Slotted optical datacenter network with sub-wavelength resource allocation

Paraskevas Bakopoulos, *Member, IEEE,* Konstantinos Tokas, Christos Spatharakis, and Hercules Avramopoulos, *Member, IEEE.*

***Abstract*—Optical switching is gaining traction for scaling datacenter networks, apace with soaring traffic demand. We experimentally evaluate a slotted optical network architecture capable of dynamically allocating network resources with sub-wavelength granularity. Network operation is demonstrated with multiple communication scenarios, using 200 µs duration optical bursts.**

***Index Terms*—Optical switching, Optical interconnects, Data center networks.**

I. Introduction

DATA center traffic is on a steep rise, stimulated by the rapid cloudification of consumer and enterprise applications. As I/O interfaces migrate to higher data rates to cope with traffic demand, power consumption of the network becomes a significant fraction of the overall IT equipment's; even more, this trend is exacerbated as the servers themselves are becoming more energy-optimized [1]. Faced with the challenge of ever-scaling network capacity without hitting the "energy wall", equipment manufacturers are seeking alternative networking concepts. In this context, optical switching is gaining traction owing to its inherent speed, energy efficiency and transparency to bitrate and protocol.

Deployment of optical switching in the datacenter entails substantial changes to the design of the network, by virtue of the technology's unique characteristics. Optical switches differ from their electronic counterparts in terms of port count, buffering capabilities and technological maturity; even more, among the various available switching technologies diverse performance characteristics are offered. This is inspiring work on novel data center network architectures tailored to prominent optical switching technologies, such as micro-electro-mechanical systems (MEMS) [2], semiconductor optical amplifier (SOA) switches [3], tunable lasers combined with arrayed-waveguide-grating routers (AWGRs) [4] and wavelength-selective switches (WSSs) [5]. One of the key challenges currently pertaining to optical datacenter networks is the combination of scalability and fast reconfigurability.

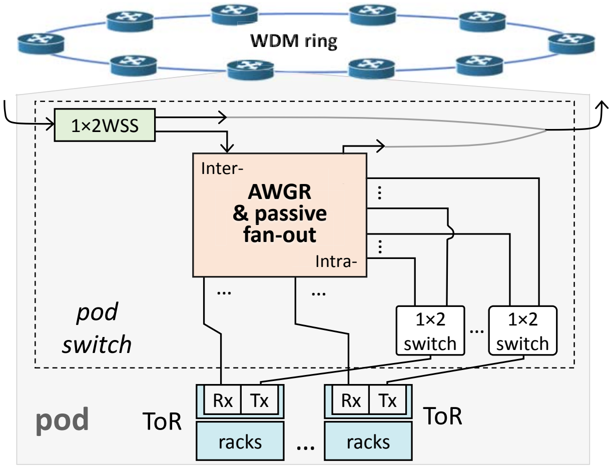

Fig. 1: The NEPHELE data plane architecture.

Recently we proposed a novel datacenter architecture developed within the NEPHELE European project, shown schematically in Fig. 1 [6]. The network architecture relies in pods, each accommodating a number of racks. Each rack is administered by a top-of-rack (ToR) switch, and the ToRs are connected to the POD switch in a star topology. Each ToR is equipped with a tunable laser and can reach the remaining ToRs in its pod by properly tuning its transmitted wavelength. Scaling network dimensions is achieved by interconnecting multiple pods in a DWDM ring. Inter-pod traffic is routed by means of WSSs placed in each POD switch, allowing wavelength reuse among pods, and thus enabling network scalability beyond the typical wavelength count of DWDM systems. Dynamic and efficient sharing of network resources and collision-free routing are facilitated with slotted operation of the network (i.e. using time-division multiple-access - TDMA). In this paper we experimentally demonstrate routing in intra- and inter-pod communication scenarios, verifying successful slotted operation of the NEPHELE data plane.

II. Experimental Setup

Fig. 2 shows the experimental setup for verifying inter- and intra-pod routing operation in NEPHELE's data plane. A ToR transmitter generated 10 Gb/s optical data packets at two wavelengths, targeting two ToR receivers at the same (intra-) or different (inter-pod traffic) POD. A Pulse Pattern Generator (PPG) was used for data generation, driving a Mach-Zehnder modulator after amplification in a broadband RF driver. A Finisar S7500 tunable laser fed the modulator, emitting alternatingly at two wavelengths (λ_1=1535.822 nm and

Manuscript received March 12, 2017. This work was supported through Horizon 2020 research and innovation program under grant agreement n° 645212 (NEPHELE).

P. Bakopoulos, K. Tokas, C. Spatharakis and H. Avramopoulos are with the School of Electrical & Computer Engineering, National Technical University of Athens, Iroon Polytechniou 9, 15780 Athens, Greece (phone: +30 210-772-4454; e-mail: hav@mail.ntua.gr).

978-1-5090-6571-4/17/$31.00 ©2017 IEEE

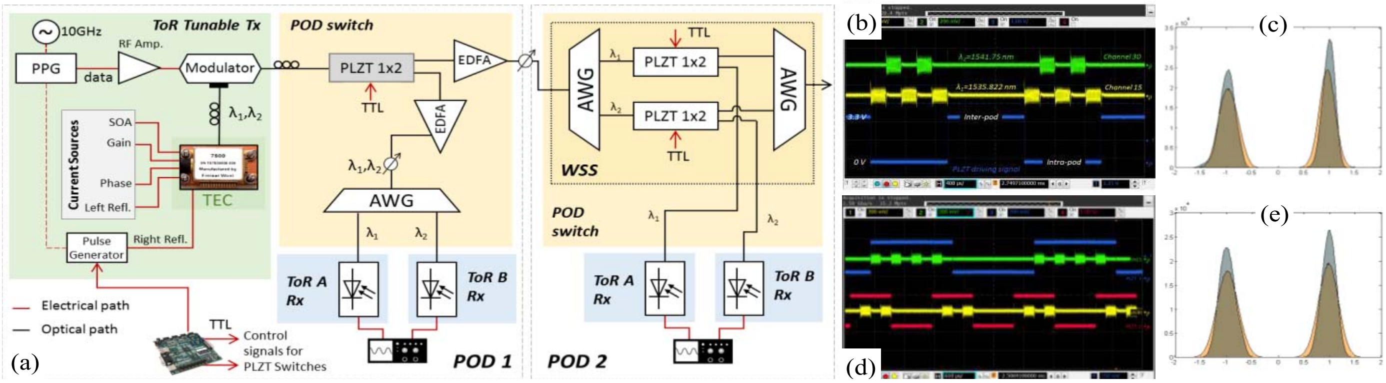

Fig. 2 (a) Experimental Setup for testing Intra- and Inter-pod communication, (b) NEPHELE packets enrolled in λ_1 (yellow) and λ_2 (green). Blue waveform represents the control signal for the 1x2 optical switch (in grey), (c) Probability Density Function (PDF) histograms, (d) Green waveform represents NEPHELE packets dropped inside the pod according to the control signals (blue and red waveforms) driving the WSS, (e) Probability Density Function (PDF) histograms.

λ_2=1541.756 nm), thus generating TDMA data packets of 200 μs duration. Laser tuning was achieved by applying programmable control currents to its tuning sections (left and right reflector and phase section). In the presented configuration, the right reflector current was dynamically controlled using a pulse generator locked to the data PPG and the remaining two sections were statically driven.

In order to select intra- or inter-pod operation, a 1x2 PLZT optical switch (shown in grey color in Fig. 2) routed the traffic inside the same POD or towards a different POD. The 1x2 switch was controlled by a Xilinx SPARTAN SP605 FPGA board, locked to the data PPG. In the intra-POD scenario, an erbium-doped fiber amplifier (EDFA) was used to amplify the optical signals before entering a 1x80 AWG, which passively routed the traffic according to its wavelength, to the respective ToR. After detection at the ToR photoreceivers, the received data was captured in an Agilent 33 GHz real-time oscilloscope for offline DSP processing (thresholding) and BER estimation.

To evaluate the inter-POD scenario, the optical traffic was routed by the 1x2 PLZT switch to POD 2, where it entered the WSS. The latter followed a "demultiplex-switch-and-multiplex" architecture: the input 1x80 AWG demultiplexed the traffic, and for each wavelength-path a 1x2 PLZT switch was employed, dropping each λ inside the POD or forwarding it towards the next POD, after multiplexing of all λs in an identical AWG. The 1x2switches were controlled by the same FPGA board as above, while the dropped optical packets were received, captured and processed in a similar manner to the intra-pod scenario.

III. Experimental Results

Fig.2 (b) depicts the packets imprinted in λ_1 (yellow) and λ_2 (green) after detection at the ToR receivers within the same pod. The blue waveform represents the TTL control signal for the grey 1x2 optical switch. In order to evaluate system performance, the first packet of each λ was captured in a real-time oscilloscope for different received optical power levels. BER measurements were carried out in MATLAB, and the acquisition window length limited the measurable BER to 10^{-7} with 95% confidence interval. The PDF histograms for received optical power of -14 dBm (minimum measurable BER - orange trace) and for -6 dBm (grey trace) are compared in Fig.2(c). A significant improvement of signal quality is observed in terms of statistical variance for each symbol distribution as the power level increases, indicating the absence of an error floor.

The control signals driving the 1x2 switches in the WSS are shown in Fig.2 (d) along with the corresponding data streams at different λs. Since the packets were transmitted alternatingly through the two wavelengths, switching resulted in 4 packets at λ_1 being dropped inside pod 2 while the following 4 were forwarded towards the next pod in the WDM ring. In the same manner two λ_2 packets were dropped inside the pod while the following two were forwarded towards the WDM ring. As in the previous experiment, offline DSP was employed and as the power level increases beyond the BER measurement limit, an ample improvement of the signal quality is observed (Fig.2 (e)).

To scale the implemented transmitter for larger pods, a laser control module was developed, allowing dynamic tuning of the laser to any desired wavelength on the 50 GHz ITU grid. The control module was based on a 4-channel current DAC connected to a Xilinx SPARTAN board and facilitated control of all laser tuning sections. Tuning to 80 wavelengths on the ITU grid was verified (thus enabling network configurations with 80 ToRs per pod), with tuning time below 18 ns.

IV. Conclusions

The NEPHELE optical network architecture was evaluated experimentally. Intra- and inter-pod routing was demonstrated using wavelength- and space-switching. Slotted operation with 200 μs optical bursts was achieved, thus enabling network resource allocation with sub-wavelength granularity.

References

[1] D. Abts et al., "Energy proportional datacenter networks," in *Proc. ACM SIGARCH Comp.Architecture News*, Saint-Malo, France, 2010.

[2] H. Liu et al., "REACToR: A reconfigurable packet and circuit ToR switch," in *Proc. IEEE Photonics Society Summer Topical Meeting Series*, 2013, TuE3.1.

[3] W. Miao et al., "Novel flat datacenter network architecture based on scalable and flow-controlled optical switch system," *Opt. Express* vol. 22, pp. 2465-2472 (2014).

[4] R. Proietti et al., "A Scalable, Low-Latency, High-Throughput, Optical Interconnect Architecture Based on Arrayed Waveguide Grating Routers," J. Lightwave Technol. 33, 911-920 (2015).

[5] G. Porter et al., "Integrating Microsecond Circuit Switching into the Data Center," in SIGCOMM, 2013, pp. 447-458.

[6] K. Christodoulopoulos et al., "Bandwidth allocation in the NEPHELE hybrid optical interconnect," in Proc. 18th International Conference on Transparent Optical Networks (ICTON), Trento, 2016, Th.B5.1.