Sub-µsec Latency High-Port Optical Packet Switch Fabrics for Disaggregated Computing: the Hipoλaos OPS Architecture

M. Moralis-Pegios(1), N. Terzenidis(1), G. Mourgias-Alexandris(1), K. Vrysokinos(2), N. Pleros(1)  
(1) School of Informatics, Aristotle University of Thessaloniki, 54624, Greece  
(2) School of Physics, Aristotle University of Thessaloniki, 54624, Greece  
Tel: +30 2310 998026, Fax: +30 2310 998019, e-mail: mmoralis@csd.auth.gr

ABSTRACT

We present a high-radix optical switch architecture, called Hipoλaos, that exploits hybrid Broadcast-and-Select/Wavelength routing along with small-size optical buffering, towards meeting the requirements of Disaggregated Data Centers. The architecture is evaluated for a 256-node system through simulations revealing sub-µs latency, while feasibility of the switch is experimentally demonstrated at 10Gb/s.

Keywords: Optical interconnects; Optical Packet switching, Broadcast and Select, Optical Buffering, Disaggregated Data Centers, Silicon Photonics

1. INTRODUCTION

Departing from traditional server-centric Data Center (DC) architectures towards disaggregated compute, memory and network systems [1], that can offer increased resource utilization at a reduced cost and energy envelope, the use of high-port switching conforming to stringent latency and bandwidth requirements becomes a necessity [2],[3]. To this end, various architectures have been investigated exploiting different optical switching technologies. Optical Circuit Switches (OCS) can yield the necessary high-port connectivity, but their increased switching time values limit their employment as slow reconfigurable backplanes [4]. Optical Packet Switch (OPS) fabrics can definitely offer nsec-scale switching time arrangements, however low latency values can hardly be maintained as the number of ports increases [5]. High-port count OPS fabric implementations relying on cascaded AWGRs [6] and on combined Delivery-and-Coupling with AWGR-based wavelength routing [7] necessitate a centralized control over the entire number of ingress ports, significantly increasing path computation times. Recent work has indicated that sub-µs latencies with high-port OPS fabric designs, as required in disaggregated DCs, can be offered only via distributed control over small clusters of inputs [5]. This has been verified in a SOA ON/OFF-based bufferless Broadcast-and-Select (BS) switch implementation, employing electronic edge buffering and retransmission and demonstrating sub-µs latencies with a maximum throughput of 70% for higher than 64x64 switch designs.

In this paper, we demonstrate a novel optical switch architecture, named Hipoλaos, for disaggregated DCs[8], exploiting FPGA-based distributed control over a combined BS and AWGR-based wavelength routing architecture in conjunction with contention resolution stages that employ small-size optical feedforward buffering in order to offer increased throughput while retaining low-latency characteristics. The architecture is evaluated for a 256-node system where simulation-based analysis reveals that even a small buffer size of 2 optical packets per contention resolution stage, yields <620nsec latency with >85% throughput for up to 100% loads. Feasibility of the proposed 256-port architecture has been experimentally validated with 10Gb/s optical data packets using 1:16 optical splitters and SOA-MZI wavelength converters (WC) along with fiber delay lines for a 2-packet buffer, followed by an additional SOA-MZI tunable WC and a 16x16 AWGR. Error-free performance in all different switch input/output combinations has been obtained with a power penalty of <2.5dB.

2. Switch Architecture and Simulation Performance Analysis

An example of the proposed optical switch architecture is illustrated in Fig. 1 where 16 rack-trays, each one incorporating 16 nodes, form a 256-node disaggregated DC Rack system interconnected via a 256x256 optical switch. The switch is organized in 16 vertical Planes followed by 16 horizontal AWGRs. On the ingress path, every node communicates with a specific Plane over a different fiber link. Input port allocation per Plane is performed so as to have node#i from every tray connecting to Plane#i, suggesting that Plane#i inputs (i,j) designate connection to node#i from tray#j, with i,j=1,2,..16. The Plane’s role is to aggregate traffic from 16 input ports and forward it using a BS scheme to the proper contention resolution stage that includes optical feed-forward buffers at a packet-size granularity. Contention resolution is performed on a per-output-tray basis, therefore providing the highest throughput under uniform traffic distribution among output trays.

The control operations of the switch are distributed among the switch planes, requiring one FPGA controller per plane and allowing independent header processing and buffering. Synchronous slotted operation is
considered at the Plane edge, where the FPGA simultaneously processes all routing requests. Finally, each Plane is connected to 16 16x16 AWGR devices, each one forwarding data to a specific tray. On the AWGR egress path, the nodes are capable of receiving a WDM signal composed by 16 wavelengths, equal to the number of nodes per tray.

To ease our design description, we classified the architecture into 3 functional stages, presented by the detailed layout of the switch at Fig. 2. Stage A embodies header processing and signal broadcasting, Stage B is responsible for tray selection and contention resolution, while destination node selection takes place at Stage C. At Stage A, part of the optical signal is sent to the FPGA for header processing following optoelectronic conversion. The remaining part of the signal is delayed to account for the required header processing time and is subsequently amplified and broadcasted via a 1:16. At Stage B, tray selection is realized by activating the appropriate tunable Wavelength Converter (WC) that forwards the packet to the respective Tray Contention Resolution (TCR) block. At the same time, the WC control signal dictates the WC input wavelength value, so that the packet is forwarded to the desired delay line of the TCR block through an AWG demultiplexer located at the WC output. Each TCR block comprises a feed-forward optical buffer with K+1 delay lines, inducing delay ranging from 0 to K packet slots (tp) and providing buffering capability up to K packets. Each delay line of TCR#i accumulates optical packets from all 16 WC#i’s using a 16/1 combiner. Afterwards, the individual delay line signals per TCR block are multiplexed, with the FPGA controller ensuring collision avoidance by allowing a single packet to exit the multiplexer at a given packet-slot. At Stage C, routing to the desired node is achieved using tunable WCs and AWGRs. Every input of a single AWGR#k connects to the WC#k’s of every Plane that are

---

**Fig. 1.** Schematic illustration of a 256-node D.C. Rack system, organized in 16 Rack-Trays with 16 nodes per tray, and interconnected via the proposed switch architecture

**Fig. 2.** Detailed Layout of the 256x256 architecture utilizing optical feed-forward buffers with a max size of K packets
destined to the same tray\#k, while AWGR\#k outputs are connected to all nodes of the tray\#k. The node that the packet will be forwarded to is defined again by the FPGA control, by selecting the input wavelength of the Stage C WCs.

The system level performance of the switch in terms of throughput and packet latency has been investigated, using the Omnet++ simulation platform. A 256-node system has been modelled, featuring 10Gbps node-switch bandwidth and 72 bytes packet size, with 64-bytes comprising the cache line content and 8-bytes standing for header, synchronization and guardband requirements. Fig 3 (a) presents the respective throughput versus the offered load results. The number of buffers per TCR ranges from 0 to 2 revealing linear throughput increase until 70% load, with the maximum throughput reaching ~85% utilizing just 2 buffers. Fig. 3 (b) presents the mean packet delay versus the offered load results. The latency reaches a value of 610ns at 100% load with the biggest part associated to the PCS/PMA latency of the FPGA that was modelled to be 450ns, according to measurements performed during the experimental evaluation of the concept, described in detail in Section 3.

3. Architecture experimental evaluation

An optical switch plane with two incoming ports, a 2-packet-size optical feed-forward buffer followed by a 16x16 AWGR plane was experimentally implemented and is shown in Fig. 4. A Stratix V FPGA has been used for generating the 2 incoming packet streams, employing SFP and XFP modules, and the WC control signals. The streams are looped back to FPGA for header processing exploiting 70/30 splitters. Each stream comprises three 405-bit-long 10.3125Gb/s NRZ data packets and two dummy packets, with an inter-packet guard-band of 35 bits. Dummy packets are used for synchronization purposes at the receiver side. Each stream is delayed and amplified before being split at the BS 1/16 splitter. The XFP data stream entering Stage B splits into two identical signals that arrive at ports D and E of SOA-MZI#1 serving as control signals. The same procedure is followed for the SFP data stream. At the same stage, three CW laser beams tuned at 1556.70nm, 1559.10nm and 1559.70nm were modulated to produce 425-bit-long 4.5 MHz 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 TCR block according to each signal’s wavelength. The signal entering Stage C, after being amplified and filtered in a 5nm OBPF, splits into two identical signals that arrive at ports A and H of SOA-MZI#3 serving as control signals. A Tunable Laser Source (TLS) is modulated to produce 415-bit-long 4.5 MHz 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 16x16 AWGR device. Signal at the AWGR outputs is recorded at a digital sampling oscilloscope and evaluated with a BER tester. All WCs relied on the differentially-
biased SOA-MZI configuration [9], while optical envelope signals at all three WCs were modulated utilizing LiNbO3 modulators controlled by the FPGA.

Fig. 4 illustrates experimental results obtained when two data streams were simultaneously injected into the switch, with the TLS of Stage C being sequentially tuned to two different wavelengths ($\lambda_{ch8}=1552\text{nm}$, $\lambda_{ch9}=1552.8\text{nm}$). The SFP input data stream, consisting of packets #A, #B and #C, is illustrated in Fig. 4 (a), while the XFP stream with packets #D, #E, #F are illustrated in Fig. 4 (b). The resulting output traces at AWGR ports #8 and #9 are depicted in Fig. 4 (c) and Fig. 4 (d) respectively, along with the optical spectra and eye diagrams of each packet, revealing proper packet routing and buffering. BER measurements were performed for each packet separately, shown in Fig. 4 (e) and depicting an average power penalty of 2dB.

4. Conclusions
We propose a high-radix low latency switch architecture, named Hipoaos, that aims to meet the requirements of sub-$\mu\text{s}$ latency and high port connectivity of disaggregated DCs, by incorporating, for the first time, optical delay-line-based feed-forward buffering in a high-port switch architecture, while exploiting a hybrid BS/wavelength routing scheme with distributed control. The feasibility of a 256×256 switch at 10Gb/s line-rates was experimentally investigated, while Network performance analysis revealed latency values up to 610nsec and linear throughput increase for up to 70% loads, for uniform traffic profiles. Given the proven credential of the Hipoaos SOA-MZI based switching scheme to operate up to 40 Gb/s [10] along with the promising first demonstrations of the architecture utilizing integrated um-SOI components [11], this work is a first step towards implementation of high-port, high bandwidth, low latency optical switches.

5. Acknowledgments
This work has been supported by the EU H2020 projects ICT-STREAMS (688172) and L3MATRIX (688544).

6. References