Multicasting in a 256-port sub-μsec latency Ηιπολαος switch architecture for disaggregated DataCenters

Miltiadis Moralis-Pegios(1), Nikos Terzenidis(1), George Mourgias-Alexandris(1), Konstantinos Vyrskinos(2), Nikos Pleros(1)

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

Abstract We demonstrate experimentally a multicast enabling Ηιπολαος optical packet switch in a 256×256 port configuration with 4 packet optical buffering capacity for 10 Gb/s data, multicasting up to 5 receiver nodes along with a broadband multicast operation analysis.

Introduction

Resource disaggregation in DC architectures is currently promoted as a novel architectural paradigm that aims to increase resource utilization and meet the high disparity in server resource requirements and diversity in network communication profiles. Disaggregated systems enforce, however, a challenging framework for network infrastructure, requiring high-port count switches with sub-μsec latency and high-throughput performance. At the same time, a large subset of DC applications that employ distributed file systems for storage, high-performance data analytics and MapReduce type of algorithms to process data, inherently require multicast traffic delivery as part of iterative machine learning approaches or as part of Paxos-algorithm-based storage mechanisms.

Multicasting to N receiving nodes via existing DC electronic switches comprises an expensive proposition as it consumes a number of N sending ports. Optics can definitely facilitate in-network multicasting with well-defined benefits in several DC workloads, however only a rather small number of experimental multicast switch prototypes for DC environments has been shown so far. In the meantime, high-port Optical Packet Switch (OPS) fabrics have managed to reach sub-μsec latency levels via distributed control algorithm designs but have still not combined high-port sub-μsec switching with multicast capabilities as would be ideally required in disaggregated DC systems.

In this paper, we demonstrate 10Gb/s multicasting in a 256-port optical switch architecture exploiting a modified layout of the recently reported Ηιπολαος optical switch. Ηιπολαος relies on a granular λ-routed modified Spanke switch design and can yield sub-μs latency performance in 256-node systems with multicast capabilities being introduced through multi-λ-routing at its final AWGR-based switching stage. We extend the, optical delay line-based packet, buffer capacity from 2 to 4 packets and demonstrate its experimental evaluation for all possible buffer/output port combinations, revealing error-free performance with a mean power penalty value of 2.19dB. Successful multicasting is then experimentally validated in a mixed-traffic scenario involving both unicast and multicast packets at 10Gb/s, with a maximum multicast-group of 5 nodes. Error-free operation is reported for all cases with a power penalty of <2dB compared to the unicast operation.

Multicast-enabling 256-port Ηιπολαος layout

Fig.1 illustrates the Ηιπολαος architecture, expanded to provide optical multicasting.

![Fig. 1: Layout of 256-port Ηιπολαος configuration, with k-buffers per S-DLB, along with schematic illustration of multicast operation.](image)
capabilities in a 256-node system that is organized in 16 rack trays. It comprises 16 independent 16×16 switching clusters denoted as Planes, exploiting an Optical Shared Delay-Line Bank (S-DLB) per plane’s output port for optical buffering purposes. The 16 Planes are interconnected to 16 AWGRs so that AWGR#i connects the i-th ports of all Planes. The detailed architecture is described in [13]. Multicasting is introduced additionally to the design reported in [13] by incorporating a Multicast Laser Bank (MLB) with #j tunable lasers at every Plane’s Output Tunable Wavelength Converter (TWC), with #j defining the multicast degree supported (j ≤ 16).

The multicasting operation is described by an example scenario where packet#A emerging at Input #1, propagates through the Hipoλ aos switch and is successfully multicasted to #j-output ports of AWGR#1, as shown in Fig. 1. At Stage A, part of the optical data stream carrying packet#A on wavelength λi is forwarded to the FPGA control plane for header processing, to identify the destination node(s) and determine the subsequent control operations required. Considering that packet#A is destined to the multicast group involving nodes #1-#j of Tray#1, the rest of the optical signal propagates through Stages A and B of the switch similar to the way described in [13]. In this example, all (k-1) buffers of S-DLB#1 are assumed to be free and packet#A gets converted to λ1 in Stage-A TWC#1, being forwarded to the “direct” line of S-DLB#1 that corresponds to zero buffering time. Stage C comprises a TWC at every Plane output port (output#1-#16) followed by sixteen 16x16 AWGRs. Output#i ports of all 16 AWGRs connect to respective ports of the same 16x16 AWGR, so that Stage C allows for collision-less wavelength routing between signals emerging from different Planes. Activating all j-lasers of the MLB in Stage-C TWC#1, which are controlled by the FPGA, TWC#1 performs a seamless “copy” of packet#A to all #j wavelengths. The multiwavelength-converted output signal is forwarded to the AWGR and every wavelength reaches a different AWGR output, so that finally 1-to-j multicasting is obtained and a copy of the packet is delivered to all #j destination nodes.

Experimental Setup and Results

The experimental setup is shown in Fig. 2. An Altera Stratix V FPGA was used for: i) Input data generation, through an SFP transceiver ii) Header extraction and processing iii) Generation of the required electrical control signals for the modulators, used for creating the optical input envelope signals of In.1 TWC and Out.1 TWC. The input stream comprised five 405-bit-long 10.3125Gb/s NRZ packets, while the envelopes had a duration of 410-bits. Five beams (λ1, λ2, λ3, λ4, λ5) were injected in five optical modulators to form respective envelope signals that were then multiplexed and launched as input signals into In.1 TWC. The S-DLB comprised five different optical delay lines, offering delays from zero (Direct) up to four (4tp) packet-slots, so that TWC#1 wavelength conversion to λ1 dictates zero-time buffering, to λ2 one-packet-slot buffering time (1tp), etc. After exiting the S-DLB, the signal enters Out.1 TWC as the control signal. Moreover, packet envelopes generated from up to 5 TLS, at different wavelengths (λn1-λn5), were combined in an optical combiner to form the input signal of Out.1 WC, allowing in this way multicasting for up to 5 nodes. Finally, an 1/16 splitter at the data input and 1/16 combiner in the S-DLB delay branches, were used to emulate the full 256×256 switch losses.

BER performance was evaluated for all
possible buffering and routing states of the switch, considering i) the wavelength of the control signal of Out.1 TWC, and ii) the wavelength of the TLS, used as Out.1 TWC input signal. BER was optimized for the combination corresponding to a buffering stage of 2 packets and the AWGR output port #6 and then all combinations were evaluated without altering the SOA-MZI2 electrical and optical driving settings. Fig. 3 depicts the measured power penalty, at 10^-9 bit-error-rate, for all the 80 possible wavelength combinations, concluding to a mean value of 2.19dB with a standard deviation of 0.31.

Multicasting to 5 recipients can be validated via Fig.4. Fig. 4(a) depicts the Out1.TWC input signal comprising 5 data packets that exit the S-DLB. Fig. 4(b) shows the optical spectrum of the output of Out1.WC when all 5 control envelope signals (\(\lambda_{n1}-\lambda_{n5}\)) were utilized in order to multicast packet A to 5 different AWGR outputs. Fig 4 (c)-(d),(e)-(f),(g)-(h),(i)-(j) and (k-l) illustrate the captured oscilloscope traces, respective optical spectra and eye diagrams of packet A as they exit AWGR channels 8,9,10,11 and 12, respectively. Finally, BER measurements were performed revealing error-free operation with an additional maximum power penalty of 2dB compared to the unicast operation measurements shown in Fig.3.

Fig.5 illustrates results for a mixed scenario involving both multicasted and unicast packets when the two more distant AWGR channels #1 and #16 were selected, corresponding to the worst-case scenario of multicasting across a 12nm channel spacing. Fig. 5(a) depicts the Out1.TWC input signal comprising 5 packets (A-B-C-D-E), while Fig.5(b) and (d) depict the two control envelope signals of Out1.TWC. Fig. 5(c) and (e) illustrate the resulting optical traces monitored at AWGR outputs #1 and #16, respectively. Successful unicasting for packets A and D at output #1 and for packets C and E at output #16 can be confirmed, while packet B is multicasted to both output ports. BER measurements for all individual output data packets are depicted in Fig. 5(f-g), along with the captured eye-diagrams, revealing error-free operation at 10 Gb/s. The power penalty between unicast and multicast packets is negligible.

**Acknowledgements**

EU H2020 projects L3MATRIX (contract nr. 688544) and ICT-STREAMS (contract nr. 688172).

**References**


