Assessment of Optical Switching in Data Center Networks

Xiaohui Ye, Yawei Yin, Dan Ding, Samuel Johnson, Venkatesh Akella, and S. J. B. Yoo
Department of Electrical and Computer Engineering, University of California, Davis, California 95616, USA.
Email: {xye, yyin, dding, samjohnson, akella, sbyoo}@ucdavis.edu

Abstract: We propose a scalable optical switching architecture towards contention-free interconnection for data center networks. The simulations show it can support higher load and throughput with significantly lower latency compared to today's commercial Ethernet switches**.**

OCIS codes: (200.4650) Optical interconnects; (060.4250) Networks

1. Introduction

Low-latency, scalable, and high-throughput interconnection of computing systems is essential for future cloud computing and high-performance computing. While massively parallel computing architectures and large data storage systems on the scale of petaFLOPs and petaBytes are being deployed today, peta-scale interconnection systems are not yet available. In fact, today's data centers and computing centers are relying on interconnection of many electronic switches concatenated in multi-stage interconnection topology of fat-tree, CLOS, torus, or other architectures with relatively poor scalability [1]. Not only these electronic switch architectures result in poor throughput, latency, and scalability, but they also catastrophically degrade the power efficiency in the data center, first by creating bottlenecks and latency in data communications and second by substantially increasing the amount of energy consumed in the interconnection networks. All-optical switching, on the other hand, can benefit from inherent parallelism of optics mapped naturally to the need for massive concurrency and scalability in parallel computing. The capability to transport and switch many parallel data flows on parallel wavelengths on an optical waveguide resolves severe I/O limitations faced by essentially all data systems today. Contention in the switching fabric and head-of-line blocking in queues commonly seen in electronic interconnection switches can also be potentially resolved by introducing optical parallelisms offered by parallel wavelengths where data packets on separate wavelengths can be queued and switched without interfering with others on other wavelengths. While the synergy between the need for parallelism in computing and the inherent parallelism offered by optics is strikingly strong, there have been few demonstrations of large-scale optical switching systems for computer interconnection networks [2]. Recently, an all-optical packet switching system scalable to 2 million x 2 million ports and 42 petabit/second throughput with nano-second switching time has been demonstrated for telecommunications networks [3, 4]. The switching system includes an all-optical switching fabric containing tunable wavelength converters (TWC), arrayed waveguide grating router (AWGR) [5], and fixed wavelength converters. This paper investigates a new class of optical interconnection switches consisting of tunable lasers, an AWGR, and electronic memory queue switches with a new arbitration scheme leveraging the wavelength domain parallelism.

2. Optical switch architecture

Fig. 1(a) shows the top-level architecture of the optical interconnection switching system for data centers. The key features pursued in the optical switching system are: (a) ultralow latency from greatly simplified arbitration, (b) free from head-of-line-blocking and contention-less optical switching, (c) optical concurrency, and (d) scalability. At the core of the switching system is an optical switching fabric involving tunable wavelength converters (TWC), an AWGR, and a shared SDRAM buffer with parallel wavelength demux and multiplexer (MUX). The network interface card (NIC) includes parallel wavelength demultiplexers (DEMUX) interface at the input and electronic aggregator at the output. In addition, the switching system includes the control block which reads the header (or label) fields of the data packet, arbitrates packets by looking at the buffer queue status in the NICs, and achieves flow control. The optical switch fabric supports a single-hop interconnection network topology shown in Fig. 1 (c) with extremely high scalability. The $N \times N$ AWGR supports non-blocking interconnection of $N \times N$ nodes simply by tuning the wavelength on the TWCs by virtue of the well-known wavelength routing characteristic [5]. Another notable feature of the AWGR based switch is that each output port interconnects with multiple input ports on separate and distinct wavelengths. Therefore placing a DEMUX in the NIC can separate different flows from each wavelength, thus achieving contention-free optical switching. The simulation studies involve interconnection of $N = 24$ servers. The labels containing information pertaining to the destination address, packet length, etc will be optically extracted and received at the control block, where arbitration takes place. While different packets can contend towards an identical destination server from different source servers at the same time, the packets will be on different wavelengths and will be split to separate queuing buffers in the NIC after being demultiplexed by 1:k

optical demux and converted to electrical signals (O/E). When k=N (=24), each buffer exclusively serves a single wavelength, requiring that every NIC to contain k individual buffers. The buffers' status – which records the number of available buffer units - is transmitted to the control block, which using a credit-based flow control mechanism [6]. The packets that fail to gain the output and buffer space (due to insufficient electronic buffer space or when k<N is employed) will be sent to the 25th output ports where it is then forwarded into the shared SDRAM buffers for temporary storage. These stalled packets will be sent back to input ports and contends for the resources later.

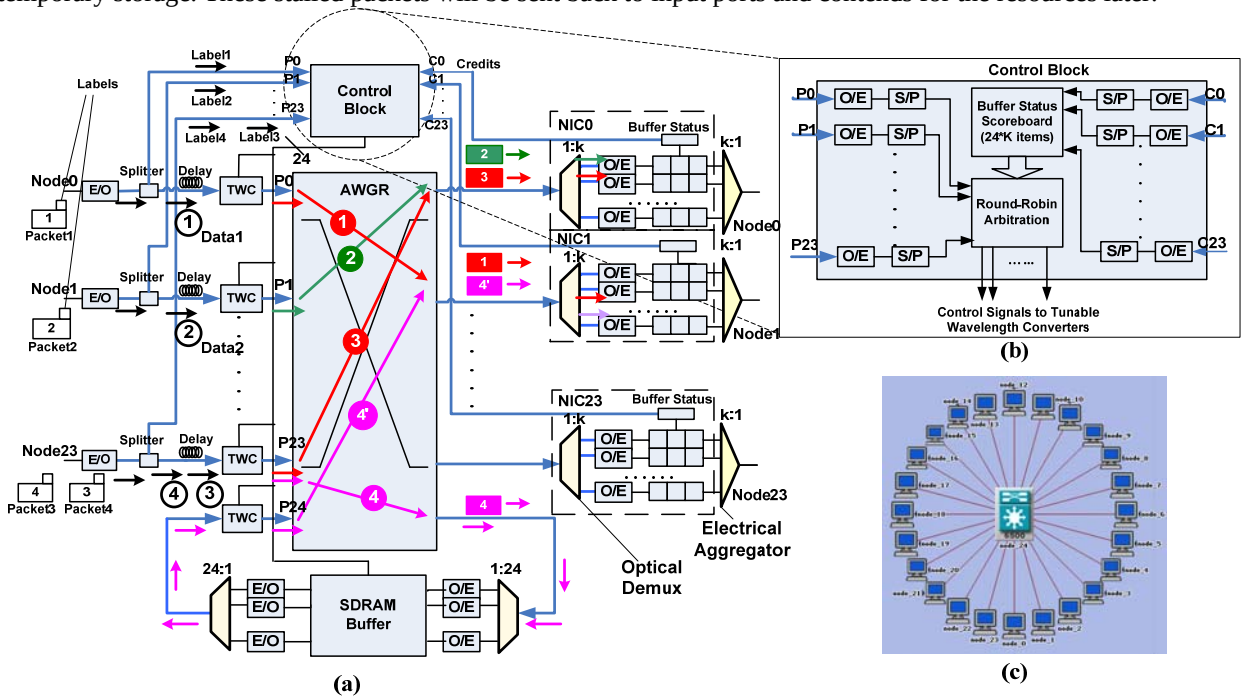

Fig. 1 (a). Optical switching architecture (blue-waveguides, black-electrical lines), (b) optical switch control block, and (c) optical interconnection network topology

Fig. 1(b) and (c) show the architecture of Control Block and the switching topology used in our simulation. The control block consists of 48 input ports, a buffer status scoreboard and arbitration logic. The first twenty four of the control block input ports receive labels from the 24 switch input ports and pass this data to the arbitration logic. The second set of twenty-four control block input ports receive the buffer status reports from the receiving NICs and present them to the scoreboard where credits are assigned. The arbitration logic implements a virtual channel-like flow control mechanism. As discussed above, each output ports can receive optical packets from at most k wavelengths simultaneously. Contention-free switching will be achieved if *k* equals to 24. If *k* is smaller than 24, the wavelengths will be divided evenly into *k* groups. In the event that a packet faces contention, all packets from different groups will be allowed to move through the switch without any blocking. But packets from the same group will be arbitrated in a round-robin fashion. The blocked packets will be switched to 25th output port towards SDRAM buffers, where they will be stored temporarily. Later, those packets that failed to gain switching will be given priority over those that had previously won the switching, thereby realizing the round-robin arbitration with fairness. After allotting the switch resources, the arbitrator checks the scoreboard to determine whether each allotted flow has available buffer space at the receiving end. Since the wavelength-specific buffers in the NIC are finite, we adopt the credit-based flow control to avoid buffer overflow. A buffer status scoreboard is kept in the control block, recording the amount of available space in each buffer. The bookkeeping of the scoreboard is done as follows: once packets are delivered from the buffers to the memory of the destination host, the buffer status in each NIC will calculate the amount of buffer space cleared and send an update to the scoreboard inside the control block.

3. Results and discussion

The following simulation results compare the performance of a commercial Ethernet switch against the proposed AWG-based optical switch. Both networks are constructed in a star topology under OPNET [7], with 24 peripheral nodes – each representing a single rack-switch -- and one centralized node that represents either the Ethernet switch or the optical switch. The Ethernet scenario uses a 1000BaseX link and the Cisco WS-C6509. The optical switch scenario uses 10Gbps optical fiber to connect the peripheral nodes with the proposed optical switch. Traffic is injected into different peripheral nodes with destination randomly selected based on a uniform distribution. Two

series of simulation are conducted using the constant packet size of 64 bytes and 1500 bytes respectively. Here we show results of the optical switch by setting k to 4.

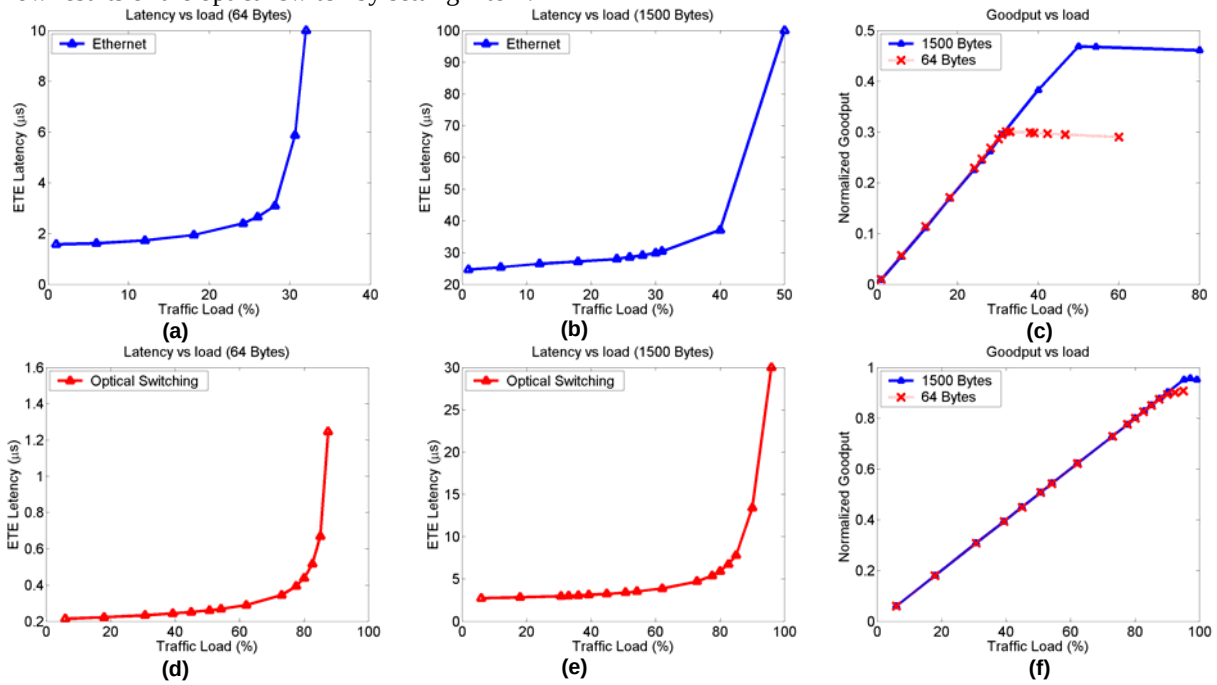

Fig. 2. ETE Latency of Ethernet switch with 64 Bytes packet traffic, (b) ETE Latency of Ethernet switch with 1500 Bytes packet traffic, (c) Normalized goodput for the Ethernet switch, (d) ETE Latency of optical switch with 64 Bytes packet traffic, (e) ETE Latency of optical switch with 1500 Bytes packet traffic, and (f) Normalized goodput for the optical switch.

Fig. 2 (a), (b), (d), and (e) present End-to-End (ETE) latencies of both Ethernet based and the proposed optical switch based data center network. As shown in the Fig. 2 (a), ETE latencies of the Ethernet switching network are under 5 μs before the traffic is saturated at the load of 32%. Comparatively as shown in Fig. 2 (d), when using the optical switch, the latencies are only one-tenth of that in the Ethernet. Moreover, the traffic load can go beyond 85% before congested at the switch. For the packet size of 1500 bytes, both switching networks saturate at a higher load, with 54% for Ethernet switch and 95% for optical switch respectively. However, ETE latencies for both scenarios increase because large packets have larger transmission latency. Fig. 2 (c) and (f) show the network goodput for both switching networks with different packet sizes. (The network goodput is identical to the network throughput in this network.) Again, it is clear that the Ethernet switch suffers from saturation at a lower traffic load level while the proposed optical switch can accommodate heavy traffic load. In addition to the performance advantages shown above, scalability is another merit of the proposed optical switching architecture. The AWG based optical switch can be easily extended to thousands of ports without sacrificing latency or throughput.

4. Conclusion

A novel AWGR-based optical switching architecture is proposed for data center interconnecting network. Multiple packets can travel on different wavelengths to the same output port simultaneously, thereby reducing the contention probability at the core switch. A credit-based flow control mechanism is adopted to solve the problem of buffer overflow at the receiver side. Overall, the simulation results show that the proposed switch architecture is advantageous in latency and throughput over Ethernet.

References:

[1] Al-Fares, M., A. Loukissas, and A. Vahdat. A Scalable, Commodity Data Center Network Architecture. in SIGCOMM'08. August 2008.
[2] Luijten, R., et al. Optical interconnection networks: The OSMOSIS project. in 2004 IEEE LEOS Annual Meeting Conference Proceedings. Rio Grande, Puerto Rico. 7-11 Nov. 2004. 2004.
[3] Yoo, S.J.B., et al., Rapidly switching all-optical packet routing system with optical-label swapping incorporating tunable wavelength conversion and a uniform-loss cyclic frequency AWGR. IEEE Photonics Technology Letters, 2002. 14(8): p. 1211-13.
[4] Yoo, S.J.B., Optical packet and burst switching technologies for the future photonic Internet. JLT, 2006. 24(12): p. 4468-92.
[5] Kato, K., et al., 32 x 32 full-mesh (1024 path) wavelength-routing WDM network based on uniformloss cyclic-frequency arrayed-waveguide grating. Electronics Letters, 2000. 36(15): p. 1294-1296.
[6] Dally, W. and B. Towles, Principles and Practices of Interconnection Networks. 2003: Morgan Kaufmann Publishers.
[7] "http://www.opnet.com/"