OMV4.pdf

40 Gb/s 8×8 Low-latency Optical Switch for Data Centers

Roberto Proietti[1], Xiaohui Ye[1], Yawei Yin[1], Andrew Potter[1], Runxiang Yu[1], Junya Kurumida[2], Venkatesh Akella[1], and S. J. B. Yoo[1]

1 Department of Electrical and Computer Engineering, University of California, Davis, California 95616
2 NPRC, National Institute of Advanced Industrial Science and Technology, Tsukuba Central 2, 1-1-1 Umezono, Ibaraki, 305-8568, Japan
Author e-mail address: rproietti@ucdavis.edu

Abstract: This paper reports on a 40 Gb/s 8×8 optical switch for data centers. Experiments demonstrate error-free operation with 118.2 ns switching latency in contention-less architecture. Simulations show 150 ns latency for consolidated architectures.
OCIS codes: (200.4650) Optical Interconnects; (200.6715) Switching.

1. Introduction

Due to exponentially growing Internet traffic, data-centers are growing to warehouse sizes with hundreds of thousands of servers. The performance (especially latency) and scalability of the network that connects these servers is becoming an important concern. Today, multi-hop networks based on Fat tree or Butterfly topologies using commodity electrical switches are primarily used in the data centers. These topologies are based on store-and-forward paradigm that incur huge latencies and high power consumption [1]. However by using optical interconnects, we can exploit the inherent parallelism and high-capacity of wavelength division multiplexing (WDM), helping to resolve I/O limitations and head-of-line blocking problems faced by most of today's systems. Data centers would then significantly benefit from the introduction of high-data-rate optical switches with very-large large port count, very high bisection bandwidth, low latency typically in 100's of nanoseconds or less, low power consumption, and agility in handling short packets (100's of bytes) [2].

While optical switches exploiting optical parallelism and high speed switching can be good candidates, the lack of practical optical buffer technology makes it difficult to seek all-optical solutions relying on optical buffer delay lines or deflection routing [3], which cannot guarantee arbitrary delays and avoid packet dropping. On the other hand, a traditional approach relying on the store-and-forward paradigm would cause most of the bottlenecks typical of electrical switches (high latency, large power consumption, limited number of ports and switching capacity), reducing the benefits offer by optical technology [4]. A hybrid approach where the data plane shares both optical and electrical technology appear to be a viable solution.

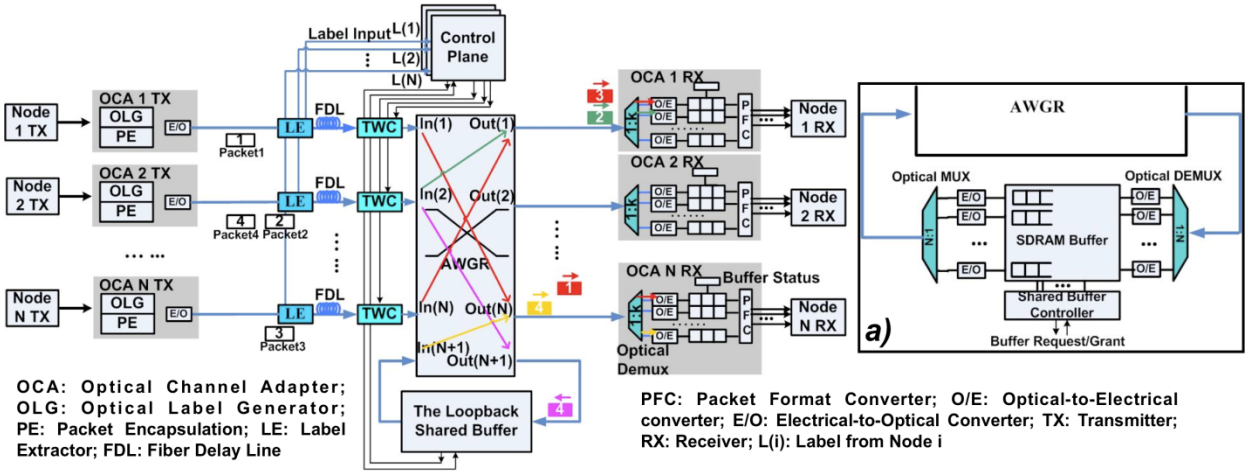

Figure 1. Optical switch architecture. Inset a): Loop-back shared buffer architecture.

Fig. 1 shows the hybrid architecture previously reported in [5], which relies on a single-hop all-optical switching fabric based on an arrayed-waveguide grating router (AWGR) and tunable wavelength converters, with a loop-back electronic shared buffer and a low-speed electrical control plane. *N* represents the switch port count, while *k* indicates the number of wavelengths that can be simultaneously present at a same output. The OCA Tx attaches a low-speed label to each incoming packet. By doing that, the control plane can run at a much lower frequency than the data plane. This can offer lower power operation and increased scalability. Moreover, by avoiding the store-and-forward paradigm, and by leveraging the optical parallelism, the optical switch can greatly reduce the average

switching latency. The loopback buffer (Fig. 1 inset) stores only the packets that cannot gain access to the desired output, thereby avoiding any packet drop. Ref. [5, 6] give a detailed explanation of this architecture, andreport on simulation proving the higher throughput and lower latency of this solution when compared to Ethernet switches and legacy flattened butterfly architectures.

This paper reports on the performance of a 40 Gb/s 8×8 optical interconnect switch based on the discussed architecture. Experimental results show error-free operation of the switch for various short packet sizes with a switching latency as low as 118.2 ns for an optical contention-less scenario (no loop-back buffer is implemented). Experiment-driven simulations show that avoiding the store-and-forward paradigm and leverages optical parallelism, a low average switching latency below 150 ns can be obtained also in case of contention.

2. Experimental Testbed

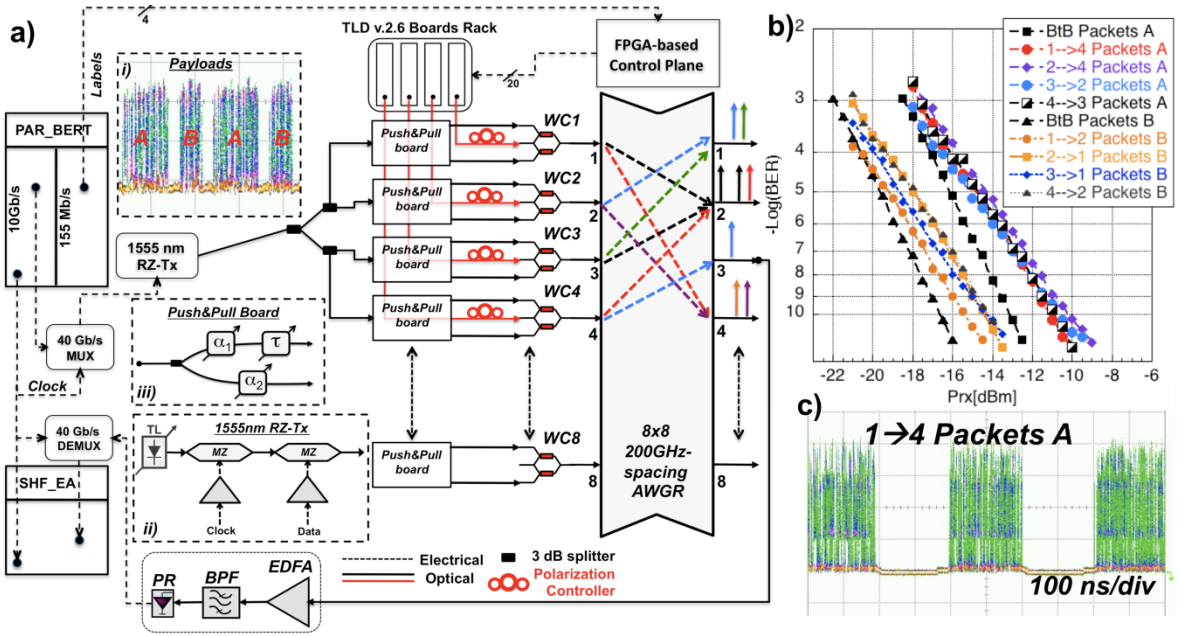

Figure 2. a) Experimental testbed. b) Dynamic switching BER measurements. c) Example of a switched packet from input 1 to output 4.

Fig. 2a shows the 40 Gb/s 8×8 Layer 1 experimental testbed for the optical contention-less scenario. A 8×8 200 GHz-spacing AWGR occupies the core of the switch architecture. It also includes four wavelength converters (WCs) based on cross-phase modulation (XPM) in a semiconductor optical amplifier Mach-Zehnder interferometer (SOA-MZI). Each WC accepts a continuous wave (CW) signal from a tunable laser diode (TLD) board, and a 40 Gb/s payload stream signal. Two payloads, A and B, of different lengths (1 kB and 512 B respectively – see inset *i* of Fig.2a) are continuously generated by using a 10 Gb/s pulse pattern generator (PPG) and a 40 Gb/s electrical multiplexer (MUX), whose output modulates a return-to-zero (RZ) optical transmitter (RZ-Tx) (see inset *ii* of Fig.2a). The output of the RZ-Tx is then split in four copies to emulate the traffic coming from the different nodes. A dedicated PPG also generates 155 Mb/s 12 bit-long labels for each of the four 40 Gb/s payload streams. The 40 Gb/s MUX output and the 155 Mb/s output represent the signals produced by the payload and label generator in the Tx-side of the OCAs shown in the architecture of Fig. 1. In this experiment, the 155 Mb/s label signals remain in the electrical domain, but in a real system they will be transmitted in the optical domain on a separate wavelength or subcarrier. An FPGA-based control plane reads the destination address carried by each packet's label and sends out a 5-bit parallel control signal to the TLD boards, which sets the TLD wavelength accordingly. This establishes an optical path between each AWGR input and output on a per-packet basis. Each payload packet coming from the RZ-Tx is then converted to the TLD output wavelength by means of the SOA-MZI WC operating in push-pull mode (inset *iii* of Fig.2a). A tunable filter with 0.6-nm bandwidth (−3 dB) and sharp roll-off is used to select the packets simultaneously present on different wavelengths and measure their signal quality in terms of bit error rate (BER). Figure 2b reports BER curves proving error-free operation of the switch for the scenario in which packets A and B coming from AWGR inputs (nodes) 1,2,3,4 are switched respectively to AWGR output ports (nodes) 4,4,2,3, and 2,1,1,2. An average power penalty of 2 dB is mainly caused by the limited bandwidth of the AWGR channels.

The measured switching latency associated with the above measurements is equal to 118.2 ns. This latency is

composed of three different components according to the following equation:

$$\tau_{CF} = \left(\frac{1}{FPGA\,clock} \times n \right) + TLD_{Switching\,Time} + AWGR_{TransitTime} \quad (1)$$

where n is the is the number of FPGA clock cycles necessary to process the label information. In our experiment, we have 5 ns for the AWGR transit time, 30 ns for the TLD switching time, and 83.2 ns given by the FPGA control plane, which runs at a clock speed of 155 MHz (clock cycle = 6.4 ns). In this case, n is equal to 13, since it takes 12 clock cycles to serially read the 8 bit preamble and the 4 bit address, and then one clock cycle to send out the control signals to the TLD boards. In a real application, the label length, label bit-rate and control plane clock speed would all be greater. Assuming a 5-Byte label, a label-rate of 1.25 Gb/s and a control plane based on a high-speed application specific IC (ASIC) running at 1.25 GHz, the control plane latency would be reduced to 32.8 ns (41 clock cycles), and the total switching latency to 67.8 ns (≈ 85 clock cycles – see red line of Fig. 3b).

Experimental values for the AWGR transit time and TLD switching time are then used in the simulator of [5] to evaluate the average switching latency in case of contention ($k = 2;4$). Fig. 3 shows the simulation results. Fig.3a shows the packet contention probability (assuming random uniformly distributed traffic) for the case k=2:4, while Fig.3b shows the average switching latency. Even a small value of k is enough to keep the switching latency at 180 cycles. This means a latency value as low as 144 ns assuming a control plane speed of 1.25 GHz and a label bit-rate equal to 1.25 Gb/s. As the time to transmit bigger packets becomes longer, delayed packets may need to spend more time in the shared buffer to wait for the requested resources to become available. This explains the switching latency dependence on the packet size. The contention-less case k=8 is also reported for reference. The latency is constant and independent from the packet size, being the contention probability null.

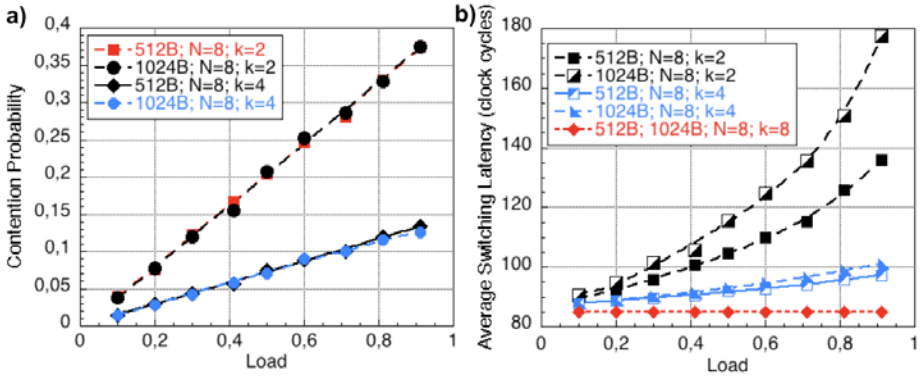

Figure 3. (a) Contention probability and (b) average switching latency as functions of the load.

4. Conclusion

This paper considered the performance of a 40 Gb/s 8×8 switch for application in data centers. Error-free operation of the switch with a low-latency of 118.2 ns in contention-less architecture has been validated with an experimental test-bed. A relatively low system power consumption below 200 pJ/bit is obtained just by using discrete components. Experiment-driven simulations show that by leveraging optical parallelism, the average switching latency can be kept at low values (<150 ns) even in the contention case. Ongoing work is focusing on the development of a complete hardware test-bed to evaluate the switch performance in a real computing environment.

5. References

[1] M. Al-Fares, A. Loukissas, and A. Vahdat, "A Scalable, Commodity Data Center Network Architecture," in *SIGCOMM'08*.

[2] T. Benson, A. Anand, A. Akella, and M. Zhang, "Understanding data center traffic characteristics," in *the 1st ACM workshop on Research on enterprise networking*, 2009, pp. 65-72.

[3] O. Liboiron-Ladouceur, A. Shacham, B. A. Small, B. G. Lee, H. Wang, C. P. Lai, A. Biberman, and K. Bergman, "The Data Vortex Optical Packet Switched Interconnection Network," *Journal of Lightwave Technology* vol. 26, July 2008.

[4] R. Hemenway, R. R. Grzybowski, C. Minkenberg, and R. Luijten, "Optical-packet-switched interconnect for supercomputer applications," *Journal of Optical Networks*, 2004.

[5] X. Ye, Y. Yin, D. Ding, S. Johnson, V. Akella, and S. J. B. Yoo, "Assessment of Optical Switching in Data Center Networks," in *OFC/NFOEC 2010*.

[6] X. Ye, P. Mejia, Y. Yin, R. Proietti, S. J. B. Yoo, and V. Akella, "DOS - A scalable Optical Switch for Datacenters," in *ANCS2010*

This work was supported in part by the Department of Defense through the project Ultra-Low Latency Low-Power All-Optical Interconnection Switch for Peta-Scale Computing under contract #H88230-08-C-0202.