Performance of AWGR-based Optical Interconnects with Contention Resolution based on All-Optical NACKs

Roberto Proietti, Christopher Nitta, Xiaohui Ye, Yawei Yin, Venkatesh Akella, and S. J. B. Yoo
Department of Electrical and Computer Engineering, University of California, Davis, California 95616
Author e-mail address: rproietti@ucdavis.edu

Abstract: This paper analyzes a 10Gb/s 64x64 optical interconnect switch exploiting an all-optical NACK technique. Simulations show that this switch architecture supports high throughput and low-latency even at 0.85 load in contrast to the flattened-butterfly architecture.

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

1. Introduction

Optical interconnects have emerged as a promising method for realizing scalable energy-efficient low-latency high-throughput networks used in high-performance computing (HPC) systems. Several research projects such as OSMOSIS [1], Data Vortex [2], and DOS [3] have already proposed architectures for optical interconnects. In particular, arrayed waveguide grating router (AWGR) based all-optical switches are attractive because they scale linearly, are non-blocking, and exploit optical parallelism to realize non-blocking and fully-connected interconnection [3]. However, the lack of practical methods for buffering light makes it difficult to implement all-optical solutions, especially in applications in which packet-loss cannot be tolerated. The fiber delay loops used in typical optical label switching (OLS) systems [4] provide a fixed amount of delay and are less effective as a contention avoidance mechanism. On the other hand, a conventional approach relying on a store-and-forward paradigm can cause bottlenecks typically seen in electrical switches with high latency, low throughput, and high power consumption. In [5], we demonstrated that, with an AWGR-based switch, it is possible to use an on-the-fly physical layer negative acknowledgement (AO-NACK) technique to notify the sending node of a contended packet. While this physical layer technique does not replace the need for a higher layer ACK (layer 4), the near instantaneous notification via the AWGR back propagation allows the sending node to retransmit the contended packet without the need for this information to propagate up the layers.

This paper demonstrates that the physical layer notification offered by AO-NACK and the short switch-host distance (< 100m) typical of HPC networks, together with wavelength domain contention resolution, enables an architecture that can sustain low latency and high throughput even under high offered loads. AO-NACK outperforms a 64-node flattened butterfly (FB) network with 10Gb/s links under both uniform random and hostspot traffic patterns. A power consumption analysis shows that AO-NACK provides higher energy efficiency than FB under high load.

2. AO-NACK Architecture

Fig.1a shows the AO-NACK architecture. *N* hosts connect to an *(N+1)×(N+1)* AWGR by means of a fiber of length *D,* the host-switch distance. Tunable wavelength converters (TWCs) at each of the N input ports perform the switching function in the optical domain. Each input port is also equipped with two optical circulators (OCs) to separate the on-the-fly packets (the packets travelling toward the AWGR input) from the counterpropagating (traveling backwards) AO-NACKs. Each node is equipped with an optical channel adapter (OCA), which generates an optical packet and its corresponding label at different wavelengths. Label extractors (LEs) separate the low-speed label signals from the high-speed payloads and send the labels to the low-speed control plane (CP).

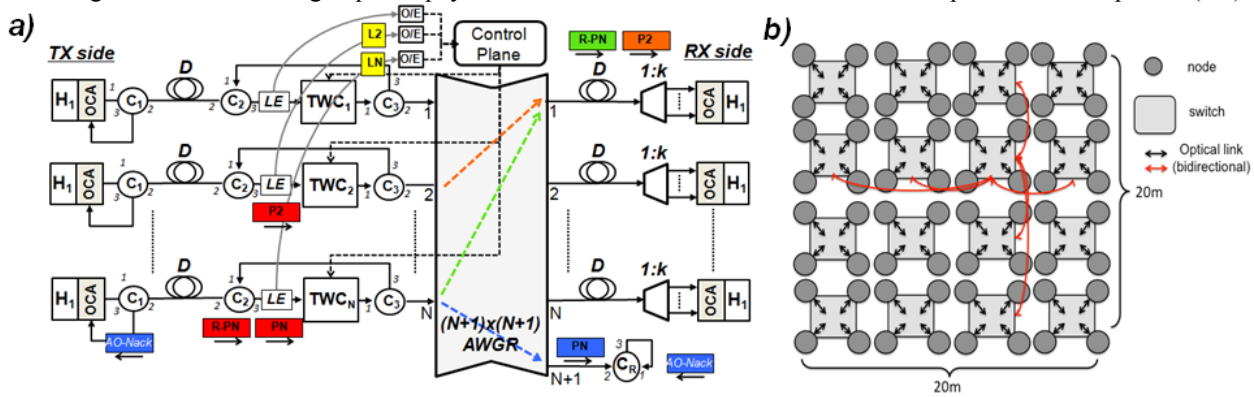

Figure 1. a) AO-NACK architecture. OCA: optical channel adapter (it includes OCA_packet_TX, OCA_label_TX, *k* BM-RXs); TWC: tunable wavelength converter; FPGA-based control plane; C_i: circulator; LE: label extractor. b) 64-node Flattened Butterfly Architecture.

The CP processes the labels and sends the control signals to the TWCs to switch the packets according to their destination. As detailed in [3], optical parallelism in AWGR can be used to reduce the contention probability-- k inputs can reach the same output port using different wavelengths. A *1:k* optical demux at each host receiver-side separates the different signals traveling simultaneously on the same fiber. Note that, when k =1 the AWGR-based switch performs similar to an electrical switch, while for $k=N$, the AWGR-based architecture is contention-free. However, a small value of k (e.g. 2, 4) represents a good trade-off between performance and cost [3].

The following is an example of how the AO-NACK mechanism works. For simplicity this example assumes that k=1. If two packets (*e.g.* P2 and PN) from different inputs are contending for the same output (output1), the CP switches one packet (*e.g.* P2) to the desired output, while the other packet (PN) is switched to the $N+1$ port (reflective port). An OC (C_R) used as shown in Fig.1a reflects the packet (PN) back to its sender (H_N). An OC at the host-site (C1) extracts the counter-propagating packet, which now acts as an AO-NACK message. A dedicated receiver is then used to detect the AO-NACK and trigger a retransmission. If $d=L/2D \geq 1$ (L is the packet length in meters), the AO-NACK reaches the sender while the transmission for the related packet is still happening or it has just finished. In this case a simple edge detector is sufficient to detect the AO-NACK since there is no ambiguity about which packet the AO-NACK refers to. If $d=L/2D < 1$, the received AO-NACK is related to a packet for which the transmission is completed. Since there may be several on-the-fly packets, an edge detector can be still used, but the sender needs to use a time-stamp for each on-the-fly packet. If the counter expires (the time counter value can be fixed since the AO-NACK arrival time is deterministic), the sender can then assume that the associated packet has reached the desired output. Otherwise, packet retransmission is triggered. Note that the passive nature of AWGR and C_R guarantees that this technique can intrinsically reflect multiple packets simultaneously without any crosstalk.

The simulated FB architecture (see Fig.1b) consists of a four by four grid of routers with four nodes connected to each router in a 20m by 20m square. Each node is connected to every other node in each row and column. The distance from each node to its "local" router is 1.76m, with each router 5m from its nearest neighbor. Table1 summarizes the main parameters used in the simulations. Note that for the AO-NACK k=4, meaning that receive bandwidth for each AO-NACK node is four times that of the FB nodes – this may appear to be an unfair comparison, however, one must keep in mind that the FB network has 2.5 times as many links as the AO-NACK.

Table 1. Parameters used in the simulations for the two different architectures.

	N	*D* [m]	*Packet Size* [B]	*d*	*k*	Bit Rate (packet)	Label Rate	OCA RX rate	Control Plane	Traffic
AO-NACK	64	10	256/64	2/0.5	4	10Gb/s	2Gb/s	4x10Gb/s	2GHz	Random/Hotspot
FB	64	n/a	256/64	n/a	1	10Gb/s	n/a	10Gb/s	2GHz	Random/Hotspot

3. Simulation results

The simulations run for the two networks involved uniform random and hotspot traffic patterns. In the hotspot traffic pattern, the hotspot node was situated at the center of the layout (potentially reducing the physical distance for the FB). Fig.2a,b shows the performance of the two architectures in terms of average throughput as a function of the offered load for different packet sizes. In particular, Fig. 2a shows the performance for uniform random traffic, while Fig.2b reports the simulation results for the hotspot port. For random traffic, the AO-NACK architecture sustained a linear throughput for traffic loads as high as 68GB/s (85% load), while the FB saturated when the offered load reaches 45GB/s (56% load). For hotspot traffic, the AO-NACK sustained a linear throughput for offered loads in excess of 4.75GB/s (5GB/s is the maximum possible load that a single node can receive), while the FB saturated below 1.25GB/s – clearly the k receivers of AO-NACK an advantage over the FB. Fig.2c,d show the average end-to-end packet latency. Fig.2c shows the results for uniform random traffic. Note that the smaller the packet size, the smaller the disparity between the AO-NACK and FB latencies, perhaps because the smaller the packet, the less important a factor the multiple hops required in FB. Fig.2d shows the results for the hotspot traffic pattern, notice that the FB latency sharply increases at less than one quarter that of the AO-NACK.

3. Power Consumption Analysis

Fig. 3a shows the power consumption as a function of the number of nodes for the two architectures. The power consumption for FB architecture is calculated from [6], while the power for AO-NACK is given by:

P_NACK= [OCA_packet_TX + OCA_label_TX + BMRX*$\boldsymbol{k}$ + CP_OE + FPGA + TWC]*$\boldsymbol{N}$;

Though AO-NACK consumes slightly more power (with values from bulky components), AO-NACK also scales linearly. Power consumption can also be considered in terms of energy per "bit of throughput". Then, since AO-NACK can sustain a much higher throughput compared to FB, for high traffic loads, AO-NACK can still reach slightly better energy efficiency in case of random traffic (see Fig. 3b), while in case of hotspot traffic, the

energy efficiency of AO-NACK is much higher (see Fig. 3c).

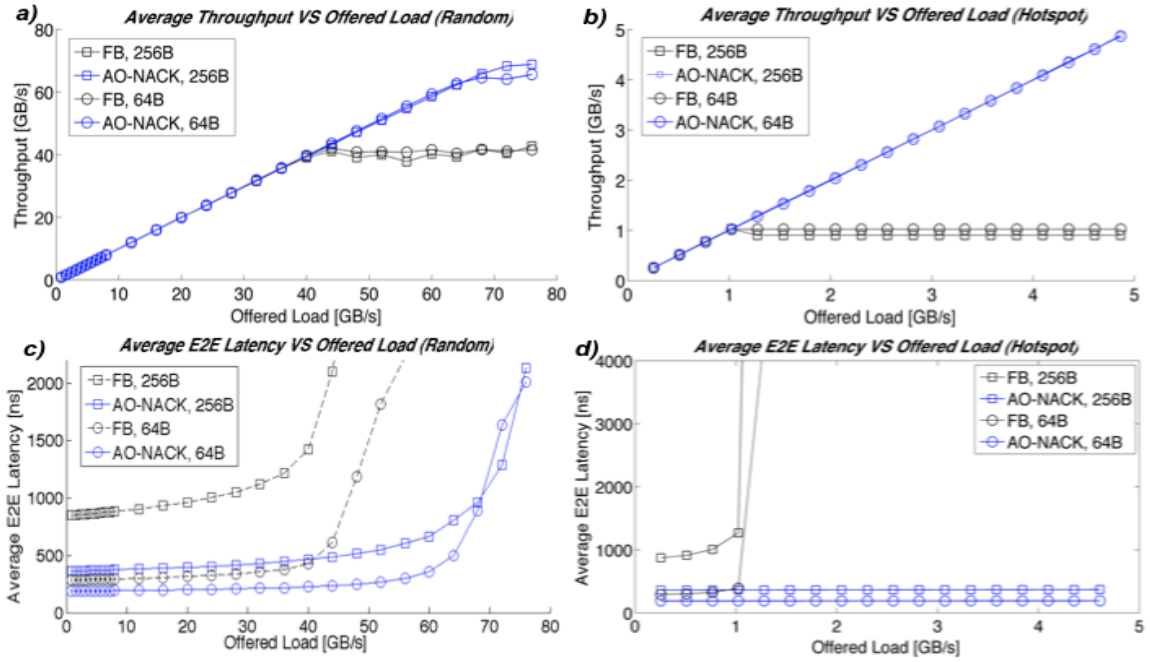

Figure 2. Average throughput and latency vs offered load for different packet sizes with random (a,c) and hotspot (b,d) traffic.

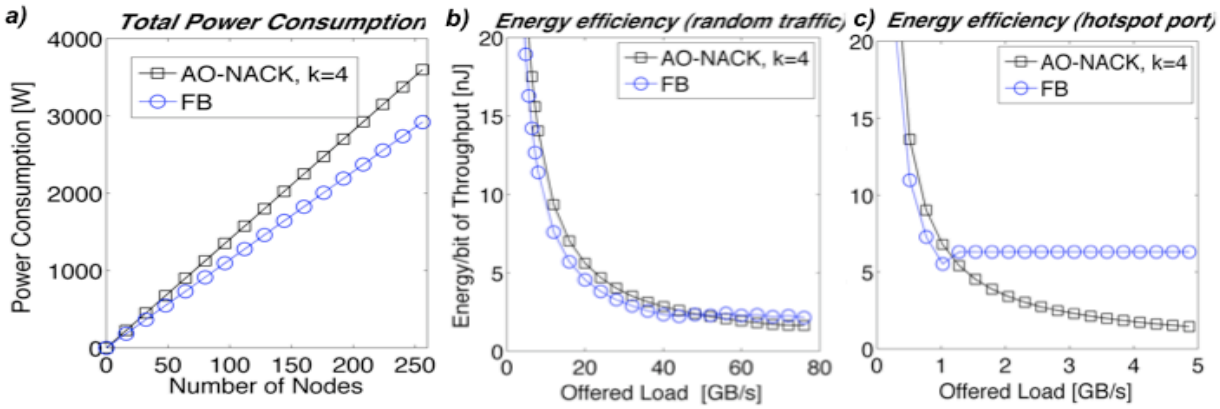

Figure 3. a) Power consumption. b,c) Energy efficiency of AO-NACK and FB for random (b) and hostspot traffic (c) for 256B packets

4. Conclusion

We proposed an AWGR-based interconnect exploiting a novel all-optical NACK technique. Simulations show that AO-NACK technique, together with the unique wavelength domain contention resolution offered by AWGR, have the potential to provide superior performance in terms of throughput, latency and energy efficiency compared to FB architecture.

5. References

[1] R. Hemenway, R.R. Grzybowski, C. Minkenberg, R. Luijten, Optical-packet-switched interconnect for supercomputer applications, Journal of Optical Networks, (2004).

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

[3] X. Ye, P. Mejia, Y. Yin, R. Proietti, S.J.B. Yoo, V. Akella, DOS - A scalable Optical Switch for Datacenters, in: ACM/IEEE Symposium on Architectures for Networking and Communications Systems (ANCS), 2010.

[4] S.J.B. Yoo, Optical packet and burst switching technologies for the future photonic Internet, Journal of Lightwave Technology, 24 (2006)

[5] R. Proietti, R. Yu, S. Yin, Y. Yin, X. Ye, V. Akella, and S. J. B. Yoo, "All-optical NACK for Fast Packet Retransmission in AWGR-based Optical Switches " in European Conference on Optical Communications (ECOC), Paper We.10.P1.50, September, 2011.

[6] Nathan Farrington, Erik Rubow, and Amin Vahdat, “Data Center Switch Architecture in the Age of Merchant Silicon,” 9 17th IEEE Symposium on High Performance Interconnects

This work was supported in part by the Department of Defense (contract #H88230-08-C-0202) and Google Research Award.