A 25.6 Tbps capacity 1024-port Hipoλaos Optical Packet Switch Architecture for disaggregated datacenters

N. Terzenidis[1], A. Tsakyridis[1], G. Giamougiannis[1], M. Moralis-Pegios[1], K. Vyrsokinos[2], N. Pleros[1]
Department of Informatics, Center for Interdisciplinary Research & Innovation, Aristotle University of Thessaloniki, Thessaloniki, Greece
Department of Physics, Center for Interdisciplinary Research & Innovation, Aristotle University of Thessaloniki, Thessaloniki, Greece
nterzeni@csd.auth.gr

Abstract: We demonstrate experimentally the feasibility of a 25.6Tb/s capacity Hipoλaos optical packet switch architecture with 1024 in/out ports operating at 25Gb/s, presenting successful contention resolution and error-free operation with a control plane latency of 97.28ns.

1. Introduction

The continuous growth of hyperscale data centers (DC), designed to handle cloud applications, social media and big data analytics, is currently driving the insatiable bandwidth demand between servers on the same DC[1], which in turn has pushed the envelope on switch capacity to the current 12.8Tb/s values offered by several switch ASIC engines like Broadcom's Tomahawk 3[2], Innovium's Teralynx 7[3] and Intel/Barefoot Tofino 2[4]. Scaling their capabilities to the next level of 25.6Tb/s capacities as prescribed by the Ethernet roadmap encounters, however, significant challenges due to several barriers like the high-power SerDes interfaces, the packaging constraints and the severe signal integrity degradation, rendering the 25.6 Tb/s deployment path still a rather unknown parameter.

At the same time, switch capacity scaling will probably have to comply with the new DC paradigm of resource disaggregation, since the huge waste of resources observed in traditional server-centric DCs that may often reach up to 50% has forced DC operators to invest in solutions that will considerably improve resource utilization and energy efficiency. Along this line, resource disaggregation [6] emerges as a groundbreaking architecture that could amortize the energy and cost impact caused by the vast diversity in resource demand of emerging DC workloads. However, disaggregated DCs are breaking apart the critical CPU-to-memory path, introducing a challenging set of requirements in the underlying network infrastructure [7]: switch configurations have to support sub-μsec latency values along with high-radix connectivity [8]. Low-latency and high-port optical packet switch deployments [9] have already emerged as promising alternatives to support resource disaggregation, with our recently demonstrated Hipoλaos optical packet switch (OPS) architecture [10]-[11] reporting on sub-μsec latency connectivity with up to 95% throughput for 10Gb/s line-rates per port for both a 256- [10] and a 1024-node DC network [11] using just a limited number of optical fiber-based feedforward packet buffering stages.

In this paper we extend our previous work on the Hipoλaos switch architecture and demonstrate experimentally, for the first time to our knowledge, the feasibility for a 25.6Tb/s capacity 1024-port optical packet switch operating with 25Gb/s NRZ data per port, scaling-up capacity by a factor of 2.5x compared to previous Hipoλaos demonstrations [11] and complying with the emerging 25.6Tb/s switch capacity requirements. A full-fledged Plane of the 1024-port Hipoλaos layout along with a 32x32 AWGR device is experimentally demonstrated at 25Gb/s, presenting successful contention resolution among contending optical packets and error-free operation for all output port combinations. Header processing and switch control are provided via an FPGA-based control-plane that yields a latency of only 97.28ns, scaled-down by a factor of 4.6x compared to previous prototypes and expected to enable drastic reductions in the overall latency, well-below the sub-μsec performance of previously demonstrated Hipoλaos fabrics.

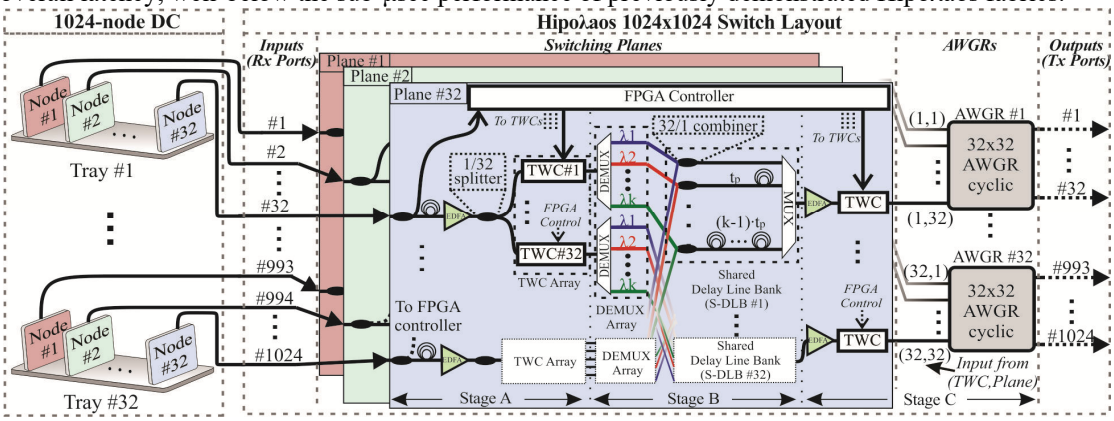

Fig. 1. Illustration of a 1024-node D.C. interconnected via the Hipoλaos switch, configured in a 1024-port layout with 32 Planes and 32 ports/Plane

2. Hipoλaos data- and control-plane architecture

The Hipoλaos optical packet switch architecture has been developed targeting maximum scalability throughout the most fundamental switch characteristics; port-count, datarate and performance. To achieve port-count scalability Hipoλaos employs a hybrid Spanke design, distributing the switch and control operations in functionally independent clusters, named as Planes, while exploiting the wavelength routing capabilities of AWGRs to interconnect the various Planes with the switch outputs. Datarate scalability is ensured by utilizing Wavelength converters (WC) on the different stages of the Spanke layout that are able to operate successfully with up-to 40Gb/s NRZ signals, exploiting the differentially biasing scheme [12]. Finally, performance scalability in terms of increased throughput and decreased latency is realized by taking advantage of AWGR's collision-less WDM routing mechanism in conjunction with the incorporation of small-scale optical feed-forward buffers towards avoiding the latency-inducing electronic buffering or packet drop-and-retransmit collision avoidance schemes. To this end, previous Hipoλaos demonstrations validated the architecture's scalability credentials reaching thousand port connectivity at 10Gb/s per port, with sub-μs p90 latency and throughput values above 97% [10],[11].

Fig. (1) illustrates a 1024-node DC system, organized in Rack-trays and interconnected through the Hipoλaos switch that comprises 32 Planes with 32 ports/Plane, followed by 32 AWGR devices. The architecture follows the design principles described in detail in [10], where header processing and data forwarding via Broadcast and Select takes place at Stage A, contention resolution via Shared Delay Line Banks (S-DLB) takes place at Stage B and wavelength routing to the destination port takes place at Stage C. On every Plane, the FPGA controller undertakes the crucial role of orchestrating the forwarding operation throughout the switching Stages, by sequentially performing the following operations [10]; lane deskewing, routing table lookup and selection of S-DLB for every packet, arbitration between packets contenting for the same S-DLB and activation of respective WC control signals.

3. Experimental evaluation at 25Gb/s per port

In order to verify the feasibility of the Hipoλaos switch at 1024-port configurations with 25Gb/s port-speed, a full-fledged Switch Plane, comprising a 2-packet-size S-DLB, followed by a 32x32 AWGR device was experimentally evaluated utilizing the setup illustrated in Fig. 2. The Plane's controller was developed in a Xilinx Ultrascale board that featured transceiver interfaces providing up-to 25Gb/s NRZ operation. In order to evaluate the contention resolution operation of the switch, 2 data streams generated by the same FPGA board were provided to respective $LiNbO_3$ modulators in order to modulate a CW laser beam at λ_0=1552.5nm and were subsequently provided to the switch input ports. The streams were looped back to the FPGA, exploiting the integrated loopback functionality of the FPGA transceivers towards achieving header processing and WC control signal generation. Each data stream comprised five 640-bit-long 25Gb/s NRZ data packets, with an inter-packet guard-band of 64 bits. The optical packet stream injected at Input #1 is split in two signals, after entering Stage B, that arrive at ports D and E of SOA-MZI#1 serving as control signals. The same procedure is followed for the packet stream at Input #2 of the switch. At the same stage, three CW laser beams tuned at 1547.8nm, 1549.4nm and 1551nm were modulated to produce 616-bit-long 7.8125MHz envelopes. All three envelopes are subsequently multiplexed in an AWG and fed as SOA-MZI#1 input signal into port G. The same procedure is followed for SOA-MZI#2. Output signals from ports C and B of both SOA-MZI#1 & #2 are demultiplexed in separate AWGs. The demultiplexed signals are combined to the appropriate delay line of the S-DLB block according to each signal's wavelength. The signal entering Stage C, after being amplified and filtered in a 5nm filter, splits into two identical signals that arrive at ports A and H of SOA-MZI#3 serving as control signals. Small-form-factor-pluggables (SFP) attached to the FPGA, were used to produce 580-bit-long 7.8125MHz envelopes that are subsequently fed into port C, serving as input of SOA-MZI#3. Finally, the SOA-MZI#3 output is injected to input #1 of a 32x32 AWGR device. The signal at the AWGR outputs is

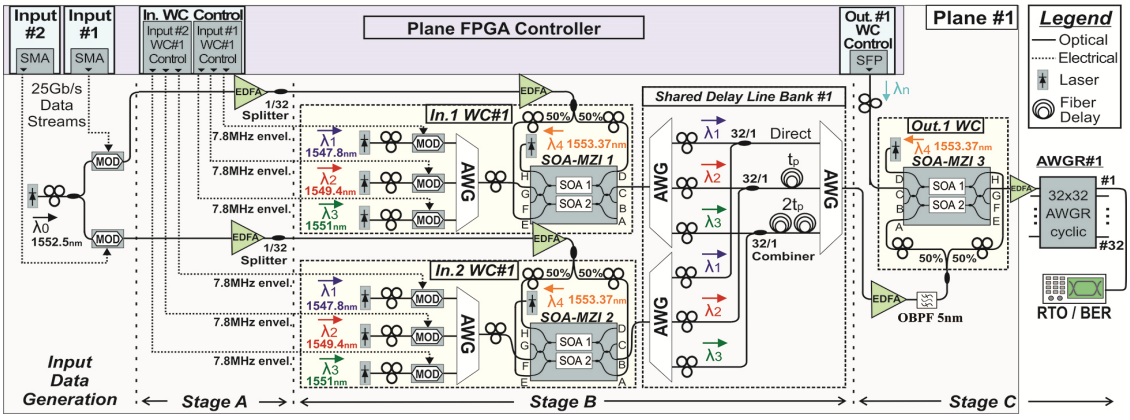

Fig. 2. Experimental setup for the evaluation of the 1024-port Hipoλaos layout at 25Gb/s, comprising a single Plane along with a 32x32 AWGR

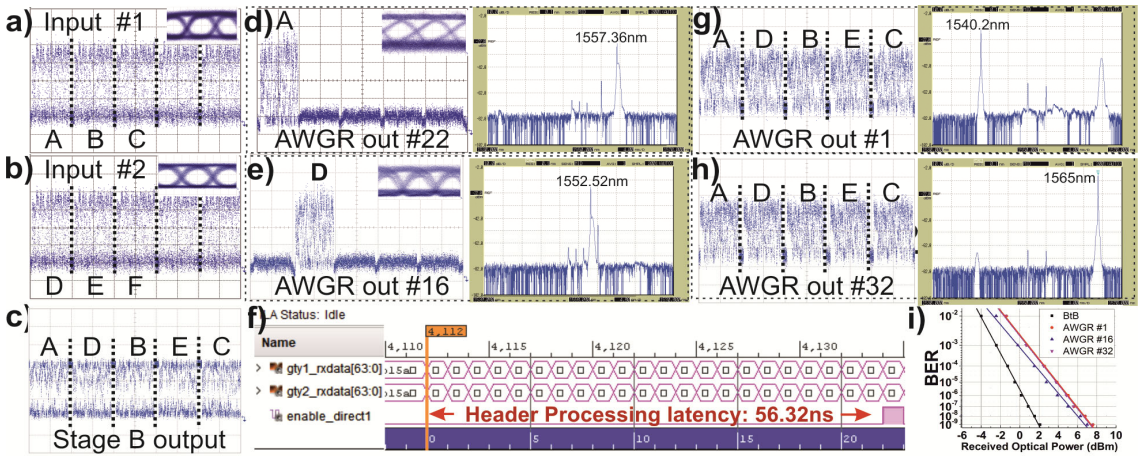

Fig. 3. Experimental results. X-axis scale for traces (a)-(e) and (g)-(h); 12.5ns/div. Y-axis scale for spectra (d)-(e) and (g)-(h); 10dB/div. (a)-(b) Data traces and eye diagrams for Input #1 and #2 signals, c) traces at Stage B output depicting successful contention resolution, d)-e) data traces, eyes and spectrums when routing packets #A and #D to output ports #22 and #16, f) trace from Xilinx ILA illustrating the header processing latency, g)-h) data traces, eyes and spectrums when routing all packets to Output ports #1 and #32, i) BER measurements.

recorded at a digital sampling oscilloscope. The optical envelope signals at all WCs were modulated using $LiNbO_3$ modulators controlled by the FPGA, while EDFAs and optical filters were utilized at different stages of the setup. Fig. 3 presents the obtained results, when the first three packets at every input comprise actual data packets to be forwarded by the switch while the last two are dummy packets designated to maintain the synchronization of the receiver CDR. Fig. 3(a) and (b) illustrate oscilloscope traces and eyes of the optical data stream injected at Inputs #1 and #2. Successful contention resolution is demonstrated at Fig. 3(c) where packets at the output of S-DLB Stage are successfully re-ordered to A-D-B-E-C, considering higher priority for packets from input#1 compared to input#2. Fig. 3(d) and (e) present the traces, eyes and spectrums for packets #A and #D respectively, when each packet is routed to a different output port #22 and #16, through the activation of different SFP+ modules in order to provide envelopes matching with the respective AWGR port resonances. Respective results were collected for all packets of the 2 data streams. Fig. 3(f) illustrates the traces collected from the FPGA Integrated Logic Analyzer (ILA) during the header processing operation for both inputs, revealing a latency of 22 clock cycles from the packet headers (gty_rxdata) arrival until the respective control signal (enable_direct1) generation. The SerDes latency was measured to be 16 cycles, concluding to a total control-panel latency of 97.28ns considering the achieved clock frequency of 390.625MHz. Fig. 3(g) and (h) present the traces, eyes and spectrums when routing all packets at output ports #1 and #32. The resonance observed at the right side of Fig. 3(g) at 1565.8nm corresponds to the ASE noise of the EDFA at a spectral distance equal to the AWGR FSR. Finally, to validate error-free performance of the system and assess the signal quality, BER measurements were obtained covering the complete set of AWGR output channels, from 1540.2nm to 1565nm. Fig. 3(i) depicts the BER curves obtained at the lower, middle and upper spectral channel of the AWGR, revealing a power penalty of 5.6dB, 5dB and 5.5dB respectively compared to the BtB scenario. Similar results were obtained for all channels with a minimum and maximum power penalty value of 5dB and 5.6dB respectively, for error-free performance. The power efficiency of the switch was calculated at 182.4pJ/bit, considering gated operation of the SOA-based WCs.

4. Acknowledgments

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

5. References

[1] "Cisco Global Cloud Index: Forecast and Methodology, 2016–2021 White Paper," (2018).
[2] S. P. Cole and D. Szabados, "At a Glance: Tomahawk® 3 is the first 12.8 Tb/s chip to achieve mass production," (2019).
[3] "Innovium Teralynx 7 Data Center Ethernet Switch Family Product Brief," (2019).
[4] "Barefoot Networks Unveils Tofino™ 2, the Next Generation of the World's First Fully P4-Programmable Network Switch ASICs," (2018).
[5] C. Minkenberg et al, "Reimagining Datacenter Topologies With Integrated Silicon Photonics," JOCN. **10**, B126-B139 (2018).
[6] A. Reale et al, "Experiences and challenges in building next-gen optically disaggregated datacenters," PSC, 1-3, (2018).
[7] P. X. Gao et al, "Network requirements for resource disaggregation," Proc. 12th USENIX, 249-264, (2016).
[8] G. Zervas, et al, "Disaggregated Compute, Memory and Network Systems: A New Era for Optical Data Centre Architectures," OFC, W3D.4, (2017).
[9] K. Ueda et al, "Large-Scale Optical Switch Utilizing Multistage Cyclic Arrayed-Waveguide Gratings for Intra-Datacenter Interconnection," PJ. 9, 1-12 (2017).
[10] 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)
[11] M. Moralis-Pegios et al, "A 1024-Port Optical Uni- and Multicast Packet Switch Fabric," JLT **37**, 1415-1423 (2019).
[12] M. Spyropoulou et al, "40 Gb/s NRZ Wavelength Conversion Using a Differentially-Biased SOA-MZI: Theory and Experiment," JLT **29**, 1489-1499 (2011)