architectures. These architectures include [9, 19, 22]. In [19], Vertical Cavity Surface Emitting Lasers (VCSELs) are used, which behave as both the source of light and the modulator. VCSELs eliminate the need for external modulators in the transmitters. The unused VCSELs are shared among transmitters to increase their transmitting bandwidth, which makes this architecture a scaling bandwidth architecture.

In [23], the architecture uses the ring topology and wavelength multiplexing to interconnect the nodes. The add/drop multiplexer and Wavelength Selective Switch (WSS) are used in the ring to select the wavelength for destination node and add the wavelength to send data by sender node. This architecture is not much scalable.

In [24], an architecture is presented in which AWGRs are used but the interconnection between racks is based on passive optical components. The mixed linear integer programming model is used for wavelength assignment, which ensures that a single wavelength is assigned between two server groups. The directionality of AWGR in ensured so

that flows are always directed from input ports to output ports.

The work in [25] presents an architecture which uses optics and commodity switches. The backplane is a switchless core made up of bus-based fiber rings. The architecture is divided into sectors and each sector is made up of ToR switches and interconnection pods. Within a sector, electronic switching is performed. The absence of switches implies full bisection bandwidth for single-hop communication.

2.6. High Radix Architectures. High port count in the architectures is achieved if the network performance is not affected by increasing the number of nodes in the network.

In [26], a distributed and scalable optical packet switch architecture is presented. It is based on Arrayed Waveguide Grating Router (AWGR) and SOA. A single AWGR can connect multiple ToRs. It has a modular structure and nonblocking nature of AWGR makes it scalable. It shows lower latency and low blocking probability.

In [9], an architecture is presented which can achieve dynamic configuration of link bandwidth. For overall connectivity, Microelectromechanical System (MEMS) switch is used, which uses micromirrors to deflect the light to the output port. Wavelengths are grouped using Wavelength Selective Switch and directed to respective destination racks.

The architecture in [10] presents a reconfigurable architecture, which is based on AWGR. Wavelength conversion mechanism is used at the input port to rout the light paths to respective destination ports. Tunable Wavelength Converters (TWC) are used for this purpose to change the wavelength of light signal. The labels are processed to extract the destination address, which is used to change the wavelength of input signal.

The architecture in [13] flattens the networks, which reduces the number of hops in the network. It is a multistage architecture, which is achieved using three modules: the input module, central module, and output module. Each module is composed of TWC and AWGR. Wavelength switching is used in each module to direct the light path to appropriate input port of next module.

The PCOI architecture was proposed to minimize the routing complexity by exploiting packet classification and use of queuing. It further reduced energy consumption by using passive optical components. The next section describes the PCOI architecture briefly.

3. The PCOI Architecture

The proposed architecture which avoids contention under various traffic patterns is shown in Figure 1. The nodes are arranged in the form of racks. Each rack is assigned a unique wavelength. This architecture exploits the benefits of input and output queuing, wavelength multiplexing, and space multiplexing. Each rack has a classify module. This module classifies the packets on the basis of destination rack and puts them in respective queues. The packets from queues are converted into optical domain by using Electrical-to-Optical (E/O) conversion. The optical flows from a single rack have a unique wavelength, thus requiring as many wavelengths as the number of nodes. The flows of a single destination rack from multiple sender racks have different wavelength, which allows simultaneous data to be received from multiple sender nodes, making the all-to-one communication possible. Optical power combiner at receiver rack collects various wavelength flows, which are later demultiplexed and given to $1 \times N$ switch at receiver rack. The optical power combiner is used because of its passive nature and low cost; it can combine various frequency optical signals without the need for external source. The packets of each source rack are kept separate using different wavelengths.

The demultiplexed wavelengths contain the packets of respective source racks. Out-of-band signalling is used to reduce the header extraction time. At the destination rack, packets of each wavelength go through a $1 \times N$ switch, which is also shown in Figure 2. The switch is designed after header detection packet goes to its destination node, where they are collected by the output queue in case of contention.

The $1 \times N$ switch in Figure 2 is a high port optical switch. The header information and payload of packet are separated. The header is sent to control unit, whereas the payload is sent to SOA after some delay, which is equal to the processing time of control unit. The delay is provided by the FDL which exits before the SOA. The payload keeps traveling in FDL until the decision is made by control unit to turn on the respective SOA of destination port. The rest of SOAs are turned off by the control unit. As a result, the payload only travels to the destination port of switch as determined by the destination address of header.

4. System Model

The PCOI architecture represents a nonblocking switch as it involves self-routing mechanism, where a packet finds its path to the desired output port. This is shown in Figure 3. The transmitter and receiver nodes represent the input and output queue system. Head-of-Line (HOL) blocking problem arises when multiple source nodes try to send to the same destination node, which results in stoppage of transmission of packets from head of input queue which are sending packets to that destination. This is shown in Figure 4. N input and output ports are assumed, and transmission is synchronous, which means that packets are sent from the input ports to different output ports at the same time. When the buffer of the output port is full, it sends backpressure signal to the respective input queue. After receiving the backpressure signal, the input queue stops transmitting the packets to that output port. The packets of different input queues that want to send packets to the same output buffer are made to wait in the input queue. The backpressure mechanism prevents overflow in the output buffer. The effect of backpressure on the performance of optical interconnect for different buffer sizes is analysed. The number of ports is assumed to be large, that is, $N \longrightarrow \infty$. The PCOI architecture achieves speed-up at the output port by receiving multiple flows using wavelength, code, and space multiplexing. It results in significant improvements in delay reduction and throughput enhancement. The throughput

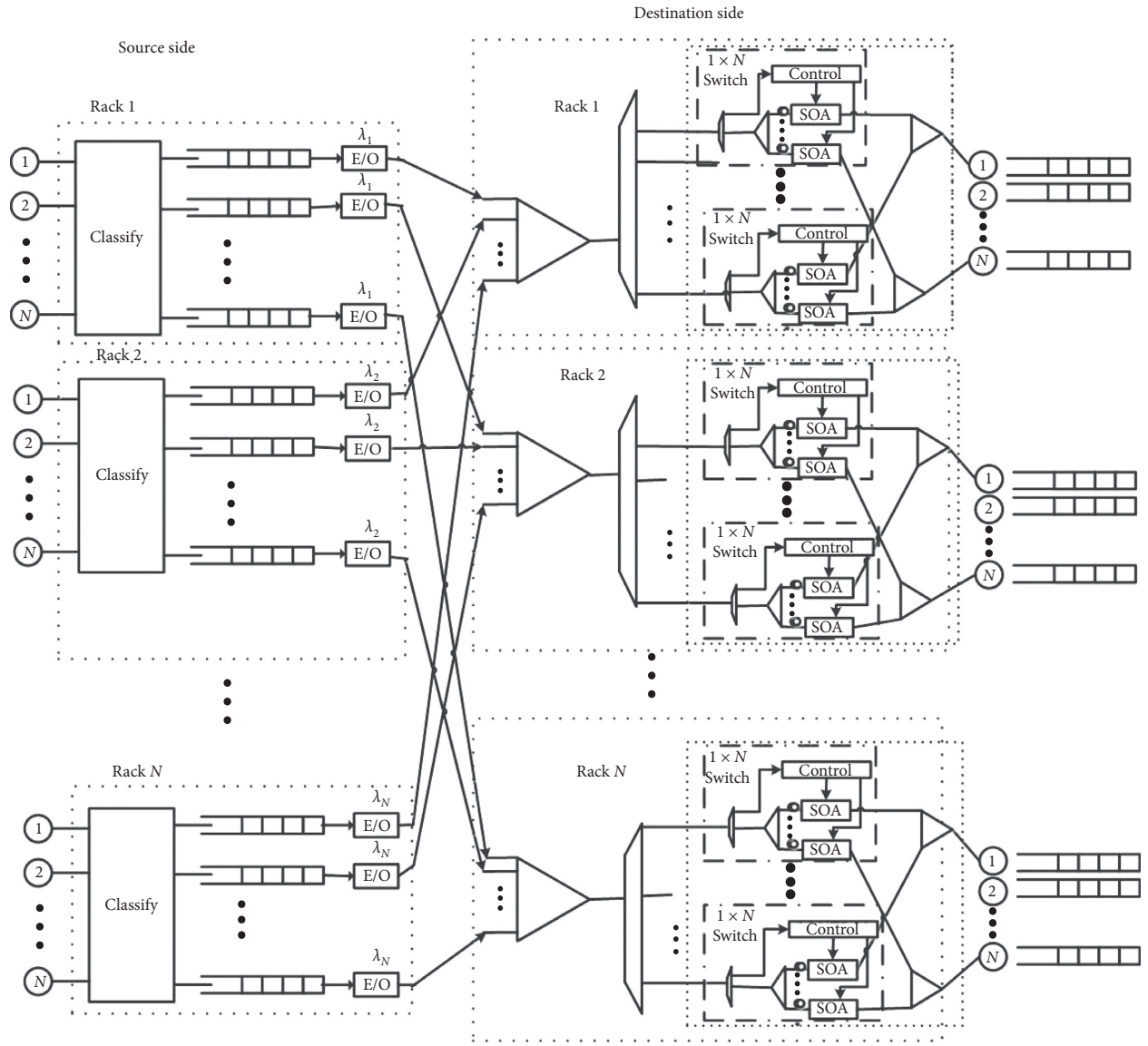

Figure 1: PCOI architecture.

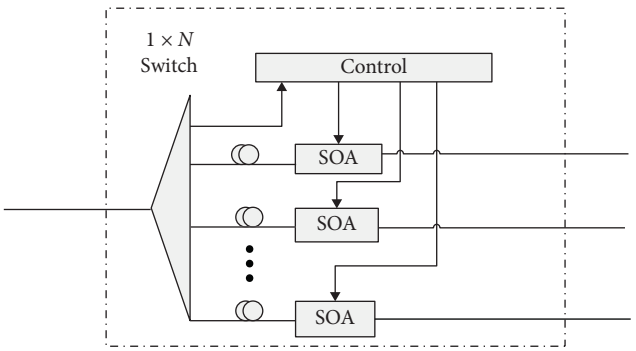

Figure 2: $1 \times N$ switch.

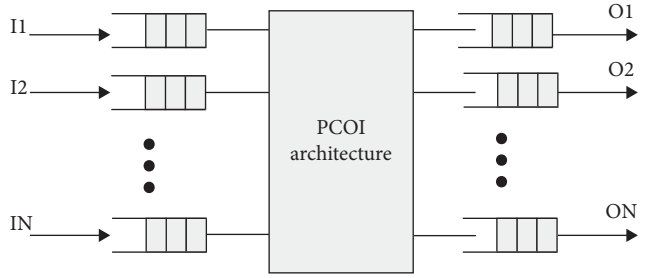

Figure 3: Switch model for the PCOI architecture throughput enhancement.

enhancement is evident from Figure 4, where each destination ports 32 servers in the queuing model. It can process 32 streams at the same time, which increases the throughput and reduces the queuing delay.

5. The Input and Output Queue Analysis of PCOI Architecture

The PCOI architecture behaves as an N\times N switch. The packets are assumed to be of fixed length and fixed transmission time. The traffic is randomly distributed, which

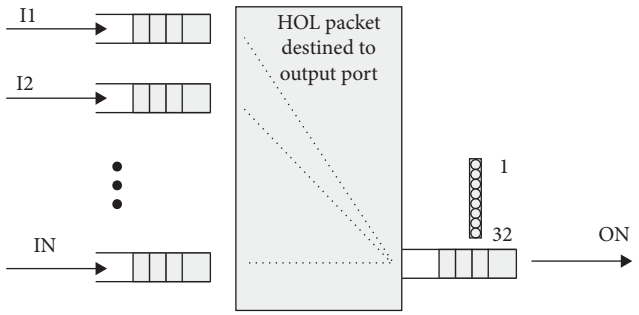

Figure 4: Queue model used to evaluate the PCOI architecture.

means that a packet from an input port can be sent to any of the N output ports with equal probability, which is $1/N$. The output queue operates on First Come First Serve (FCFS) basis. If the output port is idle, the packets go through the PCOI architecture directly. The performance of the PCOI architecture is affected by the output buffer size "b." The arrival process is Bernoulli, as the probability of arrival in a time slot defines the load. The efficiency of the PCOI architecture is measured, based on maximum throughput and average delay. The packet delay is made up of three components: (i) waiting time in input queue until the head; (ii) waiting time at the head of the input queue due to HOL, and (iii) waiting time at the output queue due to contention of the output port.

The traffic of input port on average is "p" packets per unit time. Packets of the input ports are independent of each other. There is a possibility that more than "b" packets are trying to access the same output port. The output port of the PCOI architecture can receive 32 multiplexed streams of packets. The waiting time at the input buffer until it begins the transmission to the output port represents an M/D/32 system. The M/D/32 queue can be treated as M/D/1 queue by scaling basic unit of time; the p of M/D/1 queue becomes $32p$ for M/D/32 queue [27].

The analysis is based on [28], which represents an M/D/1 queue. It is modified for the RPL architecture by replacing p with $32p$. Closed form expressions for the average delay and maximum throughput are derived. Thus, the total average delay, which is shown in Figure 5, is given by

$$D = \frac{\overline{Q_b^2} + (64p-1)\overline{Q_b} + (32p)^2}{64p\left(1-32p-\overline{Q_b}\right)} + \frac{\overline{Q}}{32p}, \tag{1}$$

$$\overline{Q_b} = \overline{Q} - b\left(1-32p_0\right) + \sum_{j=1}^{b}(b+1-j)p_j, \tag{2}$$

$$\overline{Q_b^2} = \overline{Q^2} - 2b\overline{Q} + b^2\left(1-32p_0\right) - \sum_{j=1}^{b}(b+1-j)^2 p_j, \tag{3}$$

where $\overline{Q}$ ad $\overline{Q^2}$are the first and second moments of waiting customers of M/D/c queue and p_j are the steady-state probabilities given by

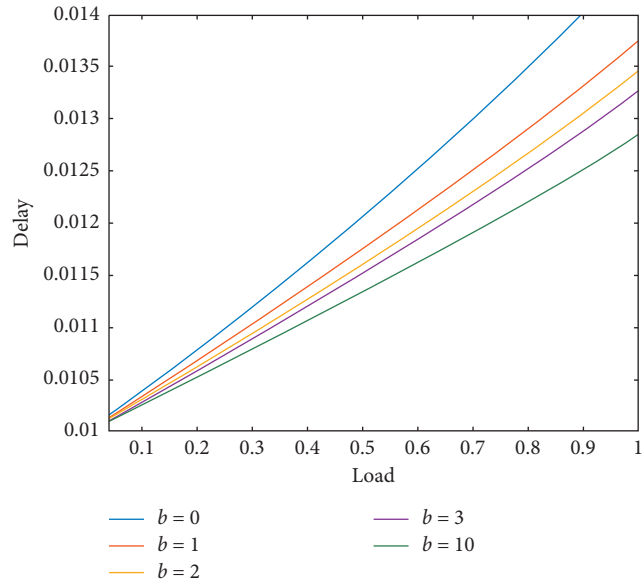

Figure 5: Delay analysis of the PCOI architecture.

$$\overline{Q} = \frac{(32p)^2}{2(1-32p)}, \tag{4}$$

$$\overline{Q^2} = \frac{(32p)^2\left((32p)^2 - 32p + 3\right)}{6(1-32p)^2}, \tag{5}$$

$$p_0 = (1-32p), \tag{6}$$

$$p_1 = (1-32p)\left(e^{32p}-1\right), \tag{7}$$

$$p_j = (1-32p)\sum_{i=1}^{j}(-1)^{j-i}e^{i(32p)}\left[\frac{(i\,32p)^{j-i}}{(j-1)!} + \frac{(i\,32p)^{j-i-1}}{(j-i-1)!}\right]. \tag{8}$$

The corresponding p. g. f for the steady-state probabilities is

$$P(z) = \frac{(1-32p)(1-z)}{1-ze^{32\,p(1-z)}}. \tag{9}$$

The probability that a packet arrives at the head of the input queue and experiences delay due to backpressure, which is shown in Figure 6, is given by

$$P_{b,s} = \frac{1}{32\,p}\left(1-\sum_{i=0}^{b+1}p_i\right). \tag{10}$$

6. Simulation

The performance of the PCOI architecture is measured for different output queue sizes using TCP. Simulation setup consists of a 512×512 network. The network is implemented in NS2, a discrete event simulator developed for research and educational use [29]. It is an open-source software. For many-to-many communication pattern, each node

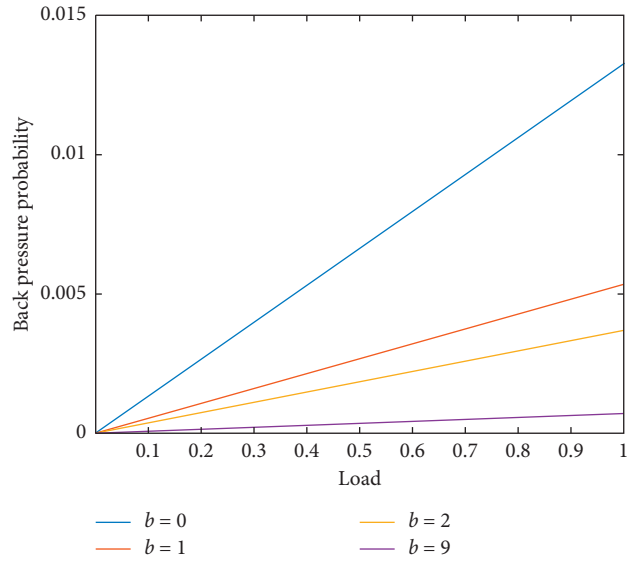

Figure 6: Backpressure analysis of the PCOI architecture.

transmits to a randomly chosen destination node. The random destination node address is generated according to a uniform random variable. Figure 7 shows the number of sender nodes trying to access the destination nodes. This number fluctuates between 1 and 5. The median of these two numbers is 3, so for many-to-many communication pattern, it is assumed that 3 nodes are trying to access a destination node.

TCP is used as the transmission protocol. The traffic has a constant bit rate. The load is varied from 1 Mbps to 10 Mbps, which accounts for the normalized load from 0.1 to 1 in the figures in the following sections. The link delay is assumed to be 10 ms. TCP is responsible for 90 percent of Internet traffic. TCP can adapt the transmission of packets according to bandwidth; it avoids congestion and retransmits lost packets. By keeping the output queue buffer size constant, the network initially shows a higher average delay due to the slow-start phase of the TCP. TCP has a slow-start phase because all the nodes try to transmit at the same time and congestion causes packets of input nodes to wait to avoid packet loss. Performance of TCP is analysed by changing the buffer size of the output queue in NS2. The Drop-Tail queue is implemented in which the last arrived packet is dropped if the queue is full. The overall throughput of TCP increases by increasing the buffer size, whereas the average delay of the packets decreases by increasing the buffer size of the output queue. Constant bit rate (CBR) application is used to generate traffic on the nodes. TCP uses reliable congestion control, in which acknowledgment is created by the destination to know whether packets have been received. Lost packets are interpreted as a congestion signal. Initially, the average delay is high but as the load increases, the average delay decreases. It uses a dynamic congestion window, which grows rapidly initially and then increases slowly as it reaches a threshold. When congestion is detected, it drops rapidly. The output of the NS2 simulation is in the form of a trace file. The required information of delay and throughput is obtained from the trace file by performing text processing using a Perl script. The small Perl programs are used as filters to extract the required information from the text. The trace file has 12 fields, the first field is of event type, the second field is the time at which the event occurs, the third field is the input node of link at which event occurs, the fourth field gives output node of link, the fifth field is of the packet type (TCP, UDP, AGT, etc.), the sixth field gives the packet size in bytes, the seventh field includes flags, the eighth field gives the flow id, the ninth field gives the address in the form "node.port," the tenth field gives the address in the same form, the eleventh field gives the packet sequence number of the network layer, and the twelfth field shows the unique id of the packet.

Figure 8 shows the TCP delay for different output buffer sizes. Initially delay is large due to slow-start phase of the TCP and it reduces later.

Figure 9 shows the throughput performance of the TCP for different buffer sizes. It is seen that for lower buffer size there is sudden drop in throughput for higher load because of sudden change in congestion window size. For lower value of output buffer size, the congestion occurs early by increasing the load.

Figure 10 shows the maximum achievable throughput as a function of output queue buffer size. All-to-one communication pattern is the worst traffic encountered in any network.

Figures 11 and 12 show that PCOI architecture shows performance benefit for this traffic pattern for all-to-one and many-to-many communication paradigm, respectively.

7. Physical Layer Analysis

The physical layer of PCOI architecture is simulated in the OptiSystem [30], to measure the BER and signal degradation. The signal degradation can be analysed from the eye diagram, which is shown in Figure 13. The simulation in

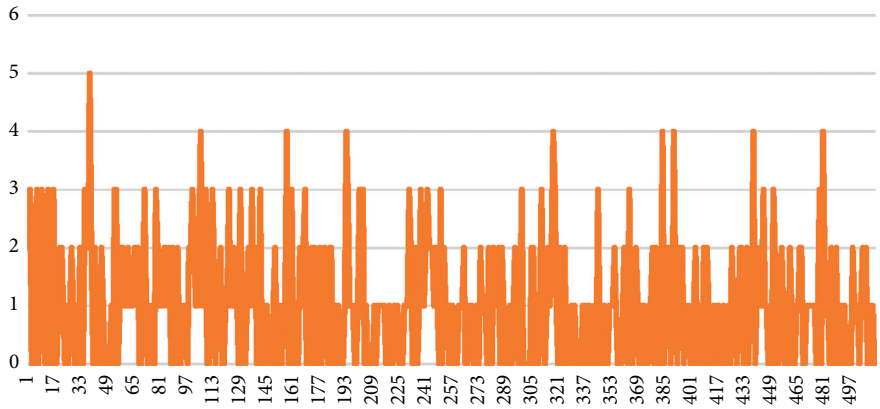

Figure 7: Number of sender nodes trying to access a destination node.

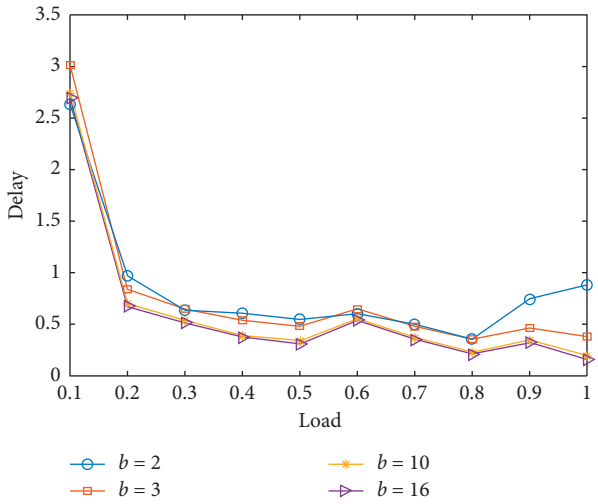

Figure 8: Delay analysis of the PCOI architecture.

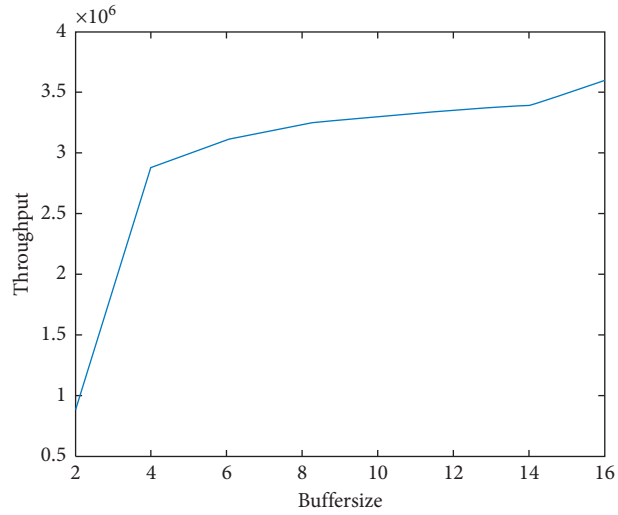

Figure 10: Maximum throughput simulation of the PCOI architecture.

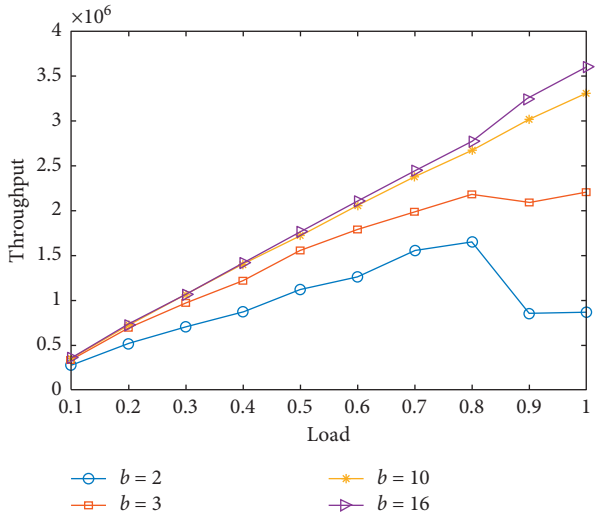

Figure 9: Throughput simulation of the PCOI architecture.

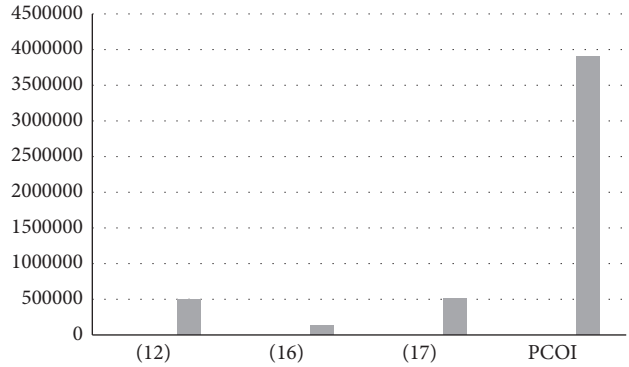

Figure 11: Throughput comparison for all-to-one communication pattern.

OptiSystem is carried out by analysing the optical path of each architecture. The optical signal consists of pseudo-random bit sequence. For bit generation, non-return-to-zero (NRZ) pulse is used. NRZ pulse is used because the pulses have more energy and have additional rest state besides zeros and ones, which gives a bigger margin between two logic levels. Mach Zehnder (MZ) modulator generates the optical

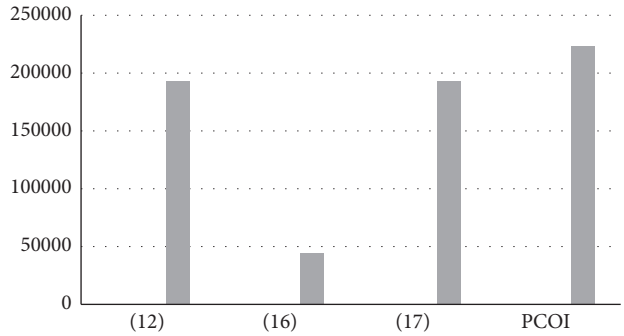

Figure 12: Throughput comparison for many-to-many communication pattern.

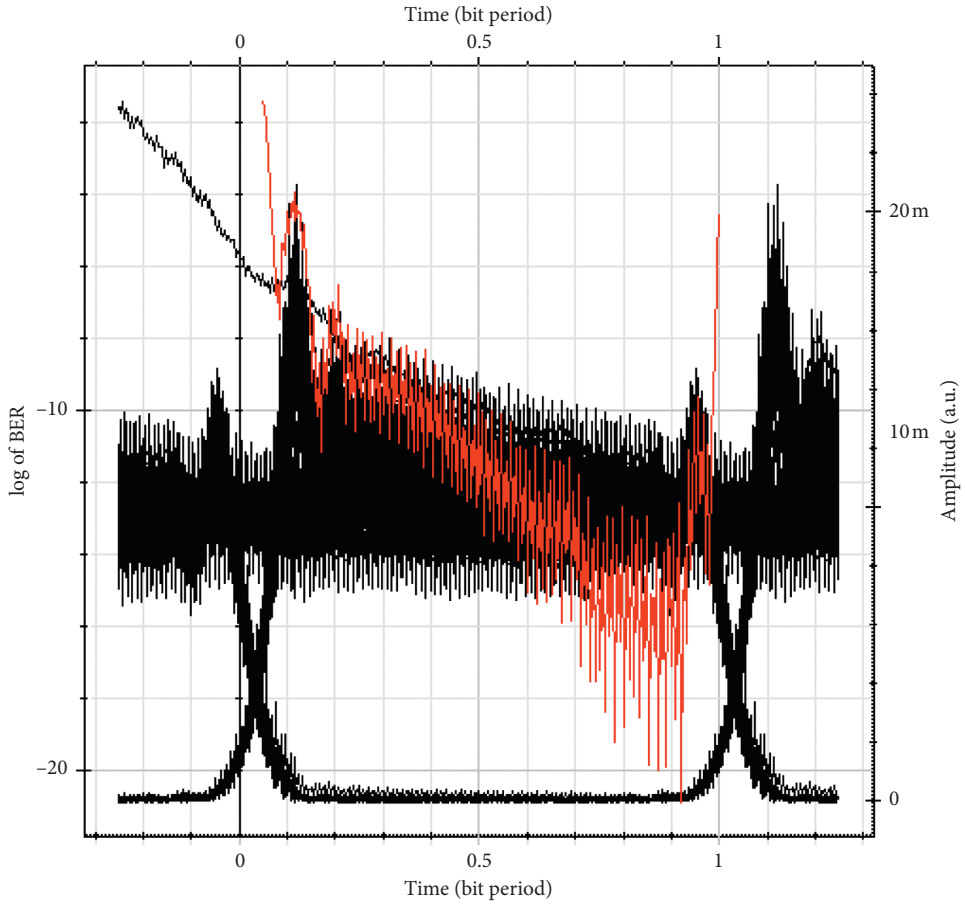

Figure 13: Eye diagram of the PCOI architecture.

signal which corresponds to the bitstream. MZ modulator is popular for low power, compact size, and monolithic integration. The transmit power is varied from −12 dBm to 5 dBm. The eye diagram of the PCOI architecture in Figure 13 is calculated using the 0 dBm transmit power. It has shown a very good eye opening. This shows that received signal in the PCOI architecture has enough signal quality to be detected. The eye opening shows that it has low jitter. The log (BER) of −10 is achieved for this eye diagram.

8. Conclusions and Future Work

The modelling of optical data centres is very important, as it helps in making important decisions about their performance. It is very important to consider at design time of optical data centres which optical components to use. The choice of optical components directly affects the quality of received signal. If the quality of received signal is bad, it can adversely affect the Bit Error Rate of the system. Ambiguity

in logic levels makes it difficult for the receiver to distinguish between bits. However, if an optical component is used that degrades the signal quality, then the mitigation techniques should be used to counter the effect of optical components. There are two ways to improve the performance of optical data centres, either reducing the signal degradation or making the design of optical data centre to reduce contention.

The main advantages of the PCOI architecture are routing simplicity, reduced energy consumption, scalability to large port count, packet loss avoidance, and all-to-one communication support. Routing simplicity is achieved in PCOI by using the packet classifier at the sender side, which classifies the packets based on destination rack and puts them in a queue. The reduced energy consumption is achieved in PCOI by use of passive optical components. Passive optical components are those which do not need any external power source for their working. The PCOI architecture is scalable to large port count due to lack of central controller. The PCOI architecture can avoid packet loss in worst communication patterns by exploiting the redundancy of optical components and queues to temporary store packets with collisions. All-to-one communication pattern is the worst communication pattern in the communication system, in which all the sender nodes try to access a destination node or rack at the same time. PCOI architecture supports all-to-one communication pattern by using queues and passive optical components.

The general problems seen in the PCOI architecture are the use of large number of optical components and the signal degradation. There are two main types of optical signal degradations in PCOI: one is caused by passive optical components which is simply the reduction in optical power as the optical signal passes through a passive optical component and the second is the change in the shape of optical signal as it goes through SOA. Due to signal degradation, it is not possible to improve the BER beyond a certain limit.

The future work is related to the introduction of new modulation formats to the performance of PCOI and the analysis of signal degradation imposed by new modulation formats.

Data Availability

No data were used to support this study. The authors have conducted the simulations to evaluate the performance of proposed protocol. However, any query about the research conducted in this paper is highly appreciated and can be answered by the principal authors Rab Nawaz Jadoon (rabnawaz@cuiatd.edu.pk) and Mohsin Fayyaz (mohsinf@cuiatd.edu.pk) upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

Dr. Rab Nawaz Jadoon personally acknowledges COMSATS University, Islamabad-Abbottabad Campus, and School of Information Science and Technology, USTC, Hefei, China, which extended their full support by providing the authors all key resources during the implementation and all afterward phases of this project. This work was financially supported by National Natural Science Foundation of China (Grant no. 61631018).

References

[1] B.-C. Lin, "Generalization of an optical asa switch," *Applied Sciences*, vol. 9, no. 6, pp. 1096–6, 2019.

[2] M. Moralis-Pegios, N. Terzenidis, G. Mourgias-Alexandris, and K. Vyrsokinos, "Silicon photonics towards disaggregation of resources in data centers," *Applied Sciences*, vol. 8, no. 1, p. 83, 2018.

[3] P.-A. Blanche, L. La Comb, Y. Wang, and M. Wu, "Diffraction-based optical switching with mems," *Applied Sciences*, vol. 7, no. 4, p. 411, 2017.

[4] M. J. Karol, M. G. Hluchyg, and S. P. Morgan, "Input versus output queueing on a space-division packet switch," *IEEE Transactions on Communication*, vol. 35, no. 12, 1987.

[5] S. Ruiz-Moreno, G. Junyent, M. J. Soneira, and J. R. Usandizaga, "Statistical analysis of nonlinear optical amplifier in high saturation," *IEE Proceedings J Optoelectronics*, vol. 135, no. 1, pp. 34–38, 1988.

[6] X. Lu, T. Xu, T. Liu et al., "Reduction of intermodulation distortion in directly modulated lasers: rf predistortion, electronics," in *Proceedings of the International Conference on Communications and Control (ICECC)*, pp. 4437–4439, Ningbo, China, September 2011.

[7] A. V. Rylyakov, C. L. Schow, B. G. Lee, F. E. Doany, C. Baks, and J. A. Kash, "Transmitter predistortion for simultaneous improvements in bit rate, sensitivity, jitter, and power efficiency in 20 Gb/s cmos-driven vcsel links," *Journal of Loghtwave Technology*, vol. 30, no. 4, pp. 399–405, 2012.

[8] A. C. Carusone, "An equalizer adaptation algorithm to reduce jitter in binary receivers," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 53, no. 9, pp. 807–811, 2006.

[9] A. Singh, K. Ramachandran, L. Xu, and Y. Zhang, "Proteus: a topology malleable data center network," in *Proceedings of the 9th ACM SIGCOMM Workshop on Hot Topics in Networks*, Monterey, CA, USA, 2010.

[10] X. Ye, Y. Yin, S. J. B. Yoo, P. Mejia, R. Proietti, and V. Akella, "DOS: a scalable optical switch for datacenters," in *Proceedings of the ANCS'10 ACM/IEEE Symposium on Architectures for Networking and Communications Systems*, La Jolla, CA, USA, October 2010.

[11] Q. Kong, S. Huang, Y. Zhou et al., "MMTDnet: a novel optical circuit switch architecture for data center networks," in *Proceedings of the Asia Communications and Photonics Conference*, Beijing China, November 2013.

[12] P. Bakopoulos, K. Christodoulopoulos, and G. Landi, "NEPHELE: an end-to-end scalable and dynamically reconfigurable optical architecture for application-aware SDN cloud data centers," *IEEE Communication Magazine*, vol. 56, no. 2, pp. 178–188, 2018.

[13] K. Xia, Y.-H. Ka, M. Yang, and H. Jonathan Chao, "A petabit bufferless optical Switch for data center networks," *Optical Networks*, Springer, Berlin, Germany, pp. 135–154, 2013.

[14] J. Gripp, J. E. Simsarian, J. D. LeGrange, P. Bernasconi, and D. T. Neilson, "Photonic terabit routers: the IRIS project, optical fiber communication (OFC)," in *Proceedings of the IEEE Conference on Optical Fiber Communication (OFC)*, pp. 1–3, San Diego, CA, USA, March 2010.

[15] R. Luijten, W. E. Denzel, R. R. Grzybowski, and R. Hemenway, "Optical interconnection networks: the osmosis project," in *Proceedings of the 17th Annual Meeting of the IEEE Lasers and Electro-Optics Society*, Rio Grande, Puerto Rico, November 2004.

[16] R. Proietti, C. J. Nitta, Y. Yin, R. Yu, S. J. B. Yoo, and A. Venkatesh, "Scalable and distributed contention resolution in awgr based data center switches using rsoa based optical mutual exclusion," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 19, no. 2, 2013.

[17] S. Di Lucente, R. P. Centelles, H. J. S. Dorren, and N. Calabretta, "Study of the performance of an optical packet switch architecture with highly distributed control in data center environment," in *Proceedings of the IEEE 16th International Conference on Optical Network Design and Modeling (ONDM)*, pp. 1–6, Colchester, UK, April 2012.

[18] B. Neel, R. Morris, D. Ditomaso, and A. Kodi, "Sprint: scalable photonic switching fabric for high performance computing (HPC)," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 4, no. 9, 2012.

[19] A. K. Kodi and A. Louri, "Energy-efficient and bandwidth-reconfigurable photonic networks for high-performance computing (HPC) systems," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 17, no. 2, pp. 384–395, 2011.

[20] S. Basu, C. McArdle, and L. P. Barry, "Scalable ocs-based intra/inter data center network with optical tor switches," in *Proceedings of the 18th International Conference on Transparent Optical Networks (ICTON)*, Trento, Italy, July 2016.

[21] O. Liboiron, I. Cerutti, R. P. Giorgio, A. Nicola, and P. Castoldi, "Energy efficient design of scalable optical multi plane interconnection architecture," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 17, no. 2, pp. 377–383, 2011.

[22] P. N. Ji, D. Qian, K. Kanonakis, C. Kachris, and I. Tomkos, "Design and evaluation of a flexible bandwidth OFDM intra data center interconnect," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 19, Article ID 3700310, 2013.

[23] F. Nathan, F. Nathan, A. Forencich et al., "A 10 μs hybrid optical/circuit electrical packet network for datacenters," in *Proceedings of the Optical Fiber Communication Conference/ National Fiber Optic Engineers Conference*, Anaheim, CA, USA, March 2013.

[24] H. Ali, E. Taisir, H. El-Gorashi, and J. M. H. Elmirghani, "High performance awgr pons in data centre networks," in *Proceedings of the 17th International Conference on Transparent Optical Networks (ICTON)*, Budapest, Hungary, July 2015.

[25] A. Kushwaha, A. Gumaste, T. Das, S. Hote, and Y. Wen, "Flexible interconnection of scalable systems integrated using optical networks (fission) data-center-concepts and demonstration," *Journal of Optical Communications and Networking*, vol. 9, no. 7, pp. 585–600, 2017.

[26] S. Chen, S. Huang, Q. Kong et al., "A distributed and scalable optical packet switch architecture for data centers," in *Proceedings of the 13th International Conference on Optical Communications and Networks (ICOCN)*, Suzhou, China, November 2014.

[27] D. Gross, C. M. Harris, J. F. Shortle, and J. M. Thompson, *Fundamentals of Queueing Theory*, John Wiley & Sons, , John Wiley & Sons, Hoboken, NJ, USA, Third edition.

[28] I. Iliadis and W. E. Denzel, "Analysis of packet switches with input and output queuing," *IEEE Transactions on Communications*, vol. 41, no. 5, pp. 731–740, 1993.

[29] Network Simulator 2 (Wiki), 2018, http://nsnam.sourceforge.net/wiki/index.php/User_Information.

[30] OptiSystem, 2018, http://www.optiwave.com.