Real time demonstration of an end-to-end optical datacenter network with dynamic bandwidth allocation

K. Tokas(1), C. Spatharakis(1), I. Patronas(1,3), P. Bakopoulos(2), G. Landi(4), K. Christodoulopoulos(5), M. Capitani(4), A. Kyriakos(1,3), M. Aziz(6), R. Pitwon(7), D. Gallico(8), D. Reisis(1,3), E. Varvarigos(5), E. Zahavi(2), H. Avramopoulos(1)

(1) Photonics Communications Research Laboratory, National Technical University of Athens, Greece; (2) Mellanox Technologies, Yokneam, Israel; (3) Electronics Laboratory, National and Kapodistrian University of Athens, Greece; (4) Nextworks, Pisa, Italy; (5) Communication Networks Laboratory, University of Patras, Greece; (6) Gesellschaft für wissenschaftliche Datenverarbeitung mbH, Göttingen, Germany; (7) Seagate, United Kingdom; (8) Interoute S.p.A, Roma, Italy;

ktok@mail.ntua.gr

Abstract *We demonstrate combined operation of a scalable optical data-plane architecture with an SDN overlay capable of slotted operation for dynamic allocation of network resources. Real-time end-to-end functionality is verified in a six-host prototype datacenter cluster.*

Introduction

Datacenter networks (DCNs) traffic is riding on a steep growth curve reaching 27% annually[1] while the industry grapples to keep pace with this skyrocketing demand that outpaces current established technologies and challenges their bandwidth and energy scalability. We have currently reached the point where the amount of new data generated is larger than our processing capabilities, a situation that is often referred to as the “data deluge”. In this backdrop, new technologies are investigated in order to avoid an imminent capacity crunch and improve power efficiency in the DCN. Optical switching is gaining traction as a promising path for sustaining the explosive growth of DCNs; however, its practical deployment necessitates extensive modifications to the network architecture and operation, tailored to the technological particularities of optical switches (i.e. no buffering, limitations in radix size and speed). European project NEPHELE is developing an optical network infrastructure that leverages optical switching within a software-defined networking (SDN) framework to overcome the bandwidth and energy scaling challenges of datacenter networks[2-4]. NEPHELE relies on commercial off the shelf (COTS) photonic components for its data plane in order to expedite its deployment cycle while on the same time leverages open source control plane frameworks to maximize compatibility with existing infrastructures. The overall architecture of NEPHELE was recently published[2]. In the current work, we report the vertical integration of a NEPHELE prototype datacenter cluster and demonstrate its real-time operation, enabling error-free server-to-server communication for various traffic scenarios, underpinning the feasibility of the architecture.

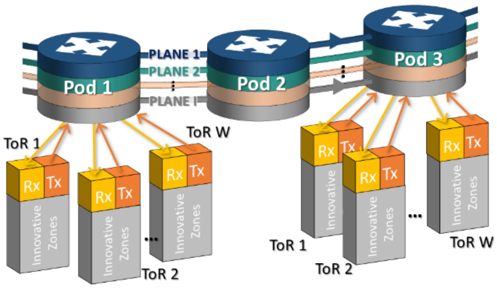


Fig. 1: The NEPHELE network architecture[5]

Overview of the NEPHELE architecture and demonstrator testbed assembly

The NEPHELE network architecture relies on optically interconnected PODs, each accommodating a number of racks. Each rack is administered by a top-of-rack (ToR) switch and consists of several hosts (i.e. disaggregated storage and compute resources, placed in “innovation zones”). The ToRs are connected to the POD switch in a star topology as shown in Fig. 1. Each ToR is equipped with tunable lasers and can address the remaining ToRs by properly tuning the transmitted wavelength. Scaling the dimensions of the network is achieved by interconnecting multiple PODs in a DWDM ring. Inter-pod traffic is routed by means of fast wavelength selective switches (WSSs) placed in each POD switch, allowing wavelength reuse among PODs, and thus enabling network scalability beyond the typical wavelength count of DWDM systems. Moreover, multiple planes serve as parallel optical paths interconnecting the PODs, further scaling the overall throughput of the network. The NEPHELE network relies on time-division multiple-access (TDMA) and the resulting slotted operation of the network enables dynamic and efficient sharing of network resources and collision-free routing within the entirety of the network. Control of the NEPHELE data plane is obtained through the OCEANiA controller[6], an extended version of the open source OpenDaylight controller comprising a set of bespoke SDN applications, TDMA scheduling algorithms[7] and OpenFlow protocol extensions to enable slotted network operation[8]. Interaction between the

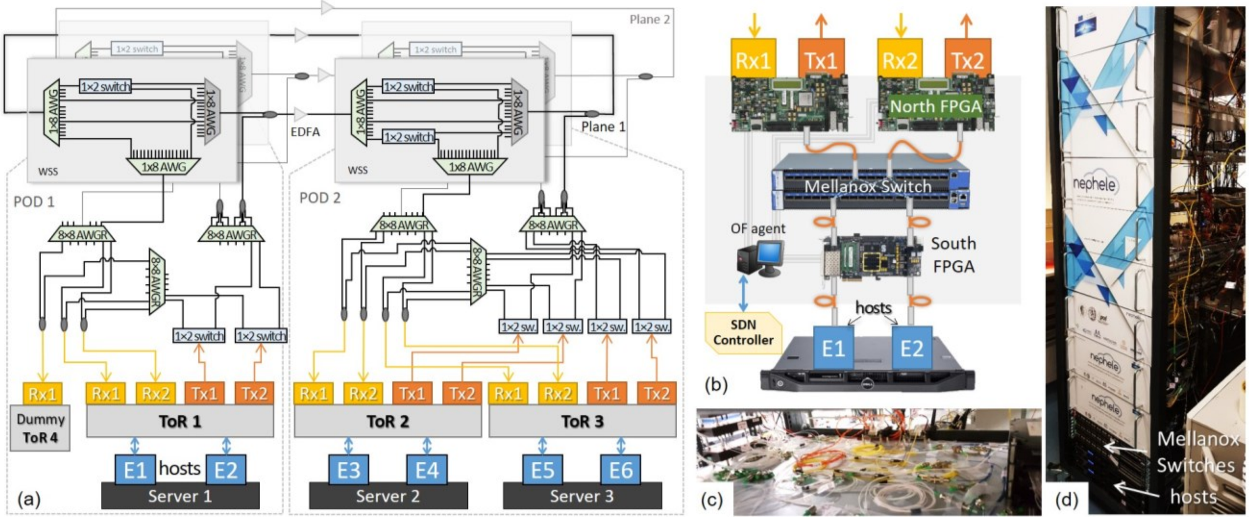


Fig. 2: (a) Schematic of the NEPHELE demo network setup, (b) Detailed schematic of the ToR implementation, (c) Photo of the actual WSS with the "demultiplex, switch and multiplex" approach inside the Nephele rack which is depicted in (d)

NEPHELE control plane and its slotted data plane is enabled with custom-made SDN agents, translating control plane commands into configuration of the photonic components (tunable transmitters and optical switches). Real time configuration of the optical data plane is facilitated with FPGA boards.

The NEPHELE demonstrator assembly is shown schematically in Fig. 2 (a). It implements two NEPHELE planes with two PODs and four ToRs (3 functional and 1 dummy ToR) serving 6 hosts that are emulated by three DELL servers, each one equipped with two independent interfaces. POD1 connects to a dummy ToR switch for signal monitoring purposes and ToR 1 which accommodates hosts E1 and E2. Furthermore, POD 2 serves 4 hosts; E3 and E4 through ToR 1, as well as E5 and E6 through ToR 3. To serve the two parallel planes, each ToR switch is equipped with two tunable transmitters and equal number of optical receivers controlled by the North FPGAs (N-FPGA). In this context, there are in total 3 fully equipped ToRs, i.e. 6 transmitter and receiver assemblies that were developed, tested and used for the demonstration purposes. Each POD switch accommodates two WSSs – (Fig. 2(c)) based on the "Demultiplex, switch and multiplex" approach, one for each plane. Moreover, optical amplifiers (EDFAs) were employed to amplify the optical signals in all optical light-paths between each POD and plane. The six hosts, located inside the racks, are directing the Ethernet traffic towards the corresponding NEPHELE ToR by means of two 10Gbps SFP interfaces. The NEPHELE ToR is shown in Fig 2(b) and its electrical part consists of a commercial Mellanox 10G Ethernet switch extended with FPGA boards responsible for controlling the tunable transmitter hardware as well as buffering and matching outbound traffic with the TDMA schedule, provided by the control plane's scheduler (North- and South-FPGA respectively). The ToR is connected via PCI/e to the SDN agents, running in PCs, enabling interaction with the control plane. Traffic generated at the hosts is forwarded to the S-FPGA extender and consequently to the Mellanox Ethernet switch that routes the frames to the available N-FPGA for transmission (Fig. 2 (b)); reception of the signals follows the opposite vertical direction. A photo of the NEPHELE cluster demonstrator assembled on a rack is depicted in Fig. 2(d).

Experimental Results

Standalone evaluation of NEPHELE subsystems has been previously demonstrated and discussed[3,5]. The current work, however, reports the successful vertical integration of all the layers and functional elements comprising the NEPHELE

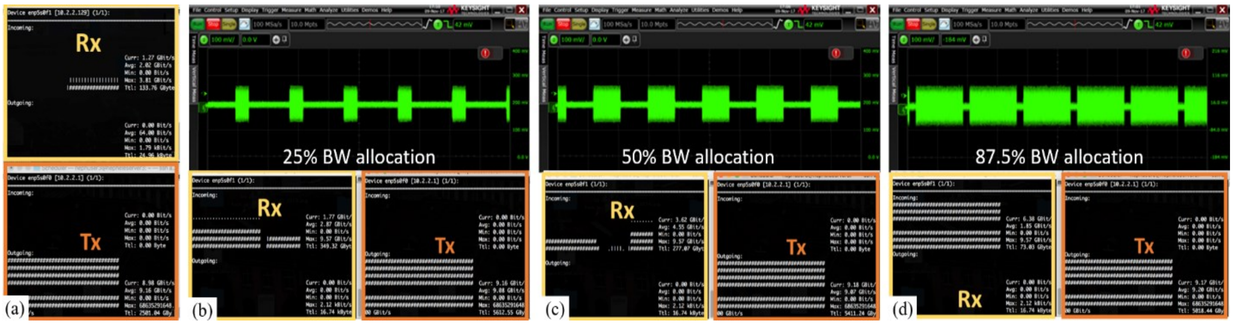
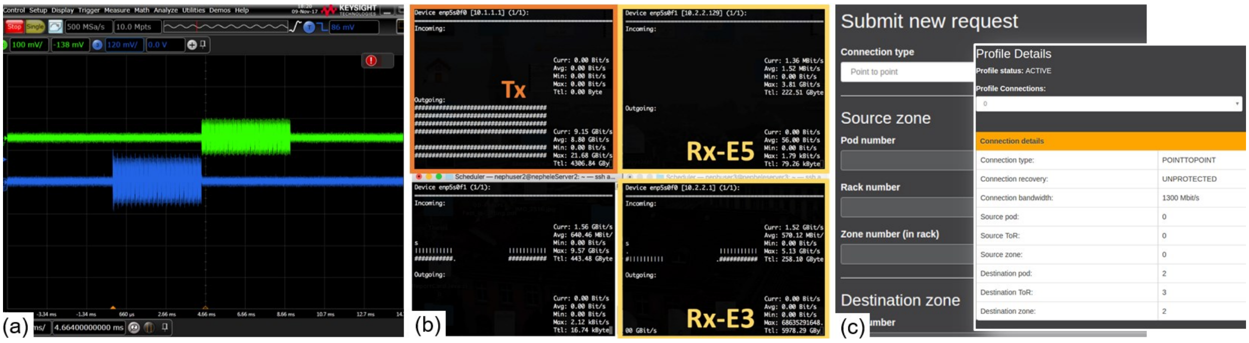


Fig. 3: (a) Screenshots of nload application for the intra-ToR scenario (2.2 GBit/s), (b) 25% bandwidth allocation for the intra-pod scenario, (c) 50% bandwidth allocation for the intra-pod scenario, (d) 87.5% bandwidth allocation for the intra-pod scenario



```
Device enp5s0f0 [10.1.1.1] (1/1):
Incoming:
                                  Curr: 0.00 Bit/s
                                  Avg: 0.00 Bit/s
                                  Min: 0.00 Bit/s
                                  Max: 0.00 Bit/s
                                  Ttl: 0.00 Byte
Outgoing:
                                  Curr: 9.15 GBit/s
                                  Avg: 8.80 GBit/s
                                  Min: 0.00 Bit/s
                                  Max: 21.68 GBit/s
                                  Ttl: 4306.84 GBy
```

```
Device enp5s0f1 [10.2.2.129] (1/1):
Incoming:
                                  Curr: 1.36 MBit/s
                                  Avg: 1.52 MBit/s
                                  Min: 0.00 Bit/s
                                  Max: 3.81 GBit/s
                                  Ttl: 222.51 GByte
Outgoing:
                                  Curr: 0.00 Bit/s
                                  Avg: 56.00 Bit/s
                                  Min: 0.00 Bit/s
                                  Max: 1.79 kBit/s
                                  Ttl: 79.26 kByte
```

Scheduler — nephuser2@nepheleServer2: ~ — ssh a...

```
Device enp5s0f1 (1/1):
Incoming:
                                  Curr: 1.56 GBit/s
                                  Avg: 640.46 MBit/
                                  Min: 0.00 Bit/s
                                  Max: 9.57 GBit/s
                                  Ttl: 443.48 GByte
Outgoing:
                                  Curr: 0.00 Bit/s
                                  Avg: 0.00 Bit/s
                                  Min: 0.00 Bit/s
                                  Max: 2.12 kBit/s
                                  Ttl: 16.74 kByte
```

```
Device enp5s0f0 [10.2.2.1] (1/1):
Incoming:
                                  Curr: 1.52 GBit/s
                                  Avg: 570.12 MBit/
                                  Min: 0.00 Bit/s
                                  Max: 5.13 GBit/s
                                  Ttl: 258.10 GByte
Outgoing:
                                  Curr: 0.00 Bit/s
                                  Avg: 0.00 Bit/s
                                  Min: 0.00 Bit/s
                                  Max: 68635291648.
                                  Ttl: 5978.29 GBy
```

Submit new request

Connection type: Point to point

Source zone

Pod number

Rack number

Zone number (in rack)

Destination zone

Profile Details

Profile status: ACTIVE

Profile Connections: 0

Connection details	
Connection type:	POINTTOPOINT
Connection recovery:	UNPROTECTED
Connection bandwidth:	1300 Mbit/s
Source pod:	0
Source ToR:	0
Source zone:	0
Destination pod:	2
Destination ToR:	3
Destination zone:	2

Fig. 4: (a) Screenshot of the traffic received by ToR2 (green trace) and ToR3 (blue trace), (b) nload application for the inter-pod scenario; bandwidth is divided equally to 1.52 GBit/s for each destination ToR, (c) OCEANiA - GUI of Application Affinity service.

architecture as well as the real-time, end-to-end operation of the DCN. Three separate traffic scenarios (intra-ToR, intra-pod, inter-pod) were implemented, applying the dynamic control of the optical switches and network elements of the DCN architecture and enabling the dynamic resource allocation.

The first scenario (intra-ToR) describes the communication between interfaces E3 and E4 (in Fig. 2(a)) that are both hanged to ToR 2. In this case, solely electronic switching is applied; E3 interface (10.2.2.1) generates Ethernet traffic that is fed to the S-FPGA and buffered in a DRAM before being forwarded to the Mellanox Ethernet switch. Consequently, the switch classifies the traffic using static MAC tables and then routes it through the S-FPGA and towards E4 interface according to the scheduling engine commands. Fig. 3 (a) shows the nload command running on the Rx (top) and Tx (bottom), an application that calculates the incoming and outward traffic of the hosts. The achieving bandwidth received by E4 (10.2.2.129) is 2.2 Gb/s.

As far as the intra-pod communication is concerned, the NEPHELE SDN controller is orchestrating the connection between interfaces E3 and E5 which are hosted in ToR 2 and ToR 3 respectively, both residing in POD 2. The Open-Flow agents are executing three separate SDN schedules, dynamically configuring the allocated bandwidth as shown in Fig. 3 (b-d). The traffic produced by E3 is forwarded through the S-FPGA and the Ethernet switch to Tx1 N-FPGA of ToR 2 where the Ethernet frames are encapsulated into NEPHELE frames. Tx1 N-FPGA, following the OpenFlow commands, configures the tunable laser to emit at the destination wavelength - ToR 3, λ$_3$=1548.515nm - and configures accordingly the optical switches. The bandwidth received by ToR 3 is 2.87Gb/s, 4.55 Gb/s and 6.38 Gb/s that corresponds to 25%, 50% and 87.5% of the total scheduling period respectively. In the presented real-time demonstration, the dynamic bandwidth allocation is achieved via the OCEANiA graphical user interface (GUI) of Application Affinity depicted in Fig. 4(c). NEPHELE has also developed a northbound interface for interfacing the OCEANiA controller with an OpenStack cloud orchestrator[8].

The final test that was demonstrated within the NEPHELE prototype testbed is the inter-pod scenario. The SDN controller establishes paths between interface E1 (ToR1) and interfaces E3 (ToR2) and E5 (ToR3) at the same time. All network elements (tunable Tx1 of ToR1, optical switches) are configured according to the routing and switching commands. ToR1 – plane1 Tx is transmitting traffic enrolled in λ$_2$=1547.715nm and λ$_3$=1548.515nm with destination ToR2 and ToR3, both located in POD2. As shown in Fig. 4 (a), the traffic produced by ToR1 is divided equally between two ToRs and the bandwidth is measured to be 1.52 Gb/s both for E3 (10.2.2.1) and E5 (10.2.3.129) interfaces as depicted in the screenshot of Fig. 4 (b).

Conclusions and Acknowledgements

We demonstrated the real time operation of the NEPHELE end-to-end optical DCN. Successful integration of a slotted (TDMA) data plane with a custom, open-source based SDN overlay enabled validation of various communication scenarios. This work has received funding from the EU's Horizon 2020 research and innovation program under grant agreement No645212 (NEPHELE).

References

[1] Cisco, "Cisco Global Cloud Index: Forecast and Methodology, 2016-2021", (Feb 2018).

[2] P. Bakopoulos et al., "NEPHELE: an end-to-end scalable and dynamically reconfigurable optical architecture for application-aware SDN cloud datacenters", IEEE Communications Magazine, (2018)

[3] K. Tokas et al., "Slotted TDMA and optically switched network for disaggregated datacenters", Proc. ICTON 2017.

[4] K. Yiannopoulos et al., "Resource partitioning in the NEPHELE datacenter interconnect", Proc. ICTON 2017.

[5] P. Bakopoulos et al., "Optical datacenter network employing slotted (TDMA) operation for dynamic resource allocation", Proc. SPIE 10538, (2018)

[6] OCEANiA DCN controller 2017, 10.5281/zenodo.1004859

[7] K. Christodoulopoulos et al., "Efficient bandwidth allocation in the NEPHELE optical/electrical datacenter interconnect", IEEE/OSA Journal of Optical Communications and Networking, (2017)

[8] Giada Landi et al., "SDN Control Framework with Dynamic Resource Assignment for Slotted Optical Datacenter Networks", OFC 2017