ARTICLE IN PRESS

Optical Fiber Technology xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Optical Fiber Technology

journal homepage: www.elsevier.com/locate/yofte

Regular Articles

Programmable OPS/OCS hybrid data centre network

Yi Shu[a,b], Shuangyi Yan[a,*], Chris Jackon[a], Koteswararao Kondepu[a], Emilio H. Salas[a], Yan Yan[a], Reza Nejabati[a], Dimitra Simeonidou[a]

[a] High Performance Networks Group, University of Bristol, UK
[b] Oxford Instruments Plasma Technology, UK

ARTICLE INFO

Keywords:
Data center networks
Data center traffic modelling
Hybrid OPS/OCS network architecture

ABSTRACT

On the basis of profound understanding of data center (DC) traffic demands and optical switching technologies, we present a hybrid optical network design for future data center network (DCN). Such design integrates optical circuit switching (OCS) and optical packet switching (OPS) schemes via hybrid Top-of-the-Rack (ToR) switches which provide flexible function-switchover between different traffic patterns in the DCN. Simulations of network behaviors under such DC traffic loads indicate that the proposed OPS/OCS DCN can effectively improve the network performance. Moreover, via a preliminary analysis of OCS and OPS network configurations, the construction of hybrid DCN is also proved as the most cost- and energy-efficient way for DCN upgrading while offering a promised quality of service. An experimental demonstration shows a complete implementation of data center virtualization in the proposed hybrid data center network.

1. Introduction

Recent trends show network applications move from private clouds to public cloud data centers. As shown in Cisco global cloud index 2016 [1], annual global cloud IP traffic will reach 14.1 ZB by the end of 2020, up from 3.9 ZB in 2015. Around 68 percent of the cloud workloads will be in public cloud data centers by 2020. Data, logic and application are migrating to the Cloud. The increasing needs for data center and cloud resources will further drive the development of large-scale public cloud data centers, i.e., hyperscale data centers. The hyperscale data center, usually operated by large Internet-driven companies, such as Google, Facebook, could reduce capital expenditure (CapEx) with a sophisticated operation and maintenance team. The large volume purchase of the key facilities, including switches, transceiver modules, gives the operators of the hyperscale data centers a big negotiation power in the market to further push down the cost of the CapEx. It's expected more and more hyperscale data centers will be built.

The traditional multi-tier DCN architectures encounter great challenges to support the ever-expanding large-scale co-located DCs [2]. Firstly, the over-subscription ratio may exceed 20:1 in core switches and 4:1 in aggregated switches [3], which will be the bottleneck for the dominant traffic inside DCNs, i.e., the east-west traffic. According to the Cisco report, the overall east-west traffic represents around 86% of the total data center by 2020 [1]. Secondly, the latency in multi-tier DCNs will increase dramatically in a large-scale DCN, as the queueing time and processing time of traffic in each hop will get even longer when DCNs scale up. The huge latency will make the DCN unable to provide latency-sensitive services, especially for the emerging 5G applications. In addition, the latency also affects the current users' satisfaction. The third challenge of the multi-tier DCN architecture comes from the increased power consumption and total cost of the high-radix electrical switches. Thus, the multi-tier DCN architecture becomes less than ideal when it comes to supporting today's low-latency, virtualized applications. In current deployment, a swift and dramatic shift to "leaf-and-spine" architecture is happening [4]. The leaf-and-spine architecture promises a better support for east-west traffic. However, the switch fabric requires a large number of fiber connections and high radix electrical switches. Currently, it is very challenging to build a single electrical switching chip with a high radix and high per port bandwidth, due to the limitations on bandwidth at the edge of chips and power constraints. Thus many low radix switching chips are connected in a Clos topology to build a high radix chassis switch [5], which require large power consumptions with low port densities. The ITRS (International Technology Roadmap for Semiconductors) predicts only modest growth in per-pin bandwidth and pin count over the next decade [6]. The leaf-and-spine DCN architecture treats all the leaf switches equally and couldn't handle traffic flow locality, such as hot Top-of-Rack (ToR) switches or servers, in an efficient way. It's very challenging to build hyperscale DCNs with leaf-and-spine architecture.

Currently, optical signaling is mainly used for point-to-point

* Corresponding author.
E-mail address: shuangyi.yan@bristol.ac.uk (S. Yan).

https://doi.org/10.1016/j.yofte.2018.01.017
Received 10 July 2017; Received in revised form 11 January 2018; Accepted 19 January 2018

interconnections in DCs. 400 Gigabit Ethernet (400G, 400GbE) and 200 Gigabit Ethernet (200G, 200GbE) standards has been officially released by the IEEE P802.3bs Task Force in December 2017 [7]. The optical fiber-based transmission technologies could provide huge bandwidths for future hyperscale DCs.

Regarding switching technologies in DCNs, optical switching technologies, such as optical circuit switching (OCS), optical packet switching (OPS), and optical time division multiplexing (TDM), could potentially provide low-cost and power-efficient optical switching for intra-DCN communications. Comparing with electrical switching technologies, optical switching is less flexible because of the unavailability of optical random-access memories (RAM). However, optical switching technologies show some advantages which are very attractive to DCN applications. Firstly, optical switching technologies could provide high-radix network switches with silicon and nanophotonic technologies. The high-radix optical switch could reduce hop and switch counts and help to reduce latency and power consumption in DCNs [5]. Secondly, optical switching technologies are transparent to optical signal formats, which provide a better compatibility with different transmission standards. Thirdly, the advances in fiber-optic technology, such as wavelength division multiplexing (WDM) and space division multiplexing (SDM), could increase fiber-link capacities with multiple parallel links. For future hyperscale data center, optical switching and transmission technologies will play a significant role to support the ever-growing traffic demands.

In order to take full advantages of optical switching technologies in data center networks, new data center architectures are required for future hyperscale DCNs. As we mentioned before, the unavailability of optical RAM makes an all-optical DCN impossible. The traditional electrical switching will be a good complementary to optical switching technologies for DCN interconnections. The ToR switch need to be redesigned to support optical switching technologies in DCNs. In the past years, there have been lots of explorations to use optical switching solutions in DCNs [8–14]. However, these proposed optical DCN solutions focused on the leveraging of a single optical technology, e.g. OCS or OPS. The lack of flexibility of these proposed network architectures lead to relatively poor performance for dynamic traffics in DCNs [15–18]. The tremendous variation and diversity in the communication matrix over space and time in DCNs require a more flexible and dynamic DCN architecture. In addition, another challenge for future hyperscale DCNs comes from data center virtualization. In future, Cloud providers require the capability to dynamically allocate cloud resources to multiple virtual DCNs. DCN embedding is one of the mandatory features for future hyperscale DCNs. However, DCN virtualization becomes even more challenging in optical switching involved DCNs [19].

In this paper, we summarize our recent research in programmable OCS/OPS hybrid DCN architecture. The proposed hybrid data center architecture could offer several features: 1) dynamic deployment of DCN network functions for unicast and broadcast traffic; 2) OPS/OCS switch-over enabled by FPGA-based NICs/ToRs support variable link bandwidth with a fine granularity; 3) topologies adaption in packet-switching based sub-network. The function programmable feature is offered based on architecture-on-demand (AoD) concept [20]. The network topologies and network functions can be reconfigured according to network traffic estimation or prediction. The key enabling technologies include FPGA-based OCS/OPS reconfigurable network interface card (NIC) [21], synchronized TDM switching [22], parallel interconnections based on space division multiplexing (SDM) or WDM [23]. A complete software stack is developed to enable virtual DCs in the test platform [24].

This paper is organized as follows. In Section II, we review the applications of optical switching technologies in recently proposed optical DCN solutions and outline the benefits of hybrid DCN architecture. The proposed architecture of the programmable optical circuit/packet switching hybrid DCN is presented in detail in Section III. In Section IV, simulation works about the proposed programmable DCN architecture are presented. In Section V, a recent experimental demonstration of data center virtualization is reported. Conclusions of our work are given as Section IV.

2. Review of optical switching technologies in DCN

Optical switching technologies promise a better solution for future hyperscale DCNs. The available optical switching technologies could be categorized to circuit- or packet- switching. Here we will review the benefits and limitations of OCS and OPS schemes by summarizing their applications in recently proposed Optical DCN solutions.

For most OCS-based DCN solutions, Micro-Electro-Mechanical Systems (MEMS) or beam-steering based lager-port-count fiber switches (LPFS) are utilized as central switches which connect all the ToRs to generate a flattened network infrastructure [8–10]. The scalability of OCS-based DCN architecture is promised by the high radix LPFS which is up to thousands of ports [25]. Optical links can be set up between ToR pairs directly through the fiber switches and variable capacity can be assigned to each link by leveraging wavelength division multiplexing (WDM) technologies (e.g. WDM transceivers, spectrum selective switches (SSS), etc.). High capacity up to terabits/s for each link is feasible with low-speed electronics on the ToR by assembling multiple optical channels. High-capacity smooth data flow between different racks can be accommodated with low latency once the optical circuit link is set up. The degree of connectivity provided by a ToR can be enhanced by dense fiber connections or advanced space division multiplexing (SDM) fiber technology [26]. Thus, the long-lived bulk data transfers with bounded degrees (e.g. data migrations, backups, inter-processor communications) can be accommodated by OCS network.

However, to accommodate short-lived traffic patterns with high communication radix, OCS-based DCN needs to either reconfigure the DCN topology frequently [8,9] or send traffic through multi-hop indirect connections to remote servers [10]. The former solution suffers from the long reconfiguration time of fiber switches, which is in the order of a few milliseconds. The latter solution requires Optical-Electrical-Optical (O-E-O) conversions on each hop which introduces channel congestions at intermediate switches as well as significant latency on the multi-hop path.

On the other hand, OPS-based DCN solutions provide packet-level switching which fits better to dynamic traffic patterns. Two major approaches have been reported for OPS-based DCN realization: Arrayed-Waveguide Grating Routers (AWGR) based scheme [11,12,14] and Semiconductor Optical Amplifier (SOA) based scheme [13,27]. The former approach allocates different connections in DCNs with different wavelengths by connecting all ToRs to AWGR switches. Tunable Wavelength Converters (TWCs) or Fast Tunable Lasers (FTLs) are deployed at each ToR for addressing a specific destination port of AWGR by selecting the appropriate wavelength. The latter approach relies on SOA-based fast switches which can reconfigure the DCN in nanoseconds [28]. This scheme provides higher flexibility, as each connection is not limited by the channel grids of AWGR. The capacity it can support is thereby adaptive by aggregating different wavelengths to a single connection, which is so called “waveband switching” in the OPS scheme.

The challenge of OPS scheme is the system complexity and scalability. Due to the lack of optical RAM, blocked packets in congestion are generally stored in electronic buffers [11,14] or optical delay fiber array [12] for retransmission. Such buffering solutions together with the utilization of TWCs, FTLs or SOAs require a large number of wire connections from the controller to buffer and switch components, which increases the complexity along with the growing of DC size. For higher scalability and resiliency, multi-stage topologies have to be exploited for OPS-based DCN architecture [29]. The end-to-end latency for each packet is thus mainly caused by the congestion and buffering at each stage. In addition, a 5%–20% overhead is required for OPS transmission, including inter-slot guard time, time for synchronization,

Table 1
Comparison of different optical network technologies for DCN solutions.

Switching Paradigm	Link Capacity	DCN Scalability	Setup Latency	Transmission Latency	Processing Overhead	Complexity	Cost	Power Consumption	Application in DCN Solutions
OCS	Adaptive	High	High	Low	Low	Low	Low	Low	Helios [8], c-Through [9], OSA [10]
OPS	Adaptive	High	Low	Depending on congestion and buffering	High	High	High	High	Petabit [11], IRIS [12], Bidirectional SOA [13], LIONS [14], OSMOSIS [27]

time for clock recovery in burst mode receivers [30], etc.

The comparison between OCS- and OPS-based DCN is summarized in Table 1. OCS-based DCN can provide direct connections with large capacity and the system can be scaled up without compromising performance or adding complexity. By contrast, OPS-based DCN works better with fast-changing and unpredictable traffics but requires expensive and power-hungry components as well as complicated control management. Therefore, hybrid OPS/OCS DCN architecture becomes an attractive solution as OCS and OPS can complement each other: OPS system offers flexible connectivity for dynamic traffics and OCS system efficiently handles long-lived bulk data transfer.

3. Proposed programmable optical/electrical data center networking

The proposed programmable optical/electrical data center network architecture is shown in Fig. 1 with several key technologies, including FPGA-based programmable switch and interface card (SIC), OPS/OCS hybrid ToR switch, programmable OCS and OPS network configuration. The main design consideration is to divide the hyperscale DCN to several clusters. Each cluster consists of tens/hundreds of racks. All the clusters are connected together through a LPFS-based inter-cluster switch. Multiple SMFs or MCFs can be used to connect all clusters to the inter-cluster switch. The inter-cluster switch configures the connection matrix between all clusters and provides adaptable link capacity between different clusters. Thus, a single hop OCS is used to serve the long-lived, large capacity data flows for inter-cluster communications.

In each cluster, a centralized LPFS, as the cluster switch, interconnects ToRs via fiber bundles or SDM links. Inside clusters, OCS and OPS are used for different traffic patterns. The OCS, implemented with a LPFS, requires a relatively large setup time, however, no extra latency for the communications. Thus, elephant flows will transmit through OCS connections. The OPS network is achieved with OPS/OCS hybrid ToRs and OPS switches that connected to the cluster LPFS. The OPS switch will be regarded as sub-functions to be deployed in the DCN. The topology of OPS can be configured through the OCS network. The OPS network will carry most of the mice flow, due to its fast setup time. In addition, OPS could offer more connections in the same time, as a complementary to the OCS network. Other optical functional elements, such as PLZT-based TDM/OPS switches, EDFAs, couplers and combiners are also connected to the cluster switch for network function programmability. According to the traffic requests, variable network functions could be deployed by configuring the cluster switch to enable network function programmability.

The key enabling technologies are introduced as follows:

3.1. FPGA-based programmable switch and interface card

The programmable switch and interface card (SIC), which is designed to replace the traditional network interface card (NIC), can be plugged into the server directly and enable intense intra-rack blade-to-blade communication [31]. Compared to traditional NIC, the SIC provides switching function to the server, which enables server-centric data architecture (e.g., BCube [32]) and also simplify the implementation of the ToR switch. With more concerning of data security [33], the SIC design attracts more interests from industries, as the SIC makes data encryption more easier in DCNs. The SIC also supports flexible OCS/OPS function switchover for each optical channel. Servers in the same rack send/receive Ethernet frames through the SIC on each server to/from the intra-rack access on the FPGA board.

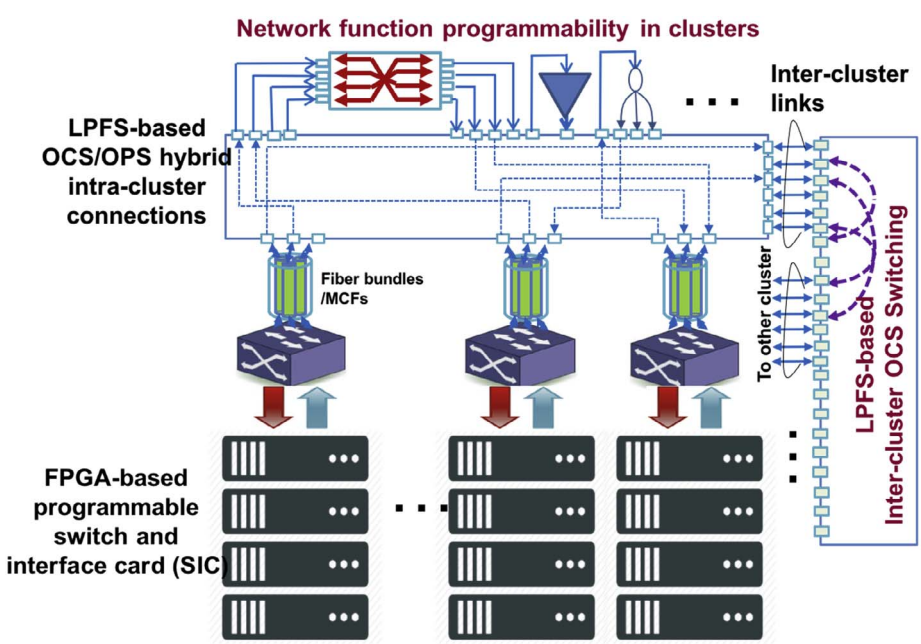

Fig. 1. Architecture of programmable optical/electrical DCN.

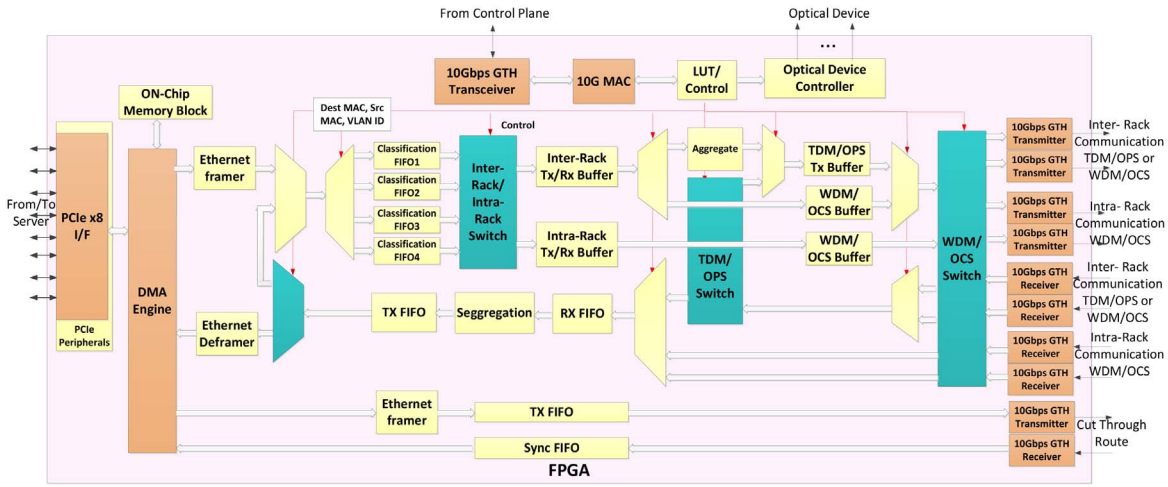

Fig. 2. FPGA-based Optical Programmable SIC design and implementation functional blocks architecture.

The SIC design and implementation architecture are shown in Fig. 2. The SIC is capable of copying the data between the memories of the blades and SIC, processing and sending out the data in particular port in TDM/OPS or WDM/OCS, based on the instruction of control plane. The SIC also acts as an OCS/WDM switch, an OPS/TDM switch, or a Layer 2 switch according to the commands of the control plane. With the switch functions, the SIC can work as a hop to supply maximum flexibility and programmability in the DCN architecture.

The SIC supports both intra-rack blade-to-blade communication and blade to optical ToR switch communication with the view to achieve high performance intra-rack evolving to inter-rack communication. The TDM-based SIC support link virtualization, which enable network virtualization in data center networks [24]. With this programmable SIC, an all-optical programmable disaggregated data center network was proposed and demonstrated successfully [21]. With a designed scheduling algorithms for disaggregated computing, the data center architecture could satisfy the high-capacity and low latency requirements [34].

3.2. OPS/OCS hybrid ToR switch

Another application of the developed OPS/OCS programmable SIC is the ToR switch [35]. The novel programmable hybrid ToR switch enables flexible OCS/OPS function switchover for each optical channel. As shown in Fig. 3, the FPGA platform performs traffic processing and traffic aggregation. For inter-rack traffic loads, the FPGA platform differentiates them to either OCS or OPS traffic with application-aware classification following the commands from the control plane. Then the FPGA platform loads/unloads variable traffic onto different wavelength channels for DCN interconnections. Extra traffic monitoring can be introduced to classify the traffic in real time [36]. Optical transceivers on the TORs can be 10Gbps SFP+, 40Gbps QSFP+ or even transceivers enabling advanced modulation formats, depending on the hardware supported on the FPGA platform. An m × n SSS are utilized as the interfaces between transceivers and the hybrid DCN. Circulators are adopted to connect transmitter (Tx) and receiver (Rx) to the SSS and enable bidirectional communication through it. Different optical channels can be aggregated by the SSS so that multi-granularity capacity for each link from the ToR is enabled. A proportion of such links are connected to OPS system according to the link requirement from each ToR to a specific OPS network configuration in the hybrid DCN. The rest links are connected to the OCS system. The maximum bandwidth a ToR can support is given by the total capacity provided on the FPGA platform, which is evolving rapidly; the node degree (link number) of each ToR is decided by the radix of SSS.

Thus, traffic switching is enabled at the ToR level: hybrid OPS/OCS switchover functions can be implemented hitless; adaptive capacity can be assigned to different links which enables flexible capacity assignment for different services in a DC and the isolation between them.

3.3. OCS network configurations

The OCS network is constructed based on programmable optical networks with the AoD concept [20]. As shown in Fig. 4, both the inter-cluster and intra-cluster switches are implemented with a LPFS (e.g., Polatis beam-steering fiber switches) based on the AoD programmable switch. Regarding inter-cluster communication, the AoD-based inter-cluster switch provide connections between different clusters based on OCS. The link capacity can be dynamically programmabled by offering variable numbers of connection links.

For intra-cluster communication, the AoD-based cluster switches provide OCS for intra-cluster communications. In addition, the cluster

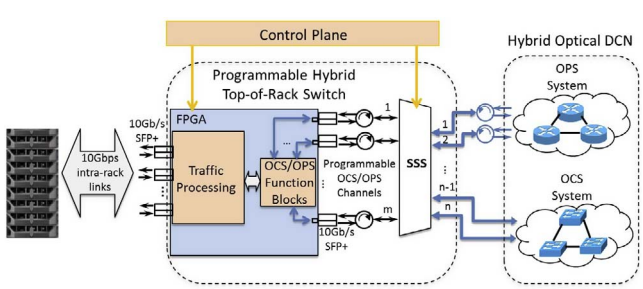

Fig. 3. Schematic of the programmable hybrid ToR switch: a high-speed FPGA platform provides the processing of both intra- and inter-rack traffic; inter-rack traffic is sorted into OCS/OPS traffic and sent to the DCN via m × n SSS interfaces.

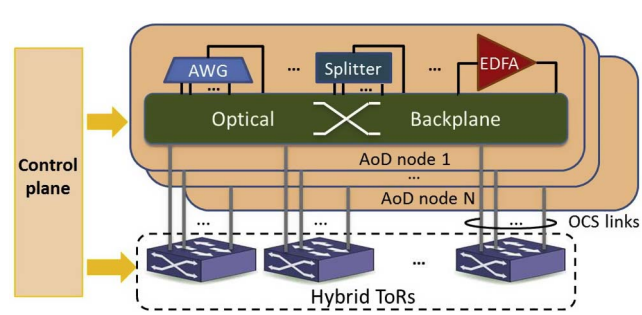

Fig. 4. Schematic of the OCS network configuration: an AoD-based optical programmable system where hybrid ToRs and a variety of optical function modules are connected via several large port-count optical backplanes.

switches also connect various optical modules (e.g. AWGs, splitters, EDFAs and etc.) to achieve network function programmability. Depending on the number of OCS-enabled links from each ToR, several AoD nodes are utilized in parallel to construct the OCS network and each of them connects one OCS link from each ToR. To make full use of the ports on the LPFS, each OCS link is set to work bidirectional, which also saves the utilization of circulators between the ToRs and the backplane.

With such configuration, arbitrary network topologies and functionalities can be delivered on demand by setting appropriate cross-connections between hybrid ToRs and optical modules in the optical backplane: required OCS links can be constructed directly between relevant ToR pairs; optical channels aggregated in the same OCS link can be separated or reassembled through AWGs; OCS broadcasting can be accomplished by utilizing splitters. And all the connections can be dynamically reconfigured according to the traffic pattern variation.

3.4. OPS network configurations

As mentioned before, OPS switching technology is only used for intra-cluster communications. All the OPS modules are connected to the cluster switch in an AoD approach. The interconnection between the OPS modules can be configured to form different network topologies. Given the challenge of practical applications with high-radix OPS modules, multi-stage topologies are exploited for OPS-based intra-cluster communications. For the convenience of study, we assume that there are 75 racks (ToRs) in a cluster and the size of OPS module is no bigger than 16 × 16. Fig. 5 illustrates different OPS network configurations with different topologies: single-rooted tree, multi-rooted tree and butterfly. The architecture of each configuration is summarized in Table 2. Here we assume only unidirectional operation for OPS modules. In other words, each OPS link works in a simplex way. Thus, circulators are required to interface the links between hybrid ToRs and the OPS-based network.

Fig. 5(a) shows a non-blocking OPS network architecture by cascading OPS nodes following the classic single-rooted tree topology: five 16 × 16 OPS modules work as branch nodes (P_1–P_5); a 5 × 5 OPS root node R_1 connects all branch nodes. Each ToR provides only one OPS link with variable capacities. Each branch node provides connectivity among 15 ToRs (from both Tx and Rx sides) and their accesses to other branches of the tree. The capacity between different branches is oversubscribed for high connectivity demand with an oversubscription rate of 15:1, which leads to the capacity bottleneck for delivering the majority of "east-west traffic" in the DC clusters.

By introducing more redundancy together with intelligent multi-path routing strategies, DCN architectures with full-bisection provision can be constructed. Fig. 5(b) gives an example leveraging multi-rooted tree topology: each ToR provides one OPS link; fifteen 10 × 10 OPS modules work as branch nodes (P_1–P_{15}) and each of them connects 5 ToRs (from both Tx and Rx sides); five 15 × 15 OPS modules (R_1–R_5) connect all branch nodes thus the oversubscription is avoided by the overprovision of root nodes. Similar topologies such as Fat-tree [37], D-Cell [38], or BCube [32] can be constructed as well, depending on the provision of OPS modules. Multi-path routing protocols, such as equal-cost multi-path (ECMP) algorithm [39], are required in these cases to efficiently allocate workloads among different root nodes in order to optimize network performance.

Moreover, a butterfly topology with 25 15 × 15 OPS nodes (P_1–P_{25}) is shown in Fig. 5(c). Each OPS node connects the Tx of 15 ToRs with the Rx from another 15 ToRs. With 5 OPS links provided on each ToR, the Tx on each ToR is connected to 5 different OPS nodes and so does the Rx on each ToR. Therefore, all the ToRs are fully connected and a unique "one hop" OPS path is set for each ToR pair. In such configuration, collocated traffic loads from the same ToR are split and loaded on different links, and sent to different OPS nodes according to their destinations.

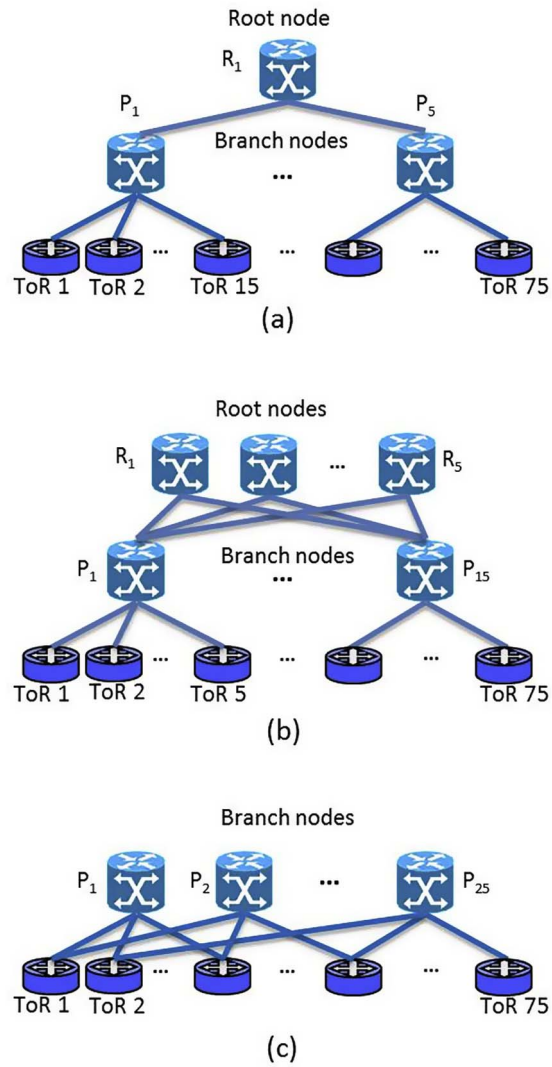

Fig. 5. Illustration of representative multi-stage OPS network topologies: (a) single-rooted tree topology; (b) multi-rooted tree topology; (c) butterfly topology.

Table 2
Summary of OPS network configurations.

Configuration	Single-rooted Tree	Multi-rooted Tree	Butterfly
Architecture Description	5 branch nodes: with 16 × 16 OPS module; 1 root node: with 5 × 5 OPS module;	15 branch nodes: with 10 × 10 OPS module; 5 root nodes: with 15 × 15 OPS module;	25 OPS nodes: with 15 × 15 OPS module;

In all the cases described above, OCS network described in the last subsection is constructed in parallel as a supplementary to the OPS network. By setting OCS links between ToR pairs where augmented capacity or tight latency is required, the network performance can be further improved.

3.5. Transmission media for DCN

In order to provide low latency interconnections in large-scale DCNs, flat DCN architectures are preferred with a reduced number of hops. Compared to the traditionally hierarchy DCN, the flat-structured DCN requires more connections, as each ToR needs to connect more

ToR directly. The connectivity will be one of the big challenges for future large-scale DCN with a flat architecture. Thanks to recent advances in fiber technologies, space division multiplexing (SDM) is now possible, allowing a large number of signals to be multiplexed not only in wavelength, but also in space, and be transmitted along a single optical fiber at the same time. Several SDM technological alternatives have been reported in the literature, and include multimode fibers (MMFs), multicore fibers (MCFs) [40], Multi-element fibers (MEFs) [41] and even their combinations. Using SDM, a spatial multiplicity as high as 36 has been demonstrated in fibers with dimensions not too dissimilar to those of a typical SMF [42]. The use of SDM technologies in DCN can help simplify the connectivity between ToRs and the centralized LPFS. In combing with wavelength division multiplexing, a dramatic increase of connections can be achieved to provide more connectivity in DCNs. In [23], we demonstrated the use of SDM in a DCN for the first time.

Another way to reduce communication latency is to use hollow-core bandgap fiber, which could reduce propagation delay by 30%. By combing with a flat DCN architecture, ultra-low latency communications could be offered for chip-level access in a disaggregated data center [43].

4. Simulation of OPS/OCS hybrid DCN

4.1. Simulation scenarios and parameters

We use the simulated DC traffic patterns [44] as traffic demands in the hybrid DCN and examine the network behaviors under such traffic loads. Two typical DC traffic patterns are simulated. In Case I, we assume that the inter-processor-like traffic dominates the whole DC, whereas in Case II, the hot-spots-like traffic is the major traffic pattern. Fig. 6(a) illustrates the modeled DC traffic pattern in Case I via the heat map of inter-rack traffic matrix of $log_{10}(Bytes)$ in a simulated 1 s interval. We can see that the communication degree of each ToR is bounded and hot ToRs exchange much of their data with only a few other ToRs (see dark and red dots in the heat map). By contrast, Fig. 6(b) illustrates the simulated 1 s traffic pattern in Case II, where hot ToRs communicate with most ToRs in the DC following a “fan-in/fan-out” pattern while the “cold traffic” pattern is popular among cold ToRs.

Firstly, we assume each ToR can provide 12 OPS/OCS programmable channels with 10Gbps capacity per channel. The maximum bandwidth a ToR can offer is 120Gbps and the capacity of each link from this ToR can vary from 10 Gb/s to 120 Gb/s. An OPS emulator is developed and programmed with Matlab, as illustrated in Fig. 7. A 2-μs slot size is selected for synchronous OPS operation and a 15% overhead is assumed for each packet. Each flow transmitted by OPS network in the simulation is firstly divided to a queue of optical packets (or frames)

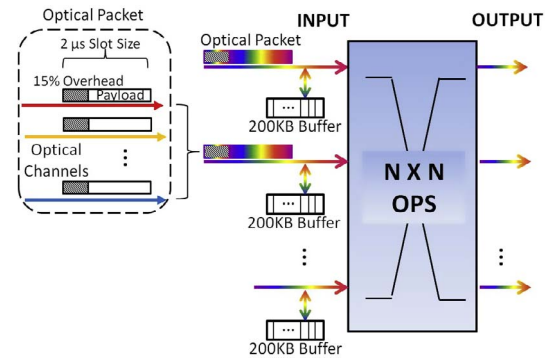

Fig. 7. Schematic of OPS emulator enabling optical waveband switching.

with an equal size, instead of Ethernet packets with uncertain sizes. To take the most advantage of such flexible capacity, we assume that each OPS node is enabled for optical waveband switching. OPS modules are transparent to optical wavelength (e.g., OPS based on semiconductor optical amplifier (SOA)). Thus, multiple wavelengths can be utilized for a single optical packet. Thus, the optical packet size (in Bytes) is decided by the efficient OPS link bandwidth:

$$pt_{size} = slotsize \times linkcapacity. \quad (1)$$

In case of congestions, random priority is given to each packet for being switched to the output port successfully. Blocked packets are immediately buffered at the corresponding input of OPS module and waiting for retransmission. The capacity of each electrical buffer is set as 200 KB. The latency of each received byte caused by buffering and the amounts of bytes dropped due to buffer overflow are counted in the simulation.

We assume that all the OPS emulators in our model are switching simultaneously. Each optical packet transferred through a multi-stage OPS link is switched in different time slots for each hop. Moreover, regarding the multi-rooted tree network, a simple multi-path routing protocol is used to distribute traffic loads among different root nodes efficiently: the relevant packet is always switched to the root node with the least buffer occupation.

Apart from the OPS emulator, traffic transmission through OCS links is also simulated. Variable capacity channels can be assigned to the OCS network while the overall channel number from each ToR is fixed as 12. Given the fact that the required capacity between any ToR pair is never beyond 10 Gb/s in our traffic model, we always assign a single 10 Gb/s channel to each OCS link so that the degree of connectivity in the OCS network, i.e. the number of OCS links from each ToR, is maximized. Once an OCS link is set between two ToRs, all the data exchanged between these two ToRs are going through the OCS link and counted without any delay or loss in our simulation.

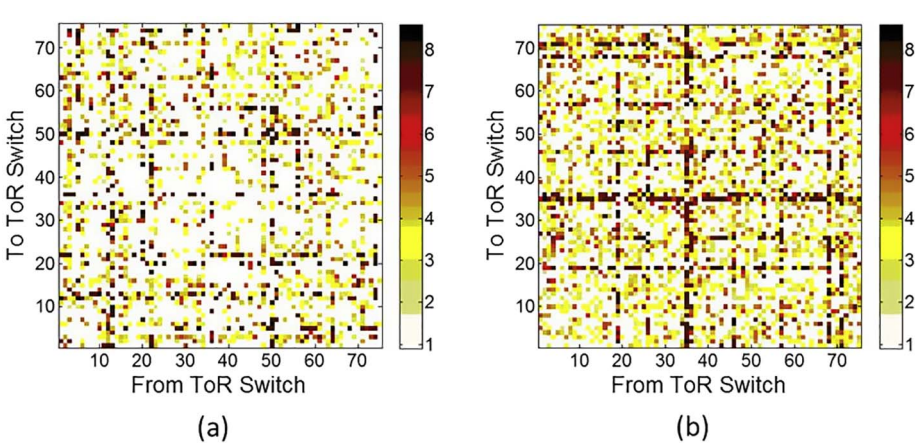

Fig. 6. Inter-rack traffic distributions among 75 ToRs in a simulated 1 s interval are illustrated by heat maps of traffic matrices of $log_{10}(Bytes)$ for (a) Case I; (b) Case II.

Table 3
Summary of extra simulation parameters for simulation.

ToR				OPS			OCS
Optical Channel Number	capacity per channel	Processing delay by FPGA	Fiber length	Reconfiguration strategy	Overhead including guard time, setting time, synchronization.	capacity of each buffer	Reconfiguration strategy
12	10Gbit/s	Refer to [21], but omitted in simulation	5 m	Synchronous reconfigurations in every 2 us	15% (300 ns)	200 KB	Setting time around ~1 ms, no more delay or loss is counted once an OCS link is established.

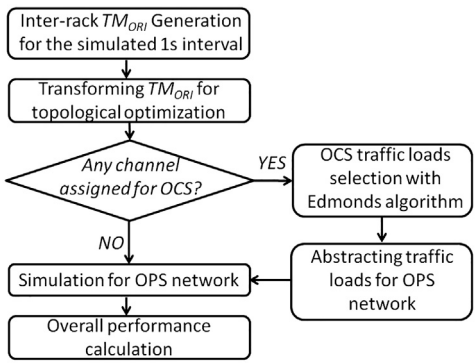

Fig. 8. Schematic of network behaviour simulation with the function-topology management for each 1 s time slot.

Some extra parameters are listed in Table 3.

It is worthy noting that all the parameters used in our simulation are chosen due to our limited computation capability. Future DC with optical inter-connects should equip with higher channel capacity and shorter slot size to support even heavier traffic workloads.

4.2. Function-topology management strategy

We simulate network behaviors for each 1 s time slot within a continuous 30 s period. The processes of simulation are schemed as Fig. 8, where the function-topology management is composed of two steps: topology optimization for OPS network, and traffic load separating between OCS and OPS networks.

Topology optimization for the OPS network aims to derive the objective matrix TM_{OPT} from the original traffic matrix TM_{ORI} by column switching and row switching. In the single-/multi-rooted tree or butterfly topologies, the objective matrix should balance traffic distributions among OPS branch nodes. Regarding the butterfly topology, the objective matrix need to distribute traffic more in OPS branch nodes. Hence, we present a greedy heuristic algorithm for constructing the objective matrix, as outlined in Algorithm 1.

Firstly, all the source ToRs and destination ToRs are sorted from the hottest to the coldest. The matrix TM_{ORI} is thereby transformed into TM_{sort} with rescheduled source list S_{sort} and destination list D_{sort}, which satisfies,

$$\forall\ m,n \in S_{sort}\text{: } if\ \ m < n, \quad then\ T(m) > T(n)$$

$$\forall\ m,n \in R_{sort}\text{: } if\ \ m < n, \quad then\ R(m) > R(n),$$

where N is the rack number in the DC, and

$$T(j) = \sum_i \delta_{ij}, \quad \delta_{ij} \in TM_{sort}, \tag{2}$$

$$R(i) = \sum_j \delta_{ij}, \quad \delta_{ij} \in TM_{sort}, \tag{3}$$

which are the traffic loads transferred/received by each ToR.

Taking the inter-rack traffic in the 25th second for instance, Fig. 9(a) illustrates the heat map of TM_{sort} with rescheduled ToR sequences which are marked with the original ToR labels. We can see that such matrix is the very objective matrix for the butterfly topology optimization, where most traffic loads are distributed among branch nodes.

Algorithm 1. Heuristic algorithm for optimizing traffic load distribution in OPS networks: *TopologyOptimization*(TM_{ORI}, *topology*), where TM_{ORI} is the original traffic matrix, and *topology* is the topological name of OPS network.

begin
  sort source ToRs in descending order;
  return new source list S_{sort}
  sort destination ToRs in descending order;
  return new destination list D_{sort}
  construct matrix TM_{sort} in the order of S_{sort} and D_{sort};
  calculate $(T, R) \leftarrow f_{ToR}(TM_{sort})$;
  if *topology* = "*butterfly-tree*" **then**
    $S_{OPT} \leftarrow S_{sort}$;
    $D_{OPT} \leftarrow D_{sort}$;
    $TM_{OPT} \leftarrow TM_{sort}$;
  else
    $k \leftarrow$ branch node number;
    $N \leftarrow$ rack number;
    for $m = 1$ to k **do**
      $S_g(m) \leftarrow$v$\varnothing$;
      $T_g(m) \leftarrow 0$;
      $D_g(m) \leftarrow \varnothing$;
      $R_g(m) \leftarrow 0$;;
    end
    $p \leftarrow 1$;
    $q \leftarrow 1$;
    $r \leftarrow 1$;
    while $p \leqslant N$ **do**
      sort T_g in ascending order;
      return branch order B_{source}
      $S_g(B_{source}(q)) \leftarrow S_g(B_{source}(q)) \cup \{S_{sort}(p)\}$;
      $T_g(B_{source}(q)) \leftarrow T_g(B_{source}(q)) + T(p)$;
      if $|S_g(B_{source}(q))| \geqslant N/k$ **then**
        $q \leftarrow q + 1$;
      end
      sort R_g in ascending order;
      return branch order B_{dest}
      $D_g(B_{dest}(r)) \leftarrow D_g(B_{dest}(r)) \cup^{\{} D_{sort}(p)\}$;
      $R_g(B_{dest}(r) \leftarrow R_g(B_{dest}(r)) + R(p)$;
      if $|D_g(B_{dest}(r))| \geqslant N/k$ **then**
        $r \leftarrow r + 1$;
      end
      $p \leftarrow p + 1$;
    end
    construct S_{OPT} by combining S_g;
    construct D_{OPT} by combining D_g;
    construct TM_{OPT} in the order of S_{OPT} and D_{OPT};
  end
end

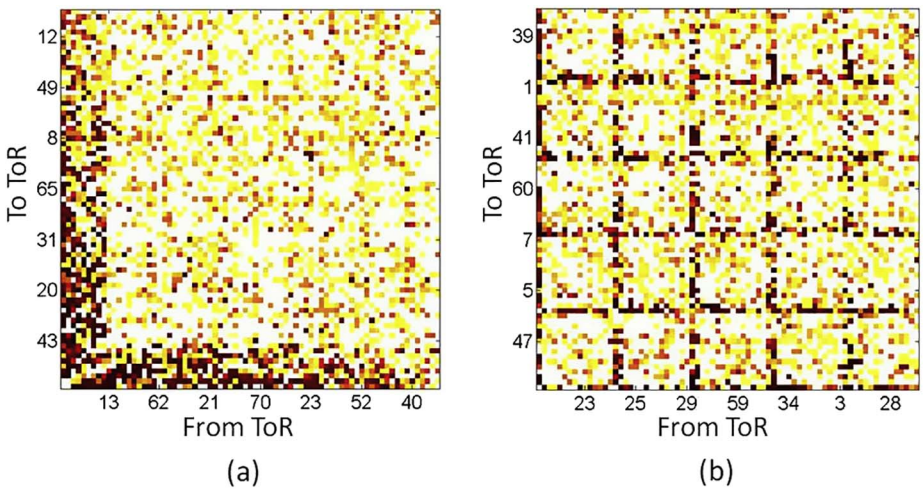

Fig. 9. Matrix transformation for topology optimization in the 25th second: (a) traffic matrix TM_{sort} with rescheduled source and destination sequences from the hottest to the coldest; (b) traffic matrix TM_{OPT} with balanced traffic loads distribution among branches in the single-rooted tree.

return TM_{OPT}

For other topologies, matrix with more balanced traffic distribution needs to be constructed. In our heuristic algorithm, sub-lists $S_g(1)$–$S_g(k)$ and $D_g(1)$–$D_g(k)$ are set to represent different groups of source ToRs and destination ToRs connected to each branch node of OPS network, where k is the number of branch nodes. The overall traffic loads on each sub-list are calculated as:

$$T_g(m) = \sum_i T_i, \quad i \in S_g(m) \tag{4}$$

$$R_g(m) = \sum_i R_i, \quad i \in D_g(m) \tag{5}$$

Then, we distribute source ToRs and destination ToRs one by one from S_{sort} and D_{sort} onto different sub-lists while minimizing the difference of traffic loads among them: during the ToR assignment, the sub-list with the least overall traffic loads always tends to have the hottest ToR from unassigned ToRs unless such sub-list is full, i.e. if i satisfies $T_g(i) = \min\{T_g\}$, and $|S_g(i)| < N/k$, then the hottest unassigned source ToR is added into $S_g(i)$. And so does the assignment for destination ToRs. Fig. 9(b) illustrates the traffic distribution in the simulated 25th second after such matrix operations for the single-rooted tree topology optimization, where hot ToRs are separated for different branch nodes to reduce traffic congestions on them.

Next, depending on the capacity assigned to OCS network, we simulate OCS links that set between certain ToR pairs. OCS network is constructed by establishing OCS links between relevant ToR pairs directly. To maximum the throughput of OCS network, those ToR pairs with OCS interconnections are selected by the Edmonds algorithm [45], which takes the inter-rack traffic matrix as a weighted graph G(V, E, W) and selects source-to-destination ToR pairs out from it. The heuristic algorithm for the OCS network construction is outlined in Algorithm 2, where OCS traffic loads are picked out and separated from OPS traffic loads. To make each OCS link bi-directional, the Edmonds algorithm is applied to a symmetric matrix TM_{sym} which is derived from the inter-rack traffic matrix TM_{Rack} and its transpose, so that the constructed OCS network is symmetric as well.

Algorithm 2: Heuristic algorithm for constructing OCS network and separating traffic loads between OCS and OPS network: *OCS_construction*(TM_{Rack}, n), where TM_{Rack} is the original inter-rack traffic matrix, and n is the channel number enabled for OCS on each ToR.

Begin
  $OCS_{link} \leftarrow \varnothing$;
  $TM_{sym} = TM_{Rack} + TM_{Rack}^T$;
  Transfer matrix TM_{sym} into graph $G(V, E, W)$;
  while $n \neq 0$ do
    $n \leftarrow n$-1;
    Apply Edmonds algorithm to graph G;
    return *mates*
    $OCS_{link} \leftarrow OCS_{link} \cup^m ates$;
    Remove *mates* from graph G;
  end
  Transfer OCS_{link} to OCS traffic matrix TM_{OCS};
  $TM_{OPS} \leftarrow TM_{Rack} - TM_{OCS}$;
end
return TM_{OCS}
return TM_{OPS}

4.3. Network performance for OPS/OCS hybrid DCN

The OPS/OCS hybrid DCN can be configured to implement different network topologies. A matlab-based DCN simulation platform is implement based on the previous assumption. Evaluations of Network performance have been done for different network topologies in terms of traffic drop rate and average latency. The statistic traffic drop rate and average delay for each Byte, instead of that for each packet, are used to describe the network behaviors.

As assumed, the OPS/OCS hybrid ToR switch could configure the 12 channels on each ToR either for OCS network or OPS network. The configuration would affect the network performance in DCN. Fig. 10. illustrates the varying of data drop rate for each configuration with different channel numbers assigned to OCS network. For both traffic patterns in Case I and II, simulation results indicate that hybrid OCS/OPS networks can perform better than either of the homogenous networks: the traffic drop rate could be reduced by at least an order of

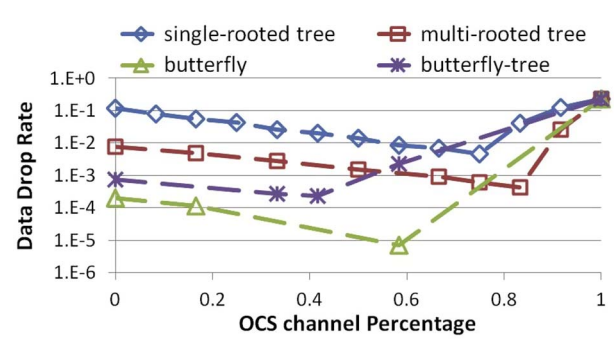

Fig. 10. Simulated traffic drop rates in the 25th second for different network configurations, with various capacity allocations between OCS and OPS networks during the function-topology management.

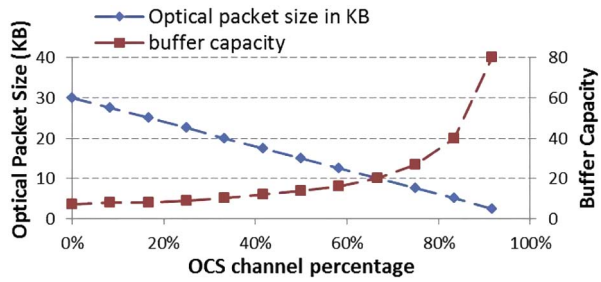

Fig. 11. The size of optical packet is decreasing with the growing of OCS capacity, which leads to the increment of OPS buffer capacity.

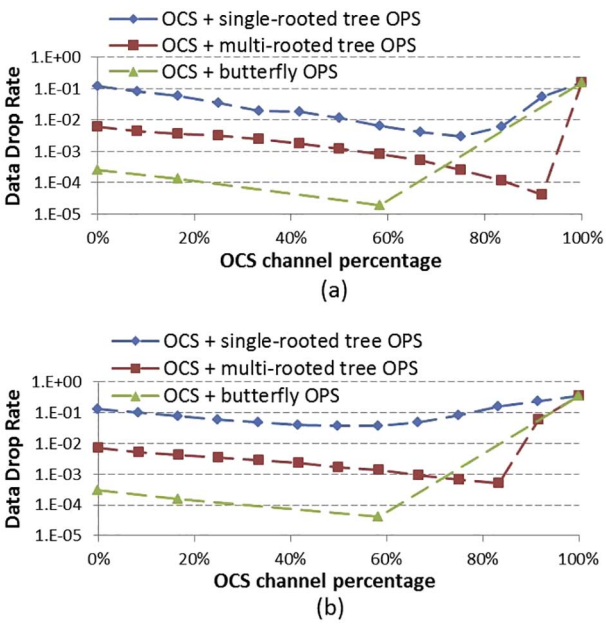

Fig. 12. Simulated traffic drop rates with different capacity allocations between OCS and OPS networks under the traffic demands in: (a) Case I; (b) Case II.

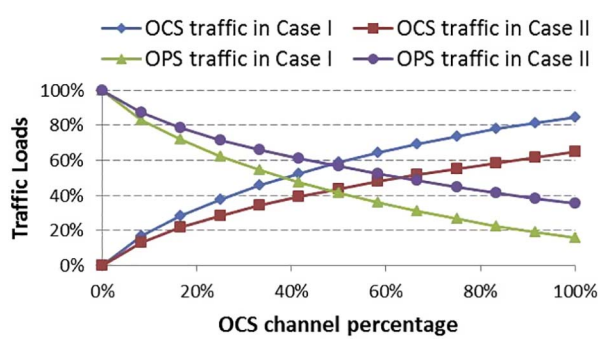

Fig. 13. Traffic loads shared by OCS and OPS network with different capacity allocations between them in the network simulation for Case I and Case II.

magnitude when introducing appropriate OCS links in parallel with the OPS networks. The detailed simulation results are shown in Fig. 12. The reason behind is that the distribution of DC traffic is highly skewed. Thus, even a small amount of OCS links can effectively split the bulk traffic loads and reduce the competing with the other traffic at the hot ToRs in the OPS network. Fig. 13. shows the share of traffic loads taken by OCS and OPS networks with different OCS channel numbers in Case I and Case II. As expected, the more skewed that the traffic pattern is, the more effectively that the OCS network can perform.

Besides, the bandwidth of OPS network is decreased with growing OCS channels since they share the same capacity provided by each ToR.

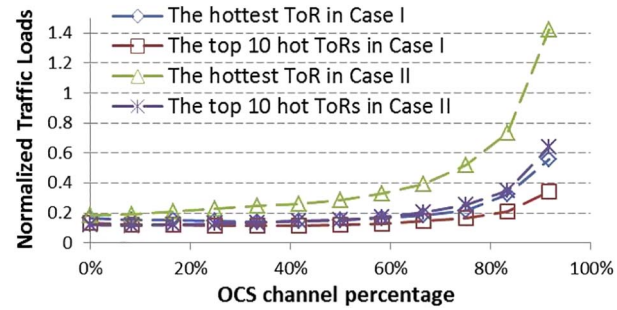

Fig. 14. The traffic loads for the hottest ToR and the average traffic loads for the top 10 hot ToRs in the OPS network, normalized by the OPS bandwidth of each ToR, are varying with the increment of OCS channel assignment.

Thus, increasing the OCS channels leads to the increment of OPS buffer capacity in the simulation: the size of each electrical buffer in the OPS network is fixed as 200 KB but the size of optical packet is reduced with fewer OPS bandwidth (see Eq. (1)), which means that more optical packets are able to be stored in each buffer. Fig. 11 illustrates the varying of optical packet size and the number of packets able to be buffered with different OCS channel numbers. Meanwhile, the performance of OPS network will start to degrade if the bandwidth drops quicker than the traffic loads it partakes. Fig. 14 illustrates the normalized OPS traffic loads for the hottest ToR and the top 10 hot ToRs in the OPS network. Thus, the overall network behavior can only be benefited when the increment of network utilization brought by OCS links exceeds the loss of it due to the reduction of OPS capacity. A trade-off has to be made for allocating capacity between OCS and OPS systems in order to optimize the network performance, which is depending not only on the network topology but also on the distribution of traffic pattern, as summarized in Table 4.

4.4. Cost and power consumption

The comparison of network performance illustrated in Fig. 12 indicates two alternative ways to improve DCN architecture: a) upgrading the OPS network topology by introducing more redundant OPS nodes; b) constructing the hybrid DCN by synthesizing the OCS network in parallel with the OPS network. A preliminary analysis on the cost and power consumption of such improvements is provided in this section.

The cost estimation of OCS network is based on the price of commercially available 320-port 3D-MEMS in [46], where 0.17/port can be derived as the cost of each OCS port in arbitrary unit. Similarly, the power consumption of OCS network is assumed as 0.47 W/port according to power requirement of the 3D-MEMS quoted in [46].

The estimation of OPS network is tricky since there has not been any mature technology developed for such scheme so far. Thus, we build our estimation model based on the switching fabric of each OPS node. Given that there are a number of other key elements in each OPS node such as buffers, label processors and controllers, the switch fabric are assumed to represent at most 50% of the total cost and power consumption of the whole OPS node, which is a rather conservative estimation compared to the case in present-day packet switching scheme [47]. For example, we use the price of the PLZT switch in [46] to calculate the cost of switching fabric which is per port in arbitrary unit, thus the cost of OPS node is 2/port in arbitrary unit (we assume that the fast switch adopted in OPS nodes should be competitive with PLZT switch on price). The power consumption of OPS switching fabric is estimated based on the Benes switch with 2×2 SOA gate arrays and 0.4 W is utilized as the power required for each SOA gate working in "ON-state" [48]. Thus, the power consumption of such switching fabric rises exponentially with port number n:

$$P_{n\times nBenes} = \frac{n}{2}(2\log_2 n - 1) \times 2 \times 0.4W \tag{6}$$

Table 4
Summary of the network performance for various configurations.

Simulated Network Performance			OCS + OPS with Single-rooted Tree Topology	OCS + OPS with Multi-rooted Tree Topology	OCS + OPS with Butterfly Topology*
Case I	pure OCS network	Data Drop rate	1.55e−1	1.55e−1	1.55e−1
		Ave. Latency (s)	null	null	null
	pure OPS network	Data Drop rate	1.21e−1	6.12e−3	2.61e−4
		Ave. Latency (s)	1.87e−5	6.32e−6	7.13e−7
	Optimized hybrid OCS/OCS network	Data Drop rate	3.04e−3	4.05e−5	1.90e−5
		Ave. Latency (s)	5.39e−6	1.23e−6	1.98e−7
		Capacity assignment on each ToR	9 OCS links with 10 Gb/s each; 1 OPS link with 30 Gb/s;	11 OCS links with 10 Gb/s each; 1 OPS link with 10 Gb/s;	7 OCS links with 10 Gb/s each; 5 OPS links with 10 Gb/s each;
Case II	pure OCS network	Data Drop rate	3.54e−1	3.54e−1	3.54e−1
		Ave. Latency (s)	null	null	null
	pure OPS network	Data Drop rate	1.31e−1	7.52e−3	2.92e−4
		Ave. Latency (s)	1.91e−5	6.77e−6	1.24e−6
	Optimized hybrid OCS/OCS network	Data Drop rate	3.70e−2	5.01e−4	4.20e−5
		Ave. Latency (s)	1.42e−5	3.92e−6	6.85e−7
		Capacity assignment on each ToR	6 OCS links with 10 Gb/s each; 1 OPS link with 60 Gb/s;	10 OCS links with 10 Gb/s each; 1 OPS link with 20 Gb/s;	7 OCS links with 10 Gb/s each; 5 OPS links with 10 Gb/s each;

* The OPS network with butterfly topology requires 5 OPS links from each ToR, thus the maximum number of OCS links that each ToR can support is 7.

Table 5
Model of cost and power consumption for hybrid DCN.

DCN with 75 ToRs	Cost (a.u.)	Power (W)
OCS network (with m OCS links/ToR)	$75 * m * 0.17$	$75 * m * 0.47$
$n \times n$ OPS node	$n * 2$	$n*(2 \log_2 n - 1) * 0.8$
Single-rooted tree OPS network	170	462
Multi-rooted tree OPS network	450	1086
Butterfly OPS network	750	2044

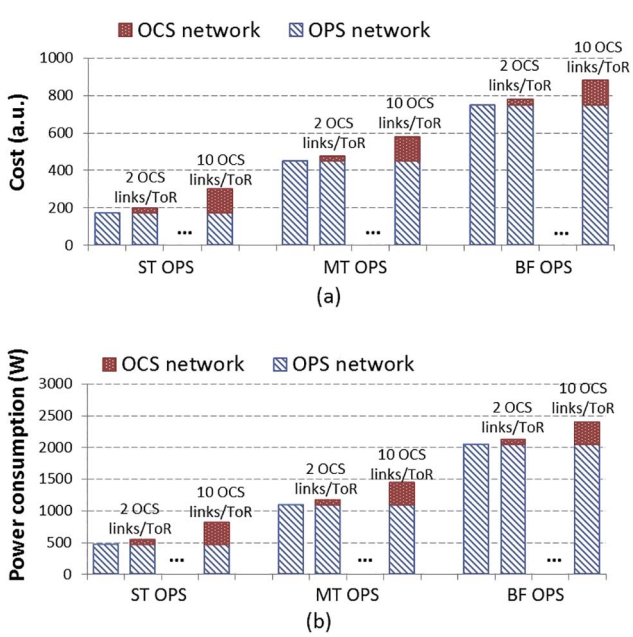

Fig. 15. Comparisons with overall cost and power consumption for different configurations of hybrid DCN. ST: single-rooted tree; MT: multi-rooted tree; BF: butterfly.

Table 5 summarizes the cost and power consumption model for each OCS and OPS configurations. The comparisons of the overall cost and power consumption between different hybrid DCN configurations are illustrated in Fig. 15. Together with the performance comparison shown in Fig. 12, we can see that the construction of hybrid DCN is more efficient than the overprovision of OPS nodes in terms of cost and power consumption while offering the similar or even better improvement on network performance.

4.5. Scalability

To scale up the proposed OPS/OCS hybrid DCN, the main challenges are the limited port numbers of the LPFS and that of the OPS switch modules. In the proposed OPS/OCS hybrid DCN, the LPFS is used to implement OCS networks and also manage the OPS switch modules based on architecture-on-demand. Currently, high radix optical switch can offer over 384 × 384 switching [49]. The recent developing silicon photonics provides a potential candidate for low loss high radix switches [50]. Furthermore, multiple LPFSs could be cascaded together to provide even higher radix optical switches [46]. Several methods could combine several LPFSs to a large LPFS. However, the port number of the LPFS still not enough for future hyperscale DCNs. Thus, in the proposed hybrid DCN, the hyperscale DCN is firstly divided to the clusters. Each cluster can be implemented with the proposed OPS/OCS hybrid DCN solution. Cluster based DCN architecture will relax the requirements for the LPFS.

Regarding to the OPS switching, multiple technologies have been adopted with a limited number of ports, such as SOAs and PLZT [51]. The OPS switching modules are still in an early stage. The efficient OPS require precise time synchronization. Thus, the OPS are only used for intra-cluster communications with a limited scale.

5. Experimental demonstration

We demonstrated virtual data centers (VDC) provision in an OCS/OPS hybrid data center. Due to the limited scale, no cluster switches are used. Fig. 16 shows the experimental setup of the hybrid data center. The hybrid data center deployed our developed time-shared optical networking (TSON) [52], FPGA-based OPS/OCS SIC and optical circuit elements [24]. Compared to OPS technology, TSON technology provides similar but a simpler solution for optical slot switching and no extra header are required. In the experiments, TSON is used to offer similar connectivity as the OPS. On top of the data plane, a software stack that consists of the Orchestrator and the OpenFlow agents is developed for the hybrid data center. The software stack enables the provisioning of VDC instances over the optical data layer of the hybrid OPS/OCS DCN. Here we treat the TSON as a simplified OPS network. The TSON is used only for intra-cluster communications.

The data plane consists of two kinds of fiber switches. The beam-steering 2 × 2 4-core MCF switch provides optical switching over 4-core MCFs. This MCF device offers a 300% increase in fibre capacity over single-mode fibre (SMF) and we envision usage for inter-DC traffic. Another LPFS is used for the OCS system and to manage the network

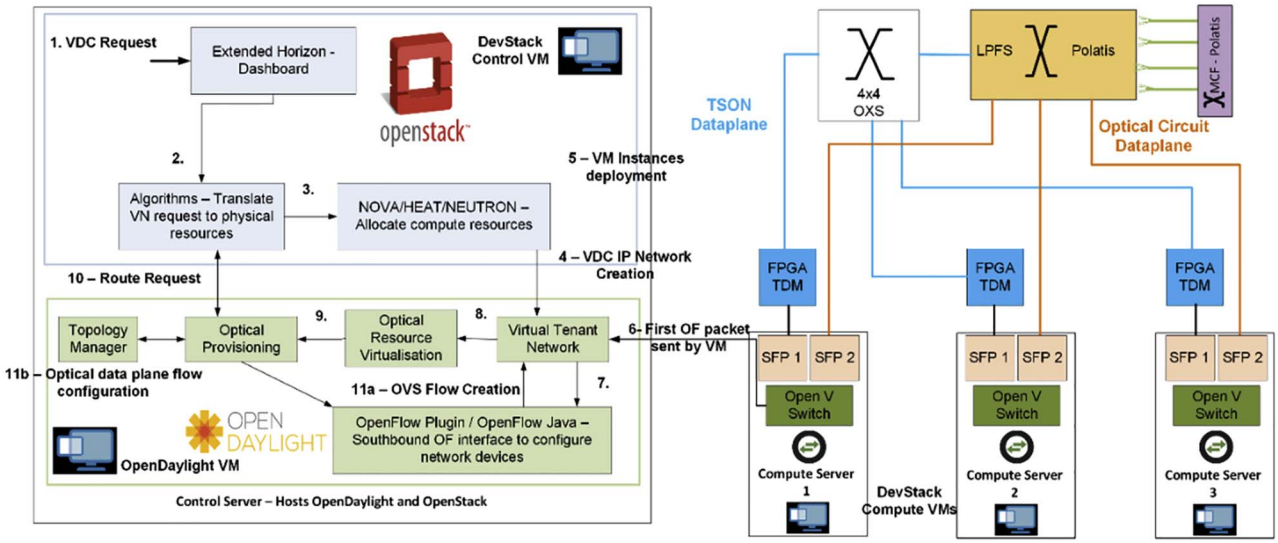

Fig. 16. Architecture and control flow for virtual data center provisioning.

function programmability. A 4 × 4 optical fast switch (OXS) with nanosecond level switching capability is used for TSON switching by incorporating with an FPGA-based controller. The 4 × 4 optical fast switch is implemented based on PLZT photonics technologies (EpiPhotonics. Inc). Compute and storage nodes are interfaced to TSON using FPGA-based SIC cards. As you can see in the experimental demonstration, multiple switching granularities are offered for different kinds of network traffic.

OpenDaylight Lithium (ODL) was enhanced with a number of extensions to support communications with the optical data plane elements. The OpenFlow protocol was also extended to support the TSON and OCS devices. The two layers of optical circuit switching, combining SMF and MCF devices, each controlled via SDN enables a flat architecture and multi-dimensional optical switching. The TSON scheme is defined and modelled with the OpenFlow protocol. The OpenFlow agents for the TSON device and the FPGA SIC are developed. The extended protocol enables the provisioning of optical resources in combination with the orchestration layer. The orchestrator is an extended OpenStack (OSK) platform. In order to optimize the provisioning of optical resources, a novel algorithm module is developed to translate tenants' bandwidth requirements into requests for TSON slots or optical circuits. The extended Northbound REST interface of ODL is used to interact with Optical Resource and Provisioning Modules to create the flow necessary to allocate the requested Virtual Network (VN).

The OpenStack (DevStack) implementation dynamically provides TSON and OCS resources via an extended and optically-enabled SDN controller. A new algorithm module is developed to determine the several logical instances, such as IP network, subnetworks and ports, to enable traffic exchanges along the VDC instance. To map the VMs and create the logical resources, the algorithm interacts with the core orchestrator services via the OpenStack Heat service. In addition to the physical route and the necessary timeslots, the algorithm also determines the particular VLAN to be employed when encapsulating the traffic of each virtual link. On each OSK compute node, an OpenVswitch (OVS) is programmed to control flows between the VM instances.

The performance of the TSON network was measured in terms of throughput and latency against allocated timeslots. These results demonstrate a sustainable maximum data rate of up to 8.6Gbps, as shown in Fig. 17. Circuit and TSON switching are combined to offer flexible and granular bandwidth provisioning. As can be observed, higher throughput and lower latency can be achieved with interleaved (or distributed) slots allocations. This is because interleaving reduces the maximum delay between data transmissions. Therefore, it is

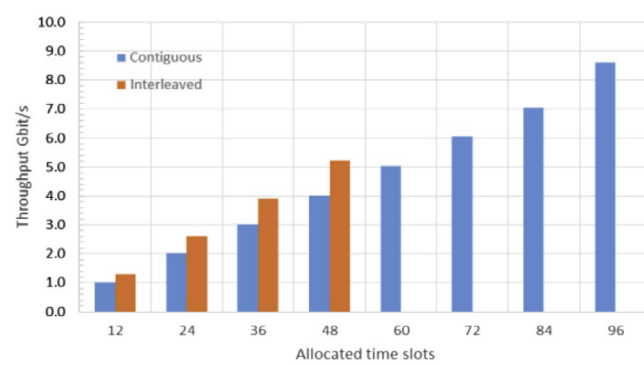

Fig. 17. Contiguous and interleaved allocated time slots vs. throughput for the TSON data plane.

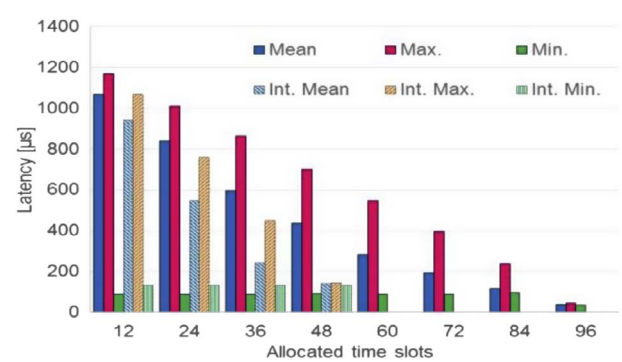

Fig. 18. TDM timeslot allocations against latency for contiguous and interleaved (Int.) slot allocation.

recommended to avoid contiguous allocation for best performance. Similarly, in Fig. 18, the maximum and mean latency measurements converge as timeslots increases because the largest gap between transmission slots reduces. The interleaved minimum is greater because unlike contiguous, there is always a no-transmit slot between transmissions.

The switching latency of the OXS was measured at both the circuit and application level. The switching time at the circuit level was measured electronically around 25 ns. Using a ping-flood method, we tested the effective reconfiguration time from an application perspective. The mean value was measured over five reconfigurations. The end-

to-end buffering and serialisation for the TSON scheme is around 38.7 μs. Same measurement is conducted for the NIC in Ethernet mode. The measured time is around 8.3 μs. An overhead of 30.1 μs is required when using the extra buffering, logic and negotiation (key characters) involved in the TSON implementation. A similar experiment measured the mean reconfiguration time of the MCF switch over several iterations as 121 μs.

6. Conclusion

In this paper, we present the design of the programmable OCS/OPS DCN architecture, a hybrid optical network solution for the future hyperscale DCs. Such design combines the advantages of OCS and OPS schemes via the adoption of FPGA-based hybrid TOR switches. Thus, traffic loads in the DC with various patterns can be effectively accommodated by different network topologies on demand.

We simulate the network behaviors for different hybrid DCN configurations under different traffic demands and evaluate the benefits brought by the flexibility of the hybrid scheme. The results indicate that the network performance can be significantly improved by configuring the hybrid OPS/OCS network topologies according to the skewed nature of DC traffic. Besides, a preliminary comparison on the cost and power consumption for different network topologies is presented, which shows that the hybrid DCN architecture is more cost- and energy-efficient than the homogenous network under the same quality of service provision. Finally, data center virtualization is demonstrated successfully based on the proposed hybrid data center architecture.

Acknowledgement

The authors acknowledge funding supports from the UK EPSRC through the project TOUCAN (EP/L020009/1) and INSIGHT (EP/L026155/2).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.yofte.2018.01.017.

References

[1] Cisco Global Cloud Index: Forecast and Methodology, 2015–2020, Cisco, 2016. < http://www.cisco.com/c/dam/en/us/solutions/collateral/service-provider/global-cloud-index-gci/white-paper-c11-738085.pdf > (accessed 19.06.17).

[2] H. Qi, M. Shiraz, J. Liu, A. Gani, Z. Abdul Rahman, T.A. Altameem, Data center network architecture in cloud computing: review, taxonomy, and open research issues, J. Zhejiang Univ. Sci. C 15 (2014) 776–793, http://dx.doi.org/10.1631/jzus.C1400013.

[3] Enterprise QoS Solution Reference Network Design Guide - Campus QoS Design [Design Zone for IPv6], Cisco. (n.d.). < http://www.cisco.com/c/en/us/td/docs/solutions/Enterprise/WAN_and_MAN/QoS_SRND/QoS-SRND-Book/QoSDesign.html > .

[4] Data Center High Speed Migration: Infrastructure issues, trends, drivers and recommendations, COMMSCOPE, 2017. < http://info.commscope.com/rs/751-QQY-459/images/Whitepaper%20-%20Data%20Center%20High%20Speed%20Migration.pdf > (accessed 15.06.17).

[5] N. Binkert, A. Davis, N.P. Jouppi, M. McLaren, N. Muralimanohar, R. Schreiber, J.H. Ahn, The role of optics in future high radix switch design, in: 2011 38th Annu. Int. Symp. Comput. Archit. ISCA, 2011, pp. 437–447.

[6] International Technology Roadmap for Semiconductors, Semiconductor Industries Association, 2009. < http://www.itrs2.net > .

[7] IEEE P802.3bs 400GbE Adopted Timeline, (n.d.). < http://www.ieee802.org/3/bs/timeline_3bs_0915.pdf > (accessed 25.06.17).

[8] N. Farrington, G. Porter, S. Radhakrishnan, H.H. Bazzaz, V. Subramanya, Y. Fainman, G. Papen, A. Vahdat, Helios: a hybrid electrical/optical switch architecture for modular data centers, ACM SIGCOMM Comput. Commun. Rev. 41 (2011) 339–350.

[9] G. Wang, D.G. Andersen, M. Kaminsky, K. Papagiannaki, T.S. Ng, M. Kozuch, M. Ryan, c-Through: Part-time optics in data centers, in: ACM SIGCOMM Comput. Commun. Rev., ACM, 2010, pp. 327–338. < http://dl.acm.org/citation.cfm?id=1851222 > (accessed 13.10.14).

[10] K. Chen, A. Singla, A. Singh, K. Ramachandran, L. Xu, Y. Zhang, X. Wen, Y. Chen, OSA: an optical switching architecture for data center networks with unprecedented flexibility, IEEEACM Trans. Netw. 22 (2014) 498–511, http://dx.doi.org/10.1109/TNET.2013.2253120.

[11] K. Xi, Y.-H. Kao, M. Yang, H.J. Chao, Petabit optical switch for data center networks, Polytech. Inst. N. Y. Univ. N. Y. Tech Rep. (2010). < http://eeweb.poly.edu/~chao/publications/petasw.pdf > (accessed 28.08.14).

[12] J. Gripp, J.E. Simsarian, J.D. LeGrange, P. Bernasconi, D.T. Neilson, Photonic terabit routers: The IRIS project, in: Opt. Fiber Commun. OFC Collocated Natl. Fiber Opt. Eng. Conf. 2010 Conf. OFCNFOEC, 2010, pp. 1–3.

[13] A. Shacham, K. Bergman, An Experimental Validation of a Wavelength-Striped, Packet Switched, Optical Interconnection Network, J. Light. Technol. 27 (2009) 841–850, http://dx.doi.org/10.1109/JLT.2008.928541.

[14] Y. Yin, R. Proietti, X. Ye, C.J. Nitta, V. Akella, S.J.B. Yoo, LIONS: an AWGR-based low-latency optical switch for high-performance computing and data centers, IEEE J. Sel. Top. Quantum Electron. 19 (2013), http://dx.doi.org/10.1109/JSTQE.2012.2209174 3600409-3600409.

[15] K.J. Barker, A. Benner, R. Hoare, A. Hoisie, A.K. Jones, D.K. Kerbyson, D. Li, R. Melhem, R. Rajamony, E. Schenfeld, S. Shao, C. Stunkel, P. Walker, On the Feasibility of Optical Circuit Switching for High Performance Computing Systems, in: Supercomput. 2005 Proc. ACMIEEE SC 2005 Conf., 2005, pp. 16–16. https://doi.org/10.1109/SC.2005.48.

[16] T. Benson, A. Anand, A. Akella, M. Zhang, Understanding data center traffic characteristics, SIGCOMM Comput Commun Rev. 40 (2010) 92–99, http://dx.doi.org/10.1145/1672308.1672325.

[17] S. Kandula, S. Sengupta, A. Greenberg, P. Patel, R. Chaiken, The nature of data center traffic: measurements & analysis, in: Proc. 9th ACM SIGCOMM Conf. Internet Meas. Conf., ACM, 2009: pp. 202–208. < http://dl.acm.org/citation.cfm?id=1644918 > (accessed 15.10.14).

[18] T. Benson, A. Akella, D.A. Maltz, Network traffic characteristics of data centers in the wild, in: Proc. 10th ACM SIGCOMM Conf. Internet Meas., ACM, 2010, pp. 267–280. < http://dl.acm.org/citation.cfm?id=1879175 > (accessed 29.06.17).

[19] L. Purushothaman, T. Truong-Huu, M. Gurusamy, Time and Bandwidth-Aware Virtual Network Embedding and Migration in Hybrid Optical-Electrical Data Centers, in: IEEE, 2017, pp. 947–954. doi: 10.1109/AINA.2017.93.

[20] N. Amaya, G. Zervas, D. Simeonidou, Introducing node architecture flexibility for elastic optical networks, IEEEOSA J. Opt. Commun. Netw. 5 (2013) 593–608, http://dx.doi.org/10.1364/JOCN.5.000593.

[21] Y. Yan, G.M. Saridis, Y. Shu, B.R. Rofoee, S. Yan, M. Arslan, T. Bradley, N.V. Wheeler, N.H.L. Wong, F. Poletti, M.N. Petrovich, D.J. Richardson, S. Poole, G. Zervas, D. Simeonidou, All-optical programmable disaggregated data centre network realized by FPGA-based switch and interface card, J. Light. Technol. 34 (2016) 1925–1932, http://dx.doi.org/10.1109/JLT.2016.2518492.

[22] B.R. Rofoee, G. Zervas, Y. Yan, D. Simeonidou, Griffin: programmable optical datacenter with SDN enabled function planning and virtualization, J. Light. Technol. 33 (2015) 5164–5177.

[23] S. Yan, E. Hugues-Salas, V.J.F. Rancano, Y. Shu, G.M. Saridis, B. Rahimzadeh Rofoee, Y. Yan, A. Peters, S. Jain, T. May-Smith, P. Petropoulos, D.J. Richardson, G. Zervas, D. Simeonidou, Archon: a function programmable optical interconnect architecture for transparent intra and inter data center SDM/TDM/WDM networking, J. Light. Technol. 33 (2015) 1586–1595, http://dx.doi.org/10.1109/JLT.2015.2392554.

[24] Chris Jackson, K. Kondepu, Yanni Ou, Arash Beldachi, A. Pagès Cruz, F. Agraz, F. Moscatelli, W. Miao, V. Kamchevska, N. Calabretta, G. Landi, S. Spadaro, R. Nejabati, D. Simeonidou, COSIGN: A Complete SDN Enabled All-Optical Architecture for Data Centre Virtualisation with Time and Space Multiplexing, in: n.d.

[25] J. Kim, C.J. Nuzman, B. Kumar, D.F. Lieuwen, J.S. Kraus, A. Weiss, C.P. Lichtenwalner, A.R. Papazian, R.E. Frahm, N.R. Basavanhally, D.A. Ramsey, V.A. Aksyuk, F. Pardo, M.E. Simon, V. Lifton, H.B. Chan, M. Haueis, A. Gasparyan, H.R. Shea, S. Arney, C.A. Bolle, P.R. Kolodner, R. Ryf, D.T. Neilson, J.V. Gates, 1100 x 1100 port MEMS-based optical crossconnect with 4-dB maximum loss, IEEE Photonics Technol. Lett. 15 (2003) 1537–1539, http://dx.doi.org/10.1109/LPT.2003.818653.

[26] T. Hayashi, T. Taru, O. Shimakawa, T. Sasaki, E. Sasaoka, Design and fabrication of ultra-low crosstalk and low-loss multi-core fiber, Opt. Express 19 (2011) 16576–16592, http://dx.doi.org/10.1364/OE.19.016576.

[27] R. Hemenway, R. Grzybowski, C. Minkenberg, R. Luijten, Optical-packet-switched interconnect for supercomputer applications [Invited], J. Opt. Netw. 3 (2004) 900–913, http://dx.doi.org/10.1364/JON.3.000900.

[28] W. Miao, J. Luo, S.D. Lucente, H. Dorren, N. Calabretta, Novel flat datacenter network architecture based on scalable and flow-controlled optical switch system, Opt. Express 22 (2014) 2465–2472, http://dx.doi.org/10.1364/OE.22.002465.

[29] R. Proietti, Z. Cao, Y. Li, S.J.B. Yoo, Scalable and distributed optical interconnect architecture based on AWGR for HPC and data centers, in: OFC 2014, 2014, pp. 1–3. https://doi.org/10.1364/OFC.2014.Th2A.59.

[30] B.C. Thomsen, B.J. Puttnam, P. Bayvel, Optically equalized 10 Gb/s NRZ digital burst-mode receiver for dynamic optical networks, Opt. Express 15 (2007) 9520–9526, http://dx.doi.org/10.1364/OE.15.009520.

[31] Y. Yan, Y. Shu, G.M. Saridis, B.R. Rofoee, G. Zervas, D. Simeonidou, FPGA-based optical programmable switch and interface card for disaggregated OPS/OCS data centre networks, in: 2015 Eur. Conf. Opt. Commun. ECOC, 2015, pp. 1–3. https://doi.org/10.1109/ECOC.2015.7341957.

[32] C. Guo, G. Lu, D. Li, H. Wu, X. Zhang, Y. Shi, C. Tian, Y. Zhang, S. Lu, BCube: a high performance, server-centric network architecture for modular data centers, ACM SIGCOMM Comput. Commun. Rev. 39 (2009) 63–74.

[33] A. Siddiqa, I.A.T. Hashem, I. Yaqoob, M. Marjani, S. Shamshirband, A. Gani, F. Nasaruddin, A survey of big data management: taxonomy and state-of-the-art, J.

Netw. Comput. Appl. 71 (2016) 151–166, http://dx.doi.org/10.1016/j.jnca.2016.04.008.

[34] A.D. Papaioannou, R. Nejabati, D. Simeonidou, The Benefits of a Disaggregated Data Centre: A Resource Allocation Approach, in: Glob. Commun. Conf. GLOBECOM 2016 IEEE, IEEE, 2016, pp. 1–7. < http://ieeexplore.ieee.org/abstract/document/7842314/ > (accessed 27.02.17).

[35] S. Yan, Y. Yan, B. Rahimzadeh Rofoee, Y. Shu, E. Hugues-Salas, G. Zervas, D. Simeonidou, Real-time ethernet to software-defined sliceable superchannel transponder, J. Light. Technol. 33 (2015) 1571–1577, http://dx.doi.org/10.1109/JLT.2015.2391299.

[36] B. Guo, S. Li, S. Yin, S. Huang, TDM based optical bypass for intra-rack elephant flow with a DPDK based online timeslot allocator, in: 2017 Opt. Fiber Commun. Conf. Exhib. OFC, 2017, p. W1H.4.

[37] M. Al-Fares, A. Loukissas, A. Vahdat, A scalable, commodity data center network architecture, in: ACM SIGCOMM Comput. Commun. Rev., ACM, 2008, pp. 63–74. < http://dl.acm.org/citation.cfm?id=1402967 > (accessed 28.07.14).

[38] C. Guo, H. Wu, K. Tan, L. Shi, Y. Zhang, S. Lu, Dcell: a scalable and fault-tolerant network structure for data centers, ACM SIGCOMM Comput. Commun. Rev. 38 (2008) 75–86.

[39] Hoops, C., Analysis of an Equal-Cost Multi-Path Algorithm, RFC Ed. (2000).

[40] S. Matsuo, Y. Sasaki, I. Ishida, K. Takenaga, K. Saitoh, M. Koshiba, Recent progress on multi-core fiber and few-mode fiber, in: Opt. Fiber Commun. Conf. Expo. Natl. Fiber Opt. Eng. Conf. OFCNFOEC 2013, 2013, pp. 1–3.

[41] S. Jain, V.J.F. Rancaño, T.C. May-Smith, P. Petropoulos, J.K. Sahu, D.J. Richardson, Multi-element fiber technology for space-division multiplexing applications, Opt. Express 22 (2014) 3787–3796, http://dx.doi.org/10.1364/OE.22.003787.

[42] K. Saitoh, S. Matsuo, Multicore fiber technology, J. Light. Technol. 34 (2016) 55–66, http://dx.doi.org/10.1109/JLT.2015.2466444.

[43] G.M. Saridis, Y. Yan, Y. Shu, S. Yan, M. Arslan, T. Bradley, N.V. Wheeler, N.H.L. Wong, F. Poletti, M.N. Petrovich, D.J. Richardson, S. Poole, G. Zervas, D. Simeonidou, EVROS: All-optical programmable disaggregated data centre interconnect utilizing hollow-core bandgap fibre, in: 2015 Eur. Conf. Opt. Commun. ECOC, 2015, pp. 1–3. < https://doi.org/10.1109/ECOC.2015.7341960 > .

[44] Y. Shu, S. Peng, Y. Yan, S. Yan, E. Hugues-salas, G. Zervas, D. Simeonidou, Evaluation of function-topology programmable (FTP) optical packet/circuit switched data centre interconnects, in: 2015 Eur. Conf. Opt. Commun. ECOC, 2015, pp. 1–3. https://doi.org.10.1109/ECOC.2015.7341871.

[45] J. Edmonds, Paths, trees, and flowers, Can. J. Math. 17 (1965) 449–467, http://dx.doi.org/10.4153/CJM-1965-045-4.

[46] M. Garrich, N. Amaya, G.S. Zervas, J.R. Oliveira, P. Giaccone, A. Bianco, D. Simeonidou, J.C.R. Oliveira, Architecture on demand design for high-capacity optical SDM/TDM/FDM switching, J. Opt. Commun. Netw. 7 (2015) 21–35.

[47] R.S. Tucker, R. Parthiban, J. Baliga, K. Hinton, R.W.A. Ayre, W.V. Sorin, Evolution of WDM optical IP networks: a cost and energy perspective, J. Light. Technol. 27 (2009) 243–252, http://dx.doi.org/10.1109/JLT.2008.2005424.

[48] R.S. Tucker, Green optical communications #x2014;Part II: energy limitations in networks, IEEE J. Sel. Top. Quantum Electron. 17 (2011) 261–274, http://dx.doi.org/10.1109/JSTQE.2010.2051217.

[49] N. Parsons, A. Hughes, R. Jensen, High radix all-optical switches for software-defined datacentre networks, in: ECOC 2016 42nd Eur. Conf. Opt. Commun., 2016, pp. 1–3.

[50] Z. Wang, Z. Wang, J. Xu, P. Yang, L.H.K. Duong, Z. Wang, H. Li, R.K.V. Maeda, Low-loss high-radix integrated optical switch networks for software-defined servers, J. Light. Technol. 34 (2016) 4364–4375, http://dx.doi.org/10.1109/JLT.2016.2601078.

[51] K. Nashimoto, N. Tanaka, M. LaBuda, D. Ritums, J. Dawley, M. Raj, D. Kudzuma, T. Vo, High-speed PLZT optical switches for burst and packet switching, in: 2nd Int. Conf. Broadband Netw., vol. 2, 2005, pp. 1118–1123. https://doi.org/10.1109/ICBN.2005.1589732.

[52] Y. Yan, G. Zervas, Y. Qin, B.R. Rofoee, D. Simeonidou, High performance and flexible FPGA-based time shared optical network (TSON) metro node, Opt. Express 21 (2013) 5499–5504, http://dx.doi.org/10.1364/OE.21.005499.