Chapter 8
LIGHTNESS: All-Optical SDN-enabled Intra-DCN with Optical Circuit and Packet Switching

George M. Saridis, Alejandro Aguado, Yan Yan, Wang Miao, Nicola Calabretta, Georgios Zervas, and Dimitra Simeonidou

8.1 Introduction

In this book chapter, a flat all-optical intra-data center network architecture is introduced and validated throughout various experimental demonstrations. The architecture, which combines novel photonic switching technologies with a fully SDN-enabled control plane, aims at delivering scalable, flexible, low-latency, and high-capacity interconnection on demand. The architecture and technology described in this chapter is built under the EU FP7 project LIGHTNESS [1–3], containing both data and control plane state-of-the-art features.

The chapter begins with the "LIGHTNESS" architecture, which provides an analytical view of the proposed data plane design while also offering an insight into the control plane architecture. The overall architecture targets for utilizing hybrid OCS/OPS principals interchangeably in all parts and levels of the DCN, from the server interface cards to the all-optical top of the rack (ToR) switch and top of the cluster (ToC) switches. On top of it, a unified SDN controller is in charge of network resource allocation, decision making, and command forwarding, making the data plane fully programmable.

The following section is about the LIGHTNESS technology enablers and the hybrid OPS/OCS interconnect. It explains the technologies that were developed and

G.M. Saridis (✉) • A. Aguado • Y. Yan • D. Simeonidou
High Performance Networks Group, Department of Electrical & Electronic Engineering, University of Bristol, Woodland Road, Bristol BS8 1UB, UK
e-mail: george.saridis@bristol.ac.uk

W. Miao • N. Calabretta
Electro-Optical Communication Systems, Institute for Photonic Integration (IPI), Department of Electrical Engineering, Technical University of Eindhoven, Eindhoven, The Netherlands

G. Zervas
Department of Electronic & Electrical Engineering, University College London, Torrington Place, London WC1E 7JE, UK

F. Testa, L. Pavesi (eds.), *Optical Switching in Next Generation Data Centers*,
DOI 10.1007/978-3-319-61052-8_8

used in the LIGHTNESS framework. It includes the design and functionality of the FPGA-based optoelectrical network interfaces, the optical ToR switch, as well as the OCS and OPS switching technologies.

Later, "experimental demonstration and evaluation" section reports the demonstration of SDN and virtualization-based capabilities (such as monitoring and file transfer or database migration) entirely integrated with an advanced all-optical physical layer. The performance of the above network is experimentally evaluated in terms of BER, end-to-end latency, and control plane functionality.

In the end, in the "Discussions" section, an overview of the proposed architecture is provided, including the pros and cons of such a design, as well as further suggestions for future work.

8.2 LIGHTNESS Data Plane Architecture

As shown in Fig. 8.1, servers at each rack are interconnected to the hybrid OCS/OPS DCN data plane via an optical ToR switch. The optical ToR switch can be implemented in many ways, as a passive optical element, such as an arrayed waveguide grating (AWG), a routing-AWG (R-AWG), or an active optical switch, such as a wavelength/spectrum selective switch (WSS/SSS) or a fiber/space switch. The functionality of the classical electrical ToR switch is moved by a large extent toward the advanced network interface cards (NICs), interfacing each server. Each NIC interface performs traffic aggregation and application-aware classification of data flows to either short- or long-lived ones. This offers an increased degree of programmability and dynamicity that is essential for modern intra-DCN capabilities.

Scalability is yet another objective of LIGHTNESS, and it is realized through the deployment of architecture-on-demand (AoD)-based [4] large-port count fiber switches (OCS) and numerous nanosecond fast switches (OPS) interconnecting racks in the DC, which allows for extending the number of input/output wavelengths between ToRs (WDM scaling) when the port count of OPS or OCS switches becomes a limitation. For higher scalability, the creation of clusters made of fixed number of racks could be an alternative, e.g., the top-of-the-cluster AoD-based switch of each cluster can be interconnected with other clusters through a large port-count fiber switch.

Based on the AoD design shown in Fig. 8.1, OCS switches support the interconnection of both short-lived and long-lived traffic flows within and between clusters of the DCN. In addition, for the case of the short-lived traffic flows, OPS subsystems can efficiently fill in the gaps of the OCS by loosening interconnection demands for the OCS with an optical packet-based approach, which offers finer network granularity while allowing the sharing of the same WDM channels among different servers. Additional links directly interconnecting servers between them can be an alternative approach (as shown in Fig. 8.1); even if it is limited by the port number, the NIC is able to support.

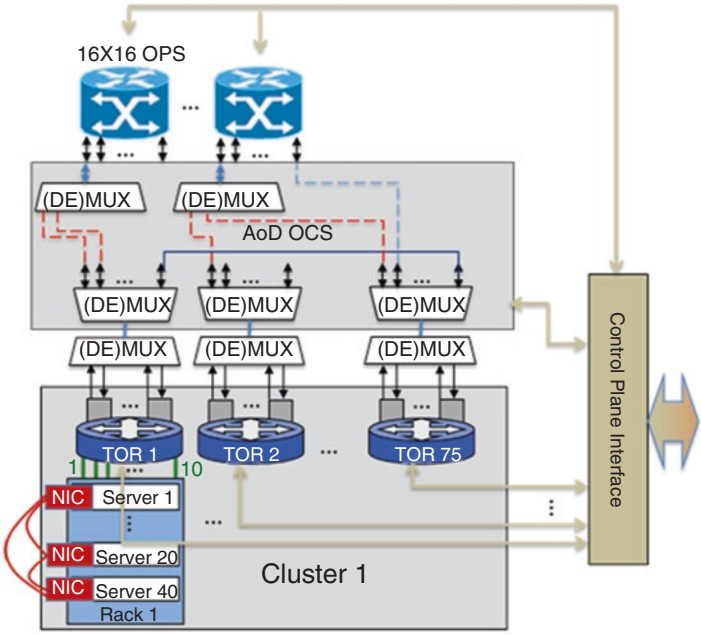

Fig. 8.1 Cluster-level LIGHTNESS AoD-based architecture

The combination of OPS and AoD-enabled OCS modules makes it possible to switch traffic in all three optical dimensions, namely, space, frequency, and time, as well as to provide a range of additional capabilities on demand. The reconfigurable optical backplane provides flexible connectivity for the AoD node. Based on the AoD-based OCS concept, the LIGHTNESS network shows the following advantages:

- Flattened network infrastructure providing the on-demand bandwidth allocation and low latency which are desirable for next generation DCN.
- Multi-granularity configuration: fiber switching, spectrum (or a single wavelength) switching, and subwavelength switching (optical packets).
- Fully reconfigurable and programmable connectivity: each NIC transponder can be directed to either OCS or OPS for transmitting information among servers.
- According to the traffic demand, OCS/OPS or both can be selected.

The proposed design for intra-cluster DCN is illustrated in Fig. 8.1. In this case, one cluster consists of several racks of servers. A ToR switch is used for interconnecting the servers in one rack to the AoD cluster backplane. Each ToR switch provides ten channels with 10Gbps/channel capacity, and each channel can support either OPS or OCS transmission, which is again directed by the ToR switch. Those ten channels are combined by a mux or a demux, creating the optical fiber input and output ports of one ToR. As shown in Fig. 8.2, an AoD-enabled OCS interconnects all the input and output ports of different ToRs, OPS modules, and traffic from/to other clusters as well.

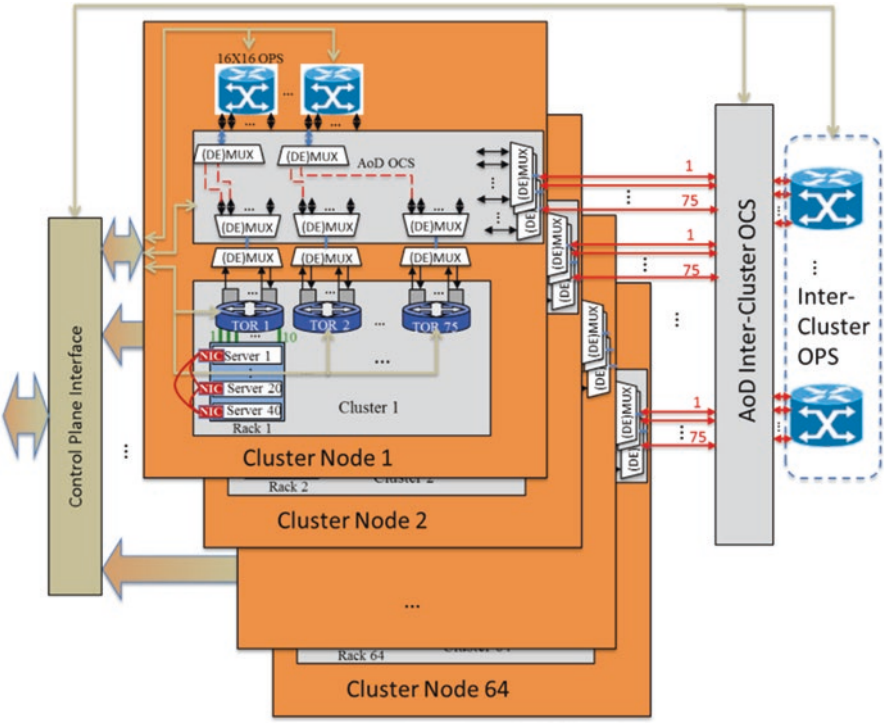

Fig. 8.2 High-level LIGHTNESS AoD-based architecture including inter-cluster connectivity

The use of the AoD structure for the OCS node makes it capable of supporting multi-granularity switching (from 10Gbps to 100Gbps). For example, by cross-connecting two ToRs directly, e.g., ToR to ToR, coarse granularity switching is achieved, whereas lower bandwidth granularities can be accomplished by separating channels from each ToR and assembling them again when required. After demultiplexing, channels from each ToR can be connected to the OPS module or interconnected with each other as OCS connections, depending on the requirement.

For this design, the OPS module could be either 4×4 or 16×16, depending on the port count and bandwidth required for the OPS. The number of links between each ToR and OPS module can be flexible, while it is also possible to scale the OPS switch by cascading several small switch modules in a Clos topology. The WDM mux/demux (e.g., AWG) interface between ToRs and AoD OCS backplane can potentially be replaced by the space division multiplexing (SDM) technologies, like multiple fibers (or ribbon fibers) or multicore fiber for further cost savings [5]. However, this might compromise the switching granularity at the OCS level.

8.3 LIGHTNESS Control Plane Architecture

The proposed LIGHTNESS control plane allows the provisioning of the hybrid optical data center network resources through a set of integrated and innovative functionalities and procedures: implementing connectivity services setup and teardown, dynamic service modification, path and flow computation, data center network resiliency and monitoring, and dynamic and automated resource optimization. The LIGHNTESS control plane is also equipped with an open, flexible, and extensible interface at the northbound for cooperation with management, VDC planning, and orchestration entities or user applications. In addition, an open, standard, vendor-independent and technology-agnostic southbound interface allows to configure and monitor the underlying physical devices composing the hybrid data center network. It is important to note that the term northbound interface (NBI) refers to any interface in a system used to communicate with higher layer systems and applications, as shown in Fig. 8.3. Similarly, southbound interfaces are defined to communicate a system with lower level systems and devices.

The finalized LIGHTNESS SDN control plane architecture is shown in Fig. 8.3. It is composed by an SDN controller natively implementing a set of basic network control functions and protocols, which are crucial to meet the requirements for data center services and applications.

A component that differentiates the LIGHTNESS SDN control plane from other SDN approaches is the virtualization manager (highlighted in red in Fig. 8.3). This module enables multi-tenancy within the hybrid data center by provisioning virtual DCNs. It is introduced as a new base service in the SDN controller, sitting on top of the resource manager and directly using the abstracted view of the hybrid optical data center network the resource manager exposes. In particular, the virtualization manager allows users to create their own virtual topologies on demand based on their specific QoS requirements.

Also, by taking advantage of the SDN approach, any features or functionalities supported by the LIGHTNESS control plane can be easily extended and implemented as network applications running on top of the SDN controller. As an example, a VDC composition application is proposed to enhance the aforementioned LIGHTNESS architecture to enable users to specify and dynamically modify their virtual DCNs (network application highlighted in red, top right of Fig. 8.3).

The VDC composition application is the component that implements the logic and the intelligence to interface with core virtualization manager for DCN virtualization. This means that any virtualization algorithm can run inside this VDC component, while the QoS provisioning and guarantee are provided by the virtualization manager. In particular, various algorithms and procedures for the VDC allocation are designed and evaluated in LIGHTNESS, e.g., a static VDC alloca-

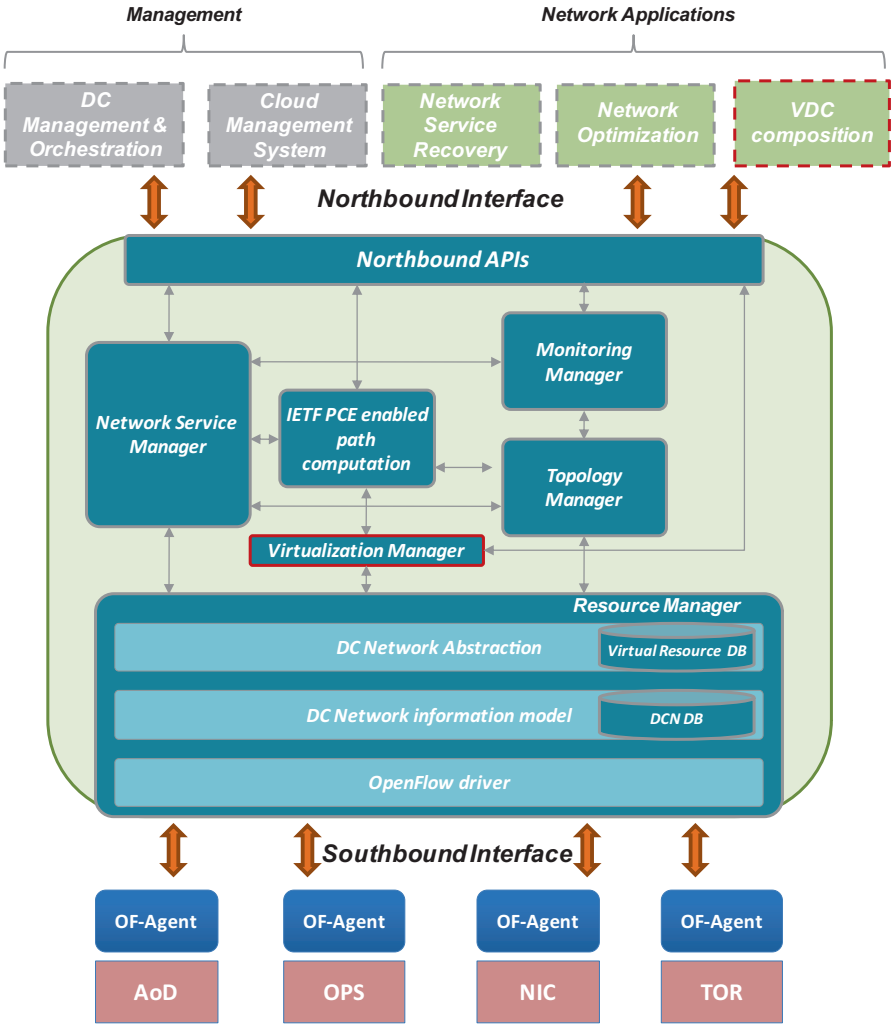

Fig. 8.3 High-level LIGHTNESS SDN-based control plane architecture

tion and a dynamic virtual slice allocation algorithm. In the first case, the minimization of the number of optical transponders is attained (DCN's CAPEX optimization), while in the second case, the number of blocked VDC allocation requests is evaluated. The main rationale behind such algorithm designs is the evaluation of the performance (in terms of allocated VDCs) of the LIGHTNESS architecture based on the hybrid data plane.

8.4 Technology Enablers for Flexible OCS/OPS

8.4.1 FPGA-Based Network Interface Card (NIC)

For DCNs, such as cloud data centers, special focus is needed on improving the intra-rack communication performance. The programmable NIC is designed and implemented to enable high-bandwidth and low-latency intra-rack communication and further empowers a flat and scalable all-optical data center inter- and intra- cluster architecture. The programmable NIC plugged directly to the server replaces the commodity NIC and enables intra-rack server-to-server full-mesh interconnection. The NIC is designed and implemented based on high-speed FPGA platforms [6] and optoelectronic transceivers. The 10x10G transceiver interfaces can be anything from ready off-the-shelf components (SFP+, CFP2, CFP4, etc.) or custom-made integrated PICs based on silicon photonics or III–V materials. In particular, hardware programming, framing, and processing methods are adopted for ultra-low latency processing and switching, as well as traffic aggregation techniques for maximum capacity handling. The traffic generated from servers is dissected over standardized network layer protocols (e.g., Ethernet) and then allocated into optical packets with the lowest possible processing delay to preserve ultra-low latency communication. It is capable of switching among multiple technologies, such as OCS and OPS, in a hitless manner, to achieve discrete bandwidth granularities. Most importantly the proposed NIC is also able to change between many-to-many aggregation mode and high-throughput point-to-point mode with little or no disruption to running applications. By using the programmable hybrid OCS/OPS NIC, the traffic features related to the OPS operation become flexible and programmable, allowing repurposing of the synchronous time-slotted mode OPS function to request and impose different levels of quality of service (QoS). These features include (a) variable optical packet size with (b) variable payload and overhead and are implemented as programmable network

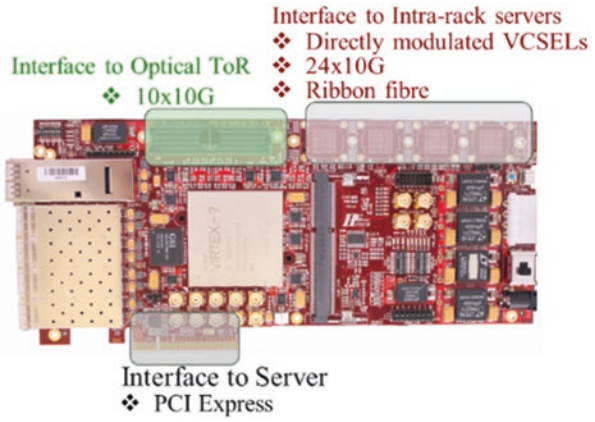

Fig. 8.4 FPGA-based network function programmable interface card (NIC)

functions rather than rigid hardware. All the above programmable features and FPGA hardware implementation are effectively exposed to the SDN-based control plane (Fig. 8.4).

8.4.2 OPS Module

A well-equipped 4 × 4 prototype integrating optical label processor (LP), optical switching fabric, and switch controller has been completed in the LIGHTNESS framework. As depicted in Fig. 8.5a, the proposed modular architecture allows the 4 × 4 OPS prototype to logically perform as two 2 × 2 OPS switches. The electrical label bits generated by each NIC are encoded in an in-band optical RF tone label [7, 8] by a prototyped label generator. The in-band optical labels are then coupled to each of the optical packets. Due to the lack of an optical buffer, a copy of the transmitted packet is stored, and a fast optical flow control between the OPS and the NIC is implemented for packet retransmission in case of contention. Figure 8.5b shows the photos of the 4 × 4 OPS prototype. The optical label of each packet is filtered out by a fiber Bragg grating (FBG) with a narrow passband. It is then detected and processed by the LP, and the recovered label bits are sent to the switch controller. The payload is fed into a 1 × N SOA-based broadcast and select stage. The switch controller checks the possible contention and, according to the lookup table (LUT), configures the 1 × N switch to forward the packets to the destination. The LUT can be remotely configured through the SDN-based control interface and can be seamlessly updated when necessary. In case of contention, the low-priority packet will be blocked, and a fast optical flow control signal (negative ACK) generated by using a low-speed

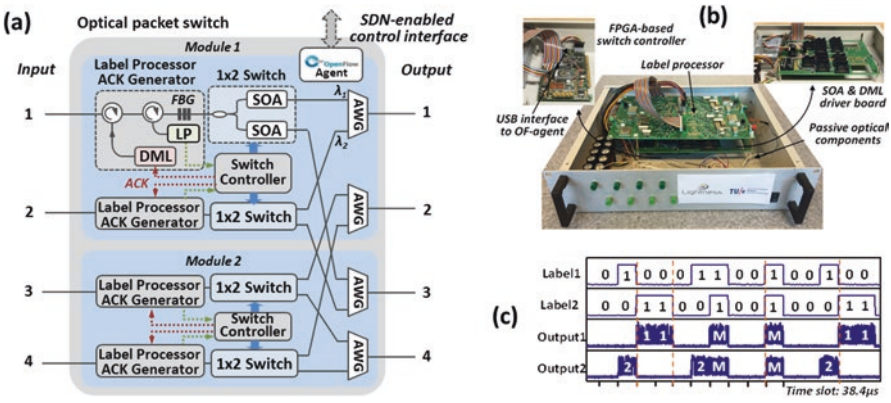

Fig. 8.5 (**a**) Schematic of the SDN-enabled OPS, (**b**) 4 × 4 OPS prototype, and (**c**) time traces of labels and OPS switch outputs for normal switching operation and multicasting. "*1*" and "*2*" indicate whether the present optical packet (*blue*) is switched in port No. 1 or 2. "*M*" indicates "multicasting," and thus the packet is present in both output ports of the switch

directly modulated laser (DML) is sent back to the NIC to request for the retransmission. If no contention occurs, a positive ACK informs the NIC to remove the packet from the buffer. As shown in the time traces of Fig. 8.5c, depending on the combination of the values of the labels, different (or no) outputs are activated for each 38.4 μsec timeslot. Multicasting is triggered when two label bits have been set as "11."

The fast response of the SOA in combination with the parallel processing of the label bits allows 20 nanoseconds switch reconfiguration time regardless of the port count. Besides the re-configuration of the LUT, the SDN-enabled control interface also enables programmable and flexible access from the SDN controller including the monitoring of the statistics.

8.5 Experimental Demonstration and Evaluation

The proposed fully dynamic all-optical circuit and packet-switched experimental data plane is able to carry out unicast/multicast switchover on request, while the powerful control plane enables the abstraction and virtualization of the networking resources. Thus, virtual data centers (VDCs) and virtual network functions (VNFs) are created on top of the data plane infrastructure. We have experimentally demonstrated practical intra-DCN interconnection use cases with deterministic latencies for both unicast and multicast, exhibiting monitoring and database transfer scenarios, each of which is facilitated by a joint software element based on the NFV and SDN principles. The outcomes validate a fully working thorough unification of the advanced optical data plane with the SDN-based control plane, committing to more efficient management of the forthcoming data center's compute and network resources.

8.5.1 Overall Experimental Architecture

The introduced high-level architecture, previously shown in Figs. 8.2 and 8.3, displays a next-generation fully reconfigurable DCN relied upon both optical circuit and optical packet switching technologies. Now for the overall experimental DCN design, as illustrated in Fig. 8.6, server blades within individual racks are connected via dedicated optoelectronic interfaces and all-optical ToR switches to the rest of the programmable DCN. The FPGA-based NICs operate as SDN-enabled hybrid OCS/OPS interfaces that support dynamic composition and transmission of Ethernet frames and/or optical packets with associated labels [9]. In order to provide direct OCS multicasting capabilities in each rack separately, optical power splitters are applied at each of the optical ToRs.

We propose a large port-count space/fiber switch as an optical ToR, because when combined with the advanced NICs on each server, they prevent the use of power-hungry electronic packet ToR switches (EPS) with unregulated latency values.

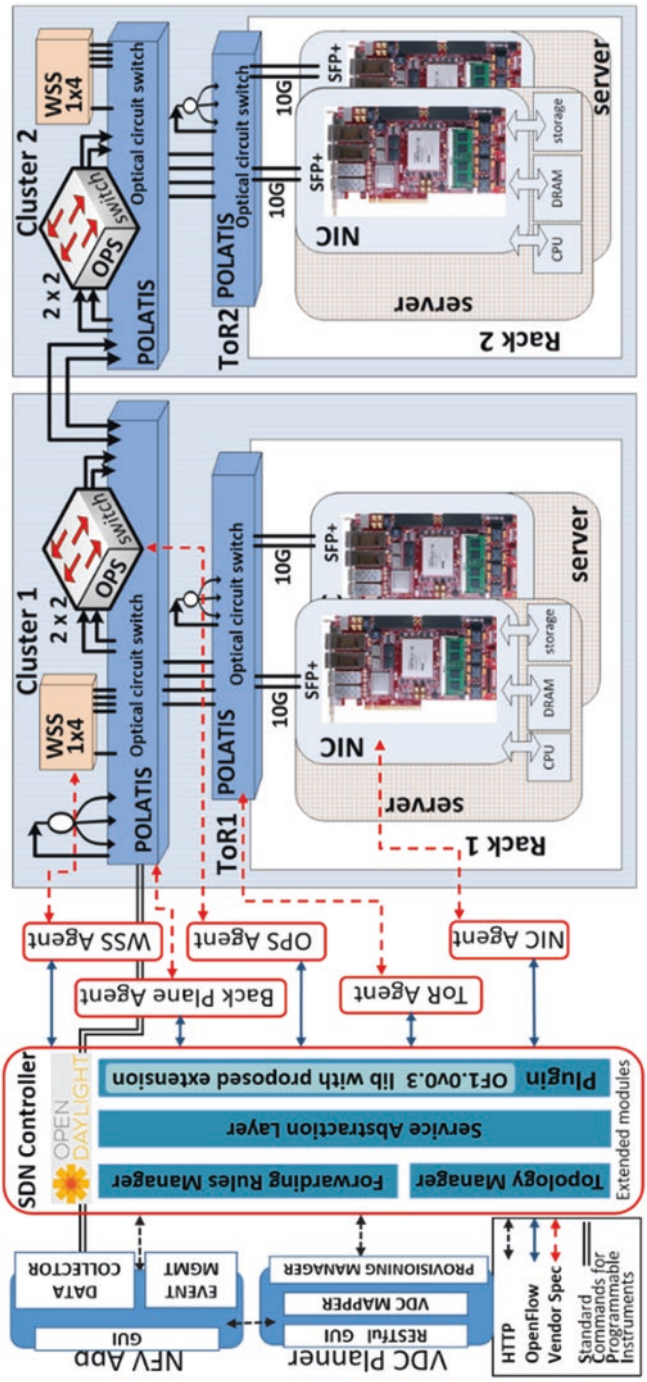

Fig. 8.6 Overall data center network architecture; experimental data plane (*center-right*), control plane (*left*), and virtualization schemes (*far left*)

Therefore, this flat design offers fixed low interconnection latency and potential cutback in power consumption of the overall network, due to the eventual lack of repeated O/E/O conversions. It also establishes full bandwidth transparency, since optical switches are entirely agnostic of the actual operated channel data rate, network protocol, or optical modulation format.

Within each cluster, the optical ToRs are linked to a top-of-the-cluster (ToC) optical switch, as shown in Fig. 8.6. The ToC incorporates a modular flexible optical network, containing a high-radix optical switch, serving as the optical backplane, optical power splitters, a wavelength/spectrum selective switch (WSS/SSS), and a 2×2 optical packet switch. The optical splitter at this network position enables OCS multicasting among various servers of separate racks within the same cluster and between remote clusters. The WSS's role is to groom/distribute multiple channels and traffic for inter-cluster communication in an elastic way. A passive optical filtering device, such as an AWG, could operate comparable jobs and be deployed instead of the WSS. However, a WSS is preferred due to its increased reconfigurability, flexibility, and bandwidth adaptability. A WSS/SSS could easily adapt to future possible bandwidth allocation needs with higher spectrum efficiency and narrower channel spacing than the present standardized ones.

The optical backplane also encompasses SDN-enabled OPS switching nodes that can perform nanosecond-fast packet switching, multicasting, as well as supporting of monitoring capabilities for optical packet reception and contention. To accomplish forwarding operation in the nanosecond scale, the OPS deals with the optical packets according to the optical label sent by the NIC [10], while the SDN controller has the role to arrange and supply the lookup tables to the OPS and NIC modules. This favors the decoupling of the fast (nanoseconds) forwarding operation of the optical data plane (to support time domain fast statistical multiplexing capability), from the slower SDN control plane (milliseconds), DCN virtualization, and VDC planner application.

Exploiting statistical multiplexing, OPS can also provide efficient and flexible bandwidth utilization therefore lowering the required optical port count at the backplane while guaranteeing the appropriate connectivity. In conjunction with the nanoseconds-fast label detection and switching control, bursty traffic demands are handled with higher degree of bandwidth granularity, lower latency, and adjustable per-packet processing agility.

Each of the NICs, optical ToRs, OCS and OPS switching nodes, and WSS switches is utterly controllable by the centralized SDN controller through a consistent control software interface, featured by each device's dedicated agent, as shown in Fig. 8.6 left. The agents abstract essential information from the hardware network modules, keeping the SDN controller up to date with every instance of the network, while they also translate and forward the management commands to the physical layer devices. Furthermore, logically on top of the SDN controller (and his left in Fig. 8.6), the VDC planner and the NFV applications provide an extra layer of abstraction and virtualization of the deployed physical computing and network infrastructure.

In summary, the programmable data plane facilitates the SDN-based DCN control plane to form and amend the physical layer topology, by flexibly arranging the relevant cross-connections in the optical backplane to suit the various applications' demands. Additionally, based on the DCN's specific requirements and data flow

durations each time, the FPGA-based hybrid OCS/OPS NIC can be set up by the SDN controller on request along with the optical ToRs, ToCs, and OPS switches, achieving also unicast and/or multicast communication among servers.

8.5.2 All-Optical Experimental Data Plane

The data plane test-bed used for the experimental studies includes four rack-mounted Dell PowerEdge T630 servers, each supplied with a state-of-the-art FPGA-based NIC board with 10G SFP+ transceivers, serving as the reconfigurable interface of the computer blades to the optical network [9]. These servers were populated by miscellaneous virtual machines (VMs), one of which also hosts the SDN controller. All servers are joined to the 192 × 192 port optical circuit switch (supplied by Polatis), which in our experiment acts both as a ToR switch and as the OCS backplane on top of each cluster. Polatis beam-steering switching technology adds around 1 dB of loss per optical cross-connection (OXC), so multiple OXCs and hops are sustainable without extensive power and signal integrity penalties. As mentioned above in the overall architecture description, a 1 × 4 optical power splitter, two 1 × 4 WSS, and one SOA-based 4 × 4 OPS [10] are also attached to the optical backplane.

The novel NIC's range of capabilities combines network interface functions, programmable aggregation and segregation duties, OCS/OPS switching, and layer 2 switching services. The FPGA-based hybrid OCS/OPS NIC has been implemented on top of the NetFPGA-SUME development board [11, 12] and has been constructed to fit directly into a server's motherboard by replacing the conventional NIC. In the original design, it has an eight-lane Gen3 PCI Express interface for DRAM communication, one 10 Gb/s dual-line optical interface for receiving instructions from the SDN control agent and forwarding any feedback, two OCS/OPS hybrid 10 Gb/s SFP+ ports for inter-server communication, and an OPS label pin interface connected to the OPS label generator. The SFP+ transceivers' frequencies are selected in the C-band and are ITU compatible, in order to be consistent with the LCoS-based WSS and SOA-based OPS modules, which both normally operate in the 1550 nm frequency region.

The 1 × 4 optical power splitter supports any OCS one-to-four multicasting schedules. The WSSs are mostly operated for merging inter-cluster (or even inter-DC) traffic carried by channels from different servers or racks into a WDM superchannel. In the destination cluster, the local WSS separates the original WDM channels and properly routes them to the receiving racks and servers.

8.5.3 SDN-Enabled Experimental Control Plane

For these experiments, OpenFlow (OF) was selected as the standard control plane protocol to communicate the network devices with the controller. OpenDaylight (ODL) was operating as the SDN controller, and OF agents for Polatis, WSS, OPS

switch, and hybrid OCS/OPS NIC were implemented to enable SDN-based programmability, as shown in Fig. 8.6. Further to the OF extensions, as previously reported in [6], the NIC OF agent is additionally enhanced to allow rearrangement of the generated OPS packets' duration. Moreover, ODL internal software elements were extended to provide some new network device-specific characteristics. For instance, in relation to the OPS and WSS ports, the switch manager and service abstract layer (SAL) were further developed to capture the supported wavelength and supported spectrum range, respectively, both of which were then used to validate the current configuration. Furthermore, the statistics of transmitted optical packet were collected and preserved by the statistics manager. In order to program accordingly the deployed optical devices, the forwarding rules manager has been also extended to build the appropriate set of configuration details. For example, for the OPS switch, label and output port details were included. For the WSS, central frequency, bandwidth, and output port information were contained in the forwarding rules manager extensions, whereas traffic matching to OCS or OPS, label, and output port information were included regarding the NIC. More specifically, the FPGA-based hybrid OCS/OPS NIC communicates with the OF agent through a bidirectional 10Gbps SFP+ Ethernet interface. The commands and information are encapsulated in a predefined 1504-byte Ethernet frame. Furthermore, through an extended ODL northbound interface, various applications can communicate directly with the hardware using the widely supported representational state transfer application programming interface (RESTful API).

8.5.4 Optical Data Center Virtualization Demonstration

As experimental objectives, two control plane operations have been produced and positioned on top of the ODL: a virtual data center planner (VDC planner) and a virtual network monitoring function (monitoring VNF).

First, the VDC planner allows the composition and arrangement of virtual network slices within the DCN, thus empowering multi-tenancy characteristics in data centers. In LIGHTNESS demonstration, the VDC planner consists of a graphical user interface (GUI) implemented in HTML/JavaScript that interfaces with a backend application developed in Python 2.7, which is capable of direct interaction with the ODL controller. The potential DC manager or user has access to the GUI with any existing browser, while he can create a VDC request, as shown in an example of Fig. 8.7. The specifiable VDC creation parameters are (i) number of servers to be allocated, (ii) optical links to be established, (iii) the preferred optical interconnection technology (OCS/OPS) for each link, and finally (iv) advice if there are any multicast properties for the allocated linked servers. Other parameters that are available (not mandatorily applied for the present VDC creation algorithm) are the required interconnection bandwidth and the bidirectionality of a given link.

Once the application has received the set of the abovementioned parameters for the VDC and has generated a group of static control flows to be forwarded to the

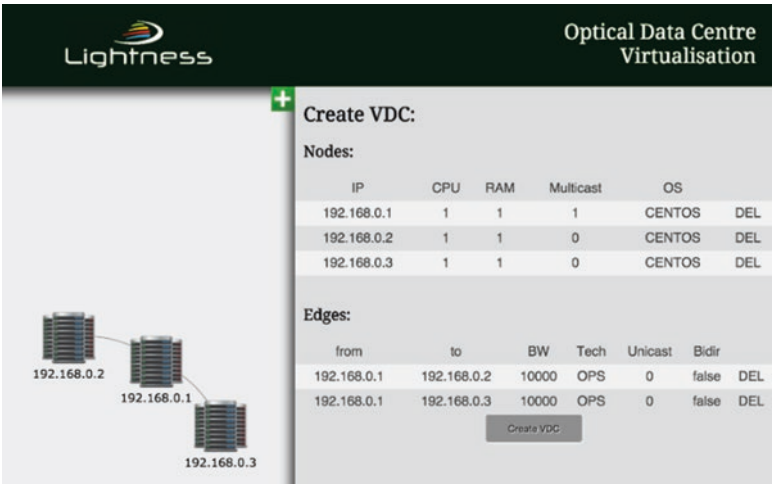

Fig. 8.7 VDC planner application with OPS or OCS multicasting options

DCN devices by ODL, it distributes them among the various data plane modules (NIC cards, OPS switches, WSS, and OCS backplane). This set of flows is produced in JavaScript Object Notation (JSON) format and sent to the ODL controller via a RESTful API.

Figure 8.7 shows an example of a VDC request using the aforementioned VDC planner. The user defines, in this example, three server hosts, one of them chosen as a multicast node, which will duplicate the content to the other two. In this case, the request also specifies OPS as the multicast technology for the VDC interconnection.

8.5.5 Experimental Results and Evaluation

For the experimental demonstration, we incorporated all the available data plane and control plane resources, as presented in the previous sections, and evaluated several intra-DCN interconnection schemes, based on VDC applications' and VNF's requests and commands.

First, for the physical layer the DCN was comprehensively evaluated for intra-rack, inter-rack, and inter-cluster unicast and multicast communication by measuring the bit error rate (BER) for both OCS and OPS switching technologies with realistic traffic (scrambled PRBS payload from the traffic analyzer). The results are presented in Fig. 8.8. The traffic analyzer provides the FPGA-based NIC with 10 Gb/s Ethernet traffic, and then the NIC forwards the data to one of its hybrid OCS/OPS ports. When OPS mode is selected, the NIC, relying on the configuration instructed by the SDN controller, sets the optical packet duration, encapsu-

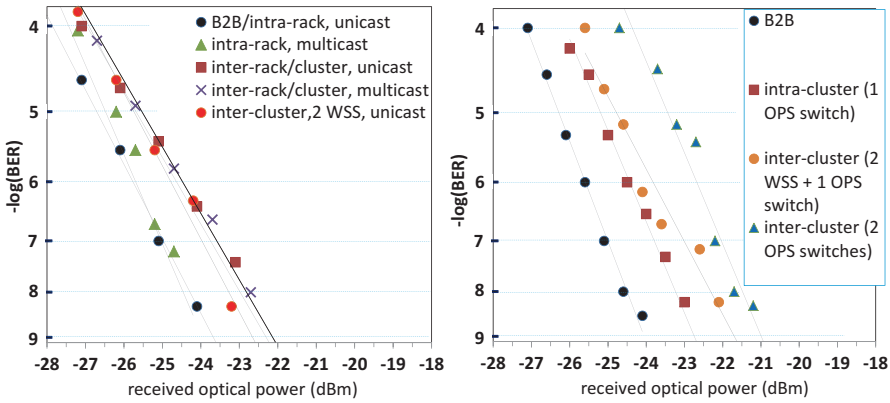

Fig. 8.8 (*Left*) OCS BER curves for intra-/inter-rack unicast/multicast and inter-cluster through 2 WSSs with 10-Gigabit Ethernet (GbE) traffic. (*Right*) OPS BER curves for intra-cluster and inter-cluster through WSS and 1 or 2 OPS switches with 10GbE traffic

lates a certain number of Ethernet frames, and issues the optical packet, while the packet label is generated and integrated in parallel. On the receiver end, in the lack of a burst mode receiver and in order to properly recover the clock and lock the data, a preamble between 10us and 30us was used, depending on the desired quality of transport.

Intra-rack communication is achieved by pushing optical flows from source to destination server via the optical ToR switch for unicast and through an optical splitter for multicast operation. Inter-rack and inter-cluster communications are identically accomplished by going through multiple Polatis OXCs and/or optical power splitters. For inter-cluster groomed interconnection, optical signals have to propagate through two extra WSSs for WDM mux/demux and switching functions. Small penalties of <2 dB are observed for all OCS interconnection scenarios (unicast/multicast), as shown in the BER curves of Fig. 8.8 (left).

OPS BER plots in Fig. 8.8 (right) indicate 1 and <3 dB penalties when signals pass through one (for intra-cluster) and two (for inter-cluster) switches, respectively.

Following the BER test of the physical links, we collected network layer 2 results regarding the chip-to-chip interconnection latency. This is the access latency between one NIC's direct memory access (DMA) and the destination NIC's DMA. DMA driver's actual delays are excluded and are separately measured later for different DMA lengths and fixed Ethernet frames. In addition, interconnection throughput is monitored and plotted, exhibiting OCS-to-OPS switchover and vice versa.

We measured the DMA-to-DMA access latency with Ethernet traffic generated from the traffic analyzer with various PRBS payloads. The traffic analyzer firstly feeds the transmitting FPGA-based NIC with the traffic. Then, the NIC pushes the data flows to the all-optical network, employing either OCS or OPS, toward the destination NIC. Optical signals traverse through the DC network and the appropriately established cross-connections. Finally, the destination NIC forwards the received

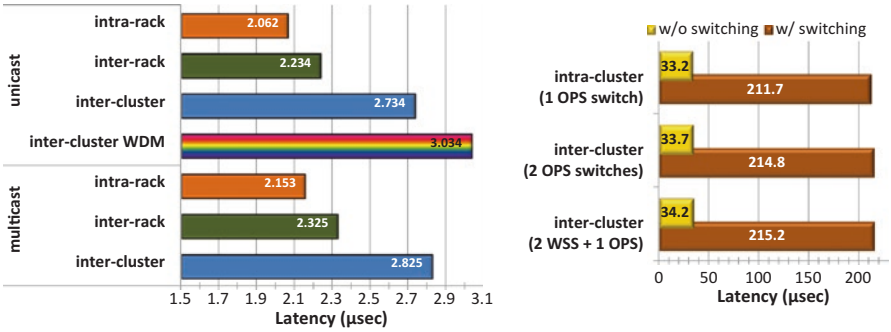

Fig. 8.9 DMA-to-DMA OCS (*left*) and OPS (*right*) latency for various intra-DC scenarios

traffic flows back to the traffic analyzer. The accumulated chip-to-chip latency was estimated by subtracting the traffic analyzer-to-NIC (and vice versa) delays and the traffic analyzers' processing delays (a few hundreds of microseconds).

The displayed investigation and measurements are based on the best possible latency with maximum bitrate, so, for OPS with switching, actual throughput bitrate is around 3 Gb/s, and for OCS it is around 8 Gb/s. All quantified latency values include FPGA physical and logic delays, which can strongly vary depending on the used frame length, selected transmission/switching scheme (OCS or OPS), and the FPGA design.

Figure 8.9 (left) shows unicast and multicast OCS access latencies for all the studied interconnection scenarios. Predictable latency values are exhibited between 2 and 3 μsec for most communication scenarios. The majority of the latency is contributed by the electronic processing (PHY and logic) of the Ethernet traffic in the source and destination FPGA-based nodes.

Figure 8.9 (right) shows intra-/inter-cluster OPS access latencies with and without switching. When no switching is performed, the clock of the receiving end of the transceiver is continuous, so there is no need for recovering it with extra payload (e.g., preamble dummy key characters).

Figure 8.10 (right) depicts the variations of interconnection throughput when a change from normal operation OCS to OPS and vice versa is initiated by the NFV application and executed in the data plane. Protocol overheads are restricting throughput in OCS whereas in OPS the dummy key characters used for receiver side and clock recovery are confining the maximum throughput, plus the fact that OPS traffic is transmitted and switched in a 50% manner for this experiment.

8.6 Conclusion: Discussions

Regarding the energy efficiency of the proposed architecture, it is widely known and proven that optical network modules deliver higher port count with fixed power consumption and non-limiting switching capacity. On the contrary,

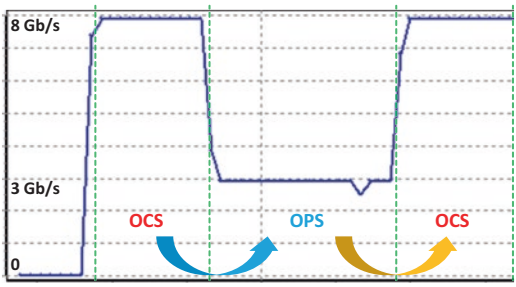

Fig. 8.10 Throughput plot over time, illustrating the OCS-to-OPS switchovers and vice versa

electrical network elements usually provide fewer ports with restricted switching capacity and significantly higher power consumption values, which can also vary counting on the switching traffic load. The predominant reason why all-optical switching is more energy efficient than electrical switching is due to the lack of optical transceivers, which count for more than 50% of the total power expenditure. Optical switching devices used in this experiment show much more modest power consumption (some tens of Watts) than regular electrical switches (several hundreds of Watts). More specifically, the 192 × 192 OCS switch consumes 75 Watts in regular operation, the 1 × 4 SSS consumes less than 10 Watts, while the total power consumption of the 4 × 4 SOA-based OPS prototype is around 50 Watts. OPS's breakdown energy contributions are FPGA controller 15 Watts, label processor 20 Watts, and SOA driver 15 Watts. Those values are based on off-the-shelf components, and further power reduction is possible by dedicated hardware photonic integration. Furthermore, recent experimental and simulation research [13, 14] have shown considerable distinction in terms of energy efficiency between architectures using all-optical switching and in others using conventional electrical switching equipment.

Lastly, with regards to the performance of the demonstrated VDC and NFV, not only the total (re)configuration times but also the contribution of each individual element was calculated. The total OCS/OPS channel configuration span includes (i) the ODL SDN controller processing time, (ii) control message transmission time (which strongly depends on the actual experiment setup), and (iii) the device reconfiguration time. Particularly in this experiment, the SDN controller needs around 210 msec to process requests arriving from the RESTful API in order to push the matching OF configuration commands to the network devices' OF agents. It approximately lasts a further 200 msec for those commands to reach the OF agents, to be processed and forwarded. At last, Polatis OCS switch, OPS switch, WSS, and NIC require approximately 16, 10, 300, and 18 msec, respectively, to properly configure themselves. The above device configurations of course can be carried out in parallel. So, assuming that in order to establish an end-to-end OCS channel we need to successfully configure the optical ToR before configuring NIC, establishing an OCS channel will need 970 msec (also using the WSS) or 690 msec without WSS, while it takes around to 420 msec for OPS connection establishment.

This chapter demonstrates an all-optical programmable DCN architecture enabling OCS/OPS multicasting for realistic monitoring, migration, and transferring scenarios. The novel networking schemes demonstrated in this chapter include an SDN-enabled, virtualize-able, and reconfigurable optical data plane integrated and supported by an extended control plane. In this work, the SDN controller and NFV server are able to offer data plane monitoring and database migration function virtualization, on top of a virtual data center environment implemented and managed by a VDC planner application.

It is apparent that there is a trend for all-optical switching in DCNs in order to tackle the disadvantages of current architectures, exactly as it was done with metro and core networks a couple of decades ago. However, the requirements of those two categories of networks are very different. Hence, WDM technologies commercially available and suitable for metro core and regional networks cannot be introduced in the intra-DCNs without further modification. The introduction of space division multiplexing (SDM) [5] and the latest advances in photonic integration will play a critical role in the future development of intra-DCN architectures and designs, by attempting to exploit the space dimension, in addition to frequency and time, and by bringing massive manufacturing costs of optical components down, while further improving their energy efficiency.

References

1. 'Lightness EU FP7 project'. [Online]. Available: http://www.ict-lightness.eu/
2. G.M. Saridis et al., Lightness: A Function-Virtualizable Software Defined Data Center Network With All-Optical Circuit/Packet Switching. J. Lightwave Technol. **34**(7), 1618–1627 (Apr. 2016)
3. G. M. Saridis et al., 'LIGHTNESS: A Deeply-Programmable SDN-enabled Data Centre Network with OCS/OPS Multicast/Unicast Switch-over', in *European Conference on Optical Communication (ECOC)*, Valencia, 2015, p. PDP 4.2
4. N. Amaya, G. S. Zervas, D. Simeonidou, 'Architecture on demand for transparent optical networks', in *2011 13th International Conference on Transparent Optical Networks*, 2011, pp. 1–4
5. G. M. Saridis, D. Alexandropoulos, G. Zervas, D. Simeonidou, 'Survey and Evaluation of Space Division Multiplexing: From Technologies to Optical Networks', *IEEE Communications Surveys Tutorials*, vol. 17, no. 4, pp. 2136–2156, Fourthquarter 2015
6. B. Guo et al., 'SDN-enabled programmable optical packet/circuit switched intra data centre network', in *Optical Fiber Communications Conference and Exhibition (OFC), 2015*, 2015, pp. 1–3
7. W. Miao, F. Yan, H. Dorren, N. Calabretta, 'Petabit/s Data Center Network Architecture with Sub-microseconds Latency Based on Fast Optical Switches', in *European Conference on Optical Communication (ECOC)*, Valencia, 2015
8. W. Miao, X. Yin, J. Bauwelinck, H. Dorren, N. Calabretta, 'Performance assessment of optical packet switching system with burst-mode receivers for intra-data center networks', in *2014 European Conference on Optical Communication (ECOC)*, 2014, pp. 1–3
9. 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 *European Conference on Optical Communication (ECOC)*, Valencia, 2015
10. W. Miao et al., SDN-enabled OPS with QoS guarantee for reconfigurable virtual data center networks. IEEE/OSA Journal of Optical Communications and Networking **7**(7), 634–643 (Jul. 2015)

11. 'NetFPGA-SUME Virtex-7 FPGA Development Board', *Digilent*. [Online]. Available: http://store.digilentinc.com/netfpga-sume-virtex-7-fpga-development-board/
12. 'NetFPGA'. [Online]. Available: https://netfpga.org/site/#/systems/1netfpga-sume/details/
13. M. Imran, P. Landais, M. Collier, K. Katrinis, 'A data center network featuring low latency and energy efficiency based on all optical core interconnect', in *2015 17th International Conference on Transparent Optical Networks (ICTON)*, 2015, pp. 1–4
14. Y. Ji et al., All Optical Switching Networks With Energy-Efficient Technologies From Components Level to Network Level. IEEE Journal on Selected Areas in Communications **32**(8), 1600–1614 (Aug. 2014)