Scalable Optical Switch System with Optical Flow Control for Flat Datacenter Architecture

Wang Miao, Stefano Di Lucente, Jun Luo, Harm Dorren, and Nicola Calabretta

COBRA Research Institute, Eindhoven University of Technology, PO Box 512, 5600MB Eindhoven, the Netherlands

An optical flat datacenter network based on scalable optical switch system with optical flow control has been proposed and studied in this paper. The validation of 4x4 switch system shows error free operation at 40 Gb/s with 1dB penalty. The minimum end-to-end latency (including 25m transmission link) is less 300ns and $<10^{-5}$ packet loss for 0.5 load with buffer size of 16 packets is reported.

Introduction

Emerging services such as cloud computing and social networks are steadily boosting the Internet traffic. For every byte transmitted over Internet to or from a data center (DC), 1 GB of data is moved through the DC network (DCN) [1], putting a tremendous pressure on DCN. In large DCs with 10.000's of servers, merchant silicon top-of-the rack (TOR) switches are used to interconnect servers in a group of 40 per rack with 1 Gb/s link (10 Gb/s expected soon). To interconnect the 100's of TORs, with 10/40 Gb/s aggregated traffic (100Gb/s is expected soon) per TOR, current DCN is built up on multiple switches, each with limited port count and speed, organized in fat-tree topology [2]. This multi-layer topology has intrinsic scalability issues in terms of bandwidth, latency, costs and power consumption (lager number of high speed links).

To improve the performance and costs, flattened DCN is currently being widely investigated. Optical technologies have the potential of exploiting the space, time, and wavelength domain for scaling the port count while switching high speed data at nanoseconds time scale with low power and small footprint photonic integrated devices. Despite the several optical switch architectures presented so far [3], no one has been proved a large number of ports while providing a port-count independent reconfiguration time for low latency operation. Moreover, the lack of a practical optical buffer demands complicated and unfeasible system control for store-and-forward operation and contention resolution.

In this work, we propose and experimentally investigate a novel flat DCN architecture for TOR interconnect based on a scalable optical switch system with hardware flow control. Experimental evaluation of a 4x4 optical switch system shows dynamic operation including multicasting and only 1dB power penalty at 40Gb/s date rate. A buffer size of 16 packets sufficiently guarantees $<10^{-5}$ packet loss for 0.5 input load and less than 300ns minimum end-to-end latency could be achieved within 25m distance. Scalability investigation also indicates that the optical switch can potentially scale up to more than 64×64 ports with less than 1.5dB penalty while the same latency is retained.

System operation

The proposed flat DCN architecture based on N×N highly distributed control OPS architecture is shown in Fig. 1. Traffic load is balanced and configured by aggregation controller at the transmitter side. Each cluster groups N TORs, which transmit with

different wavelength λ_1, λ_2…λ_M. Switching is performed based on per packet attached label information. At OPS node, the highly distributed control of the OPS [4] allows processing the M channels of each cluster in parallel, minimizing the reconfiguration time of the switch and thus the latency. Label extractor separates the optical label from the optical payload by using a fiber Bragg grating (FBG). The optical payload is then fed into the SOA based broadcast & select 1xN switch while the extracted label is split into two parts. One of them is detected and processed by the switch controller after optical-to-electrical conversion (O/E). The switch controller checks possible contentions, and configures the 1xN switch to block the contended packets with low priority and to forward packets with high priority [5]. Moreover, the switch controller generates the ACK signals used to acknowledge the aggregation controller on the reception or re-transmission of the packets. The other part of label power is re-modulated in an RSOA driven with the base band ACK signal generated by the switch controller and sent back to cluster side. This fulfills the efficient optical flow control in hardware which minimizing the latency and buffer size. Baseband ACK contributes to easy process by using a low pass filter and no interference with label information.

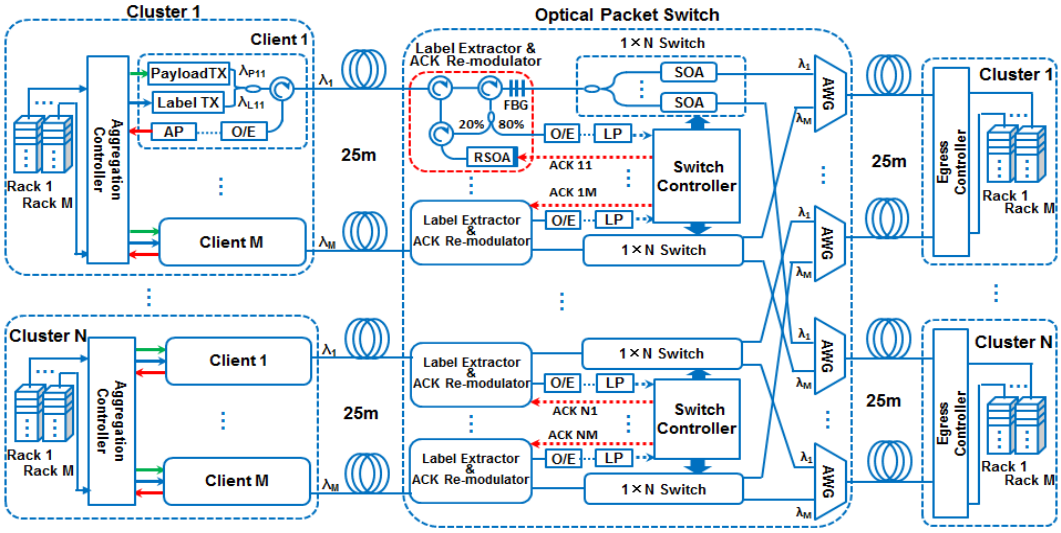

Fig. 1. WDM OPS architecture with highly distributed control.

Experimental set-up and results

For the validation of the DCN, we experimentally investigate the full dynamic operation including flow control of a 4×4 system. Packetized 40Gb/s NRZ-OOK payloads are generated with 540ns duration and 60ns guard time. An FPGA acts as aggregation controller that generates for each packet the label according to the port destination, and simultaneously provides a gate signal to enable the transmission of the payload with a certain load. Buffer manager inside FPGA stores the label information in a FIFO queue with a size of 16 packets and removes the label from the queue in response to a positive ACK. Otherwise the label and payload are retransmitted after re-signaling the input optical gates, implementing the packets retransmission.

RF tone labeling technique and bi-directional optical system are deployed to efficiently transmit the labels and ACK in a single fiber. Such labeling technique allows the parallel processing of the label bits which will greatly reduce the OPS processing time [4]. Here we use two RF tones (f1=284.2MHz, f2=647.1MHz) for coding the 2-bit label

information. Payload wavelengths are placed at λ_{P11}=λP_{21}=1544.9nm and λ_{P12}=λ_{P22}=1548.0nm. The label wavelengths, each carrying two RF tones, are centered at λ_{L1}=1545.1nm and λ_{L2}=1548.2nm. The average optical power of the payload and the label at the OPS input is 2.5dBm and -2dBm, respectively. FBG is centered at label wavelength and has a -3dB bandwidth of 6 GHz which avoid spectral distortion of the payload. Optical spectra of the packets before and after label extractor for Cluster1 are shown in Fig. 2a and Fig. 2b.

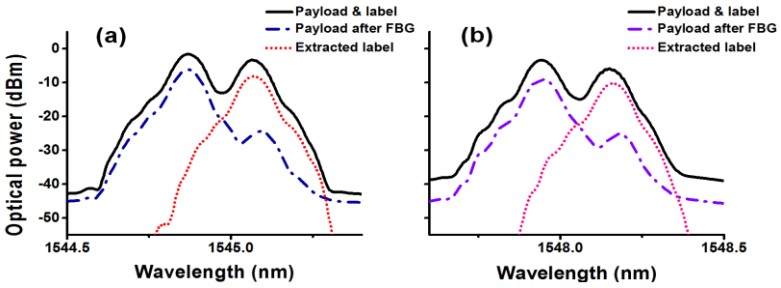

Fig. 2. Optical spectrum before and after label extractor for (a) Client1 (b) Client2.

Fig. 3 shows the dynamic operation of label and payload from both clusters. The time traces of the label detected by switch controller and the ACK feed-back detected by the aggregation controller at transmitter side are reported at the top. 2-bit label brings up 3 possibilities of switching since "00" represents no packet. "01" stands for output1, "10" for output2 and "11" for multicasting the payload to both ports. To clearly show the contention and switching mechanism, fixed priority has been adopted in the contention resolution algorithm. If two packets from different clients have the same destination, Client 1's packet will be forwarded at the output while the packet from Client 2 will be blocked and a negative ACK will be sent back requesting packet retransmission. If Client1 is multicast, any data in Client2 will be blocked. Client2's multicasting will only be approved if Client1 has no data.

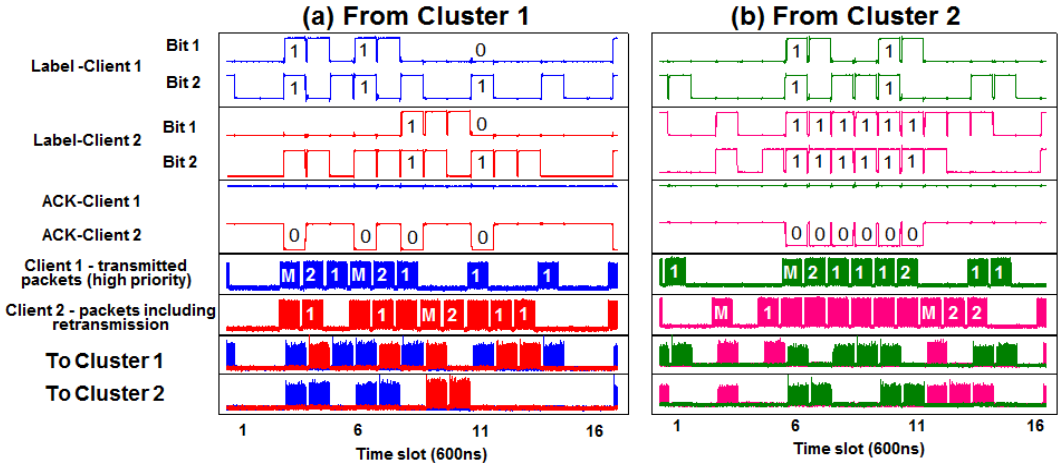

Fig. 3. Dynamic operation of labels and payloads from (a) Cluster1; (b) Cluster2

One or both of the SOA will be switched on to forward the packets to right destination. The waveforms of the transmitted packets (including retransmitted packets for Client 2) and the switch outputs are shown at bottom. "M" packet stands for multicast packet, which should be forwarded to both output ports. If Client 2 contends with Client 1 the packets will be blocked which are shown with unmarked packets in Fig. 3. In this case, a negative ACK is generated to inform buffer manager that the packets have to be

retransmitted. Fig. 3 clearly shows the successful optical flow control and multicasting operation. The minimum end-to-end latency (no retransmission) is 300ns including 250ns propagation delay introduced by 25m link.

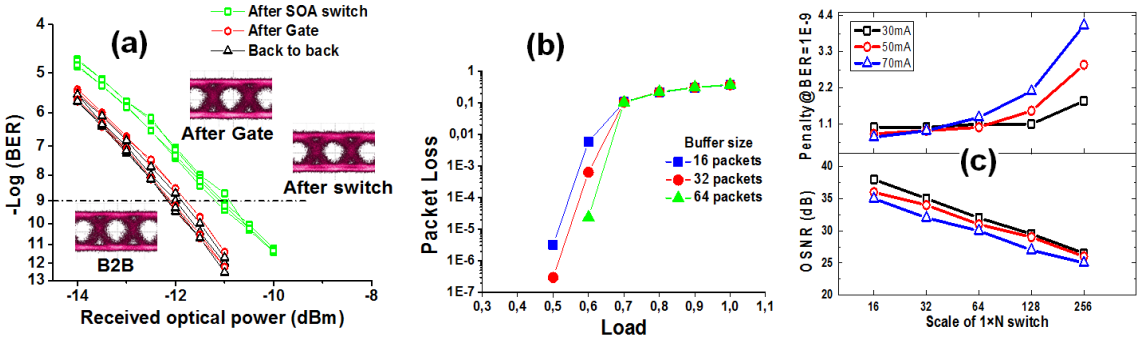

Fig. 4. (a) 40Gb/s payload BER curves and eye diagrams; (b) Packet loss vs. load with different buffer size (c) Penalty and OSNR vs. scale of 1×N switch.

Fig. 4a shows the measured BER performance for the system. Error free operation with only 1dB power penalty after the switch was measured. Less that 10^{-5} packet loss has been measured at 0.5 load and 16-packet buffer size as shown in Fig. 4b. Larger buffer size would improve the packet loss performance for the load smaller than 0.7. We further investigate the 1×N switch scalability of the proposed system by using an attenuator before the 1xN switch to emulate the coupling loss of the broadcast and select. Fig. 4c shows the penalty (measured at 1E-9) and OSNR of switch output as a function of N within different SOA bias current. It clearly shows that at N=64, less than 1.5 dB penalty is obtained which indicates that the OPS can be scaled to a large number of ports at the expense of limited extra penalty.

Conclusions

We experimentally demonstrate a 4x4 DCN architecture for 40 Gb/s packets. Exploiting the highly distributed control architecture, the efficient optical flow control, and the fast SOA switches, we report 300 ns minimum end-to-end latency (including 250 ns offset provided by the 25m transmission link) and packet loss of 10^{-5} for 0.5 input load and limited buffer size. Investigation on the switch scalability indicates that scaling up to 64×64 ports is possible at cost of 1.5dB extra power penalty.

Acknowledgements

This work has been supported by the FP7 European Project LIGHTNESS (FP7-318606).

References

[1] G. Astfalk, “Why optical data communications and why now?,” Applied Physics A, vol. 95, 933-940, 2009.

[2] L. A. Barroso and U. Hölze, “The datacenter as a computer: an introduction to the design of warehouse-scale machines,” Morgan and Claypool Publishers, Los Angeles, 2009.

[3] C. Kachris and I. Tomkos, “A survey on optical interconnects for data Centers,” IEEE Communication Surveys & Tutorials, vol. 14, 1021–1036, 2012.

[4] J. Luo et al., “Optical Packet Switch with Distributed Control Based on InP Wavelength-Space Switch Modules,” IEEE Photon. Technol. Lett., vol. 24, 2151–2154, 2012.

[5] S. Di Lucente et al., “Numerical and experimental study of a high port-density WDM optical packet switch architecture for data centers,” Optical Express, vol. 21, 263–269, 2013.