End-to-end optical packet switching with burst-mode reception at 25 Gb/s through a 1024-port 25.6 Tb/s capacity Hipoλaos Optical Packet Switch

A. Tsakyridis[1], N. Terzenidis[1], G. Giamougiannis[1], J. Van Kerrebrouck[2], M. Verbeke[2], G. Torfs[2], M. Moralis-Pegios[1] and N. Pleros[1]
[1] Department of Informatics, Center for Interdisciplinary Research & Innovation, Aristotle University of Thessaloniki, Thessaloniki, Greece
[2] IDLab, INTEC, Ghent University - imec, 9052 Ghent, Belgium
atsakyrid@csd.auth.gr

Abstract *We demonstrate end-to-end 25Gb/s true optical packet switching featuring burst-mode reception with <50ns locking time through a 1024-port 25.6Tb/s capacity Hipoλaos Optical Packet Switch architecture. Error-free performance at 10^{-9} was obtained for all validated port-combinations.*

Introduction

At the dawning of the exaflop era, the growing demand for ubiquitous high-bandwidth processing and cloud computing applications, has stimulated an unprecedented increase on the data traffic residing within Data Centers (DC)[1], revealing the necessity for novel DC architectures capable of providing increased resource utilization and reduced energy consumption. The realization of disaggregated DC architectures appears as a promising solution to meet these requirements[2], but in turn urges for high-capacity, sub-μs latency and high-radix connectivity. Present electrical DC switches are able to provide capacities up-to 12.8Tb/s[3],[4], with the next milestone of 25.6Tb/s switches being expected before 2022. However, this objective introduces major challenges that mainly stem from the high-power SerDes interfaces and the packaging constraints[5]. Demarcating from electrical towards Optical Packet Switch (OPS) architectures, a number of novel layouts has been proposed to support high-port connectivity and high-capacity[6]-[11] with the recently demonstrated Hipoλaos architecture allowing for 1024-port and 25.6Tb/s capacity configurations, along with sub-μs latency performance[12]. However, all these demonstrations have been limited to the realization of the optical forwarding plane assuming synchronized source and destination nodes, ignoring the requirement for asynchronous packet traffic between the OPS network nodes, which would in turn necessitate the employment of a Burst-Mode Clock & Data Recovery (BM-CDR) circuitry with ns-scale locking time in order to handle the phase-mismatch of the packets emerging from different nodes. OPS demonstrations with end-to-end asynchronous packet operation have been recently reported[13],[14], but with limited port count and capacities well below the 25.6Tb/s target for next-generation switches.

In this paper, we demonstrate for the first time, to the best of our knowledge, an end-to-end optical packet switch link over the recently reported Hipoλaos architecture[12], by incorporating a BM-CDR at the receiver end in order to establish true packet-level communication between non-synchronized source and destination nodes through a 1024-port OPS layout with 25.6Tb/s capacity. We demonstrate successful routing of 25Gb/s optical burst-data packets through a fully functional Hipoλaos Plane, followed by a 32x32 AWGR, that are finally received by a BM-CDR[15] in order to recover the data with locking time <50ns. The architecture's performance was assessed via Bit Error Rate (BER) measurements, revealing error-free operation with a mean power penalty (PP) of 2.88dB compared to Back-to-Back.

Hipoλaos data and control-plane architecture

The Hipoλaos OPS architecture comprises an

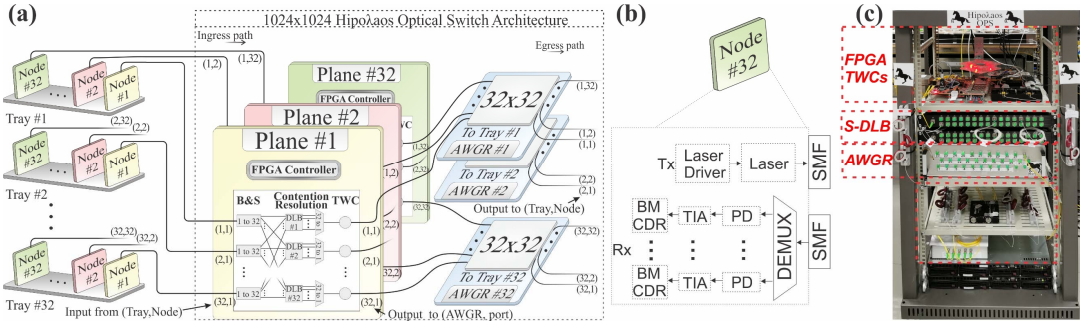

Fig. 1. (a) Generic layout of the 1024x1024 Hipoλaos switch architecture, (b) Node transceiver interface, (c) Hipoλaos prototype.

optimized Spanke layout, overcoming the scalability limitations by distributing the control and switching functions in separate clusters, named as Planes, while the incorporation of optical feed forward buffering enables the realization of high-throughput and low-latency performance. Moreover, in order to achieve high-data rates, the Hipoλaos switch employs a differentially-biased scheme[16]-[18] at the wavelength conversion (WC) stage that allows operation up-to 40Gb/s[19]. Finally, the architecture exploits the cyclic routing properties of AWGRs, in order to extend the switch port-count through a collision-less WDM routing mechanism.

The 1024-port Hipoλaos OPS is illustrated in Fig.1(a), interconnecting a DC rack system, that is composed of 32 trays with each tray hosting 32 nodes. The switch is organized in 32 Planes with 32 ports/Plane, followed by 32 32x32 AWGR devices, for traffic aggregation and wavelength routing, respectively. Fig.1(b) shows the interface of a node, that is connected to the switch by a pair of fibers, utilizing a single optical link at a fixed wavelength on the transmitter side. At the receiver side each node is considered to have the capability of receiving multiple packets at different wavelengths by employing a demultiplexer along with a photodiode (PD), a transimpedance amplifier and a BM-CDR. The architecture follows the design principles described in detail in[11], where header processing was performed by an FPGA, contention resolution by Shared Delay Line Banks (S-DLB) and wavelength routing to the destination port was achieved by an AWGR. On every Plane, the FPGA controller undertakes the crucial role of orchestrating the forwarding operation throughout the switching stages, by sequentially performing the operations described in[11]. Fig.1(c) depicts a photo of the deployed Hipoλaos switch prototype, with the first rack level including the FPGA control-plane along with the Tunable WCs that are realized by SOA-MZI devices. The second rack level hosts the S-DLB comprising three fiber delay lines that offer up to three packet buffering capacity, with system-level simulations revealing that high throughput values of >90% can be achieved with this buffering configuration[10]. Finally, the third rack level comprises a 32x32 AWGR, undertaking the wavelength routing, while the fourth rack level includes the EDFAs and the other optical components required (Pol.Ctrls, ODLs etc).

Experimental validation of end-to-end packet switched performance at 25Gb/s per port

In order to demonstrate an end-to-end true packet switched link through the 1024-port Hipoλaos architecture operating with 25Gb/s burst-mode optical packets, a fully functional Plane has been experimentally validated, comprising an S-DLB with three delay lines (direct, tp, 2tp), interconnected to a 32x32 AWGR along with a BM-CDR. The experimental setup used for the evaluation of the proposed switch is illustrated in Fig. 2(a). A Pulse Pattern Generator (PPG) was used for data generation at 25Gb/s while a Xilinx FPGA was employed for controlling the switching functionalities with a processing latency of 97.28ns. A CW beam at λ_0=1552.5nm was launched into a $LiNbO_3$ modulator driven by the PPG to produce 26.112μs NRZ data packets at 25Gb/s, comprising a 153.6ns preamble, a 25.9584μs payload and an inter-packet guardband of 153.6ns. The modulated signal was injected to an EDFA for amplification, then filtered by a 5nm Optical Bandpass Filter (OBPF) and fed to the Hipoλaos switch where a 1/32 splitter was used to emulate the actual losses of a 32x32 Plane within a 1024x1024 switch. The resulting signal after a second stage of amplification and filtering, was launched in a 50/50 splitter and split into two identical signals feeding the ports D and E of the SOA-MZI#1, realizing the differentially-biased scheme, while a CW beam at λ_4=1553.37nm was injected into port H as an auxiliary holding beam. The output signal C was then injected to the S-DLB, where 1/32 combiners were used to emulate the 1024x1024 combining ratio. The signal exiting the S-DLB, after being amplified by an EDFA and filtered by a 5nm OBPF, was injected into the Out.1 WC and

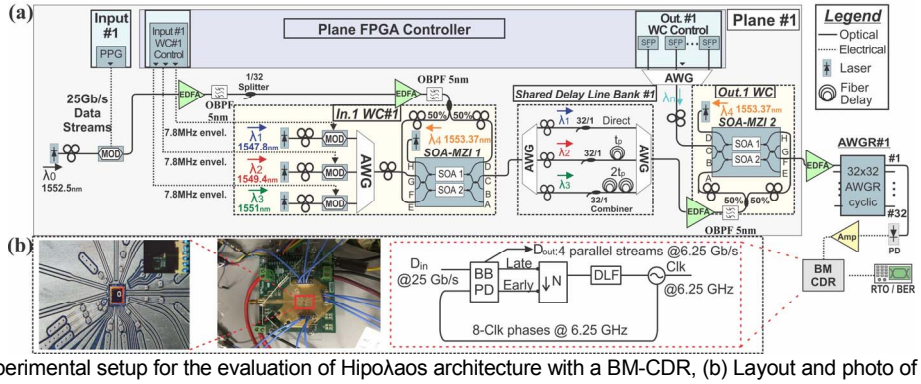

Fig. 2. (a) Experimental setup for the evaluation of Hipoλaos architecture with a BM-CDR, (b) Layout and photo of the BM-CDR.

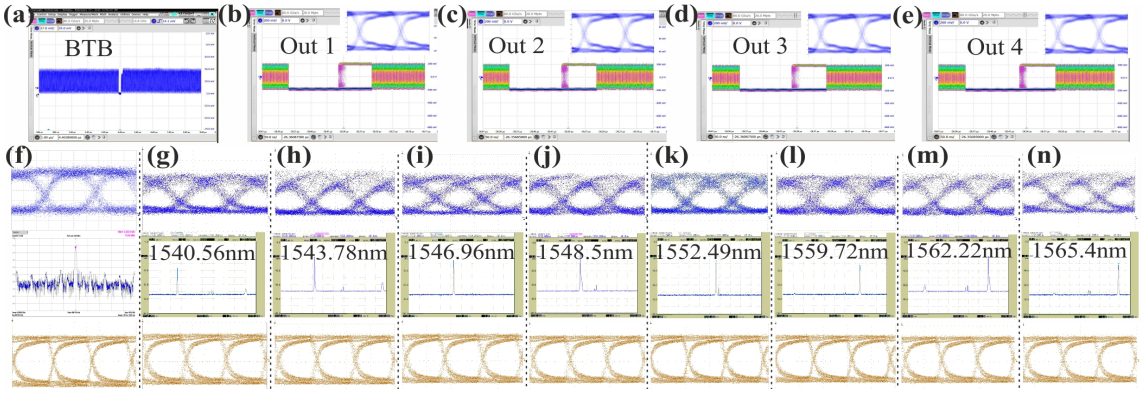

Fig. 3. Experimental results. (a) Modulated output signal (BTB) @25Gb/s with 153.6ns guardband, (b)-(e) the 4 outputs of the BM-CDR with ~50ns mean locking time and respective eye diagrams, (f) BTB optical eye diagram (top) - electrical spectrum of the BM-CDR locked @6.25GHz (mid) -electrical eye diagram of a BM-CDR output (bottom), (g)-(n) Optical eye diagrams (top) and optical spectrums (mid) of AWGR outputs #1, #5, #9, #11, #16, #25, #28 and #32 - electrical eye diagrams of a BM-CDR output (bottom). X-axis scale for (a): 1μs/div, (b)-(e): 50ns/div, (f)-(n): optical eye diagrams: 10ps/div, electrical spectrum 10MHz/div, optical spectrums: 4nm/div, electrical eye diagrams: 50ps/div, Y-axis scale for (a)-(e): 200mV/div, (f)-(n): optical eye diagrams: 5mV/div, electrical and optical spectrums 10dB/div.

at ports A and H of the SOA-MZI#2. A CW holding beam at λ_4=1553.37nm was launched into port D, while SFPs were used to generate the input signals injected at port C. The signal emerging at output G, after amplification, was injected to input port#1 of the AWGR and sequentially switched to each AWGR output by enabling a different SFP at the FPGA. The signal at the AWGR outputs was transferred to the electrical domain by a PD and after amplification, was launched into the BM-CDR to handle the variation in optical power and phase-mismatch on a packet-by-packet basis. Fig.2(b) depicts the BM-CDR that was used, comprising a bang-bang phase detector (BB-PD), a subsampling block, a digital loop filter (DLF) and a digitally controlled oscillator (DCO). In order to relax the circuit requirements, the output of the BM-CDR was demultiplexed in 4 data signals at a clock rate of 6.25GHz that were evaluated separately. A more detailed description of the BM-CDR can be found in[15].

The experimental results for 25Gb/s burst-mode operation are presented in Fig.3(a)-(e). More specifically, Fig.3(a) depicts the output trace of the modulator (BTB) at 25Gb/s, showing two consecutive 26.112μs long data packets that are subsequently routed through the Hipoλaos switch and received at the BM-CDR circuit. Fig.3(b)-(e) illustrate the obtained traces from a real-time scope and eye-diagrams from the 4x6.25Gb/s demultiplexed signals at the BM-CDR output, revealing a mean locking time of ~50ns. The captured traces were evaluated off-line via the Matlab software, revealing error free operation for 10^6 bits. The same performance has been obtained for all possible routing paths through the Hipoλaos switch, i.e. for all possible wavelength combinations that correspond to the 3 delay stages and the 32 AWGR ports. A more in-depth evaluation of the BER performance has been carried out by employing a continuous 25Gb/s PRBS7 signal instead of packet-level traffic to allow for monitoring the BM-CDR output at an error detector. Fig.3(f) shows the eye diagram at the modulator output, along with its electrical spectrum at the BM-CDR output revealing

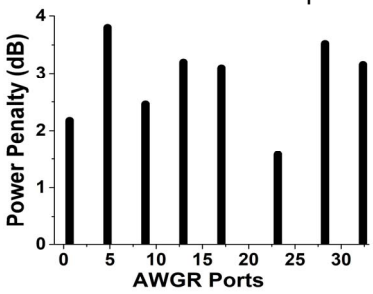

Fig. 4. Power Penalty at error-free operation for 8-channels of the AWGR.

successful locking at 6.25GHz, with the respective electrical eye diagram at 6.25Gb/s at 1 of the 4 outputs of the BM-CDR being depicted at the bottom. Fig.3(g)-(n) depict the eye diagrams and spectrums of the optical signal after being routed to outputs #1, #5, #9, #11, #16, #25, #28 and #32 of the AWGR, along with the corresponding electrical eye diagrams obtained by the BM-CDR. In this case, different SFPs+ were used to evaluate the broadband operation of the switch, by selecting equally distributed wavelengths from 1540.56nm-1565.4nm to cover the wavelength range supported by the AWGR channels. Error-free operation (10^{-9}) was obtained for all channels, with PP values for these 8-indicative output-channels being depicted in Fig.4 and revealing a mean PP of 2.88dB with a max value of <4dB compared to the BTB.

Acknowledgements

This work was supported by the EU projects 5G-PHOS (761989) and NEBULA (871658).

References

[1] “Cisco Global Cloud Index: Forecast and Methodology, 2016–2021 White Paper,” (2018).

[2] G. Zervas, et al, "Disaggregated Compute, Memory and Network Systems: A New Era for Optical Data Centre Architectures," OFC, W3D.4, (2017).

[3] S. P. Cole and D. Szabados, “At a Glance: Tomahawk® 3 is the first 12.8 Tb/s chip to achieve mass production,” (2019).

[4] “Innovium Teralynx 7 Data Center Ethernet Switch Family Product Brief,” (2019).

[5] A. Ghiasi, “Large data centers interconnect bottlenecks”, Opt. Express (2015).

[6] Y. Yin, et. al., "LIONS: An AWGR-Based Low-Latency Optical Switch for High Performance Computing and Data Centers," in IEEE Journal of Selected Topics in Quantum Electronics, vol. 19, no. 2, pp. 3600409-3600409, March-April 2013.

[7] K. Ueda et al, "Large-Scale Optical Switch Utilizing Multistage Cyclic Arrayed-Waveguide Gratings for Intra-Datacenter Interconnection," PJ. 9, 1-12 (2017).

[8] M. Moralis-Pegios et al, "A 1024-Port Optical Uni- and Multicast Packet Switch Fabric," J. of Lightw. Techn. 37, 1415-1423 (2019).

[9] M. Moralis-Pegios et. al., "Multicast-Enabling Optical Switch Design Employing Si Buffering and Routing Elements," in IEEE Photonics Technology Letters, vol. 30, no. 8, pp. 712-715, 15 April15, (2018)

[10] N. Terzenidis et. al, “High-port and low-latency optical switches for disaggregated data centers: The Hipoλaos switch architecture,” JOCN 10, 7, B102-B116, 2018.

[11] N. Terzenidis et al, "High-port low-latency optical switch architecture with optical feed-forward buffering for 256-node disaggregated data centers," Opex 26, 8756-8766 (2018).

[12] N. Terzenidis et al., “A 25.6 Tb/s capacity 1024-port Hipoλaos Optical Packet Switch Architecture for disaggregated datacenters”, accepted at OFC 2020.

[13] A. Forencich et al., "A Dynamically-Reconfigurable Burst-Mode Link Using a Nanosecond Photonic Switch," in JLT (2020).

[14] P. Bakopoulos et al., "NEPHELE: An End-to-End Scalable and Dynamically Reconfigurable Optical Architecture for Application-Aware SDN Cloud Data Centers," in IEEE Communications Magazine, vol. 56, no. 2, pp. 178-188, Feb. (2018).

[15] M. Verbeke et al., "A 25 Gb/s All-Digital Clock and Data Recovery Circuit for Burst-Mode Applications in PONs,", JLT 36, 1503-1509(2018).

[16] A. Tsakyridis, et. al., “10 Gb/s optical random access memory (RAM) cell,” Opt. Lett. 44, 1821-1824 (2019).

[17] G. Mourgias-Alexandris, et. al., "An all-optical neuron with sigmoid activation function", Opt. Express 27, 9620-9630 (2019).

[18] A. Tsakyridis et. al, “Theoretical and experimental analysis of Burst-mode Wavelength Conversion via a Differentially -biased SOA-MZI”, Journal of Lightwwave Technology in early access. 10.1109/JLT.2020.2995471 (2020).

[19] M. Spyropoulou et. al., "40 Gb/s NRZ Wavelength Conversion Using a Differentially-Biased SOA-MZI: Theory and Experiment," JLT 29, 1489-1499 (2011).