Slotted TDMA and Optically Switched Network for Disaggregated Datacenters

Konstantinos Tokas[1], Ioannis Patronas[1,2], Christos Spatharakis[1], Dionysios Reisis[1,2], Paraskevas Bakopoulos[1], and Hercules Avramopoulos[1]

[1]*Photonics Communications Research Laboratory, National Technical University of Athens, Athens, Greece*

[2]*Electronics Laboratory, Faculty of Physics, National and Kapodistrian University of Athens, Athens, Greece*

ABSTRACT

The relentless traffic growth in datacenter networks is stimulating the adoption of pioneering optical interconnect technologies as well as their integration with novel network and switching architectures. Even more, the need for disaggregation of data storage and processing resources significantly increases the capacity and dimensioning requirements of such networks. In this context, a novel datacenter network architecture that combines space and wavelength switching functionalities is demonstrated experimentally. The architecture leverages slotted TDMA/WDM switching to realize dynamic resource allocation with sub-wavelength granularity, thus realizing a low cost and power consumption, scalable datacenter network. Dynamic reconfiguration of the slotted network vouches for low latency operation of the data plane and hence, it fulfils the requirements of the envisaged disaggregated datacenter infrastructure. The current paper reports the experimental evaluation of the optical subsystems and demonstrates the proof of concept for combined space- and wavelength-switching with optical bursts of 200 μs duration in different network scenarios. The generation and reception of slotted traffic, as well as the control of the optical switching subsystems is performed by means of addressable FPGA boards.

Keywords: datacenter network, optical switching, slotted network, scheduling.

1. INTRODUCTION

The massive deployment of cloud applications is causing datacenters to expand exponentially in size, bringing data traffic on a double-digit growth curve [1]. To keep pace with soaring traffic demand, scalable networking technologies are urgently sought, aiming at sustaining bandwidth delivery in the datacenter without hitting the energy wall, whilst ensuring low latency to meet the needs of the application layer. The conundrum is put in its full perspective when considering the architectural limitations of current datacenters: hierarchical fat-tree topologies favour north-south communication and strive to comply with the east-west traffic profile of modern datacenters, whereas the number of switches scales superlinearly with the number of hosts. As a means to overcome these challenges, flatter, optically-switched datacenter network architectures are put in the spotlight, benefiting from the inherent energy efficiency and bitrate-transparency of optical switches. Switching of bypass traffic at the optical layer is not just a fascinating research topic; the concept is currently enjoying broad deployment in commercial telecom networks based on reconfigurable add-drop multiplexing (ROADM) nodes. However, application of the concept inside the datacenter stumbles upon several nontrivial challenges, pertinent to the larger scale and higher dynamicity of datacenter traffic, as well as the different target cost envelope. What is arguably one of the most salient challenges towards scalable optically-switched datacenter architectures is the trade-off between switching speed and number of ports, which is faced by current optical switch technologies: Although large switches with hundreds of ports exist, their speed is typically in the millisecond regime and cannot follow the dynamic fluctuations of datacenter traffic. On the other hand, very fast (nanosecond-scale) switches are constrained to a few I/O ports. While efforts to develop switching components with fast reconfiguration speed and large number of I/Os are underway [2], particular focus is given on new architectures that can make the most out of existing, well-established technologies, allowing rapid deployment in real networks. In this context, optically-switched network architectures tailored to the particularities of the most prominent optical switching technologies are being researched, such as space-switching (e.g. with micro-electro-mechanical systems – MEMS [3], semiconductor optical amplifiers – SOAs [4] or electroabsorption modulators – EAMs [5]), wavelength-switching (through combination of tunable lasers with arrayed-waveguide-grating routers – AWGRs [6][7]) or combination thereof (e.g. using wavelength-selective switches – WSSs [8][9]).

A new optically-switched network infrastructure is under development within the NEPHEE European project, and the baseline architecture was proposed recently along with the mechanisms and methodologies for obtaining dynamic assignment of the network resources through an overarching SDN framework [8]. The architecture is scalable to the order of 100,000 network hosts whereas the number of switches scales linearly with the number of hosts. In this communication we demonstrate the experimental implementation and evaluation of the NEPHELE data plane, proving the feasibility of the proposed architecture. Commercial off-the-shelf (COTS) optical components are used in liaison with field-programmable gate array (FPGA) boards to demonstrate end-to-end switching of traffic. Different networking scenarios are investigated and error-free operation is demonstrated at all cases.

2. OVERVIEW OF NEPHELE ARCHITECTURE

The NEPHELE network architecture is shown schematically in Fig. 1a. The network consists of pods, with each pod accommodating a number of racks, and with each rack hosting the datacenter resources (i.e.. storage or compute) in so-called innovation zones. As a result, each pod comprises in essence a small self-contained datacenter, whereas scaling to larger system dimensions is achieved by interconnecting multiple pods in a DWDM fiber ring. Inside the pods, each rack is administered by a top-of-rack switch (ToR), and each pod is controlled by a pod switch (POD). The ToRs are interconnected to the respective PODs in a star topology.

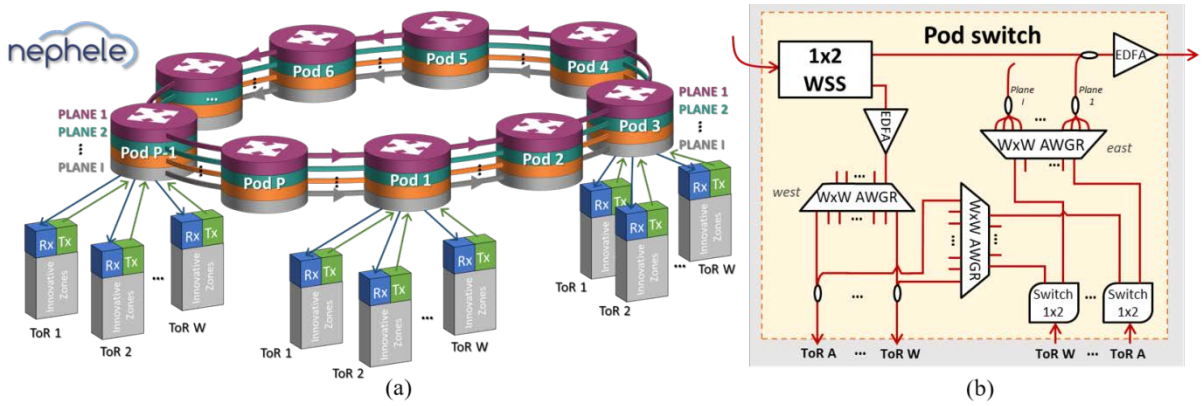

Figure 1: (a) NEPHELE architecture; (b) Block diagram of NEPHELE pod switch.

Routing of traffic flows in the NEPHELE network relies on different switching mechanisms, depending on traffic locality. For source and destination ToRs located in the same POD (i.e. "intra-pod" traffic), wavelength switching is applied. The ToRs are equipped with tunable lasers and burst-mode transmitters and receivers, whereas the PODs consist of passive filtering elements, namely arrayed waveguide grating routers (AWGRs), as illustrated in Fig. 1b. Thus, by properly tuning the transmitter's wavelength at the source ToR, traffic is passively routed through the POD's AWGR to the destination ToR inside the POD. For source and destination ToRs located in different PODs (i.e. "inter-pod" traffic), a combination of wavelength- and space-switching is applied. Inter-pod traffic flows generated at the source ToR arrive at the source POD, where they are forwarded towards the DWDM ring by means of 1×2 switches. The DWDM ring consists of multiple fibers and added traffic is passively distributed among them according to its wavelength, by means of a dedicated AWGR at the east port of the POD. Traffic transversing the DWDM ring enters the west port of the neighboring PODs, which is equipped with wavelength selective switches (WSSs); thus traffic flows are processed per wavelength and are forwarded to the next POD until they reach the destination. At the destination POD, traffic is dropped at the WSS and is distributed to the destination ToRs inside the POD according to the wavelength of each flow, by means of another AWGR.

Considering the typical wavelength count of DWDM systems in the C-band, it is possible to accommodate more than 80 ToRs in each POD, with each ToR connected to the corresponding POD via a single port. To scale network capacity, additional optical planes are deployed as shown in Fig. 1. This involves installation of parallel PODs interconnected through parallel WDM rings, as well as populating additional ports in the ToR switches to connect to the newly added PODs.

Allocation of network resources in NEPHELE is facilitated through slotted (i.e. time-division multiple access – TDMA) operation of its data plane, offering dynamic reconfiguration with sub-wavelength granularity. Slots (and therefore network resources) are assigned dynamically to communicating racks using bespoke network allocation algorithms that optimize network utilization [10]. On the other hand, slotted operation poses stringent requirements to the data plane for smooth operation without collisions, which are investigated in this manuscript.

3. EXPERIMENTAL SETUP

Figure 2 shows the experimental setup used to evaluate the operation of the NEPHELE architecture described above. The highlighted sections represent the implemented part of the setup while the faded out sections were not realized and are shown for completeness. The setup comprises two POD and three ToR switches so that both intra- and inter-pod communication scenarios as well as combinations thereof can be realized and tested. POD 1 contains all the optical circuitry for handling intra-pod traffic as well as outgoing flows of inter-pod traffic, (incoming flows are handled by the shaded section which was not implemented). As such, ToR A is serving as transmitter for intra- or inter-pod traffic, while ToR B is only serving as a receiver for intra-pod traffic. POD 2, as well as its underlying ToR C switch, are equipped only with the subsystems necessary to handle incoming intra-pod traffic (highlighted section).

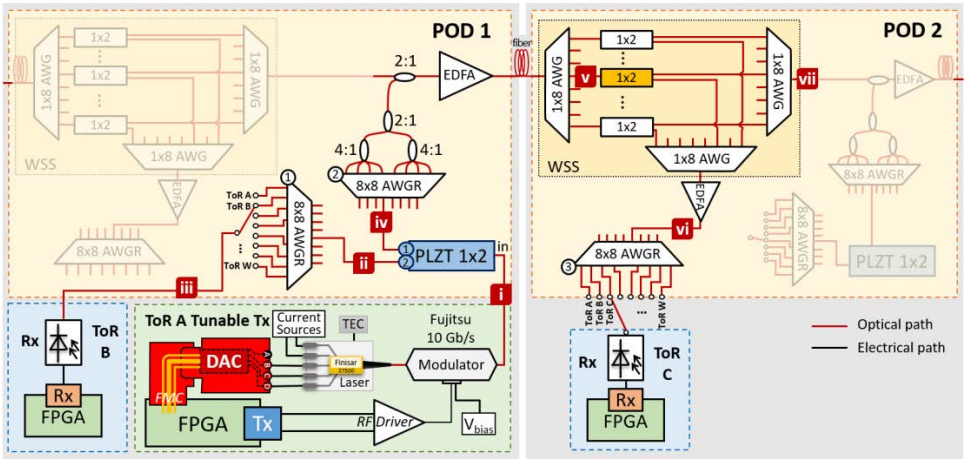

Figure 2. Experimental setup.

3.1 Optical data plane

ToR A transmitter is based on a XILINX Kintex KC707 FPGA board, which is programmed to generate data packets of 200 μs duration and separated by 9.6 μs guard-time. The differential electrical signals from the FPGA board are amplified in a Maxim 3941 10 Gb/s RF driver. The single ended amplified signal is introduced to a Fujitsu 10 Gb/s Mach-Zehnder Modulator (MZM) while a DC bias voltage is applied simultaneously to the respective port to ensure operation at the optimal biasing point. The optical carrier is provided to the modulator by a Finisar S7500 tunable laser, which is controlled by 5 input currents. Two of these currents concern the gain and SOA sections of the laser and are provided by respective current sources and kept constant. The remaining three currents are applied to the phase and left/right reflector sections and serve to adjust the laser's emission wavelength. To enable fast wavelength tuning, the latter three currents are controlled by a Texas Instruments current Digital-to-Analog-Converter (DAC) evaluation module. The DAC is dynamically controlled from the transmitter FPGA, changing accordingly the value of the laser tuning currents in every packet so as to tune the laser to the desired wavelength (according to a look-up table stored in the FPGA for the C-band 50 GHz grid).

The optical packets generated at the ToR 1 transmitter are introduced to the POD. The signal is first fed into a 1×2 PLZT switch which classifies inter- from intra-pod traffic, directing the packets either to a ToR inside the same POD or towards a different POD. The PLZT switch is in principle controlled by a dedicated FPGA located at the host POD; however, due to unavailability of resources, in the current experiment it is controlled by the same FPGA board that was used at the ToR 1 transmitter. In the case of intra-pod communication, the optical packets are switched to output port 2 of the PLZT switch and fed into an 8×8 AWGR (AWGR 1), where each packet is passively routed to the desired destination ToR according to its wavelength as well as the AWGR port that it was introduced to. The optical signals are detected at the ToR B receiver by means of a U^2t photoreceiver connected to a XILINX Virtex VC707 FPGA, employed for data reception and BER estimation. In the case of intra-pod communication, the optical slots are directed to output 1 of the PLZT switch and introduced into an 8×8 AWGR (AWGR 2). Note that in a full scale implementation, AGWR 1 and 2 receive signals from other ToR transmitters as well, whereas larger ToR counts can be accommodated in the POD by upgrading the 8×8 AWGRs with larger units, or by cascading multiple AWGR stages as in [11]. AWGR 2 serves to distribute packets coming from different ToRs inside the POD, among the multiple fibers of the NEPHELE DWDM ring. In the experiment, given that the DWDM ring is implemented with one fiber, optical couplers are employed to combine all outputs of AWGR 2 into a single fiber before entering the NEPHELE ring. An additional 2×1 optical coupler is placed at the connection to the ring as per the NEPHELE architecture of Fig. 1 (serving to combine traffic transversing the POD with traffic added to the DWDM ring in the current POD. The combined optical traffic is amplified in an erbium-doped fiber amplifier (EDFA) and propagates in the DWDM ring – emulated by 12 m of standard single-mode fiber- before reaching POD 2.

The optical traffic entering POD 2 is first introduced into a WSS module, consisting of 3 1×8 AWGs and n 1×2 PLZT switches where n is the number of supported wavelengths. In the current experiment $n = 2$ and the 1×2 PLZT switches are again controlled by the transmitter ToR FPGA due to unavailability of resources. The WSS follows the "demultiplex, switch and multiplex" approach, so that incoming optical traffic is split into its DWDM tributaries and each wavelength is switched on-demand by the respective 1×2 switch. Depending on the schedule and the destination of each packet, it will be either forwarded towards the next POD or dropped to the current POD; in both cases, an AWG is employed to recombine the optical traffic into a single flow. Dropped traffic is amplified in an additional EDFA and is introduced into an 8×8 AWGR (AWGR 3) that passively routes the DWDM traffic to the target destination port (the 8×8 AWGR can accommodate the outputs of multiple WSSs in case the WDM ring consists of multiple fibers). For the experiment ToR C served for receiving optical packets and was implemented in a similar manner as described above for ToR B.

3.2 FPGA data generation and reception

In order to control the data plane elements and evaluate slotted operation of the NEPHELE infrastructure, FPGA boards were employed and programmed. The format of the packets used to test the bursty communication of the system and the functionality of the FPGAs realizing the transmitter and the receiver are outlined in this section. The packets consist of the following four fields: a) Preamble, b) Synchronization Pattern, c) Packet Sequence Number and d) Payload. The preamble is the first part of the packet, used for the correct configuration of the optical path as well as to facilitate the receiver's Clock Data Recovery (CDR) operation before the reception of valid data. The preamble duration was set at 9.6 μs during the presented experiments. The Synchronization Pattern is a sequence of 64 bits (0xc5e51840fd59bb49) that indicates the beginning of valid data within the packet. The next field, Packet Sequence Number, is a serial number that is used as packet ID. The Payload is generated by a pseudo-random binary sequence (PRBS) generator and the data is scrambled before transmission in order to facilitate Clock Data Recovery.

The Receiver FPGA calculates the packet loss and the Bit Error Rate (BER) of the received packets. The received bit stream is scanned until a match with the Synchronization pattern is identified. Next, the received packet is forwarded to the Scrambler unit for unscrambling. The Receiver FPGA contains an identical PRBS generator with the Transmitter FPGA. Thus, the receiver FPGA calculates the actual BER by comparing the unscrambled data with the locally generated PRBS. The packet loss is calculated by comparing the number of received packets with the serial number of the currently received packet, i.e. the Packet Sequence Number.

4. RESULTS AND DISCUSSION

4.1 Evaluation of intra-pod communication

Intra-pod communication is evaluated first, involving connectivity between ToRs inside the same POD. In this experimental scenario, the tunable Tx transmits NEPHELE packets enrolled in $\lambda_1 = 1546.91$ nm and $\lambda_2 = 1551.72$ nm continuously and alternatingly (Fig. 3i). The 1×2 PLZT optical switch, shown in dark blue in Fig. 2, separates intra- from inter-pod packets, as shown in Fig. 3i and 3iv. In this scenario, 8 packets are switched towards a different POD and 8 packets remain within POD 1 alternatingly, via outputs 1 and 2 of the PLZT switch. The packets of the intra-pod path (output 2 of the 1×2 optical switch) reach the destination ToR through AWGR 1. Figure 3iii depicts the packets of λ_2 received in ToR B receiver and Fig. 3b the respective BER curve (black dots) as a function of the received optical power. For received power higher than -6.5 dBm no errors were observed, corresponding to a BER better than 3×10^{-13}.

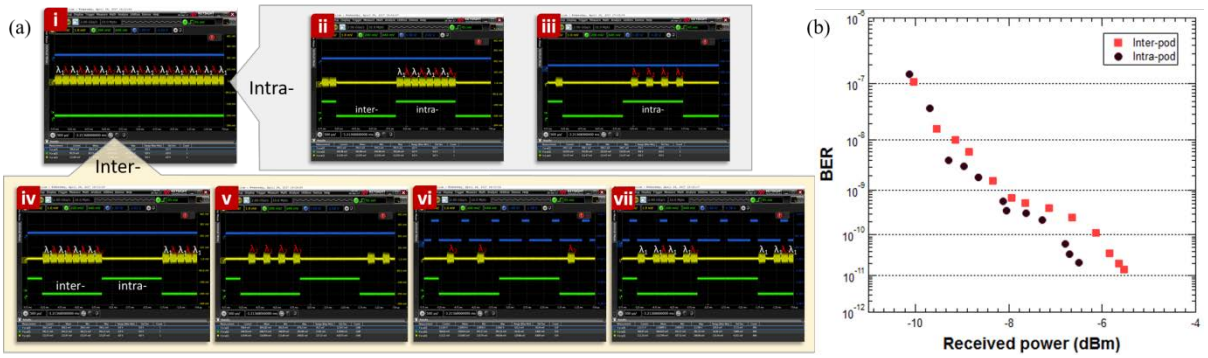

Figure 3: (a) Snapshots of NEPHELE packets in different spots of intra- and inter-pod communication. The red numbers (i)-(vii) corresponds to the same ones depicted in the experimental setup of Fig. 2; (b) BER vs. received optical power (optical modulation amplitude - OMA) for intra- and inter-pod scenarios.

4.2 Evaluation of inter-pod communication

In the case of inter-pod communication (evaluated simultaneously with the inter-pod case), the NEPHELE packets destined to POD 2 (enrolled in $\lambda_1 = 1546.91$ nm and $\lambda_2 = 1551.72$ nm) are switched to output 1 of the 1×2 PLZT switch, as depicted in Fig. 3iv. The optical traffic is amplified by the EDFA located at the output of POD 1, before entering POD 2 through its respective WSS module. Due to limitations of available resources, in this scenario only one 1×2 optical switch is available, coloured in orange in Fig. 2. In Fig. 3v, a snapshot of the λ_2- packets entering the orange optical switch is depicted. Figures 3vi and 3vii illustrate the optical packets at the outputs of the WSS module, destined either to a ToR inside the current POD (dropped traffic) or to a subsequent POD (forwarded traffic) respectively. The blue trace in these figures shows the driving signal for the 1×2 optical switch, controlling which of the λ_2- packets will be dropped in the current POD and which will continue to the next POD. The packets dropped inside POD 2 are routed to ToR C where they are detected in the optical receiver and fed to the FPGA. The BER curve as a function of the received optical power is shown in Fig. 3b

(red squares). The results show very similar performance to the intra-pod scenario, with error free operation observed for received powers higher than -5.8 dBm.

An additional inter-pod scenario was evaluated individually, focusing on the functionality of the WSS module. Figure 4a shows the modified experimental setup based on the "demultiplex, switch and multiplex" approach, where in this case two 1×2 PLZT optical switches were used to allow dynamic control of two wavelengths in the WSS. A continuous flow of NEPHELE packets enrolled in four wavelengths (λ_1 = 1550.116 nm, λ_2 = 1548.515 nm, λ_3 = 1550.918 nm, λ_4 = 1552.11524 nm) enters the left 1×8 AWG of the WSS module, as shown in Fig. 4i. After demultiplexing, λ_1 and λ_2 are hard-wired to the corresponding inputs of the south 8×1 AWG whereas λ_3 and λ_4 are controlled by the two 1×2 switches, so that the corresponding packets can be dropped inside the POD or forwarded to the DWDM ring, according to their driving signals. Figures 4ii and 4iii depict the two outputs of the WSS module. After the 8×8 AWGR, the packets enrolled in λ_3 and λ_4 are received separately by the ToR receiver, which is connected to the appropriate outputs of AWGR 3. Figure 4b shows the BER curve as a function of the received optical power; for optical power higher than -6.8 dBm no errors were observed (BER better than 3×10^{-13}).

Figure 4: (a) The experimental setup for inter-pod communication scenario, focusing on the WSS functionality. (i)-(iii) snapshots of NEPHELE packets in the spots highlighted in (a); (b) BER vs. received power (optical modulation amplitude - OMA) for λ3 (channel 53) and λ4 (channel 57).

5. CONCLUSIONS

In this paper we presented and experimentally implemented the NEPHELE network architecture for disaggregated datacenters. This scalable architecture combines slotted TDMA and DWDM switching techniques to provide dynamic resource allocation and sub-wavelength granularity. The pillars of this architecture are the POD and ToR switches, dynamically generating and switching traffic flows according to the scheduler's commands. In the experiment, FPGA boards are used for controlling the POD and ToR optical subsystems according to pre-defined commands. Different slotted network scenarios are implemented, experimentally verifying successful intra- and inter-pod communication as well as combined operation for several wavelengths. The experimental results ensure error free operation of all tested scenarios and communication paths for a wide range of received optical powers, proving the feasibility of the overall architecture and its potential scalability to larger network dimensions without jeopardizing the robust performance of the system operation.

ACKNOWLEDGEMENTS

This work has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 645212 (NEPHELE).

REFERENCES

[1] Cisco, "Cisco Global Cloud Index: Forecast and Methodology, 2014-2019 White Paper".
[2] W. M. Mellette *et al*., "A scalable, partially configurable optical switch for data center networks, " *J. Lightwave Technol.* vol. 35, pp. 136-144, Jan 2017.
[3] G.Wang *et al*., "c-Through: Part-time optics in data centers," in *Proc. ACM SIGCOMM '10*, pp. 327-338.
[4] G. M. Saridis *et al*., "Lightness: A function-virtualizable software defined data center network with all-optical circuit/packet switching," *J. Lightwave Technol.* vol. 34, pp. 1618-1627, Apr. 2016.
[5] T. Segawa, Y. Muranaka, and R. Takahashi, "High-speed optical packet switching for photonic datacenter networks," *NTT Technical Review*, vol. 14 no. 1, Jan. 2016.
[6] K. Xia *et al*., "Petabit optical switch for data center networks," *Tech. Rep., Polytechnic Inst. of NYU*, 2010.
[7] R. Proietti *et al*., "Scalable optical interconnect architecture using AWGR-based TONAK LION switch with limited number of wavelengths," *J. Lightwave Technol.* vol. 31, pp. 4087-4097, Dec. 2013.
[8] V. Kamchevska *et al*., "Experimental demonstration of multidimensional switching nodes for all-optical data center networks", *J. Lightwave Technol.*, vol. 34. pp. 1837-1843, Apr. 2016.
[9] G. Porter *et al*., "Integrating microsecond circuit switching into the data center," in *Proc. SIGCOMM 2013*, pp. 447-458.
[10] K. Christodoulopoulos *et al*., "Bandwidth allocation in the NEPHELE hybrid optical interconnect", in *Proc. ICTON*, Trento, Italy, Jul. 2016, paper Th.B5.1.
[11] K. Sato *et al*., "A large-scale wavelength routing optical switch for data center networks," *IEEE Communications Magazine*, vol. 51, pp. 46-52, Sept. 2013.