Scalability and Performance of a Distributed AWGR-based All-Optical Token Interconnect Architecture

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

Abstract: This paper studies an interconnect architecture with distributed all-optical control plane. A physical layer analysis shows scalability up to 128 ports. Simulations for an 128-port switch show low latency and high throughput at 0.75 load.
OCIS codes: (200.4650) Optical Interconnects; (200.6715) Switching.

1. Introduction

Optical interconnects have emerged as a promising method to realize high-port-count, low-latency, and high-throughput networks in high-performance computing (HPC) systems and data centers. Several research projects have already proposed architectures for optical interconnects [1-3] for HPC. 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 fully-connected interconnection [3].

In general, a bottleneck to the scalability of any switch architecture can arise from the centralized electrical control plane, where the maximum number of I/O resources of the integrated circuits can limit the optical switch port count. A distributed control plane is highly desirable from both scalability and architecture considerations.

Recently, the demonstration of the all-optical token (AO-TOKEN) technique [4] led to a fully distributed control plane for an AWGR-based architecture. This technique exploits the saturation effect in semiconductor optical amplifiers (SOAs) and the wavelength routing in the AWGR. In this paper, we propose a variant of the architecture in [4] to eliminate any scalability limitation given by polarization crosstalk. A physical layer scalability analysis (SOA gain bandwidth, four-wave mixing (FWM) in SOAs, optical noise) shows that the AO-TOKEN architecture can scale up to 128 ports. Network simulation results, obtained under uniform random distribution, show that a 10Gb/s 128-port AO-TOKEN switch can achieve low latency and high throughput under high offered loads. This can be obtained thanks to the short host-switch distance typical of HPC networks (~1 m for board-to-board communication in a rack) and the unique wavelength-domain contention resolution offered by AWGR technology.

2. AO-TOKEN architecture with distributed all-optical control plane.

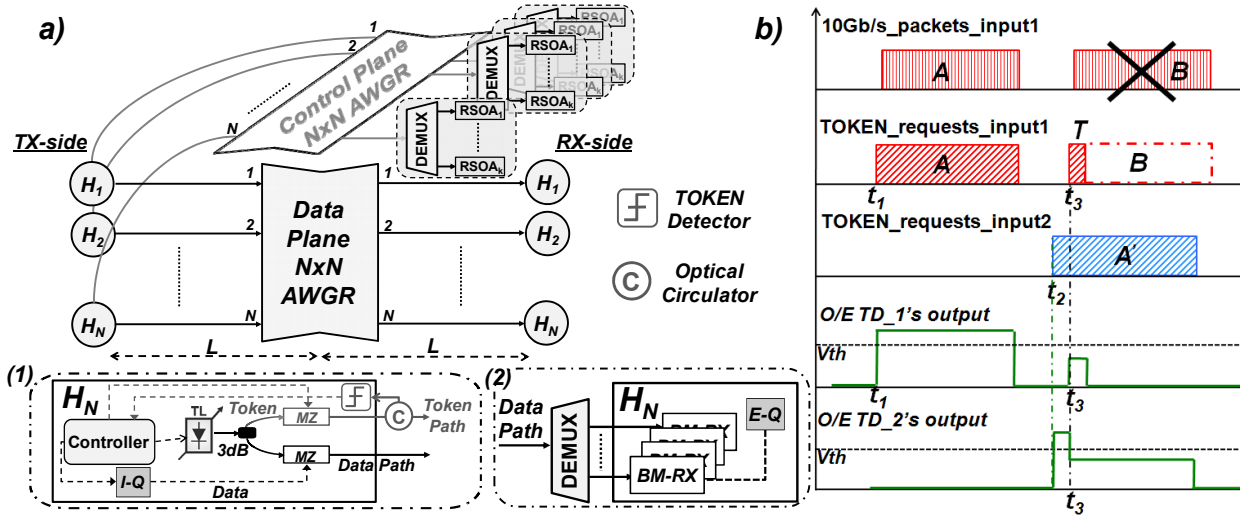

Figure 1. (a) Distributed AO-TOKEN architecture. *L* is the distance between hosts (H_i) and AWGRs input ports. Inset 1: host TX interface with ingress buffer queue (I-Q), fast tunable transmitter (TL) and token detector. Inset 2: host RX interface with 1:*k* demux, *k* burst-mode receivers (BM-RX) and egress buffer queues (E-Q). Each control plane AWGR output port connects to an optical demultiplexer and *k* RSOAs. (b) Timing diagram explaining how the all-optical control plane can detect contention.

Figure 1(a) shows the AO-TOKEN architecture. Insets 1 and 2 show the host TX and RX interface respectively. TX equipped with a fast tunable laser (TL) [5] transmits the packets in the TX ingress buffer (I-Q). A packet transmission starts only after sending a token request and receiving of a positive grant by the token detector (TD). While the solution in [4] used a polarization-diversified (PD) scheme to transmit the token-based control plane (CP)

messages and data packets on the same shared media, here the CP messages use a separate optical path. In this way, no polarization maintaining components are necessary and scalability limitation given by polarization crosstalk is avoided. Inset 1 of Figure 1(a) shows the optical transmitter generating both packets and token requests. Figure 1(b) illustrates the token-based distributed contention resolution. Let us assume that at $t=t_1$, host_1 sends a packet to host_*N*. As first, host_1 tunes its fast TL [5] to λ_{1N} (the wavelength to reach output *N* from input 1, according to AWGR routing table) to generate a token request **A** which reaches the CP AWGR input port 1. Here **A** is routed to output *N,* where it enters in a reflecting SOA (RSOA) after going through a 1 by *k* optical demultiplexer. In general, there are *k* RSOAs for each CP AWGR output in order to exploit the wavelength parallelism and reduce the contention probability [3]. The RSOA amplifies and reflects the token request **A**, which travels back to the host TX interface, is extracted by an optical circulator, and is converted in the electrical domain by a TD. The TD generates an electrical signal with Vp= V_{TO1} proportional to the optical power (P_{TO1}) of the reflected token request and above a certain threshold V_{th}. This condition means that output *N* is available and that the transmission of packet **A** on λ_{1N} can start. The same situation arises at $t= t_2$ when host_2 generates a token request **A'** directed to output *N*. Note that the token requests stay active for the entire packets transmission to hold the token and to prevent collision. The reader should take note of the behavior at $t= t_3$, when the transmission of packet **A'** has not yet completed, but when host_1 wants to transmit another packet to output *N*. The RSOA at output *N*, which is already fed with the token request signal **A'** at λ_{2N}, amplifies and reflects back the new token request **B** at λ_{1N}, which reaches the TD with optical power P_{TO3}. The TD generates an electrical signal with Vp= V_{TO3}. Assume that the RSOA was saturated by the token request **A'** at $t = t_2$. Due to the gain saturation effect [6], P_{TO3} becomes then ~ Psat/2 and V_{TO3} becomes~ $V_{TO1}/2$, where Psat is the output saturation power of the RSOA. By setting V_{th} between V_{TO1} and $V_{TO1}/2$, it becomes possible for host_1 to recognize that the token for output 1 is not available and that it must retry at a later time (see [4] for more details). The AO-TOKEN technique does not require a centralized control plane and the acquisition of the token is handled all-optically in a fully distributed manner.

3. Physical Layer Scalability: RSOA gain bandwidth, ASE noise, and four-wave mixing.

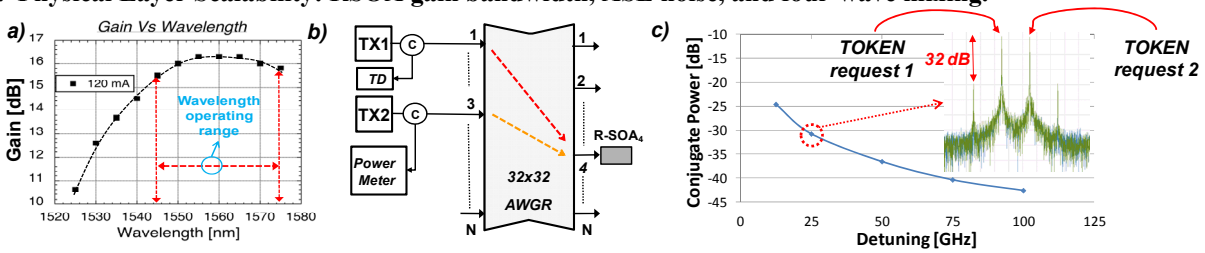

Figure 2. (a) Wavelength operating range for the RSOA used in [4]. (b) Experiment setup to measure the RSOA noise power contribution at the TD input. (c) Simulation results showing the normalized conjugate power of the FWM products.

Figure 2(a) shows experimental data for the wavelength operating range of the RSOA used in [4]. The range is determined by the SOA gain 1-dB bandwidth. Within this range, the power of the reflected token request is almost independent of the wavelength value, making it possible to use a constant TD's threshold voltage V_{th}. With a 1-dB bandwidth of ~30 nm and a 0.2 nm channel spacing for the AWGR, a port count ≥ 128 is possible.

Figure 2(b) shows a simple setup to measure the RSOA noise contribution at the TD input. We used a 50-GHz-spacing AWGR. The RSOAs noise is fed back into the 128 receivers after being narrowly filtered by the AWGR. When adding the noise power of all the 128 RSOAs, the total noise power added to the received token is not negligible. We measured a ratio of 30dB between the power of one reflected token request and the noise contribution given by one RSOA. Adding all the contributions from 128 ports (in reality is 127), each TD would see a noise level equal to -30dB+ $10\log_{10}(127)$= -8.96dB. Thus, for 128 ports, the RSOA noise is not a limiting factor.

The RSOAs work in saturation. Thus, there will be FWM products acting as crosstalk at the TD inputs. Using VPI photonics software, we simulated FWM when two token requests saturate the RSOA. Figure 2(c) shows the results as function of the detuning between the token requests (detuning is a multiple of the channel spacing; *i.e.* 0.2 nm). The normalized FWM conjugate power is < 30 dB for detuning ≥ 0.2 nm. The number of interfering FWM products highly depend on the traffic pattern. However, because of the constraints given by the AWGR routing table, the number of FWM products at the TD inputs will be < N/2 (more details will be presented at the conference). Thus, FWM is not the main limiting factor for the scalability of the AO-TOKEN architecture.

4. Network performance

We conducted network simulations for an AO-TOKEN switch with 64 and 128 ports and compared against the centralized AO-NACK architecture in [7], which is also based on wavelength contention resolution in AWGR with *k* receivers per output port (the reader can refer to [7] for more details). Table 1 shows the simulation parameters.

Table 1. Main simulation parameters. *k* is the number of receivers for each host; *d* is the distance effect parameter quantifying the ratio between the packet size (transmission time) and host-switch distance (propagation time).

	N	***L* [m]**	***Packet Size* [B]**	***d***	***k***	**Bit-rate**	**Traffic**
AO-NACK	64/128	1	256/64	40/10	4	10Gb/s	Random
AO-TOKEN	64/128	1	256/64	40/10	4	10Gb/s	Random

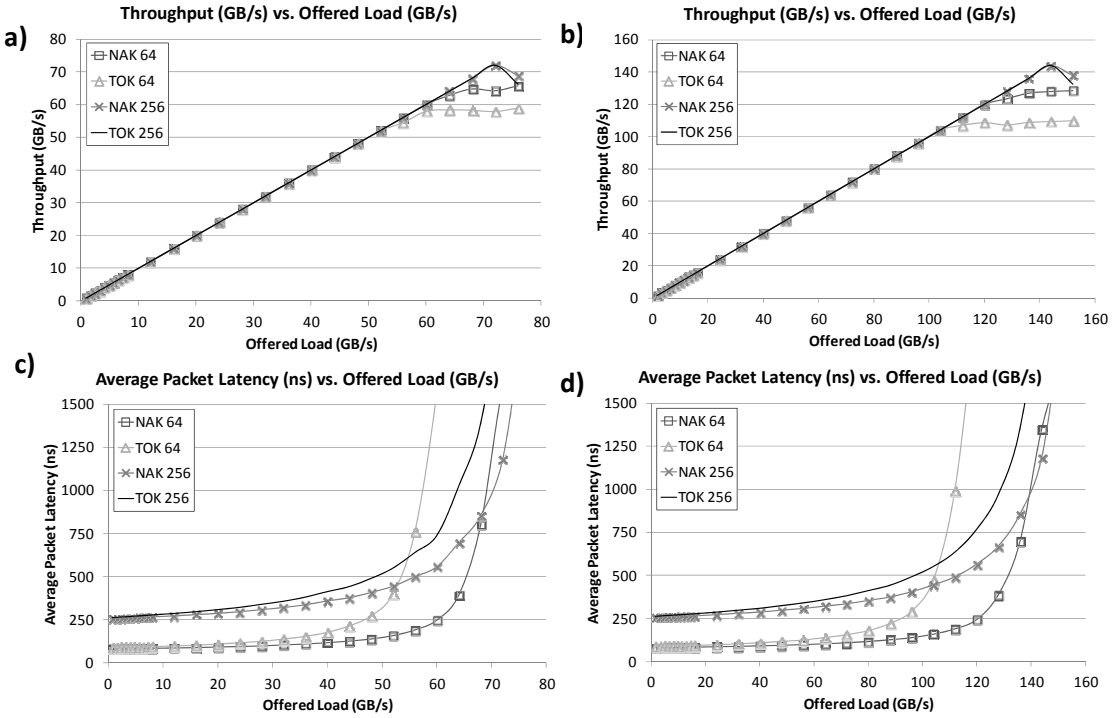

Figure 3. Throughput and average packet latency vs offered load for 64 (a,c) and 128 nodes (b,d).

Figure **3** shows the performance of the two architectures in terms of throughput and average packet latency as a function of the offered load. The host-switch distance is fixed to 1m while the average packet size changes to study how the parameter *d* (see Table **1**) affects the performance. Both the AO-TOKEN and AO-NACK (denoted as TOK and NAK in Figure 3) perform better with a larger *d*; however, the AO-TOKEN is more affected by node count than is the AO-NACK. Notice that for larger values of *d* the difference in performance of the AO-NACK and AO-TOKEN is smaller even when there is a higher port count, meaning that the AO-TOKEN may be best suited for switches with large port count and large average packet size.

5. Conclusions

We proposed an AWGR-based interconnect architecture (AO-TOKEN) with a distributed all-optical control plane. Network simulations show that AO-TOKEN can guarantee low latency and high-throughput at high traffic load when the packet transmission time is sufficiently low compared to the token requests propagation time.

[1] R. Hemenway, R. Grzybowski, C. Minkenberg, and R. Luijten, "Optical-packet-switched interconnect for supercomputer applications [Invited]," *Journal of Optical Networking,* vol. 3, pp. 900-913, 2004.

[2] C. Hawkins, B. A. Small, D. S. Wills, and K. Bergman, "The Data Vortex, an All Optical Path Multicomputer Interconnection Network," *IEEE Transactions on Parallel and Distributed Systems,* vol. 18, pp. 409-420, 2007.

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

[4] R. Proietti, C. Nitta, Y. Yin, R. Yu, S. Yoo, and V. Akella, "Scalable and Distributed Contention Resolution in AWGR-based Data Center Switches Using RSOA-based Optical Mutual Exclusion," *IEEE Journal of Selected Topics in Quantum Electronics,* vol. PP, pp. 1-1, 2012.

[5] G. Sarlet, G. Morthier, and R. Baets, "Control of widely tunable SSG-DBR lasers for dense wavelength division multiplexing," *Journal of Lightwave Technology,* vol. 18, pp. 1128-1138, 2000.

[6] M. J. Connelly, *Semiconductor optical amplifiers*: Springer, 2002.

[7] R. Proietti, C. Nitta, X. Ye, Y. Yin, V. Akella, and S. Ben Yoo, "Performance of AWGR-based Optical Interconnects with Contention Resolution based on All-Optical NACKs," *Optical Fiber Communication Conference (OFC),* 2012.

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