Hybrid OCS/OBS interconnect in intra-data-center network

Qian Kong (孔 谦)[1,2,*], Yafeng Zhan (詹亚锋)[1,2,], and Peng Wan (万 鹏)[1,2]**

[1]*Space Center, Tsinghua University, Beijing 100084, China*
[2]*Beijing National Research Center for Information Science and Technology (BNRist), Beijing 100084, China*
**Corresponding author: kongqian01@yeah.net; **corresponding author: zhanyf@tsinghua.edu.cn*
Received March 26, 2019; accepted May 17, 2019; posted online July 18, 2019

We report on a data center network (DCN) architecture based on hybrid optical circuit switching (OCS) and optical burst switching (OBS) interconnect for dynamic DCN connectivity provisioning. With the combination of the centralized and distributed control of the software-defined optical networks, the proposed interconnect can achieve unprecedented flexibility in dealing with both mice and elephant flow in the DCN. Numerical simulation is employed to investigate the performance of the proposed architecture. The results show that the OBS module has preferable performance in dealing with a larger burst packet, and the throughput is constrained by the capacity of the server random access memory module.

OCIS codes: 060.4510, 060.6719, 060.6718.
doi: 10.3788/COL201917.080605.

The rise of the cloud computing and emerging big-data applications has significantly increased the traffic within the data center (DC). The bandwidth bottleneck and growing power requirements have become central challenges for high performance DC network (DCN) interconnects. Various optical interconnects[1,2] and optical switching devices[3,4] have been proposed to take advantage of the high bandwidth capacity and low power consumption offered by optical switching. C-Through[5] and Helios[6] are two major representatives of hybrid optical/electrical switching networks. C-Through adopts a hybrid packet and circuit switched DCN architecture where top of rack (ToR) switches are connected to the Ethernet and an optical circuit-based network. Similarly, Helios brings optical microelectromechanical systems (MEMS) switches and wavelength division multiplexing (WDM) links into DCs and integrates them with existing DC infrastructures. Chen *et al.*[7] proposed the optical switching architecture (OSA) optical interconnect, which can dynamically change its topology and link capacities through the optical circuit switching (OCS) module, where each ToR can communicate simultaneously with any other ToR through the configuration of the optical switching module. Huang *et al.*[8] proposed a DCN combining optical packet switching (OPS) and agile OCS in a unified platform. Hybrid optoelectronic packet routers, which combine optical switching fabric and electronic buffers, are used, supporting 100 Gbit/s (25 Gbit/s × 4 wavelengths) optical packets. However, OPS is constrained by its system complexity and scalability, which is hard to satisfy the demand of the growing large number of ports in the DC. Shu *et al.*[9,10] presented a hybrid optical network design for future DCNs. Such a design integrates OCS and OPS schemes via hybrid ToR switches, which provide flexible function–switchover between different traffic patterns in the DCN. However, OPS switches do not have any buffering capability to reduce losses, and they therefore suffer much higher losses than their electronic packet switching (EPS) counterparts. Yan *et al.*[11] proposed an all-optical function programmable OCS-based DCN architecture, combining the benefits of both space division multiplexing (SDM) and time division multiplexing (TDM) technologies. Beam-steering large-port-count fiber switches are used as the centralized cluster switches and the inter-cluster switch. Nejabati *et al.*[12] described the envisioned node architectures and technologies for the all-optical DCN data plane and the unified software-defined networking (SDN)-enabled control plane to provide the required flexibility, manageability and customizability. Imran *et al.*[13] proposed a software-defined optical burst switching (OBS) DC interconnect, which employs a single-hop topology with a two-way reservation protocol that results in zero burst loss. However, burst and fast-changing inter-rack traffic is still handled by the core switches, while the relatively stationary traffic is handled via the optical burst rings.

In summary, how to cope with various traffic granularity[14] (both mice and elephant flow) efficiently becomes a more and more crucial issue in the DCN. If the data transmission duration is short relative to the set-up time, bandwidth may not be efficiently utilized in the OCS system. To cope with the various traffic granularity and enhance the bandwidth utilization, we consider taking advantage of the merit of both OCS and OBS. In particular, OCS-based DCNs can effectively handle large, stable elephant flows, and OBS-based DCNs can adapt to flexible bandwidth granularity requirements and reduce packet losses through buffering the traffic in the electronic random access memory (RAM) module. Therefore, hybrid OCS/OBS DCN architecture can complement each other to accommodate the dynamics and burstiness of DC traffic. In addition, an OBS system can achieve

interconnection of thousands of ports, flatten the DC topology, and avoid frequent power-hungry optical-electronic-optical (OEO) conversion. Due to the OBS system having lower service latency and device configuration time, it can achieve high speed and low latency data stream switching.

In this Letter, based on our previous work[15,16], we propose a hybrid OCS/OBS interconnect for a DCN. The services are differentiated and aggregated by the switch interface card (SIC) embedded in the server, thus completing the separation of OCS and OBS services. Software-defined optical network (SDON) control is implemented in the proposed interconnect. The performance of the interconnect in terms of throughput, delay, and packet loss rate is also investigated. To the best of our knowledge, such a hybrid OCS/OBS intra-DC interconnect has never been investigated before.

The hybrid OCS/OBS network design for the inter-cluster DCN is illustrated in Fig. 1. The server is directly connected to the network through the OCS switches. The servers connected to the same OCS switches are called intra-cluster servers, and servers connected to different OCS switches are called inter-cluster servers. Each of the intra-cluster OCS switches is connected by an inter-cluster OCS switch. OBS switches [which can be implemented by a waveguide array grating (AWG) or a semiconductor optical amplifier (SOA)] are connected in both intra- and inter-cluster OCS switches. Each optical component in the network is connected through a proxy and a central controller. In particular, the OBS switches internally implement distributed control and allow resource information to interact with the central controller.

The servers in the racks are directly connected to the intra-cluster optical circuit switches by replacing the traditional network interface card with an SIC [which can be implemented by a programmable logic device (PLD)], and each server is connected to the optical circuit switches through such an interface. Each of the SICs is connected to the central controller through an agent. The agent is responsible for reporting the service status information of the server, interacting with the traffic request information, and configuring the control command. OBS switches are connected to each of the OCS clusters, and the OBS can be shared by all intra-cluster servers. Each of the intra-cluster optical switches has a corresponding interface to an inter-cluster optical switch, and the inter-cluster optical switch is responsible for communication between the inter-cluster servers. In addition, the inter-cluster optical switch is also connected with an inter-cluster OBS switch, which is responsible for the switching of optical burst packets between the inter-cluster servers. All optical switching components are connected by a proxy and a central controller, which are in charge of providing resource information and configuration of the optical modules.

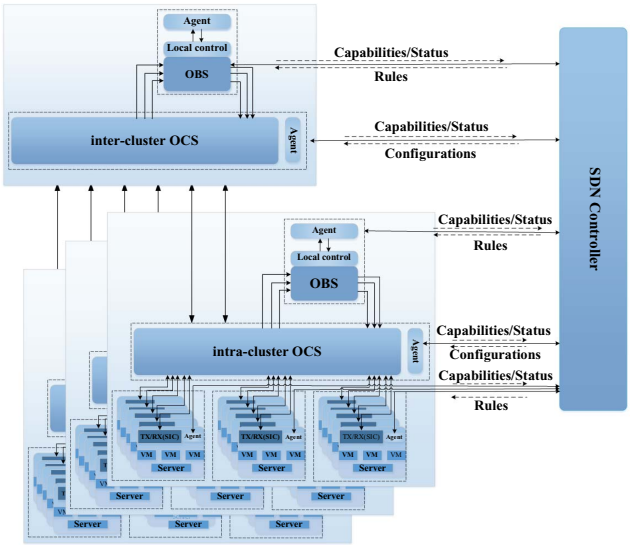

Fig. 1. Hybrid OCS and OBS intra-data-center interconnect.

In current DCNs, upper-layer applications would result in more elephant flows between servers, and such flows would be accompanied by services with tight latency and short duration. OCS is implemented to cope with elephant flow. Meanwhile, OBS is implemented to cope with delay-sensitive and mice flow. In order to provide unified OBS functionality, PLDs are used to package and aggregate delay-sensitive traffic.

The ToR switches are not employed in the proposed optical interconnect, which means the traffic does not have to aggregate on the electrical layer. Therefore, the SIC fully utilizes RAM in the server to buffer the traffic and implements routing based on the media access control (MAC) address. In addition to the function of the traditional network interface card, the SIC can send and receive mixed OCS/OBS traffic by reading and writing data on the server or sending and receiving data according to the communication protocol. Different switching scenarios (such as the OCS scenario, OBS scenario, hybrid OBS and OCS scenario) can be realized by programmable logic control devices [field programmable gate array (FPGA) and erasable PLD (EPLD)]. The designed SIC needs to include the following interfaces: server internal interface, control plane SDN proxy interface, and intra- and inter-cluster server communication interface.

The server interface directly reads the data in the server RAM module so that the SIC can duplicate and transfer data between the servers by directly interacting with the server RAM module. For control plane interface, the SDN proxy sends commands in Ethernet frames. Similarly, the interface card on this interface can also send status information to the SDN proxy. After receiving the commands sent by the SDN agent, the switching interface card updates its lookup table (LUT), while other modules implement the corresponding functions according to the resource information in the LUT.

The SIC designed two 10 Gbit/s links for hybrid OCS and OBS. Based on the LUT, data interaction between servers can select either the OCS or OBS mode.

The received traffic can be sent directly to the corresponding port without processing. In addition, there are two 10 Gbit/s links that serve as the internal communication interface of the OCS racks. This interface enables communication between servers in the same rack. When used as optical circuit switches, the received traffic can be directly sent to other OCS interfaces without returning to the server.

The control plane based on the SDONs is illustrated in Fig. 2. Additional functions are extended to meet the requirements of DC traffic and applications. The intelligent logic is contained in the SDN controller (the corresponding network application runs on it), while maintaining complete, dynamic DCN intra- and inter- racks, through a series of basic network control plane functions such as setting up end-to-end connection, traffic monitoring, and path computation. The essential part of this SDON control architecture is the resource management unit, which abstracts DC optical network resources that rely on carrier technology by providing specific functions. The highly flexible programmability can be used to overcome the lack of autonomy and flexibility of traditional automatically switched optical networks. Automated dynamic connection establishment, protection recovery, and rerouting mechanisms are available through a programmable mechanism in the management and control of DCs. A series of different functions are provided on the southbound interface, which include the collection of bandwidth resource information, dynamic configuration and connection establishment, and monitoring of various granularity routing requirements. For the northbound interface, a series of open application interfaces (APIs) can be run through the management plane of the DC, such as virtualization and network resource optimization. On the other hand, this also allows the control plane to have a more open interface and an external management system (such as a cloud-based OpenStack system).

The SDON control architecture can increase the flexibility of DCN control and provides a versatile modular framework to support future new technologies and network devices. In addition, the northbound open API and the southbound standard interface can support DC operators and even users to customize the features based on their specific needs.

Hybrid OCS and OBS can be controlled and managed by an SDON controller [such as the OpenDaylight[17] and open network operating system (ONOS)[18]], and corresponding protocol extensions need to be made to support the performance of hybrid OCS and OBS. In addition, the southbound interface also provides an SIC driver, which is responsible for the interaction of the control and configuration information and the reporting of network status information. Meanwhile, the network configuration (NETCONF) driver, path computation protocol (PCEP) driver, openflow protocol (OF) driver, and simple network management protocol (SNMP) driver are also provided to support various communication protocols.

In an OBS network, various types of client data are aggregated at the source node and transmitted as data bursts that later will be disassembled at the destination node. During burst assembly/disassembly, the client data is buffered at the source node, where electronic RAM is cheap and abundant. While the OCS network is connection oriented, which means a path must be set up from the source node to the destination node before the data arrives, this paragraph describes how the combination of the centralized and distributed control works and how the traffic process flows in detail. As shown in Fig. 3, when a new connection request arrives, the network controller is cognitive of the requested bandwidth and latency. After evaluation, the corresponding modules establish a demanding optical link according to the requested bandwidth and latency. The SIC can choose to connect

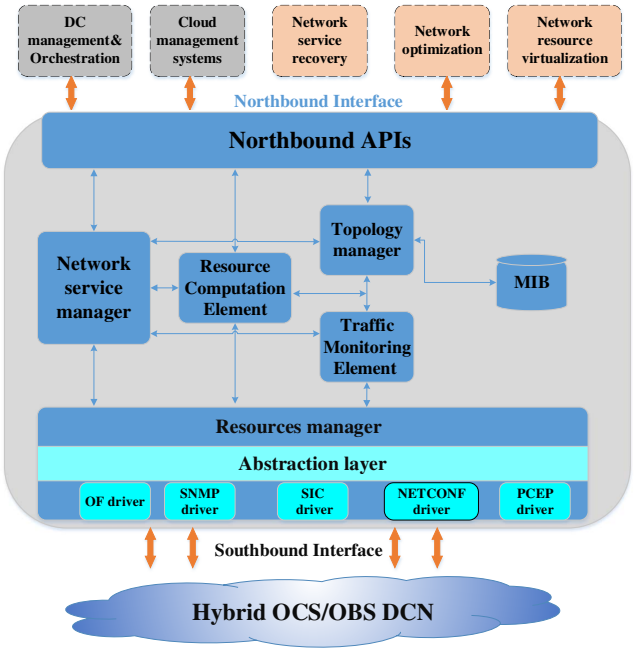

Fig. 2. SDON-enabled control scheme.

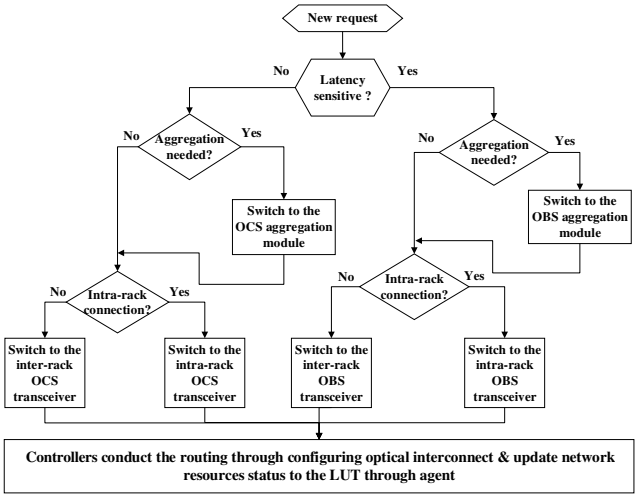

Fig. 3. Flow chart of traffic processing in the SIC.

the request service to the OCS port or the OBS port to establish the request connection from the Ethernet to the optical path. Meanwhile, SIC also supports the aggregation function of the Ethernet service, which includes the traffic to the same destination being packed together, different latency of the incoming traffic being divided into mice and elephant flow, respectively, and implementing OCS or OBS functionality according to the traffic granularity.

For delay-sensitive services, SIC supports Ethernet frame transmission, where the Ethernet frame is configured to the OBS module through the OCS module, while OCS switches provide a channel for OBS switches through centralized control, and client data performs the switching through distributed control in the OBS module and then switches back through the OCS switches to the destination server. In this process, centralized control and distributed control should cooperate with each other to ensure the data moves from the source server to the destination server.

The small form-factor pluggable transceivers (SFP+) module is implemented to perform electrical–optical (E-O) conversion in order to modulate the data from the Ethernet frame to the optical domain. Traffic can also be aggregated to the same destination port to form a burst flow. Meanwhile, SICs allocate delay-insensitive services to the OCS output module and can package the traffic to the same output port or directly switch to the OCS transmitting port without convergence processing.

The SIC can read data from the server RAM module, complete the aggregation function, and then classify the traffic to process the arrived Ethernet packet. The service is divided into a fixed time slice with flexible frame numbers or fixed frame size with flexible time slices. The requirements of different routing modules (OCS or OBS) are supported by isolating traffic into different buffers based on the source MAC address and the destination MAC address. The control and management modules in charge of establishing a link connection between the server and the SIC and updating the LUT in the SIC after the optical circuit switches and the optical burst switches are configured. The LUT is responsible for updating resource information and configuration status in the switches.

We investigated the proposed hybrid OCS/OBS interconnect through numerical simulation, and employed OPNET simulation software to evaluate the performance of the entire optical interconnect. The generated traffic arrival rate obeys the Poisson distribution and has a random destination address. Burst sources are individually simulated with the on–off model. In order to capture the burstiness of data at the source nodes, the traffic from a user to the destination server is generated by Poisson packet arrivals from 50 to 100 independent traffic sources that are either transmitting packets at a mean rate during the on period or idle during the off period. The traffic load is 100 Erlang at each source node. The switching time of OCS is typically 50 ms, and the OCS switches are configured before the traffic arrives. The switching time of OBS depends on the burst size of the traffic, ranging from 1 to 10 μs. We investigated the scalability of the proposed interconnect through enlarging the input/output ports of optical burst switches. Assuming the ports of optical circuit switches are large enough to support the connection of the corresponding servers and the OBS units, we change the scale optical burst switches. The input/output ports adopt 4×4, 8×8, 16×16, and 32×32, respectively, the number of connected servers increases linearly, and the number of servers in the cluster is 10, 20, 40, and 60, respectively. To the OCS unit in the cluster, each of the cluster and the inter-cluster OCS switch is connected to the OBS switch of the port size, as described above. In this simulation, there are four intra-cluster optical circuit switches and one inter-cluster optical circuit switch. The total number of servers is 40, 80, 160, and 320. The principle of control and routing is described; we assume that the traffic that needs to perform OCS is configured by the central controller before the traffic enters the network. That means that client data is connection oriented, and switching is performed only on the OBS unit. Therefore, the service latency discussed here does not include the configuration time of the optical circuit switches. Since the OCS performance has been analyzed in our previous work[12,13], this section will focus on the impact of the proposed interconnect on OBS.

Distributed control is implemented in the OBS module, and the processing unit inside the switching module completes switching according to the MAC address. For the burst switching process, we use the improved just-enough-time (JET) protocol as the transmission protocol. The traffic switching is constrained in the optical domain, and its control command is implemented through the burst electrical packet processed in the electrical domain. Control packets are sent before burst data packets, so traffic has an offset delay latency when transmitting at the source. The traffic needs half of the total service delay to reach the output port, and the destination node needs to return an acknowledgement message to inform the source node that the packet has been received. If the source end does not receive the acknowledgement message, the packet will be automatically retransmitted. The traffic within the transmission latency range will be automatically stored in the server RAM module. If the transmission waiting latency exceeds the transmission delay, the SIC automatically enables packet dropout processing. In our simulation, the average burst size is set to 100 kbytes, and the average packet size is fixed to 256 and 1500 bytes, respectively, in the Poisson case. Comparing Figs. 4 and 5, we can conclude that the throughput of the network is higher when the 1500 bytes burst packets are implemented compared with 256 bytes burst packets, which illustrate that the OBS module has better switching performance for traffic with larger packet sizes. The reason for this is that the short burst packets incline to cause congestion of the control channel, thus blocking the control information, where

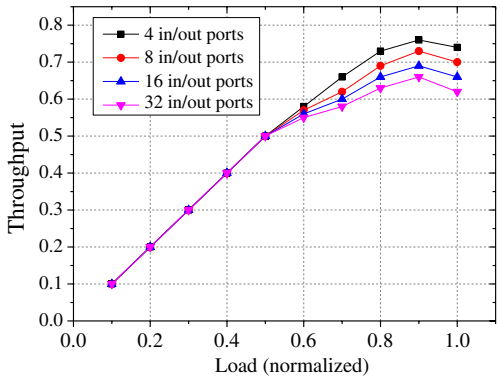

Fig. 4. Throughput of the system as a function of payload when the average packet size is 256 bytes.

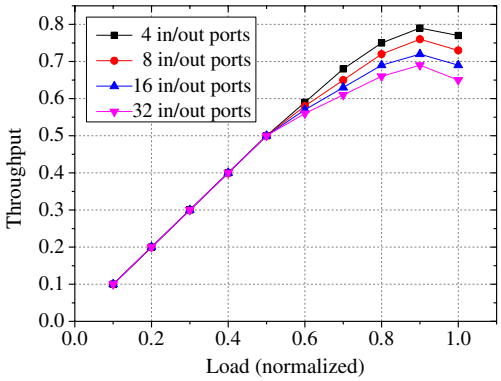

Fig. 5. Throughput of the system as a function of payload when the average packet size is 1500 bytes.

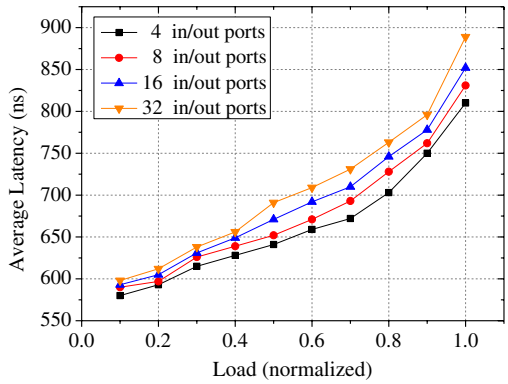

Fig. 6. Performance analysis of the system in terms of average latency.

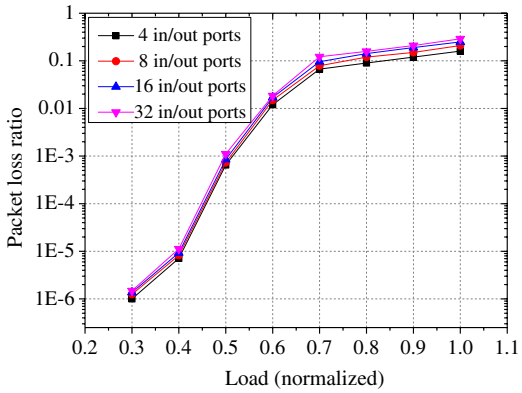

Fig. 7. Performance analysis of the system in terms of packet drop ratio.

the congestion of the control information results in a decrease in network throughput.

As can be seen from Figs. 4 and 5, the throughput changes linearly before the traffic load (normalized) reaches 0.5. After the load reaches 0.5, the throughput growth becomes slow. When the load reaches 0.9, the ratio of received total traffic shows a downward trend. This is mostly caused by the overflow of the server RAM module when the throughput reaches 0.9. As illustrated in Figs. 4 and 5, when the OBS input/output ports increase, the relative throughput tends to decrease. This is due to the distributed control scheme. As the scale of the switching module enlarges, the traffic in the network increases. The distributed control is employed to process the control command, and the frequent configuration of the switches increases the resource occupation rate, which improves the blocking rate of the traffic and reduces the throughput.

In the following numerical results, the average burst size is set to 100 kbytes, and the average packet size is fixed to 256 bytes. The performance of the delay and packet loss rate of the proposed interconnect is investigated. As illustrated in Fig. 6, the delay of the system increases with the expansion of the burst switching input/output ports. As the traffic load increases, the delay increases more sharply. The reason is that higher traffic load leads to more burst traffic in the system, thus increasing the probability of the traffic competing for the same port. The greater probability of port competition leads to an increase in the traffic blocking rate, which results in more packets retransmitted or lost and an increase in service delay. On the other hand, when the load of the service exceeds 0.8, the overflow of server RAM module is unavoidable.

In this case, many packets will count to the network packet loss, and other packets will be retransmitted depending on the demand. This process greatly increases the service delay. As can be seen from Fig. 7, after the traffic load reaches 0.6, the packet loss rate increases significantly. This is due to the significant usage of the RAM module as the traffic load increases. When the traffic load reaches 0.7, the server RAM module overflows, and the packet loss rate increases significantly.

In conclusion, we propose a hybrid OCS and OBS interconnect in intra-DCNs. An SIC is introduced in the server, which enables it to directly communicate between servers without the ToR switches and fully utilize the internal RAM of the server to buffer the traffic. After the process of the SIC, traffic is packaged and aggregated to OCS or OBS modules according to the granularity. A combination of the centralized and distributed control scheme is employed in the control plane. The central controller is based on the SDON, which is in charge of the configuration of

the whole OCS network. Distributed control is implemented in the OBS network. The resource information in the OBS module is uploaded to the central controller in real time so that the whole network resource is transparent to the central controller. We also investigated the performance of the OBS module in the proposed interconnect in terms of throughput, delay, and packet loss rate with traffic load and evaluated the influence of different port sizes. The results show that when the payload is not more than 0.6 (normalized to one), the proposed interconnect indicates preferable performance.

This work was supported by the National Natural Science Foundation of China (No. 61671263) and the Tsinghua University Independent Scientific Research Project (No. 20194180037).

References

1. F. Testa and L. Pavesi, *Optical Switching in Next Generation Data Centers* (Springer International Publishing, 2018).
2. C. Kachris and I. Tomkos, IEEE Commun. Surv. Tut. **14**, 1021 (2012).
3. X. Chen, J. Gao, and B. Kang, Chin. Opt. Lett. **16**, 081202 (2018).
4. X. Guo, X. Li, and R. Huang, Chin. Opt. Lett. **15**, 110604 (2017).
5. G. Wang, D. G. Andersen, M. Kaminsky, K. Papagiannaki, T. Eugene Ng, M. Kozuch, and M. Ryan, in *Proceedings of the ACM SIGCOMM Conference on SIGCOMM* (ACM, 2010), p. 327.
6. N. Farrington, G. Porter, S. Radhakrishnan, H. Bazzaz, V. Subramanya, Y. Fainman, G. Papen, and A. Vahdat, in *Proceedings of the ACM SIGCOMM 2010* (ACM, 2010), p. 339.
7. K. Chen, A. Singla, A. Singh, K. Ramachandran, L. Xu, Y. Zhang, X. Wen, and Y. Chen, IEEE Trans. Network. **22**, 498 (2014).
8. Y. Huang, Y. Yoshida, K. Kitayama, S. Ibrahim, R. Takahashi, and A. Hiramatsu, J. Opt. Commun. Netw, **7**, 1109 (2015).
9. Y. Shu, S. Yan, C. Jackson, K. Kondepu, E. H. Salas, Y. Yan, R. Nejabati, and D. Simeonidou, Opt. Fiber. Technol. **44**, 102 (2018).
10. Y. Yan, G. Zervas, B. R. Rofoee, and D. Simeonidou, in *Optical Fiber Communications Conference & Exhibition* (2014), p. 1.
11. S. Yan, E. Hugues-Salas, V. Rancano, Y. Shu, G. Saridis, B. Rofoee, Y. Yan, A. Peters, S. Jain, T. Smith, P. Petropoulos, D. Richardson, G. Zervas, and D. Simeonidou, J. Lightw. Technol. **33**, 1586 (2015).
12. J. Perelló, S. Spadaro, S. Ricciardi, D. Careglio, S. Peng, R. Nejabati, G. Zervas, D. Simeonidou, A. Predieri, M. Biancani, J. S. Dorren, S. Lucente, J. Luo, N. Calabretta, G. Bernini, N. Ciulli, J. Sancho, S. Iordache, M. Farreras, Y. Becerra, C. Liou, I. Hussain, L. Liu, and R. Proietti, IEEE Network **27**, 14 (2013).
13. M. Imran, M. Collier, P. Landais, and K. Katrinis, J. Opt. Commun. Netw. **8**, 610 (2016).
14. S. Kandula, S. Sengupta, A. Greenberg, P. Patel, and R. Chaiken, in *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference* (ACM, 2009), p. 202.
15. Q. Kong, S. Huang, B. Guo, X. Li, M. Zhang, Y. Zhao, J. Zhang, and W. Gu, Opt. Eng. **55**, 076111 (2016).
16. Q. Kong, Y. Zhan, and C. Duan, IET Commun. **12**, 2623 (2018).
17. OpenDaylight project, https://www.opendaylight.org/.
18. ONOS project, https://onosproject.org/.