Scalable, High-Throughput, Low-Latency AWGR-based Optical Switches with Distributed Control Plane for Future Computing Systems

Roberto Proietti, Christopher J. Nitta, Yawei Yin, Xiaohui Ye, Runxiang Yu, Zheng Cao, Venkatesh Akella, and S.J.B. Yoo

Department of Electrical and Computer Engineering, University of California, Davis, CA, 95616, USA
Author e-mail address: rproietti@ucdavis.edu

Abstract: We summarize our research on LION switches exploiting wavelength routing in arrayed waveguide grating routers (AWGRs). We discuss the loopback-buffer, all-optical-NACK and all-optical-TOKEN architectures, presenting their effectiveness in terms of network performance and experimental studies.

OCIS codes: (200.4650) Optical Interconnects; (200.6715) Switching.

1. Introduction

The critical performance bottleneck of large computing systems has shifted from the processors to the communications infrastructure [1]. While optics is commonly used to carry high-bandwidth information between top of the rack (ToR) switches, the use of optics for routing and switching is still at the research level and electronic-crossbar switches with store-and-forward architectures still represent the only commercially available solution. Large-scale systems cascade many such switches to interconnect a large number of servers, leading to limited scalability, capacity, throughput, and power-efficiency [2].

As **Figure 1**(a) shows, an arrayed waveguide grating router (AWGR) [3] inherently supports all-to-all communication among N nodes in a flat topology with N wavelengths, but the large number of wavelengths (N) and transceivers (TRXs) (N^2) is a challenge when N is high. In order to reduce the number of TRXs, it becomes then necessary to add some active elements and a contention resolution scheme to the passive AWGR-based cross-connect. This paper reviews the research carried at UC Davis on low-latency interconnect optical network switch (LIONS) based on AWGRs. We discuss the loopback-buffer [4], all-optical-NACK [5] and all-optical-TOKEN architectures [6], presenting their effectiveness in terms of network performance and showing a testbed study for the loopback buffer architecture.

2. LIONS: architectures and network performance.

Figure 1(a) shows how a single passive N-port AWGR can implement an all-to-all network, assuming that each of the N nodes have N TX-RX pairs at N different wavelengths. Clearly the N^2 TRXs will not scale. We can define an architecture with k_{TX} TXs and k_{RX} RXs (k_{TX}, $k_{\mathrm{RX}} < N$) by introducing an arbitration or scheduling mechanism to avoid packet contention. Figure 1(b) shows a silicon

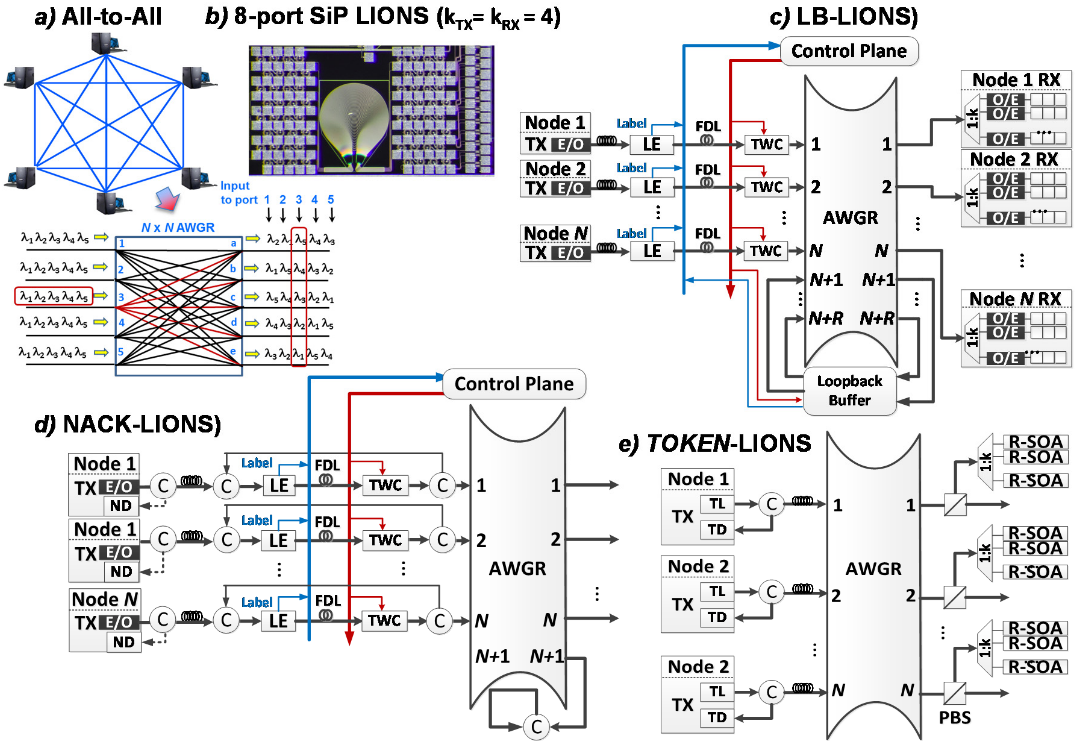
a) All-to-All
b) 8-port SiP LIONS (k_{TX}= k_{RX} = 4)
c) LB-LIONS)
Input to port 1 2 3 4 5
N x N AWGR
$\lambda_1\lambda_2\lambda_3\lambda_4\lambda_5$
1 2 3 4 5
a b c d e
$\lambda_2\lambda_1\lambda_5\lambda_4\lambda_3$
$\lambda_1\lambda_5\lambda_4\lambda_3\lambda_2$
$\lambda_5\lambda_4\lambda_3\lambda_2\lambda_1$
$\lambda_4\lambda_3\lambda_2\lambda_1\lambda_5$
$\lambda_3\lambda_2\lambda_1\lambda_5\lambda_4$
Control Plane
Node 1
Node 2
Node N
TX E/O
Label
LE
FDL
TWC
AWGR
1 2 N N+1 N+R
Node 1 RX
Node 2 RX
Node N RX
1:k
O/E
Loopback Buffer
d) NACK-LIONS)
Control Plane
Node 1
Node N
TX E/O
ND
C
Label
LE
FDL
TWC
AWGR
1 2 N N+1
e) TOKEN-LIONS
Node 1
Node 2
TX
TL
TD
C
AWGR
1 2 N
1:k
R-SOA
PBS

Figure 1. (a) All-to-all interconnection using AWGR. (b) 8-port Silicon Photonics LIONS. (c) Loopback Buffer LIONS. (d) All-Optical NACK LIONS. (e) All-Optical TOKEN LIONS.

photonics 8-port LIONS with $k_{TX} = k_{RX} = 4$, which could be suitable for on-chip interconnection [7]. For off-chip inter-server, inter-rack and inter-cluster interconnection, we focused on the architectures with $k_{TX}=1$ and $k_{RX} = \boldsymbol{k}$, shown **Figure 1**(c-d-e). **Figure 1**(c) shows the loopback buffer (LB) low-latency interconnect optical network (LB-LION) switch consisting of an *N* port AWGR, *N* Tunable Wavelength Converters (TWCs), an FPGA-based electrical control plane, electrical loopback buffers, label extractors (LE), and Fiber Delay Lines (FDLs). Between the switch and each end-node, an Optical Channel Adapter (OCA) serves as the media interface. The LION switch uses a forward-and-store strategy for packets, as opposed to the store-and-forward strategy employed in an electrical switch. Only the contending packets that fail to get grants are stored in the LB (see [4] for details on the different buffer architectures). Note that, the use of ***k*** RXs per node, together with the unique wavelength routing property in AWGR, and wavelength division multiplexing (WDM), naturally implement output queuing, which is very challenging in electronics at high bit-rates. Note that, the LB memory read/write speed can be a bottleneck. In addition, the LB requires a large amount of O/E/O conversion. In [5] we proposed the all-optical negative acknowledgement (AO-NACK) technique allowing to replace the whole LB with a simple optical circulator acting as reflector, improving also the line-rate scalability of the LION switch. Exploiting the duplex nature of the AWGR, the "dropped" packets are simultaneously directed to the circulator port and reflected back to the TX nodes, where a simple edge detector senses that a packet has been sent back (NACK). The node receiving the NACK retransmits the packet based on certain retransmission policies. The simulation results in **Figure 2**(a), shows that, given the short host-switch distances typical of datacom systems, the NACK-LIONS can achieve nearly the same performance in terms of average throughput and latency as the LIONS with distributed LB, which has the best throughput, far exceeding the throughput of electronic switches, including the flattened butterfly architecture.

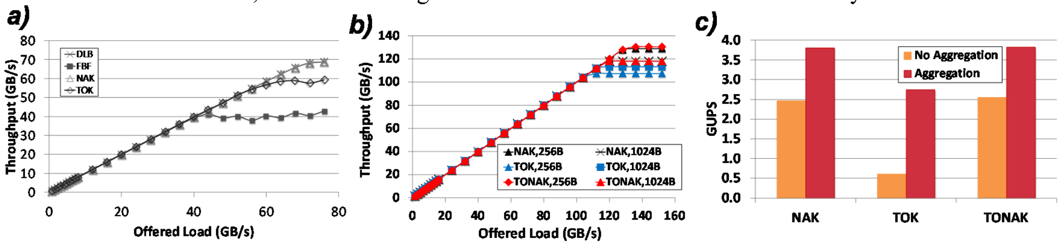

Figure 2. (a) Throughput performance comparison among LB-LIONS (DLB), NACK-LIONS (NAK), TOKEN-LIONS (TOK) and electrical Flattened Butterfly (FB) for 64 nodes, 10Gb/s line-rate, and 256B-long packets with uniform random distribution. (b) Throughput performance of combined TOKEN and NACK LIONS (TONAK-LIONS) and comparison with TOKEN-LIONS and NACK-LIONS for 128 nodes. (c) Giga Updates per Seconds (GUPS) benchmarking results.

To further simplify the LIONS, we designed the all-optical token (AO-TOKEN) LIONS (see **Figure 1**(c)), which exploits the saturation effect in the Reflective Semiconductor Optical Amplifier (RSOA) [8]. AO-TOKEN removes the centralized control plane (now optical and distributed) and the TWCs at the switch site. The basic idea is the use of one or more RSOAs as the mutual exclusion (mutex) type of arbiter at each output port of the AWGR (up to ***k*** RSOAs per output port). The TOKEN signaling and data can share the same physical layer by using polarization diversity, as demonstrated in [9]. The transmission of a packet happens only after a node applies for and receives a grant from the specific output port arbiters. The major advantage of the TOKEN technique is that it distributes the contention resolution in the control plane without the requirement of a global coordination scheme. However, the delay caused by the wait for the token response can negatively affect the switch performance, as can be observed in **Figure 2**(a). Overall, the host-switch distance and the packet size affect the switch performance. It is possible to overcome the above limitation by combining TOKEN and NACK technique in the same architecture, named TONAK LIONS [8]. **Figure 2**(b) shows how TONAK significantly outperforms AO-TOKEN (TOK) for the reasons mentioned above. Note that the simulation results in **Figure 2**(a-b) are obtained for uniform random traffic with Bernoulli distribution, 10 Gb/s line rate, 10 m fiber from the hosts to the switch, and 8 ns laser tuning time. A buffer of 40 packets was used in the TXs, RXs, and LB.

Table 1. Comparison among the different LIONS architectures.

	All-to-All	LB-LIONS	NACK-LIONS	TOKEN-LIONS	TONAK-LIONS
Pros	100% throughput	High throughput and low latency	High throughput and low latency	Distributed CP; no TWCs	Same as NACK and TOKEN
Cons	Poor scalability	Power hungry LB and TWCs; centralized CP;	Power hungry TWCs; centralized CP	Performance affected by distance and packet size	More complex than TOKEN

Figure 2(c) shows also results for GUPS benchmarking, which is of particular interest in high performance computation. Traffic in GUPS is typical of in-memory database applications that implement transactional query processing. Each "update" requires a node to read a random memory location, modify the value and then write back to the same memory location. The GUPS benchmarking simulated a 64-bit address space distributed across 128 nodes. Each update was applied to 64-bit data values and each node

was allowed up to 1024 outstanding requests. The results are shown for the case in which requests and replies are aggregated into larger packets and for requests/replies being sent independently.

4. Experimental Testbed Studies

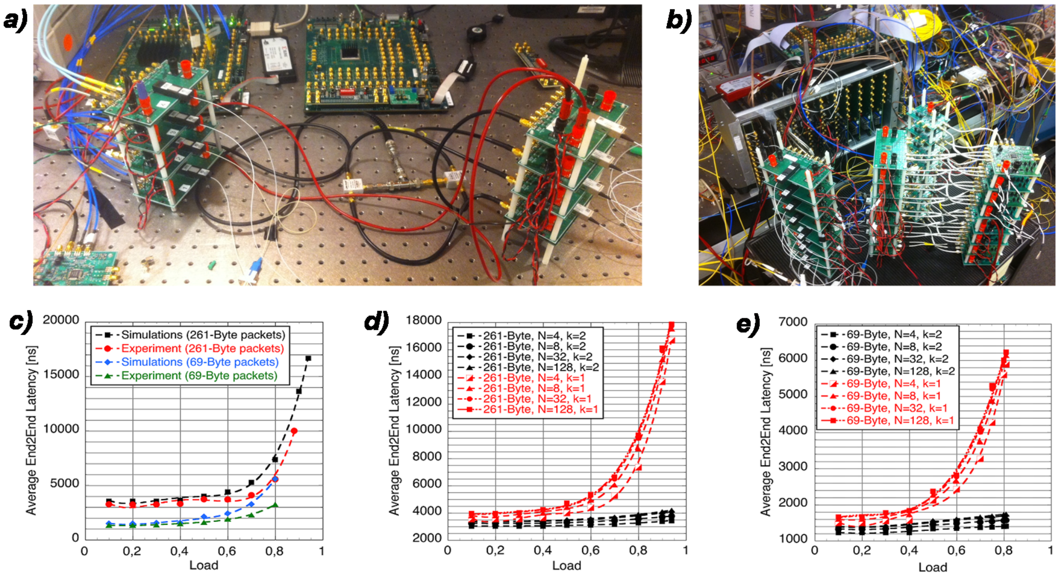

Figure 3. 4-node LB-LIONS Testbed studies. (a) FPGA-emulated nodes with optical channel adapters (OCAs). (b) FPGA-based control plane and loopback buffer with O/E/O conversion and burst-mode RXs. (c) Experimental latency statistics and comparison with simulation-based statistics. (d-e) Projection of switch performance for higher port count.

Figure 3(a-b) shows some pictures of the four-node LB-LIONS testbed we implemented to verify the accuracy of our simulator by comparing network statistics collected by experiments and simulations. The testbed was running at 1.25 Gb/s only limited by the commercial availability of burst-mode RXs (higher line rate will only change the packet transmission time and the arbitration process stays the same no matter what line rate is used in the simulation). The control plane and loopback buffers are implemented using a Xilinx Virtex 5 FPGA ML523 with RocketIO GTP transceivers. Another Virtex 5 FPGA emulate the four nodes by instantiating four MicroBlaze Soft Processor Cores generating synthetic traffic. **Figure 3**(c) shows the comparison of latency statistic results from the experiments and simulations with k=1 and packet size of 256 and 64 Bytes (plus 5 byte header). As shown, the comparison of the results shows a close match between the experimental data and the simulation data, which verifies the correctness and accuracy of the simulator we developed. The other curves in **Figure 3**(d-e) show the projection of the results to high port count, and also to k=2 case. A higher port count does not significantly affect the end to end latency, while k=2 can strongly reduce it since it reduces the contention at each output port.

4. Conclusion

This paper presented an overview of the research activity on optical switches carried out at UC Davis. We described the main features of the proposed AWGR-based architectures, discussed and showed their network performance via simulations, and report a testbed study for the LB-LIONS architecture.

5. References

[1] M. Al-Fares, A. Loukissas, and A. Vahdat, "A scalable, commodity data center network architecture," presented at the Proceedings of the ACM SIGCOMM 2008 conference on Data communication, Seattle, WA, USA, 2008.

[2] L. A. Barroso and U. Hölzle, "The datacenter as a computer: An introduction to the design of warehouse-scale machines," *Synthesis Lectures on Computer Architecture,* vol. 4, pp. 1-108, 2009.

[3] B. Glance, I. P. Kaminow, and R. W. Wilson, "Applications of the integrated waveguide grating router," *Lightwave Technology, Journal of,* vol. 12, pp. 957-962, 1994.

[4] X. Ye, R. Proietti, Y. Yin, S. Yoo, and V. Akella, "Buffering and Flow Control in Optical Switches for High Performance Computing," *Optical Communications and Networking, IEEE/OSA Journal of,* vol. 3, pp. A59-A72, 2011.

[5] R. Proietti, Y. W. Yin, R. X. Yu, X. H. Ye, C. Nitta, V. Akella, and S. J. Ben Yoo, "All-Optical Physical Layer NACK in AWGR-Based Optical Interconnects," *IEEE Photonics Technology Letters,* vol. 24, pp. 410-412, Mar 2012.

[6] R. Proietti, Y. Yin, R. Yu, C. J. Nitta, V. Akella, C. Mineo, and S. J. B. Yoo, "Scalable Optical Interconnect Architecture Using AWGR-Based TONAK LION Switch With Limited Number of Wavelengths," *Journal of Lightwave Technology,* vol. 31, pp. 4087-4097, 2013/12/15 2013.

[7] R. Yu, S. Cheung, Y. Li, K. Okamoto, R. Proietti, Y. Yin, and S. J. B. Yoo, "A scalable silicon photonic chip-scale optical switch for high performance computing systems," *Optics Express,* vol. 21, pp. 32655-32667, 2013/12/30 2013.

[8] R. Proietti, Y. Yawei, Y. Runxiang, C. J. Nitta, V. Akella, C. Mineo, and S. J. B. Yoo, "Scalable Optical Interconnect Architecture Using AWGR-Based TONAK LION Switch With Limited Number of Wavelengths," *Lightwave Technology, Journal of,* vol. 31, pp. 4087-4097, 2013.

[9] R. Proietti, Y. Yin, R. Yu, C. Nitta, V. Akella, and S. J. B. Yoo, "An All-Optical Token Technique Enabling a Fully-Distributed Control Plane in AWGR-Based Optical Interconnects," *Lightwave Technology, Journal of,* vol. 31, pp. 414-422, 2013.