All-Optical Physical Layer NACK in AWGR-Based Optical Interconnects

Roberto Proietti, *Member, IEEE*, Yawei Yin, *Member, IEEE*, Runxiang Yu, *Student Member, IEEE*, Xiaohui Ye, Christopher Nitta, Venkatesh Akella, *Member, IEEE*, and S. J. Ben Yoo, *Fellow, IEEE*

***Abstract*— This letter, proposes and experimentally demonstrates an all-optical physical layer negative acknowledgment (AO-NACK) technique to handle contention in array waveguide grating router (AWGR)-based optical interconnects. By using back-propagation in AWGR, the packets experiencing contention are reflected back to the senders in the optical domain to serve as a physical layer negative acknowledgement to trigger the retransmission. A host-switch distance of ≈20 m and a packet length of 204.8 ns are used in this proof-of-principle demonstration. Notification of AO-NACKs messages and successful packet retransmission and switching is demonstrated with error-free operation at 10 and 40 Gb/s.**

***Index Terms*— All-optical, array waveguide gratings, data centers, negative acknowledgement, optical interconnects.**

I. Introduction

OPTICAL interconnects have emerged as a promising method to realize scalable, low-latency, and high-throughput networks in datacenters and high-performance computing. Several research projects such as OSMOSIS [1], Data Vortex [2], and DOS [3] have proposed architectures for optical interconnects in data center applications. In particular, arrayed waveguide grating router (AWGR) based all-optical switches are attractive because they are non-blocking, scale linearly, and exploit the optical parallelism to reduce contention [3]. However, the lack of optical buffering makes seeking an all-optical solution difficult, especially in datacenter applications where packet-loss must be avoided. The fiber delay loops used in optical label switching (OLS) systems [4] cannot provide arbitrary delay, thus failing to prevent packets loss. The DOS architecture of [3] uses an electrical loopback buffer and flow control scheme to store the contending packets and prevent packet drop. Although DOS can support low latency switching under high input loads, its loopback buffer requires complex and power-hungry components ($e.g. N$ tunable lasers, N high-speed TX/RX pairs running at 10Gb/s or higher, $2N$ high-speed serializer/deserializer, and high-speed memories, with N being the switch radix). The architecture proposed in [5] eliminates the buffers in the switch through physical layer acknowledgments to notify the senders when the packets reach the desired output. However, its acknowledgement scheme, which uses a semiconductor optical amplifier (SOA) at each switch output, is not designed for AWGR-based interconnects. This Letter proposes an all-optical technique in an AWGR-based switch capable of promptly notifying the end-nodes whenever one of their packets cannot reach the desired output due to contention. This technique eliminates the need for the complex loopback buffer of [3]. The instantaneous reflection, together with the short switch-host distance typical of datacenter networks (tens of meters) and the reduced contention probability offered by wavelength domain contention resolution [3], can guarantee a low retransmission-latency penalty. This technique does not replace the higher layer acknowledgements (ACK/NACK - layer 4) but handles packet contention and retransmission at the physical layer. Since these all-optical notification messages are generated only for packets experiencing contention, the technique has been designated "all-optical physical layer negative acknowledgement (AO-NACK)".

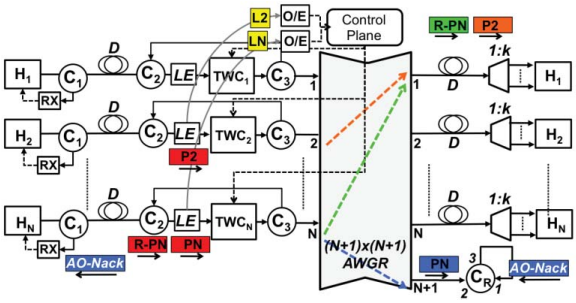

Fig. 1. Interconnect architecture using AO-NACK. H: host, C: optical circulator, LE: label extractor, TWC: tunable wavelength converter; AWGR: array waveguide grating router, and RX: AO-NACK receiver.

II. Principle of Operation

Fig.1 shows an architecture using the AO-NACK technique. N hosts connect to an $(N+1) \times (N+1)$ AWGR by means of a fiber of length D, representing the host-switch distance. The tunable wavelength converter (TWC) at each input port 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

Manuscript received July 21, 2011; revised November 14, 2011; accepted December 8, 2011. Date of publication December 15, 2011; date of current version February 15, 2012. This work was supported in part by the Department of Defense under Contract H88230-08-C-0202 and in part by Google Research Awards.

The authors are with the Department of Electrical and Computer Engineering, University of California, Davis, CA 95616 USA (e-mail: rproietti@ucdavis.edu; yyin@ucdavis.edu; rxyu@ucdavis.edu; xye@ucdavis.edu; cjnitta@ucdavis.edu; akella@ucdavis.edu; yoo@ece.ucdavis.edu).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LPT.2011.2179923

the AWGR input ports) from the counterpropagating (traveling backwards) AO-NACKs. Label extractors (LEs) separate the low-speed label signals from the high-speed payloads and send the labels to the low-speed electronics control plane (CP). The CP processes the label and sends the control signals to the TWCs to switch the packets according to their destination. As explained in detail in [3], optical parallelism in AWGR can be used to reduce contention probability. In fact, in AWGRs, multiple inputs can reach the same output using different wavelengths. Then, with k receivers per output port, the contention probability can be strongly reduced through exploitation of the wavelength contention resolution. Then a $1:k$ optical demultiplexer at each host receiver-side separates the different signals traveling simultaneously on the same fiber. Note that the reflective port does not need a demultiplexer since each wavelength (as many as $N-k$ when all the inputs send packets to the same output and only k get granted) will reflect backwards through the AWGR to reach the original transmitting node. Hence, one reflective port can handle multiple packets. The following example shows the AO-NACK mechanism, with, for simplicity, $k=1$. If two packets (P2 and PN) from different inputs are contending for the same output (output1), the CP switches P2 to output1, while PN is switched to the reflective port $N+1$. An OC used as shown in Fig.1 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 the AO-NACK. A dedicated receiver is then used to detect the AO-NACK and trigger the retransmission. If $d = L/2D \geq 1$, where L is the packet length (in meters), the AO-NACK reaches the sender while the transmission for the related packet is still taking place or when it has just completed. 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 packet has reached the desired output. Otherwise, packet retransmission is triggered. Another solution could consist of including in the packet header a small on-the-fly packet sequence number field and then receiving only the first few bytes of the AO-NACKs. In this case, the AO-NACK technique must preserve the packet content. When d <1, the performance of AO-NACK architecture will be more sensitive to the host-switch distance, since it will take longer to receive the AO-NACKs. Finally, the passive nature of AWGR and OC guarantees that this technique can reflect packets simultaneously without any crosstalk. This aspect, together with the fact that an AO-NACK cannot contend with other packets or AO-NACKs, makes this technique robust, since it is unlikely that an AO-NACK will get lost or corrupted.

III. Experimental Demonstration

Fig.2 shows the experimental setup used for the proof-of-principle demonstration of the AO-NACK technique. In this

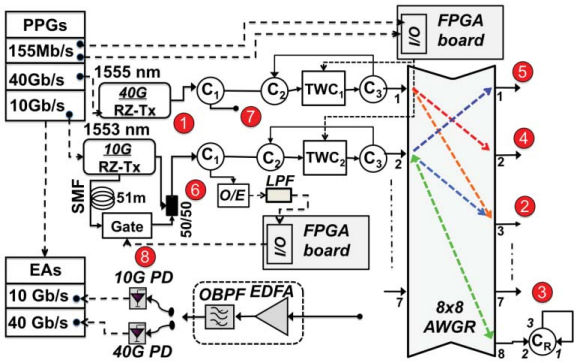

Fig. 2. Experimental setup. PPG: pulse pattern generator, EA: error analyzer, LPF: low-pass filter (electrical), EDFA: erbium-doped fiber amplifier, OBPF: optical band-pass filter, and PD: photodiode.

experiment L is 40.96m (204.8 ns) and D (measured from ports 1 of C_1 to the switch input ports 1 and 2) is $\approx$ 20m. Since $L/2D > 1$, a simple edge detector can be used as the AO-NACK receiver. Two optical return-to-zero (RZ) transmitters at 1555 nm and 1553 nm generate two streams of packets at 40 Gb/s (A packets) and 10 Gb/s (B packets), which enter at AWGR inputs 1 and 2, respectively, as shown in Fig. 3a and 3b (the number at the top-left corner of the packets represents the destination output). The decision to use two different data rates has been mainly made to demonstrate that this technique is not bit-rate-limited. Note that R-B packets (see Fig. 3b) are a copy of B packets (R stands for retransmitted) used to implement packet retransmission as follows. Whenever an A and B packet contend, the A packet is granted, while the B packet needs to be retransmitted. A copy of the B packet (called R-B packet) is then retransmitted. R-B is interleaved in the time-domain with the original B packet (see Fig3b). As shown in Fig.2, the optical copy goes through an optical gate realized with a Mach-Zehnder modulator biased at the minimum point (gate is closed). An R-B packet is transmitted (the gate opens) only upon the reception of an AO-NACK corresponding to the related B packet. Note that, in an actual system, it is possible to store copies of the on-the-fly packets in the end-node Tx buffers and avoid the use of optical gates and optical copies. Each packet contains a portion of 2^{31}-1 PRBS plus a two-byte preamble including a 2-bit on-the-fly packet sequence number (used here just to show the correct position of the switched packets) as shown in insets i, ii, and iii of Fig.3. A 8×8 200 GHz-spacing AWGR occupies the core of the switch architecture. The AWGR insertion loss is 8 dB. Two TWCs are used at AWGR inputs 1 and 2. Each TWC includes a WC based on cross-phase modulation (XPM) in a semiconductor optical amplifier Mach-Zehnder interferometer (SOA-MZI) and a fast tunable laser diode (TLD) board. A field-programmable gate array (FPGA)-based control plane running at 155 MHz is used here. Two pulse pattern generators (PPGs) generate 155 Mb/s 12 bit-long labels for each of the two packet streams. Packets are delayed with respect to labels in order to give time for the CP to tune the TWCs according to the label content. In this experiment, the 155 Mb/s label signals remain in the electrical domain, but in an actual system they would be transmitted in the optical domain on a separate wavelength.

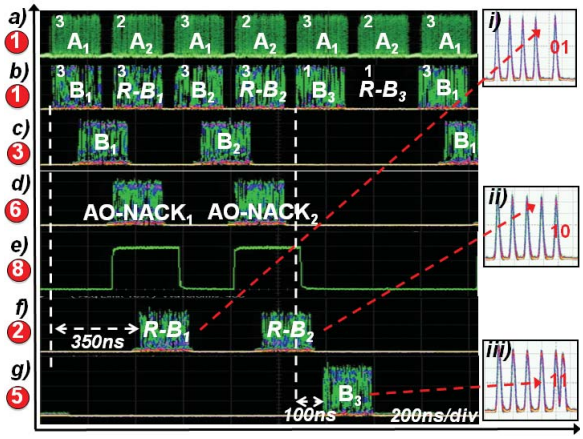

Fig. 3. Timing diagram for the packets traveling through the switch. Numbered dots refer to measured traces at various points in the setup.

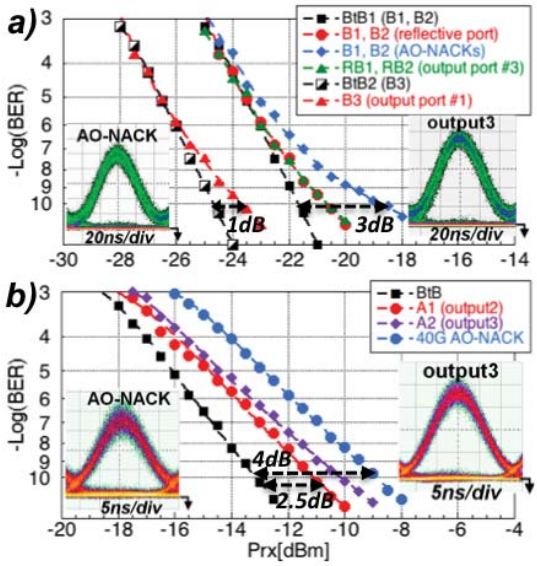

Fig. 4. BER curves for (a) 10-Gb/s packets and (b) 40-Gb/s packets.

As shown in Fig.3, the 10 Gb/s B_1 and B_2 packets contend with 40 Gb/s A_1 packets to reach output 3. The CP grants A_1 packets, while B_1 and B_2 packets are switched to port eight (reflective port) as shown in Fig.3c. C_R reflects B_1 and B_2 packets, which, traveling backwards, act as AO-NACKs, which reach C_1, where they are extracted and enter the AO-NACK receiver with an average optical power of -19.5 dBm. A 2.5 GHz photo-receiver and a low-pass filter (250MHz) works as an edge-detector to sense the AO-NACKs and trigger an FPGA running at 155MHz to generate a 300ns-long gate signal (see Fig. 3e) with $V_{PP} = V\pi$, which opens the gate for the R-B packets. Once the FPGA is triggered, its latency is one clock cycle (6.4ns). Since B_3 packets do not contend with A_1 packets, the R-B_3 packets are not transmitted because no AO-NACKs related to R-$B_3$are received (see Fig. 3b). The gate does not open for R-B_3, which explains why R-B_3 is missing in Fig. 3b. R-B_1 and R-B_2, which do not contend this time with the A_2 packets, are switched then to output 3.The retransmission latency (measured from the transmitter output to the AWGR output) is $\approx$ 350ns (see Fig. 3f). Since the latency for non-contended packets is $\approx$ 100ns (see latency for B_3 in Fig.3), the latency penalty is $\approx$ 250ns. About 200ns are accounted for by the round-trip time (2D) necessary to receive the AO-NACK, while the remaining 50ns is accounted for by the guard-time between B and R-B packets.

Figs. 4*a* and 4*b* report bit error rate (BER) measurements, as a function of the average optical power at the input of the photodiode (PD), for the 10 Gb/s and 40 Gb/s packets, respectively. All the BER measurements have been taken using an optically pre-amplified front-end (see Fig.2). For the same average optical power at the input of the photo-detector (PD), the peak power of switched B3 packets is higher than that of switched B1, B2, RB1, RB2 packets, because of the longer interval between the B3 packets (see traces c, f, g in Figure 3). Since the BER curves are plotted as a function of the average received optical power, the BER curves for B3 packets are left-shifted with respect to the BER curves for the other packets. It is also important to verify that AO-NACK technique preserve the reflected packets' content. Figs. 4*a* reports BER measurements for the AO-NACKs. The power penalty at BER $= 10^{-10}$ for the 10Gb/s packets (B1, B2, RB1, RB2, and B3) is $\approx$ 1 dB. The power penalty for the AO-NACKs is $\approx$ 3 dB due to the lower optical signal-to-noise ratio (OSNR) at the PD input. The power penalty at BER $= 10^{-10}$ for the 40 Gb/s packets (A1, A2) is $\approx$ 2.5 dB. Fig. 4*b* reports also a BER curve to show that the AO-NACKs are error-free at 40 Gb/s as well. A penalty of $\approx$ 4 dB is observed, still due to lower OSNR at the PD input.

IV. Conclusion

This letter demonstrates a technique for the provision of a physical layer all-optical NACK to the end-nodes connected to an AWGR-based optical interconnect. The technique allows fast packet contention notification and retransmission and can reduce the latency increase associated with retransmission. The average latency will depend by the distance between the switch and the end-nodes, which will determine the minimum retransmission latency for the contending packets.

References

[1] R. Hemenway, R. R. Grzybowski, C. Minkenberg, and R. Luijten, "Optical-packet-switched interconnect for supercomputer applications," *J. Opt. Netw.*, vol. 3, no. 12, pp. 900–913, 2004.

[2] O. Liboiron-Ladouceur, *et al.*, "The data vortex optical packet switched interconnection network," *J. Lightw. Technol.*, vol. 26, no. 13, pp. 1777–1789, Jul. 1, 2008.

[3] X. Ye, *et al.*, "DOS: A scalable optical switch for datacenters," in *Proc. ACM/IEEE Symp. ANCS*, La Jolla, CA, Oct. 2010, pp. 1–12.

[4] S. J. B. Yoo, "Optical packet and burst switching technologies for the future photonic internet," *J. Lightw. Technol.*, vol. 24, no. 12, pp. 4468–4492, Dec. 2006.

[5] A. Shacham and K. Bergman, "An experimental validation of a wavelength-striped, packet switched, optical interconnection network," *J. Lightw. Technol.*, vol. 27, no. 7, pp. 841–850, Apr. 1, 2009.