A 1024-port sub-μsec latency Optical Packet Switch using the Hipoλaos λ-routed modified Spanke switch architecture

Nikos Terzenidis(1), Miltiadis Moralis-Pegios(1), George Mourias-Alexandris(1), Kostantinos Vyrskinos(2), Nikos Pleros(1)

(1) School of Informatics, Aristotle University of Thessaloniki, 54624, Greece, nterzeni@csd.auth.gr
(2) School of Physics, Aristotle University of Thessaloniki, 54624, Greece

Abstract We scale-up the Hipoλaos switch architecture and demonstrate experimentally feasibility of a 10Gb/s-line-rate 1024-port configuration, with 4-packet buffering capacity. Error-free performance with <3.2dB power penalty is obtained for the Hipoλaos prototype, while network-level analysis reveals sub-μsec p90-latency and >90% throughput.

Introduction
The transition from traditional server-centric Data Center (DC) architectures towards disaggregated compute, memory and network systems1 can offer increased resource utilization at a reduced cost and energy envelope, enforcing however, the use of high-port switching with stringent network latency and bandwidth requirements2. Optical Circuit Switches (OCS) can yield the necessary high-port connectivity and have already been employed in disaggregated systems3, but their ms switching time values limit their employment as slow reconfigurable backplanes3. Optical Packet Switch (OPS) layouts can definitely offer ns-scale switching, however retaining sub-μs latency values, as required by disaggregated memory systems1,3, comprises a major challenge as the port-number increases4.

We have recently demonstrated the Hipoλaos switch architecture in a 256-port OPS layout5 that retains sub-μs latency values by exploiting a Spanke-based, λ-routed optical buffering and forwarding scheme. The Hipoλaos architecture overcomes the increased path computation time restrictions of high-port OPS fabrics that rely on centralized-control schemes6,7, by taking advantage of the distributed control switching enabled by Spanke designs4. At the same time, it avoids the latency burden imposed by packet-drop-and-retransmit operations in buffer-less Spanke OPS fabrics4, by incorporating a wavelength-routed optical delay line buffering stage, which allows also for increased throughput values, exceeding the 70% upper bound of buffer-less configurations4.

In this paper we extend our previous work5 and scale-up the Hipoλaos OPS architecture into a 1024-port configuration, that still maintains sub-μs latency performance along with >90% throughput values. The 1024-port Hipoλaos layout utilizes a number of 32 vertical 32x32 switch Planes interconnected to 32 horizontal 32x32 AWGRs, where every Plane employs a 1:32 Broadcast-and-Select (BS) switching scheme along with λ-routed optical delay line buffering stages. Feasibility of the 1024-port Hipoλaos layout has been validated at 10Gb/s line-rates for a 4-packet optical buffering stage, revealing error-free performance with a maximum power penalty value of 3.22dB for all 160 possible wavelength combinations.

Network-level simulation analysis for a 1024-node system reveals a p90 latency of <0.8μsec

Fig. 1: 1024-port Hipoλaos layout with k-1 packet buffering capacity and experimental traces for contending packet sequences entering through inputs #32 and #1024 and leaving the switch via output#8 or #9 after resolving contention in S-DLB#1, with k=3
with >90% throughput even when using a small-scale optical buffering of only 4 packets.

1024-port Hipoλaos layout and results

The Hipoλaos architecture\textsuperscript{5} relies on a latency-optimized design, through its modified Spanke layout, that offers distributed control while supporting efficient contention resolution via the employment of optical delay line buffering. The generic layout of the 1024-port Hipoλaos configuration is depicted in Fig. 1. To overcome the inherent scalability limitation of the BS schemes, the control and switching functions are distributed in small independent clusters denoted as Planes. Every 32 input ports enter a discrete Plane, where an optical coupler is used at every input port for splitting the incoming signal and directing one part to the FPGA for processing its header. Following this, the incoming signal gets amplified and then split in a 1:32 optical splitter (Stage A) to enter 32 respective SOA-MZI-based Tunable Wavelength Converters (TWCs). The “Select” function is performed by means of activating only one out of the 32 TWCs, with the TWC input wavelength being defined by the FPGA. Depending on the TWC element that is activated, the incoming signal will be directed to one of the 32 available outgoing ports of the respective switch Plane. The Select stage is followed by contention resolution blocks (Stage B) that comprise AWG demultiplexers and 32 Shared Delay-Line Banks (S-DLBs) serving as feedforward buffering setups. One S-DLB per outgoing port of the Plane is employed, with every S-DLB comprising, in general, \( k \) optical delay lines that offer a buffering time that ranges from 0 up to \( k-1 \) packet slots (\( t_p \)). Every delay line includes a 32:1 combiner at its front-end to collect the packets from all Plane’s incoming ports that have to experience the same buffering time and be directed to the same Plane’s outgoing port. Once the packets have been properly arranged in time via the S-DLB, they combine through an AWG and enter stage C of the switch. Stage C comprises a TWC at every Plane output port followed by 32 32x32 AWGRs. The i-th output ports of all 32 planes connect to respective ports of the same 32x32 AWGR, so that Stage C allows for collision-less routing between signals emerging from different Planes.

A representative example of the switch operation, when optical packets enter through two different input ports of the same Plane (Plane #1) and contend for the same output, is shown via the pulse-trace insets in Fig.1. Packet traces and respective eye diagrams of the optical signals entering Plane#1 through switch input#32 and #1024 are illustrated on the bottom-left part of Fig. 1. An S-DLB, featuring \( k=3 \) delay lines, was considered for this specific example, along with a static priority scheme, where packets from Input#32 were assigned higher priority compared to packets from Input#1024. As a consequence, an A-D-B-E-C packet sequence is obtained at the S-DLB output, with packet F being dropped due to unavailable empty delay lines. The envelope signals generated by the FPGA to control TWC#1 of Input#32, in order to ensure this packet sequence, are illustrated on the top-left part of Fig.1. The respective FPGA signals, controlling TWC#1 of Input#1024, comprise a packet-envelope during the 1\textsuperscript{st} packet-slot.

Fig. 3: a) Experimental setup of a 1024-port Hipoλaos Plane with 4 packet buff. capacity (\( k=5 \) S-DLB), b) photo of the Hipoλaos prototype
forwarding packet D to the \( t_p \)-delay line, along with a packet-envelope during the 2\(^{nd} \) packet-slot forwarding packet E to the 2\( \cdot t_p \)-delay line. No envelope signal was generated during the 3\(^{rd} \) packet-slot, enforcing in this way the drop-operation for packet F. Successful forwarding of the complete A-D-B-E-C packet sequence to the same switch output port, as illustrated in the pulse trace insets above and below AWGR#1, is accomplished by selecting the appropriate input wavelength of the Plane#1-Output#1 TWC. The pulse-traces correspond to two different cases where the packet sequence emerges either at switch output port #8 or #9, respectively.

The throughput and packet latency performance of the switch has been investigated using Omnet++. A 1024-node system has been modelled featuring 10Gbps node-switch bandwidth and 72-bytes packet size, with 64-bytes comprising the payload and 8-bytes standing for header, synchronization and guardband requirements. Fig 2 (a) presents the respective throughput versus load results. The number of buffers per S-DLB ranges from 0 to 8 (1 to 9 delay lines), revealing linear throughput increase until 80% load, with the maximum throughput reaching ~95% utilizing just 8 buffers. Fig. 2(b) presents the mean packet delay versus the offered load results. The mean latency ranges between 550 and 650 ns for loads until 80%, with the biggest part associated to the PCS/PMA latency of the FPGA that was measured to be 456ns during the prototype experimental evaluation, and tops at ~800ns. Fig. 2(c) presents the p90 latency that remains below 1\( \mu \)s even for 8 packet-buffers per S-DLB.

Experimental Performance Analysis

To demonstrate feasibility of the 1024-port HiPoAaos layout and assess performance across all possible delay line/output port combinations, a prototype Plane with 4 packet buffering capacity was experimentally evaluated. Fig. 3 (a) depicts the experimental setup, while Fig. 3 (b) presents a photo of the actual deployed HiPoAaos prototype. An Altera Stratix V FPGA was used for: i) Input data generation, utilizing an SFP transceiver ii) Header extraction and processing iii) Generation of the required electrical control signals of the modulators, used for creating the optical control signals of In.1 TWC#1. The input stream comprised three 405-bit-long 10.3125Gb/s NRZ data packets and two dummy packets, while the envelopes had a duration of 410-bits. A total number of 5 laser beams at different wavelengths were injected in respective optical modulators, and the resulting signals were multiplexed to form the input signal of In.1 TWC#1. The S-DLB comprised a total number of 5 optical branches, offering delays of 0, 1, 2, 3 and 4 packet-slots (\( t_p \)) respectively.

BER performance was evaluated for all possible buffering and routing states of the switch, taking into account that i) the packet arriving as the control signal at SOA-MZI#2 can be carried by 5 different wavelengths depending on its prior buffering time, and ii) the TLS can be tuned at 32 different wavelengths depending on the AWGR outgoing port requested by the packet. The experimental procedure was carried out by optimizing, at first, the BER for the combination of wavelengths that correspond to the 2\( \cdot t_p \)-delay line and the AWGR output port #8 and then evaluating all possible combinations without altering the SOA-MZI#2 electrical/optical driving settings. Fig. 4 depicts the measured power penalty, at 10\(^{-9} \) BER, for all the 160 possible combinations, revealing a mean power penalty value of 2.42dB with a standard deviation of 0.40 and minimum/maximum values of 1.5dB and 3.22dB, respectively.

Acknowledgements

EU-H2020 L3MATRIX (688544), ICT-STREAMS (688172).

References