Scalable and Distributed Contention Resolution in AWGR-Based Data Center Switches Using RSOA-Based Optical Mutual Exclusion

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

(Invited Paper)

***Abstract*—We describe a mutual exclusion element using a reflective semiconductor optical amplifier (RSOA) and a simple scheme for contention resolution in arrayed waveguide grating router (AWGR)-based optical switches in data centers. We describe a hardware demonstration and detailed performance analysis of an AWGR-based optical switch based on the proposed concept. We show that the proposed RSOA-based contention resolution significantly reduces latency compared to existing methods and that it does not require any global or centralized coordination, which makes it inherently scalable and suitable for emerging data center networks.**

***Index Terms*—Contention resolution, data center networks, mutual exclusion, photonic interconnects.**

I. Introduction

OPTICAL communication links have advantages such as the ability to transport *high bandwidth* data across *long distances* with little distortion, over their electrical counterparts. Over the past few decades, optical communication has been widely deployed in long-haul telecom networks and serves as the backbone of the Internet. Today, there is great interest in a different question—*Can optical interconnects be effectively used to connect computers in a data center?* Clearly, the potential to provide high-bandwidth at a lower energy per bit compared to an all-electrical network is the primary motivation for the question. However, a data center that runs computing applications is different from a general telecom network, and this introduces new requirements for a data center interconnection architecture that in turn influences the potential of optics in a data center. In fact, the network inside a data center is quite different from wide-area and local-area networks (see Table I) that have been the subject of intensive research over the past few decades both in the contest of optical packet and burst switching [1].

Data center networks have to be scalable to hundreds of thousands of nodes (where each node is typically a server) and capable of handling bursty traffic comprised of small packets [3]. Both the network performance (in terms of latency) and the power consumption have become critical in the overall performance/power consumption of the entire data center [4].

TABLE I
Requirements of Telecom and Datacom Switches

	Telecom	Datacom
Switch Port Count	5~6 with WDM	1000s
Switching Latency	100 micro ~ 1ms	100s ns
Packet Losses	Up to 2%	< 0.0001%
Hop Counts	Multi-hop	Single or few hops
Distances	100s km	10s m ~ 2 km
Contention Resolution	Wavelength-time-space-packet drop [1] [2]	Time and wavelength domain w/o packet drop
Control Plane	Centralized	Distributed

Wavelength division multiplexing (WDM) allows optical networks to provide very high bandwidth with relatively low wiring complexity and significantly lower energy cost per bit [5], [6]. However, thus far, there has been no compelling reason to reduce latency, especially by dealing aggressively with contention. Similarly, scalability to very large number of nodes and the issue of very high input loads for short intervals of time has also not received much attention in optical data communication networks.

We argue that reducing latency and improving scalability by several orders of magnitude are the keys to bringing optical interconnects into the data center. Interestingly, there is a common problem that underlies both these requirements—namely, the ability to address contention, i.e., what to do when multiple packets need to go the same destination? This is a particularly nasty problem in optical networks because of the lack of *random-access memory*, where a packet can be stored for an arbitrary amount of time.

In recent work [2], we showed that it is possible to use a small electrical buffer called a loopback buffer to store contending packets and retransmitting them at the appropriate time without contention. However, we found that the loopback buffer quickly becomes the bottleneck and a major impediment to scalability both in terms of data rate and number of ports. The electronic loopback buffer has to operate at the data rate and result in

Manuscript received June 1, 2012; revised July 10, 2012; accepted July 11, 2012. Date of publication July 17, 2012; date of current version April 3, 2013. This work was supported by the Department of Defense under Contract H88230-08-C-0202 and under 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; cjnitta@ucdavis.edu; yyin@ucdavis.edu; rxyu@ucdavis.edu; yoo@ece.ucdavis.edu; akella@ucdavis.edu).

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

Digital Object Identifier 10.1109/JSTQE.2012.2209113

significant design complexity and power consumption. In [2] and [7] more details have been provided, while in [8], a new way to eliminate the loopback buffer without affecting the performance has been proposed. More importantly, in all cases, the contention resolution and arbitration had to rely on centralized control plane. A centralized electrical control plane is another major reason for limited scalability for not only optical switches, but also for any switch architecture, since it limits port count and increases latency. In fact, the maximum number of I/O resources of currently available integrated chips [9] can pose an upper limit to the number of ports that a single control plane can handle. Considering the limitations of a centralized control plane, it should be clear that a distributed control plane is highly desirable.

In this paper, we introduce a new way to deal with contention in an optical network that is suitable for a data center. It does not have the complexity or the power consumption of the loopback buffer solution, it is distributed, and yet, at the same time, it exhibits low latency and high scalability. We introduce the use of a reflective semiconductor optical amplifier (RSOA) [10], a widely used optical active component, as a distributed mutual exclusion element and propose a simple protocol to detect contention and retransmit packets without incurring a significant latency overhead. Most importantly, we demonstrate that the proposed solution is simple to implement and makes the control plane of an arrayed waveguide grating router (AWGR)-based optical switch fully distributed and hence arbitrarily scalable.

The rest of this paper is organized as follows. In Section II, we will describe how to use a RSOA to realize mutual exclusion in the optical domain and discuss the design of a simple retransmit scheme. In Section III, we show an experimental demonstration of the concept. In Section IV, we discuss the development of a simulation model for the contention resolution scheme and present detailed performance analysis of a switch that uses the proposed RSOA-based contention resolution by comparing it with the loop-back buffer-based switch [7]. We summarize the impact of various parameters on the performance of the new scheme in Section V. In Section VI, we discuss related works from the literature and then conclude in Section VII.

II. Optical Mutual Exclusion With RSOA

A. What is Mutual Exclusion?

Mutual exclusion [11] is a basic necessity in distributed computation—it is required for independent and concurrent nodes to share a resource without resulting in an incorrect operation. In its simplest form a 2:1 mutual exclusion element (mutex) has two inputs called R_1 and R_2 and two outputs called G_1 and G_2. Every node is associated with a dedicated input/grant pair. For example, input R_1 and grant G_1 could be associated with node N_1 and input R_2 and grant G_2 with node N_2. When node N_1 needs to access the shared resource (such as memory), it makes a request by asserting R_1 and similarly when node N_2 requires the shared resource it makes its request by asserting R_2. The mutex element, however, grants access to either node N_1 or node N_2 exclusively by asserting G_1 or G_2. A mutex element realized using CMOS transistors is showed in Fig. 1 (on the right-hand side). The implementation is based on an

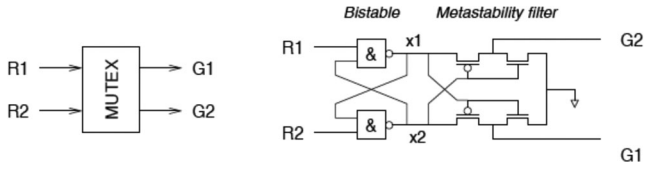

Fig. 1. CMOS mutex realization. Metastability filter ensures mutual exclusion.

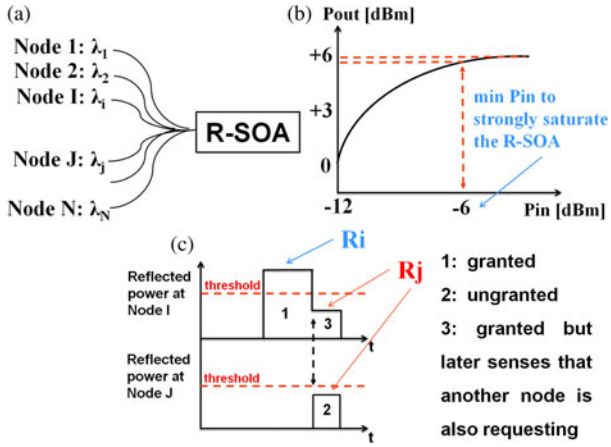

Fig. 2. (a) N:1 Mutual exclusion scheme with RSOA. (b) Typical RSOA pout/pin characteristic and minimum input power to operate the RSOA in strong saturation regime. (c) Three different possible cases in the RSOA-based mutual exclusion scheme.

NAND-based latch. Since the latch is prone to metastability, if requests R_1 and R_2 arrive simultaneously, a metastability filter placed at the output of the latch makes sure that eventually only either G_1 or G_2 (and not both G_1 and G_2) is asserted.

In general an N-input mutex receives at most N inputs and grants access to at most 1 of the requestors. A negative Acknowledgment (NACK) is sent to the remaining N-1 requestors so they can retry later. The benefit of a mutex element realized this way is that every requestor knows immediately (without a timeout or other latency introducing mechanism) whether they have been granted access to the shared resource or not. So, the question is, *is it possible to realize a general mutex element with optical devices*?

B. RSOA—Reflecting Semiconductor Optical Amplifier

The gain saturation effect in an RSOA [12] can be used to realize mutual exclusion behavior as shown in Fig. 1. N different nodes can make requests $R_1, R_2, \ldots, R_n$ to the RSOA associated with a given port using different wavelengths λ_1, $\lambda_2, \ldots, \lambda_n$ (see Fig. 2). The first request, say R_i, that arrives at the RSOA saturates it, which results in P_{tot} power reflected back to the sender node I. The RSOA stays saturated as long as the request on λ_i is held. A detector that is set to trigger at P_{tot} produces the grant signal. If another request R_j (on λ_j) from node J arrives while R_i is still active, the power reflected at λ_j will be $\approx P_{\text{tot}}/2$ (because of the saturation effect in the RSOA), which is not enough to set the trigger condition; hence, the second request will be excluded [see Fig. 2(c)].

If two requests arrive at approximately the same time at the RSOA (with a time interval comparable or lower than the RSOA

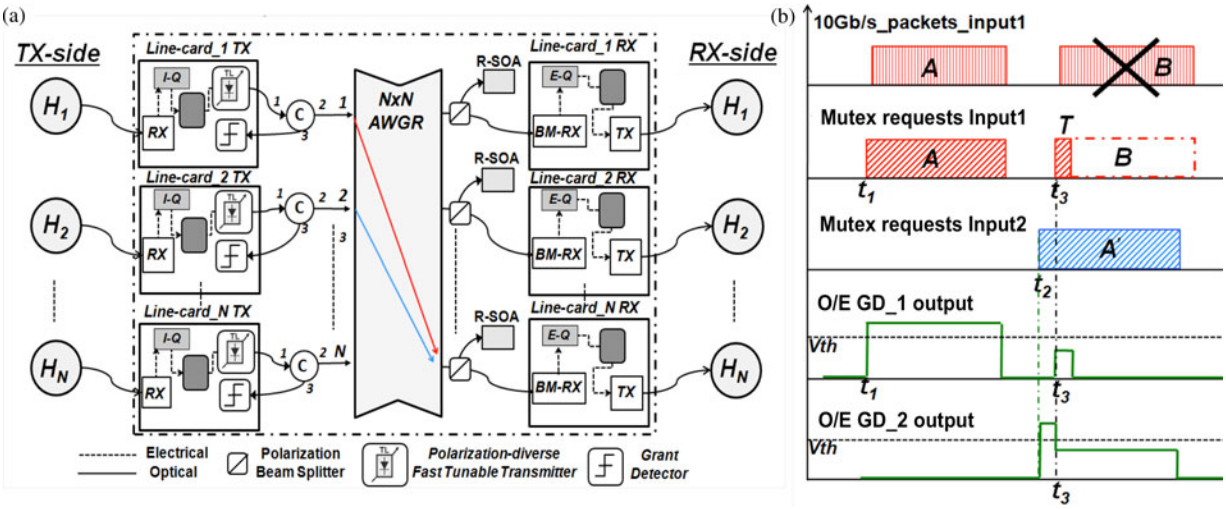

Fig. 3. (a) RSOA-based optical mutex architecture. H: host; I-Q: ingress queue; E-Q: egress queue; C: optical circulator. (b) Timing diagram that illustrates the working principle of the optical mutex contention resolution scheme.

gain dynamics, i.e., few hundreds of picoseconds), both the requestors receive approximately $P_{\rm tot}/2$ reflected power and hence the detectors at neither node triggers, which corresponds to a situation that neither requestor has been granted. Note that this is different from a classic electronic mutex element where eventually one of the requestor gets a grant. In the RSOA-based mutex element, it is possible for none of the requestors to get a grant—that is okay, because the requirement of mutual exclusion element is that *at most one* of the requestors be granted the resource—zero grant is okay from the correctness of the protocol perspective.

RSOA-based mutual exclusion has an interesting property. Suppose, a request R_j arrives at the RSOA after request R_i has been granted, i.e., while R_i is still asserted, then the reflected power to node *I* drops from $P_{\rm tot}$ to $P_{\rm tot}/2$, which could serve as a signal to node *I* that some other node has made a request to the same resource. This information can be used to ensure fairness, while removing the overhead for additional arbitration. Note that the simple mutex element shown in Fig. 1 cannot provide this information and in fact, it is very difficult to do this in an electrical implementation in general.

Another interesting aspect of RSOA-based mutual exclusion is that it is inherently scalable to N requestors, i.e., a single RSOA can function as a mutual exclusion element (with appropriate incident power) for N requestors, while in an electrical implementation a set of 2:1 mutex elements have to be interconnected (typically in a tree structure) to realize a N:1 mutex element, which increases the latency and control complexity significantly. With the RSOA-based implementation, no additional control is required, so it is *fully distributed.* A given RSOA makes the decision solely based on the requests it receives and does not need to know the state of other RSOAs to make its decision, which makes contention resolution (or arbitration) arbitrarily scalable. Note that if an optical coupler is actually used to realize the mutex as illustrated in Fig. 2, the technique itself would not really be scalable because of the attenuation of the coupler being equal to $\log_2 N$ [dB], where N is the number of nodes. In Section III, we will show how the RSOA-based optical mutex (OMX) can be used together with the unique property of wavelength routing of AWGR (with an insertion loss independent from the number of AWGR ports) to construct a scalable $N \times N$ optical switch with distributed all-optical arbitration.

C. Contention Resolution Protocol Using RSOA-Based OMX Element

The RSOA-based technique described earlier can be used to realize a simple distributed contention resolution protocol as shown as follows:

Step 1 Node *I*, asserts its request R_i on λ_i
Step 2 Node *I* senses the reflected power P_{ri}
Step 3 If $(P_{ri} > P_{\rm tot})$ then grant $= 1$ else grant $= 0$;
Step 4 If (grant $==1$), transmit packet U_i; If there are no additional packets to transmit, then deassert Request R_i;
Step 5 If (grant $==0$), RETRANSMIT (U_i, T_i) where T_i is the time at which U_i is retransmitted.
Step 6 If (grant $==1$), $P_r <= P_{\rm tot}/2$; There is another potential requestor; deassert after current transmission.

The choices of $P_{\rm tot}$ and RETRANSMIT (U_i, T_i) function are critical design parameters of the contention resolution and depend primarily upon the physical characteristics of the switch. We will discuss the impact of this on the system level performance later in this paper. However, the first question is, *how well does this scheme work*? *What are the critical design parameters*? These questions are answered in the next section.

III. RSOA-Based $N \times N$ Optical Switch Architecture, Proof of Principle, and Hardware Demonstration

A. Optical Switch Architecture

This section explains how the optical N:1 mutual exclusion block can be used to build an $N \times N$ optical switch with all-optical and distributed control plane. Fig. 3(a) shows the optical

interconnect architecture. An $N \times N$ AWGR is at the core of the system. Each input port is connected to a line-card (*Line-card_i* TX). Each line-card receives packets from a host H_i and buffers them in an input-queue (I-Q). The packets are then transmitted in the optical domain by means of a transmitter equipped with a fast tunable laser (TL) [13]. A grant detector (GD) initiates a packet transmission upon the reception of the mutex grant. A polarization-diversified scheme is used to avoid interference between the control plane requests and data. An optical circulator C is used to extract the counterpropagating control plane messages. Each AWGR output is then connected to a polarization beam splitter (PBS) to separate the control plane and data path. One PBS output connects to an RSOA, which is the key component in this all-optical contention resolution scheme, as already explained in the previous section. The PBS data output connects to a linecard (*Line-card_i RX*), which buffers the received packets in an egress queue (E-Q), and transmits them to the destination. Fig. 3(b) illustrates the working principle of the mutex technique. At time t_1, Node1 needs to send a packet to *output_N* and it tunes its TL to λ_{1N} to generate a request A. The RSOA at output N reflects the signal extracted by the PBS, which reaches the *line-card_*1's GD with an optical power P_{TO1}. The O/E converter in the GD generates an electrical signal with $V_p = V_{TO1} \geq V_{\rm th}$. The request is granted and packet transmission can start. At time t_2, Node 2 has also a packet for *output_N*. Then, *line-card_2* tunes its TL to λ_{2N} to generate a request A'. The RSOA reflects the mutex request extracted by the PBS, which reaches the *line-card_2's* GD with an optical power P_{TO2}. As was the case for request A, the O/E converter in the GD generates an electrical signal with $V_p = V_{TO2} \geq V_{\rm th}$. The packet transmission from *line-card_2* can begin because the request has been granted. At time t_3, Node 1 needs to send another packet to *output_N* and it tunes its TL to λ_{1N} to generate a request B. The RSOA at output N reflects the signal extracted by the PBS, which reaches the *line-card_*1's GD with an optical power P_{TO3}. However, at t_3, the RSOA is strongly saturated by the currently active request A'. At t_3, due to the saturation effect in the RSOA, we will have $P_{TO3} = (P_{TO1} - 3)$ dBm. Then, $V_{TO3} = V_{TO1}/2 < V_{\rm th}$. This voltage value is insufficient to trigger the packet transmission. Node1 is then notified that the desired AWGR output is not currently available and it will need to retry after a random selected time interval (the details of which are discussed in Sections IV and V). At this time, Node1 also needs to stop the mutex request B in order to prevent resource starvation. Note that the voltage drop in O/E GD2 output at $t = t_3$ can be used to sense whether or not other nodes need to access the same output. This information can be used to notify Node2 of others nodes desiring the output port so that the mutex can be released after the packet transmission has completed, removing the potential for node starvation.

The minimum end-to-end latency associated with this implementation includes the TL tuning time [12], the round-trip time (for the request to reach the RSOA and go back to the GD input—this can be considered $\leq$5 nanoseconds since the line-cards with TLs are placed at the switch), the field-programmable gate array (FPGA) processing time (assuming an FPGA running at 500 MHz this time is $\leq$4 ns since it takes not more than two clock cycles to recognize whether or not a grant is given), and the packet transmission time. Then, to this minimum value, it is necessary to add the latency associated with the retransmission in case the token request is not granted, as discussed in Section IV-C.

To generalize, all the mutex requests for output N occurring during the transmission of packet A will be denied. However, multiple mutex requests for different outputs can be satisfied simultaneously since there is an RSOA for each AWGR output. Hence, no centralized control is required for contention resolution, which means the contention resolution can be done in a fully distributed manner, removing one of the biggest barriers to scalability in networks.

B. Hardware Demonstration

The experimental demonstration of the RSOA-based contention resolution is shown in Fig. 4. Two polarization diverse TXs (PD-TX) are connected to input ports 1 and 4 of a 200GHz-spaced 8×8 AWGR (8dB uniform insertion loss). Polarization controllers (PCs) at AWGR inputs align the signal polarization with the PBSs at the AWGR outputs. Alternatively, all polarization maintaining (PM) components could be used. Each PD-TX includes a PBS and polarization beam combiner (PBC) to multiplex in the polarization domain the data and control plane request paths. The mutex arm of the PD-TX includes a Mach Zehnder (MZ) modulator. Two MZs are used in the data arm as data modulator and gate. The gate is controlled by an FPGA and remains open unless the request is not granted. The FPGA also generates the control plane requests, while the 10 Gb/s 406.9-ns-long packets are generated with a pattern generator, with each packet containing a portion of 2^{31}–1 pseudorandom bit sequence. A PBS is placed at AWGR output 3. The PBS extracts the requests, which enter an RSOA implemented here with an optical circulator and an SOA. The PC at the SOA output maximizes the optical power going back through the PBS and reaching the GDs. The second PBS output connects to an O/E converter for BER measurements on the data path.

Fig. 4(c) shows the measured traces for the packets at AWGR input1, mutex requests at AWGR inputs 1 and 4, GD1's and GD2's O/Es and comparators outputs and gate1 output. Numbered dots refer to the points at which, in the experiment setup of Fig. 4(a), each trace was measured. As we see, the timing diagram obtained from the experiment shown in Fig. 4(c) is an almost replica of Fig. 3(b) which demonstrates that the proposed scheme works exactly as described in the previous section. Note that in the experiment there is a delay between the leading edges of the mutex requests and the relative GD outputs. This delay is simply due to the propagation delay caused by the fiber pigtails of the bulky components used in the experiment. There is also a small delay between GD1's O/E converter leading edge (related to mutex request B) and the GATE input (TX1) signal. This small delay is $\leq$ two FPGA clock cycles ($\leq$12.8 ns in this experiment since the FPGA was running at 156 MHz). Note that the GD1's O/E converter output shown in Fig. 4(c) does not go to zero when B does [see Fig. 4(c)]. This is due to the fact that

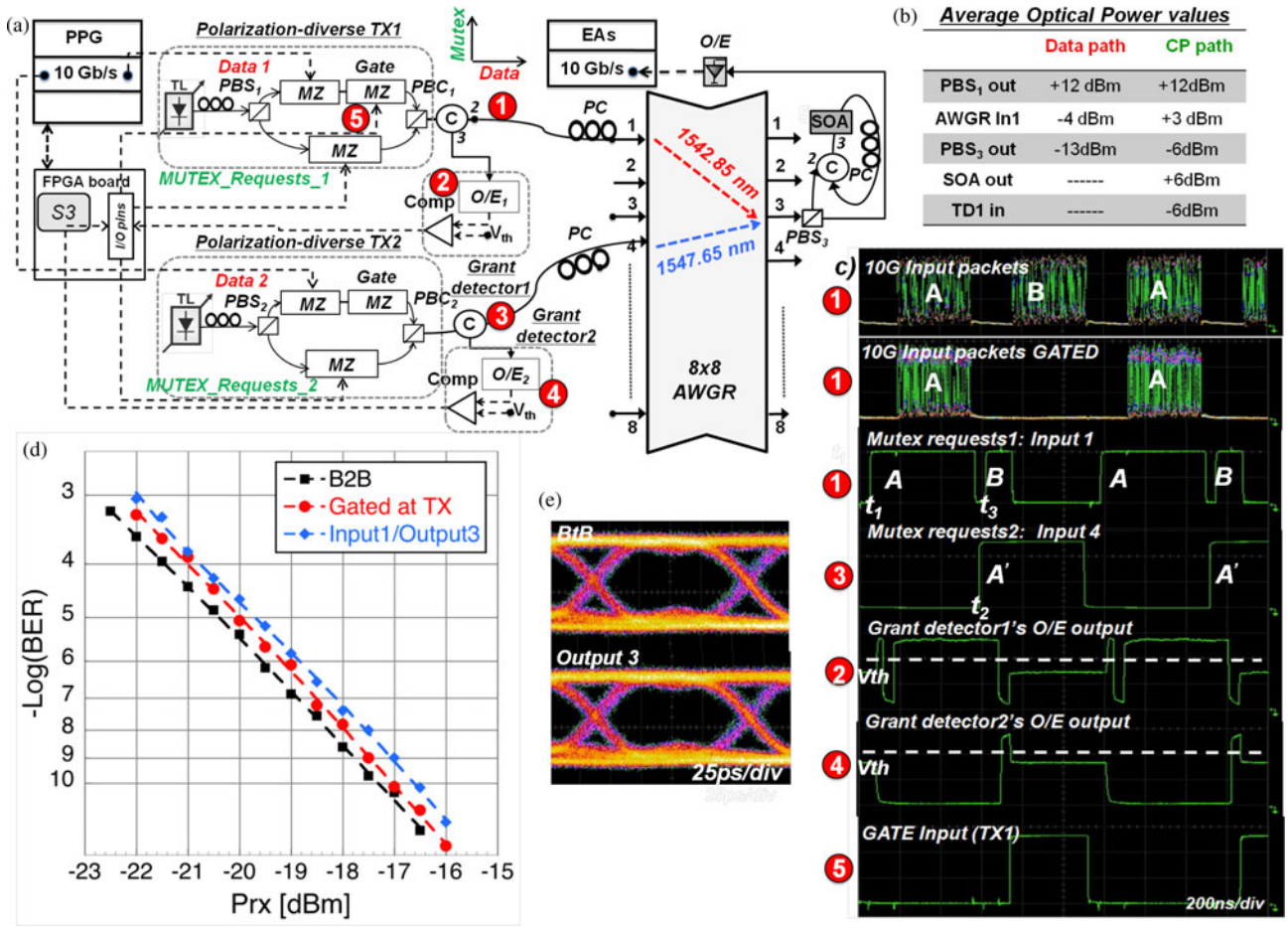

Fig. 4. (a) Experimental testbed: PC: polarization controller; PBS: polarization beam splitter; PBC: polarization beam combiner; MZ: Mach Zender modulator. (b) Optical power value for the testbed. (c) Measure traces. (d) BER measurements. (e) Eye diagrams.

the feedback loop between GD1 and the FPGA has been opened to measure the O/E converter output trace.

Fig. 4(d) shows BER measurements for the data packets at AWGR output 3, with one of every two packets coming from Node1 blocked by the RSOA-based all-optical mutex. The power penalty, compared to the back-to-back (BtB) curve (black squares) is negligible. These results demonstrate that the mutex technique works properly, granting the transmission of A packets only upon successful request, while transmission of B packets are always denied. These results also demonstrate that the coherent crosstalk penalty caused by the mutex requests on the related packets under transmission is not a serious problem for the proposed implementation. In fact, the polarization extinction ratio of the PBS (≈30 dB) and the power values used in the experiment (−13 and −6 dBm are the power values at the PBS3 outputs for data and control plane requests, respectively) guarantee a signal to coherent crosstalk ratio ≈25 dB [14].

As explained earlier, the RSOA-based OMX exploits the saturation effect in SOAs. Therefore, the technique is subject to the wavelength dependence of the RSOA gain, which can pose a higher and lower bound to the wavelength operating range of the technique, under the assumption that $V_{\rm th}$ in the GD is kept constant, as in this experiment. In practice, the technique can work over a wider wavelength range, which can be considered approximately equal to the 1-dB bandwidth of the SOA used in this experiment, i.e., ≈40 nm.

More experimental details and analysis about the wavelength operating range crosstalk impairments in this system and the minimum interval between two successive requests that guarantees the earliest request are granted, and can be found in "submitted for publication" [15].

IV. Performance Analysis of RSOA-Based Contention Resolution

In the previous section, we showed that RSOA-based contention resolution works on principle. In this section, we will evaluate the performance of this scheme and compare it to the best-known alternative today, namely, the distributed loopback-buffer-based scheme [7].

A. Simulation Framework and Parameters

We developed a cycle-accurate architecture level simulator that models Fig. 3(a). The simulator has a wide range of user selectable parameters such as: the number of nodes in the network N, the number of wavelengths per output port k, the data rate, distance between the switch and nodes, and tuning time of the TLs in the wavelength convertors. The parameters assumed for

the simulations were $N = 64$, $k = 4$, 10 Gbps data rate, 10-m switch distance and 8 ns for the tuning time [16]. Note that in order to implement the proposed architecture with $k = 4$, it is necessary to use four RSOAs at each AWGR output port. The four RSOAs connect to each AWGR output port through a 1:4 optical demux, as explained in [2].

The packet size was assumed to be 256B (corresponding to a cache line), and a Bernoulli distribution was used to determine when a packet should be injected. In Section V, we analyze the impact of varying the packet size.

The baseline architecture against which we compare is the distributed loopback buffer (DLB) described in [7]. As noted earlier, the DLB implements contention resolution by diverting all but one of the contending packets to an electrical buffer and then retransmitting them without contention subsequently when the target port is free. In [7], we evaluated different buffering schemes in terms of power, performance, and component cost (such as number of additional TXs, RXs, Buffers, etc.). Our analysis showed that the DLB is the best in terms of performance albeit at a very high component cost. Since we are interested in benchmarking the performance of the RSOA-based scheme, we decided to compare it against the DLB-based implementation. We also compared the performance of RSOA-based scheme to a state-of-the-art electrical switch based on the flattened butterfly (FBF). The FBF was shown by Google researchers [17], [18] to be more appropriate for data centers, as it can be made energy proportional.

B. Workloads

Two workloads were chosen for the evaluation of the RSOA-based switch—synthetic traffic and the GUPS benchmark. Uniform random (packet destination is chosen at random among the N nodes using a uniform distribution) and hot-spot (all packets are destined to the $N/2$ node) traffic patterns were used for the synthetic workloads. The hot-spot pattern is parameterized to simulate high-contention scenarios in a data center. The GUPS benchmark is of particular interest in high-performance computation and typical of in-memory database applications that implements *transactional* nature of 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 benchmark assumed a 64-bit memory address space and updates were performed on 64-bit data values. Since the system being modeled is a distributed system, it was assumed that 1024 outstanding memory requests per node were allowed.

As described in the previous sections, both latency and throughput are important in a data center; therefore, we show the performance in terms of both throughput and latency in the results.

C. Retransmission Function

As noted in Section II-C, RETRANSMIT(*Ui,Ti*) determines when (*Ti*) packet *Ui* should be retransmitted. We chose to use an exponential backoff algorithm for the RETRANSMIT function used in this RSOA-based switch analysis. The RETRANSMIT function randomly chooses a backoff up to the maximum window size, and then the maximum window size is doubled for next call. Upon successful acquisition of the mutex, the maximum window size is reset to the minimum "default" size. The exponential backoff was chosen because of its use in many channel access algorithms, such as is done in Carrier Sense Multiple Access (CSMA) networks. While the exponential backoff algorithm may be an obvious choice for this form collision avoidance, several potential optimizations are discussed in Section V-C.

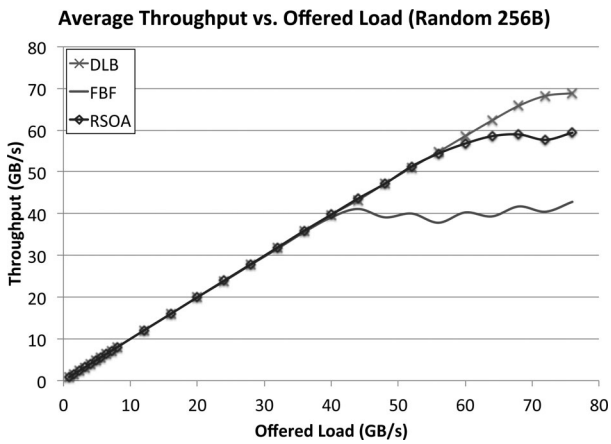

Fig. 5. DLB, FBF, and RSOA network throughput versus offered load for 256B packets on uniform random traffic.

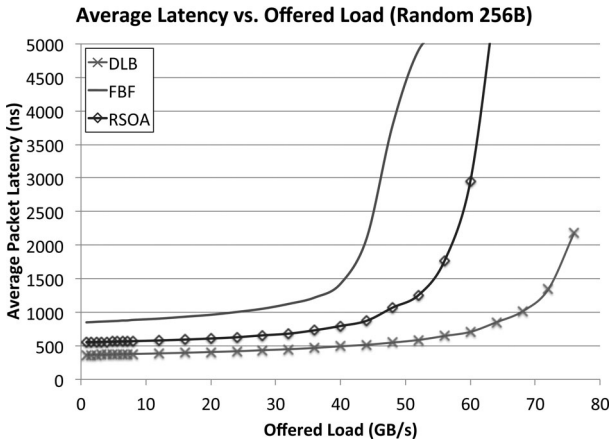

Fig. 6. DLB, FBF, and RSOA average packet latency versus offered load for 256B packets on uniform random traffic.

D. Results

The throughput and latency results of the synthetic traffic patterns are shown in Figs. 5–8. Fig. 5 shows the throughput in GB/s as a function of offered load in GB/s for the uniform random traffic. The RSOA-based architecture performs quite well compared to the FBF, saturating at ~60 GB/s or capable of ~150% of the load of the FBF. When compared to the DLB the RSOA comes within 17% of the DLB. Fig. 6 shows the average packet latency in nanoseconds (ns) as a function of offered load in GB/s. Again, the performance of RSOA is between that of the FBF and the DLB. The hot-spot throughput results are shown

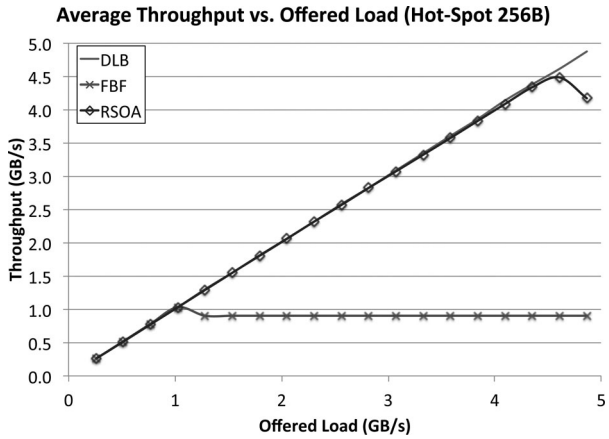

Fig. 7. DLB, FBF, and RSOA network throughput versus offered load for 256B packets on hot-spot traffic.

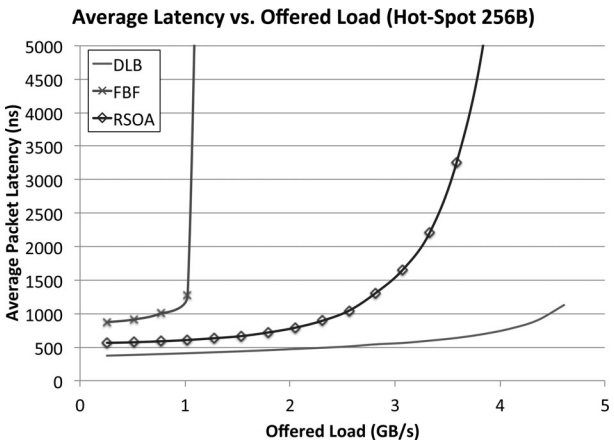

Fig. 8. DLB, FBF, and RSOA average packet latency versus offered load for 256B packets on hot-spot traffic.

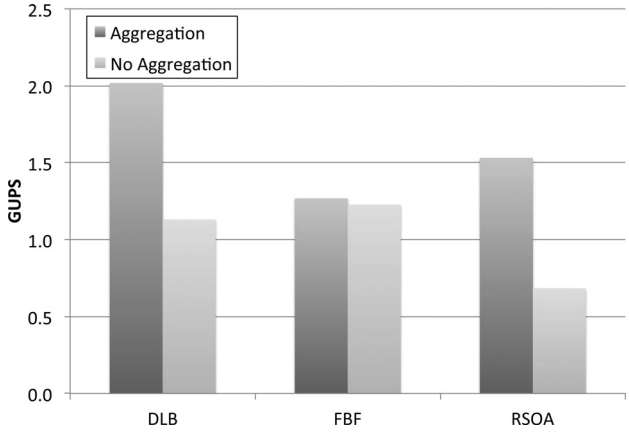

Fig. 9. DLB, FBF, and RSOA GUPS.

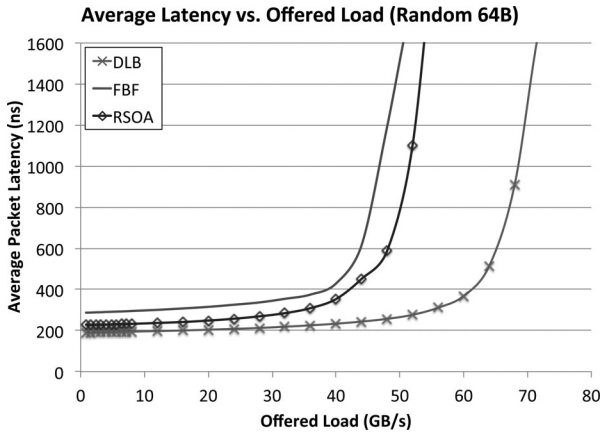

Fig. 10. DLB, FBF, and RSOA average packet latency versus offered load for 64B packets on uniform random traffic.

in Fig. 7. Note that the offered load is limited to 5 GB/s (four wavelengths at 10 Gbps). The RSOA achieves ∼90% of the throughput of the DLB on hot-Spot traffic, which is an impressive result considering that the hot-spot pattern represents the worst-case situation (maximum number of nodes attempting to acquire a single mutex). Notice that the AWGR-based switches greatly outperform the FBF; this is due to the multiple receivers groups of k placed at each AWGR output port.

In fact, DLB requires many complex and power-hungry components [7] (e.g., $(m+1) \times N$ tunable lasers—m is the number of TXs per output queue in the loopback buffer—$m \times N$ high-speed TXs, and N BM-RXs running at 10 Gb/s or higher, $m \times N$ high-speed serializer, N high-speed deserializer, and high-speed access memories, with N being the switch radix). Moreover, the DLB architecture relies on a single centralized control plane, which has to handle up to $(m+1) \times N$ incoming requests and control $(m+1) \times N$ tunable lasers. This will obviously consume a lot of I/O resources in the control plane, which will quickly run out of I/O pins as N increases.

Fig. 9 shows the simulation results for the GUPS runs. The GUPS simulations were run using two configurations, one where all outstanding requests to a single destination were aggregated into a larger packet (Aggregated), and the other were each request/reply must be sent as its own packet (No Aggregation). The AWGR-based networks did not perform as well as the FBF without aggregation due to the small average packet size and the relatively large (8 ns) tuning time [16]. The AWGR-based networks outperformed the FBF when aggregation was allowed, and the DLB actually approaches the theoretical maximum GUPS for the network configuration. These results show that the packet size and tuning time for the AWGR-based networks can be critical to overall system performance, and that data aggregation can be a tool that dramatically improves performance.

V. Discussion

A. Impact of Packet Size on Latency

The observation that the AWGR-based switches did not perform as well as the FBF in the GUPS benchmark when the requests were not aggregated led us to evaluate the impact of packet size on latency. Figs. 10 and 11 show the average packet latency in nanoseconds as a function of offered load in GB/s for the uniform random traffic with 64 B and 16 B packet sizes respectively. The difference in the average packet latency shrinks, as the packet size gets smaller. The FBF actually outperforms

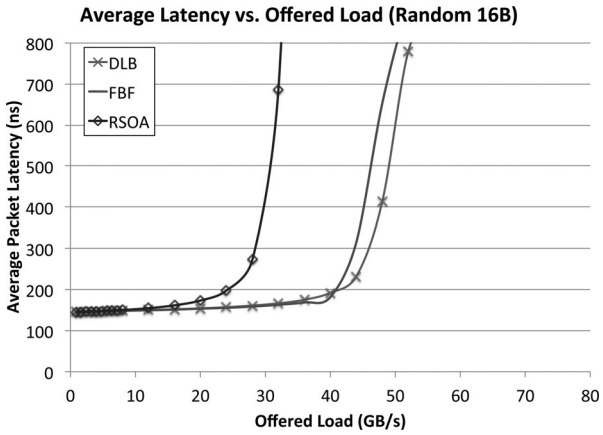

Fig. 11. DLB, FBF, and RSOA average packet latency versus offered load for 16B packets on uniform random traffic.

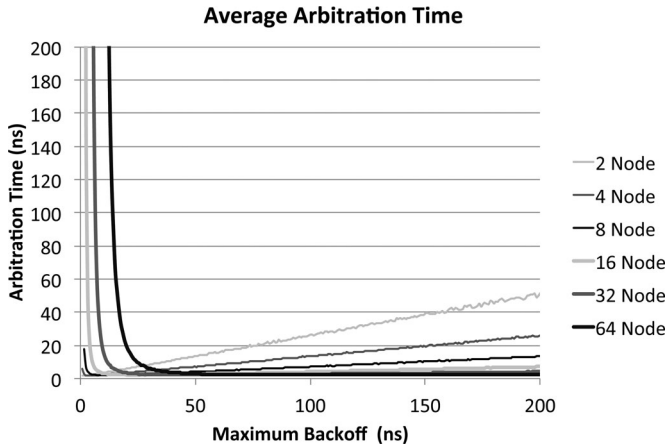

Fig. 12. Average arbitration time versus maximum backoff time.

the RSOA architecture on 16 B packets; this is due to the amount of time spent arbitrating for the output, which is now dominated by the TL tuning time, compared to the amount of time spent transmitting. Note that the choice of placing the ingress and egress buffers at the switch, as shown in Fig. 3(a), was driven by the observation that the minimum mutex request duration is bound by the RTT for the mutex request to reach the RSOA, be reflected, and arriving at the GD. This means that if the distance between the nodes and the switch was small enough (say all the nodes are in a single rack), then the ingress and egress buffers could be placed at the nodes directly. This would eliminate an O/E-E/O conversion and would remove a store and forward of each packet.

B. Impact of Number of Nodes

The performance of the RSOA-based switch is not directly impacted by the number of nodes of the switch, but is impacted by the number of nodes contending for a particular output. In order to determine the impact of the number of contending nodes, we determined the average arbitration time given a maximum backoff time and the number of nodes competing for the mutex. Fig. 12 shows the average arbitration time in nanoseconds as a function of the maximum backoff time. The different lines in Fig. 12 represent the number of nodes competing for the mutex, as the maximum backoff is shortened, and the average arbitration is reduced to a point. When the maximum backoff is too short for the number of competing nodes, the arbitration time dramatically increases due to the birthday paradox [19]. The birthday paradox from probability theory concerns the probability that a pair (from a set of n individuals) share a birthday. This is analogous to the random backoff time where the number of days in the year is replaced with the maximum backoff time (in slots) and the set of individuals is replaced with the number of contending nodes. In this case "sharing a birthday" is equivalent to a request collision.

C. Impact of the Retransmit Function

The RETRANSMIT function will greatly impact the performance of the RSOA-based switch. For example, assume a hot-spot traffic pattern and a RETRANSMIT function that randomly chooses a backoff time within a fixed window that happens to be too small (for example, 10 ns when $N = 64$). As illustrated in Fig. 12, the nodes will very likely keep colliding due to the birthday paradox. This situation will not cause deadlock since eventually progress will be made, but it is possible the progress will be very slow. It was the observation of the potential for a birthday paradox that motivated us to use an exponential backoff in the case of need for retransmission. The maximum size of the backoff window grows exponentially with every collision, but is reset again whenever the mutex is granted. The exponential backoff works well when there are many nodes competing for the single mutex; however, this approach leads to the window growing rapidly whenever a single node already has been granted the mutex (as occurs when a node is sending a long packet).

Clearly, there is balance that needs to be struck between backing off to avoid the birthday paradox and maintaining a smaller window to gain the mutex as soon as the granted node releases it. It is possible to remove the potential for zero nodes being granted the mutex if the requests are properly scheduled into slots. This scheduling would require N slots each of duration equal to the time required to saturate the RSOA. This approach would require tight clock synchronization among all the nodes and would result in an average arbitration delay of $N/2$ slot times. A relaxed version of this approach where multiple nodes would share a slot time is more likely practical and we plan to analyze this approach in the future.

The problem of the distributed nature of the RSOA-based mutex is that a requesting node cannot distinguish between an already granted mutex or a collision resulting in zero nodes receiving a grant. In the former case, the backoff time should be short so that the mutex can be gained as soon as the previous owner releases it, in the later case the backoff time should be increased to reduce the possibility of another collision. If it were possible to differentiate between the situations when there are exactly two requestors and when there are three or more requestors, then it may be possible to use different backoff schemes depending on the situation. For example, one possible scheme would keep the maximum window size fixed if the

request feedback indicates only two requestors (assuming the mutex is owned by a node currently transmitting a packet), and the maximum window size would be doubled if the request feedback indicates three or more requestors (since obviously there are multiple requestors regardless if the mutex is owned or not). We plan to analyze this alternative RETRANSMIT function among others in the future.

VI. Related Work

Much effort has been made to use optical technologies in the network design for computing applications. Webb and Louri propose, SOCNs/SYM-NET [20], [21], a multilevel hierarchical architecture for a large-scale optical crossbar network and a tree-based address distribution subnetwork. Although SOCNs/SYMNET can connect to a large number of nodes, parallel communication is not utilized and the system throughput is limited, since the optical token controlled address broadcasting scheme allows for transmitting messages only serially. Gemini [22] is an optical/electrical dual banyan network. The optical banyan network delivers long messages, while the electrical network transmits control signals and short messages. The drawbacks of Gemini are that the banyan network is a blocking network, and that the optimal scheduling for a large-scale banyan network is complicated.

The OSMOSIS utilizes semiconductor optical amplifiers (SOAs) to realize a synchronous optical crossbar switching fabric with the use of a broadcast-and-select data path combined with both space and WDM [23], [24]. Strictly speaking, the OSMOSIS still uses the store-and-forward mechanism and adopts input and output queue structures, which are commonly used in electrical switches. Although the optical switching fabric allows the OSMOSIS to have two receivers at each output, thus sustaining high input load, the power requirements of the OSMOSIS can be very high because of its broadcast-and-select architecture—signals are delivered to every select unit, even though, only one unit selects the signal. The Data Vortex is a distributed interconnection network architecture [25]–[27] based on deflection routing. To prevent packet drop when contention occurs, the packet is deflected to another output and an access control is adopted to ensure that the network will not carry traffic beyond its capability. In other words, Data Vortex treats the deflection route as temporary network storage. Due to the deflection routing and access control, Data Vortex saturates before the offered load exceeds 50% [27]. In addition, as the number of nodes increases, the end-to-end latency becomes large and nondeterministic.

AWGR-based optical switches and optical routers with packet switching capability have been investigated for a number of years. Previous work [1], [28]–[32] mainly focused on the application of the AWGR in access networks and in telecommunication/IP networks. An AWGR serves as a nonblocking switching fabric in many switch architecture designs. However, the wavelength parallelism on AWGR outputs is not explored in those designs. Because each AWGR output is connected with a fixed wavelength converter to convert the signal to a particular wavelength in order to ensure wavelength consistency on the input and output fibers, the occurrence of the multiple packets is not allowed. Since no practical optical buffer is yet available, the store-and-forwarding scheme, which is commonly used in the electrical switch, cannot be duplicated in the optical domain. The FDL is commonly used to resolve the contention, provide temporary storage, and allow packets that cannot gain the resource to compete for the resource at a later time. Use of the FDL in resolving the contention helps to significantly reduce the dropping probability, but packet loss is still possible and cannot be eliminated. In addition, the FDL cannot provide arbitrary delays, which is more critical in asynchronous switching. The resource may be available, but the delayed packet cannot access it, since the packet is still traveling through the FDL.

The architecture proposed and studied in [2] and [7] is an optical hybrid switch that adopts the AWGR as the switching fabric and utilizes wavelength parallelism to achieve output queuing. The loopback buffer (LB) with N parallel transmitters and N parallel receivers can simultaneously receive contended packets from multiple input ports and transmit packets to multiple output ports when resources are available. However, as discussed in Section I, the LB limits the scalability of the architecture due to the bandwidth requirement, complexity and cost of the loopback buffer itself.

VII. Conclusion

We showed that for optical interconnects to be viable in data centers, both latency and scalability issues have to be addressed. Existing approaches to contention resolution in optical switches such as dropping packets, using fixed fiber delay lines, wavelength conversion, or deflection routing increase latency significantly and hence are unsuitable for data center networking. We argue that detecting contention and resolving it locally (without the need of a global coordination scheme) is key to making optical interconnects scalable in a data center.

We showed that mutual exclusion can be implemented directly in the optical domain by taking advantage of an RSOA, showed that RSOA-based contention resolution can be realized using simple hardware. Furthermore, it has ideal properties such as the ability to notify the sender implicitly when a new requestor emerges for the same destination, something that is extremely difficult to do in an electrical domain. Also, a single RSOA can serve as an N-input mutex where N can be 64 or even higher. This further reduces the delay in acquiring a grant when the number of nodes is large. Finally, an RSOA-based contention resolution scheme does not require any global coordination, which makes it inherently scalable.

However, we showed that there are several critical parameters such as the relative packet size (ratio of the absolute packet size compared to the TL tuning time), the RETRANSMIT function. Our future work will involve optimizing these parameters and the overall performance of the network on more realistic workloads.

References

[1] 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.

[2] X. Ye, P. Mejia, Y. Yin, R. Proietti, S. J. B. Yoo, and V. Akella, "DOS—A scalable optical switch for datacenters," in *Proc. ACM/IEEE Symp. Archit. Netw. Commun. Syst.*, Oct. 2010, pp. 1–12.

[3] T. Benson, A. Anand, A. Akella, and M. Zhang, "Understanding data center traffic characteristics," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 40, no. 1, pp. 92–99, 2010.

[4] M. Al-Fares, A. Loukissas, and A. Vahdat, "A scalable, commodity data center network architecture," *SIGCOMM Comput. Commun. Rev.*, vol. 38, no. 4, pp. 63–74, 2008.

[5] A. V. Krishnamoorthy, R. Ho, X. Zheng, H. Schwetman, J. Lexau, P. Koka, G. L. Li, I. Shubin, and J. E. Cunningham, "Computer systems based on silicon photonic interconnects," *Proc. IEEE*, vol. 97, no. 7, pp. 1337–1361, Jul. 2009.

[6] D. Miller, "Device requirements for optical interconnects to silicon chips," *Proc. IEEE*, vol. 97, no. 7, pp. 1166–1185, Jul. 2009.

[7] X. Ye, R. Proietti, Y. Yin, S. J. B. Yoo, and V. Akella, "Buffering and flow control in optical switches for high performance computing," *IEEE/OSA J. Opt. Commun. Netw.*, vol. 3, no. 8, pp. A59–A72, Aug. 2011.

[8] R. Proietti, Y. Yin, R. Yu, X. Ye, C. Nitta, V. Akella, and S. J. B. Yoo, "All-optical physical layer NACK in AWGR-based optical interconnects," *IEEE Photonics Technol. Lett.*, no. 99, pp. 410–412, 2012.

[9] F. Abel, C. Minkenberg, I. Iliadis, T. Engbersen, M. Gusat, F. Gramsamer, and R. P. Luijten, "Design issues in next-generation merchant switch fabrics," *IEEE/ACM Trans. Netw.*, vol. 15, no. 6, pp. 1603–1615, Dec. 2007.

[10] H. C. Shin, J. S. Lee, H. I. Kim, I. K. Yun, S. W. Kim, and S. T. Hwang, "Reflective semiconductor optical amplifier," Google Patents 8 149 503, 2006.

[11] E. Dijkstra, "Solution of a problem in concurrent programming control," *Commun. ACM*, vol. 8, no. 9, pp. 569–579, 1965.

[12] M. J. Connelly, *Semiconductor Optical Amplifiers*. Berlin, Germany: Springer, 2002.

[13] C. K. Chan, K. L. Sherman, and M. Zirngibl, "A fast 100-channel wavelength-tunable transmitter for optical packet switching," *IEEE Photonics Technol. Lett.*, vol. 13, no. 7, pp. 729–731, Jul. 2001.

[14] H. Kim and S. Chandrasekhar, "Dependence of coherent crosstalk penalty on the OSNR of the signal," in *Proc. Opt. Fiber Commun. Conf.*, 2000, pp. 359–361.

[15] R. Proietti, Y. Yin, R. Yu, C. Nitta, V. Akella, and S. J. B. Yoo, "Fully-distributed control plane by all-optical token technique in AWGR-based optical interconnects," *J. Lightw. Technol.*, submitted for publication Jun. 19, 2012.

[16] S. Matsuo and T. Segawa, "Microring-resonator-based widely tunable lasers," *IEEE J. Sel. Top. Quantum Electron.*, vol. 15, no. 3, pp. 545–554, May/Jun. 2009.

[17] D. Abts, M. R. Marty, P. M. Wells, P. Klausler and H. Liu, "Energy proportional datacenter networks," *ACM SIGARCH Comput. Archit. News*, vol. 38, no. 3, pp. 338–347, 2010.

[18] J. Kim, W. J. Dally, and D. Abts, "Flattened butterfly: A cost-efficient topology for high-radix networks," *Comput. Archit. News*, vol. 35, no. 2, pp. 126–137, May 2007.

[19] K. Suzuki, D. Tonien, K. Kurosawa, and K. Toyota, "Birthday paradox for multi-collisions," in *Information Security and Cryptology* (Lecture Notes in Computer Science Series). Berlin, Germany: Springer, 2006, pp. 29–40.

[20] B. Webb and A. Louri, "A class of highly scalable optical crossbar-connected interconnection networks (SOCNs) for parallel computing systems," *IEEE Trans. Parallel Distrib. Syst.*, vol. 11, no. 5, pp. 444–458, May 2000.

[21] A. Louri and A. Kodi, "An optical interconnection network and a modified snooping protocol for the design of large-scale symmetric multiprocessors (SMPs)," *IEEE Trans. Parallel Distrib. Syst.*, vol. 15, no. 11, pp. 1093–1104, Dec. 2004.

[22] R. Chamberlain, M. Franklin, and C. Baw, "Gemini: An optical interconnection network for parallel processing," *IEEE Trans. Parallel Distrib. Syst.*, vol. 13, no. 10, pp. 1038–1055, Oct. 2002.

[23] C. Minkenberg, F. Abel, P. Muller, R. Krishnamurthy, M. Gusat, P. Dill, I. Iliadis, R. Luijten, R. R. Hemenway, R. Grzybowski, and E. Schiattarella, "Designing a crossbar scheduler for HPC applications," *IEEE Micro*, vol. 26, no. 3, pp. 58–71, May/Jun. 2006.

[24] 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.

[25] C. Hawkins, B. A. Small, D. S. Wills, and K. Bergman, "The data vortex, an all optical path multicomputer interconnection network," *IEEE Trans. Parallel Distrib. Syst.*, vol. 18, no. 3, pp. 409–420, Mar. 2007.

[26] O. Liboiron-Ladouceur, A. Shacham, B. A. Small, B. G. Lee, H. Wang, C. P. Lai, A. Biberman, and K. Bergman, "The data vortex optical packet switched interconnection network," *J. Lightw. Technol.*, vol. 26, no. 13, pp. 1777–1789, Jul. 2008.

[27] K. Bergman, D. Keezer, and S. Wills. (2010). Design, demonstration and evaluation of an all optical processor memory-interconnection network for petaflop supercomputing," *ACS Interconnects Workshop*, Available: http://lightwave.ee.columbia.edu/?s=research&p=high-performance_computing_systems#dv

[28] H. Yang and S. J. B. Yoo, "Combined input and output all-optical variable buffered switch architecture for future optical routers," *IEEE Photonics Technol. Lett.*, vol. 17, no. 6, pp. 1292–1294, Jun. 2005.

[29] S. Bregni, A. Pattavina, and G. Vegetti, "Architectures and performance of AWG-based optical switching nodes for IP networks," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 7, pp. 1113–1121, Sep. 2003.

[30] M. Maier, M. Scheutzow, and M. Reisslein, "The arrayed-waveguide grating-based single-hop WDM network: An architecture for efficient multicasting," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 9, pp. 1414–1432, Nov. 2003.

[31] W. D. Zhong and R. S. Tucker, "Wavelength routing-based photonic packet buffers and their applications in photonic packet switching systems," *J. Lightw. Technol.*, vol. 16, no. 10, pp. 1737–1745, 1998.

[32] D. Banerjee, J. Frank, and B. Mukherjee, "Passive optical network architecture based on waveguide grating routers," *IEEE J. Sel. Areas Commun.*, vol. 16, no. 7, pp. 1040–1050, Sep. 1998.

Roberto Proietti (M'11) received the M.S. degree in telecommunications engineering from the University of Pisa, Italy, in 2004, and the Ph.D. degree in electrical engineering from Scuola Superiore Sant'Anna, Pisa, Italy, in 2009.

He is currently a Postdoctoral Researcher with the Next Generation Networking Systems Laboratory, University of California, Davis. His research interests include optical switching technologies and architectures for supercomputing and data center applications, high-spectrum efficiency coherent transmission systems, and elastic optical networking.

Christopher J. Nitta (S'06–M'11) received the Ph.D. degree in computer science from the University of California, Davis, in 2011.

He is a Postdoctoral Researcher and a Lecturer at the University of California, Davis. His research interests include network-on-chip technologies, embedded system and RTOS design, and hybrid electric vehicle control.

Yawei Yin (M'11) received the B.S. degree in applied physics from the National University of Defense Technology, Changsha, China, in 2004, and the Ph.D. degree in electrical engineering from the Beijing University of Posts and Telecommunications, Beijing, China, in 2009.

He is currently a Postdoctoral Researcher with the Next Generation Networking Systems Laboratory, University of California, Davis, where he is involved in the low-latency, scalable all-optical switches for peta-scale computing, as well as flexible bandwidth elastic optical networking algorithms, simulations, and experiments.

Runxiang Yu (S'10) received the B.Eng. degree in electrical engineering from Peking University, Beijing, China, in 2007. He is currently working toward the Ph.D. degree in the Department of Electrical and Computer Engineering, University of California, Davis.

His research interests includes advanced switching technologies and system level integration for next-generation optical networks.

S. J. B. Yoo (S'82–M'84–SM'97–F'07) received the B.S. degree in electrical engineering with distinction, the M.S. degree in electrical engineering, and the Ph.D. degree in electrical engineering with minor in physics, all from Stanford University, California, in 1984, 1986, and 1991, respectively.

He is currently a Professor of electrical engineering in the University of California at Davis (UC Davis). His research at UC Davis includes optical switching devices, systems, and networking technologies for the future computing and communications. Prior to joining UC Davis in 1999, he was a Senior Research Scientist at Bellcore, where his technical effort included optical networking research and systems integration. He participated in ATD/MONET testbed integration and a number of standardization activities including GR-2918-CORE, GR-2918-ILR, GR-1377-CORE, and GR-1377-ILR on dense WDM and OC-192 systems.

Dr. Yoo is a Fellow of the Optical Society of America, and is a recipient of the DARPA Award for Sustained Excellence in 1997, the Bellcore CEO Award in 1998, the Outstanding Mid-Career Research Award, UC Davis, in 2004, and the Outstanding Senior Research Award, UC Davis, 2011.

Venkatesh Akella received the Ph.D. in computer science from the University of Utah, Salt Lake City, in 1992.

He has been a Professor of electrical and computer engineering at the University of California, Davis, since 1992. His current research interest includes various aspects of embedded systems and computer architecture with special emphasis on embedded software, hardware/software codesign, and low power system design.

Dr. Akella is member of the Association for Computing Machinery and received the National Science Foundation CAREER award.