An All-Optical Token Technique Enabling a Fully-Distributed Control Plane in AWGR-Based Optical Interconnects

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

***Abstract*—This paper proposes and experimentally demonstrates a fully-distributed All-Optical TOKEN (AO-TOKEN) contention resolution technique for AWGR-based optical interconnects. The AO-TOKEN technique is implemented by exploiting the saturation effect in SOAs placed at the AWGR outputs. A polarization-diversity scheme allows the data and control planes to share the same physical link. The AO-TOKEN is more scalable than alternative electrical/optical solutions since it eliminates the need for a centralized electrical control plane. Our experimental results show that the technique can work over a wavelength-range of $\approx$ 23 nm using off-the-shelf components. We also successfully demonstrate all-optical contention resolution, packet transmission, and switching with error-free operation at 10 Gb/s.**

***Index Terms*—All-optical, arrayed waveguide gratings, contention resolution, data centers, optical interconnects.**

I. Introduction

THE growing demand for cloud-based services and high-performance computing has spurred interest in the architecture of data centers. The network inside a data center is quite different from wide-area and local-area networks that have been the subject of intensive research over the past few decades. Datacenter 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 [1]. Both the network performance (in terms of latency) and the power consumption have become critical in data centers [2]. Fig. 1 shows an example architecture of a 64,000 node data center with a two level hierarchical network [3]. Each basic computational unit (blade or server) is connected inside a rack, with multiple racks organized as clusters, or pods [2], [4]. Each cluster connects 1000 servers based on a hierarchical fat-tree topology. These clusters then connect to the core-level switches, which represent the high-end in both port density and bandwidth (typically 32–128 10 GigE ports). The entire architecture can then be considered as a fat-tree, with the links between pods and core level switches requiring the higher bit rate. It is inevitable to use a large number of expensive power-hungry core switches to guarantee connectivity among the different clusters because of the poor scalability of high-bandwidth single-stage electrical switches [5].

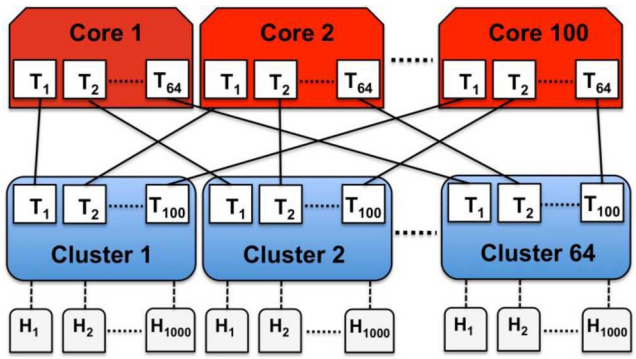

Fig. 1. Example of data center architecture. T: transponder; H: host; Core 1, Core 2,. . . Core 100: core switches; Cluster 1, Cluster 2,. . . Cluster 64: cluster switches.

Single-stage, high port count, and high-data rate optical switches could help to flatten this topology, replacing the electronic switches both at the core and cluster level. In fact, unlike an electronic switch, an optical switch can support 10 and 40 Gb/s easily and (as shown in [3], [6], [7]) offer much higher throughput and lower latency for high traffic loads. Optical switches can also use Wavelength Division Multiplexing (WDM) to create additional parallel data paths, which provides yet another opportunity to increase the overall data center performance.

Over the past decade, several research groups have proposed optical switch fabrics based on Arrayed Waveguide Grating Router (AWGR) and/or Semiconductor Optical Amplifiers (SOA) [6]–[11]. However, the inability to buffer light creates a major problem since solutions that rely on optical delay lines or deflection routing [8] cannot guarantee arbitrary delays or prevent packets from being dropped. It is for these reasons that previously proposed optical switch architectures typically rely on electrical input/output queues [6], [10] and electrical loopback buffers [7], [12].

In [6], [10], the architectures make use of a centralized control plane and input/output queues at the switch. The nodes, at a certain distance from the switch, send the packets to the switch inputs. The packets are stored in the input queues, switched, stored again in the output queues, and finally sent to the destination

Manuscript received June 19, 2012; revised October 15, 2012, November 26, 2012; accepted November 28, 2012. Date of publication December 05, 2012; date of current version January 02, 2013. This work was supported in part by the Department of Defense (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).

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/JLT.2012.2231664

node. These multiple store and forward operations will cause additional latency. Further, the architectures in [7], [12] switch the packets on the fly (employs no additional input/output buffers at the switch) while buffering only the packets that experience contention. The architectures in [7], [12] reduce the average latency compared to the architectures in [6], [10]. Both cases, however, utilize centralized control planes requiring a large number of wires and connections from the controller to the buffers and the switch components. As a result, the controller will quickly exhaust the maximum number of I/O pins of currently available integrated chips [5]. This will clearly limit the scalability of the switch. To partly overcome this problem, in [11], [13] we proposed an all-optical technique that allows removal of the loopback buffer in the LIONS architecture [12], without affecting the network-level performance. However, the control plane in the architecture still remained centralized.

Wavelength routing offered by AWGR technology can offer all-to-all communication without contention among N nodes, assuming that each node has N transmitters and N receivers. However, this requires N^2 transmitters and N^2 receivers, which can be prohibitive in terms of cost and power consumption when N is high. Research reported in [7] shows that with one transmitter and few receivers per node, it is still possible to improve significantly the switch performance compared to the electrical counterpart. However, even in this case contention is still possible and a control plane that handles arbitration is necessary.

Centralized electrical control planes are another major reason for limited scalability for not only optical switches, but for any switch architecture, since they limit port count and increase latency. In fact, as explained above, the maximum number of I/O resources of currently available integrated chips [5] 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, a distributed control plane is highly desirable. References [8], [14] propose two architectures with distributed control plane. However, to the best of our knowledge, no other AWGR-based architectures with distributed control plane have been proposed.

Reference [15] introduces for the first time the proposed technique and report some preliminary experimental results. In [16] we studied the networking performance of the proposed architecture. This paper focuses on a detailed physical layer demonstration and analysis. The proposed technique, named All-Optical TOKEN, exploits the saturation effect in SOAs [17] placed at the AWGR outputs. A polarization-diversity scheme allows the control and data planes to share the same physical links. The mutual exclusion function [18] implemented here exploits the SOA-saturation and the unique wavelength routing property offered by AWGR to enable the realization of a fully distributed all-optical control plane. In this architecture, there are no wavelength converters necessary in contrast to the previous LIONS architecture. This allows the possibility of using advanced modulation formats with high spectral efficiency if it is necessary to achieve line-rates beyond what is achievable with standard ON-OFF-Keying modulation within the passband of the AWGRs. Compared to [11], the main advantage preserved here is the use of wavelength routing properties in AWGR, while the new additional advantages are the removal of the complicated electrical loopback buffer in [12] and the new introduction of the distributed all-optical control plane.

The remainder of this paper is organized as follows. Section II describes the working principle of the proposed technique and the related interconnects architecture. Section III describes the proof-of-concept experimental demonstration of the proposed all-optical contention resolution scheme. In particular, Section III.A discusses the experimental setup, while Section III.B reports the experimental results. Section III.C, III.D, III.E, III.F discusses how the port count of AO-TOKEN architecture is affected by the polarization crosstalk, SOA wavelength operating range, SOA optical noise, and four wave mixing effect. Finally, Section IV concludes this paper and discusses possible variations and improvements of the proposed scheme and further studies.

II. All-Optical Token: Architecture and Working Principle

The All-Optical TOKEN (AO-TOKEN) interconnect architecture is shown in Fig. 2(a). At the core of the AO-TOKEN architecture is an $N \times N$ AWGR. Each input port is connected to a line-card (*Line-card_i TX*), which receives packets from a Host (H_i) and buffers them in an Ingress-Queue (I-Q). The link between each H_i and *Line-card_i* can be either electrical or optical. The packets are then transmitted and switched in the optical domain. Each transmitter includes a fast Tunable Laser Diode (TLD) [19] and makes use of a polarization-diversified scheme to transmit the token messages and data packets on two orthogonal polarizations. Each AWGR output is then connected to a Polarization Beam Splitter (PBS) to separate the token and data. One PBS output connects to a Reflective SOA (RSOA) [20], which is the key component in this AO-TOKEN based contention resolution technique. The other PBS output (data output) connects to a line-card (*Line-card_i RX*), which buffers the received packets in an Egress-Queue (E-Q) and transmits them to the final destination. A packet transmission is initiated with a token request, which is directed to the desired AWGR output port. Upon the reception of the token at the Token Detector (TD), the packet transmission begins on a polarization orthogonal to the token request. An optical Circulator (C) placed at each input port is used to extract the counter-propagating token messages. The timing diagram in Fig. 2(b) illustrates the concept of the AO-TOKEN technique. Assume that *line-card_1* wants to send a packet to output N. *Line-card_1* tunes its TLD to λ_{1N} (the wavelength to reach output N from input 1 according to the AWGR routing table) and generates a token request A, which reaches output N at time instant t_1. The RSOA at output N reflects the signal extracted by the PBS, which reaches the *line-card_1 TD* with optical power P_{TO1}. The O/E converter in the TD generates an electrical signal with $Vp = V_{TO1}$ which is above the voltage threshold *Vth*. This condition means that the token for output N is available and then the transmission of packet A can begin. The same situation arises when *line-card2* generates a token request A' arriving at output N at time instant t_2. In this case the electrical signal generated by the O/E converter in the TD is above the threshold *Vth*, activating the transmission of the related packet A′ (not shown for the sake

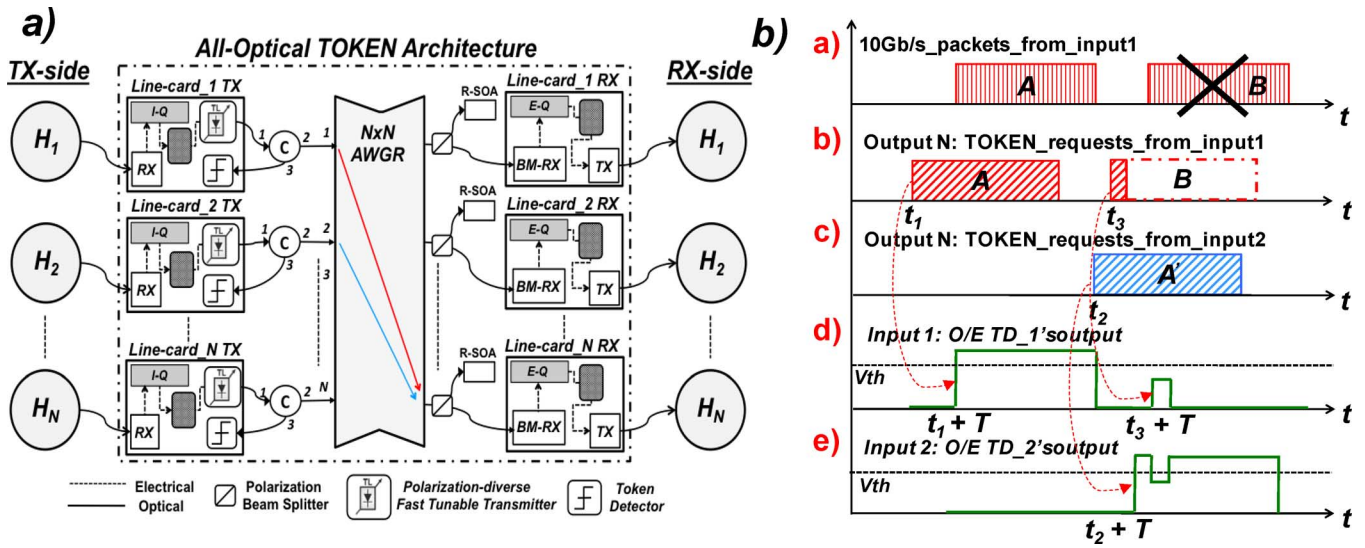

Fig. 2. (a) RSOA-based Optical Token architecture. H—Host; I-Q—Ingress Queue; E-Q: Egress Queue; C—Optical Circulator. (b) Timing diagram that illustrates the working principle of the Optical Token contention resolution scheme. (a) Transmitted packets from input 1 (node1); (b) Token requests from input 1 (node1) at output N; (c) Token request from input 2 at output N; (d) output of O/E converter in the TD at input 1; (e) output of O/E converter in the TD at input 2; T is the time that it takes for the token requests to reach the token detectors from output N.

of brevity). Note that the token request signal stays active for the duration of the entire packet transmission in order to keep the token busy (RSOA latched). This prevents collision in the case of another line-card attempting to send a packet to the same output. The reader should take note of the behavior at time instant t_3, when the transmission of packet A' has not yet completed, but another token request arrives from *line-card_ 1*. The RSOA at output N, which is already saturated with the token request signal A' at λ_{2N}, amplifies and reflects back the new token request B at λ_{1N}, which reaches the *line-card_1 TD* with optical power P_{TO3}. As described above, the O/E converter in the TD generates an electrical signal with $Vp = V_{TO3}$. Assume that the RSOA was strongly saturated by the token request A' at $t = t_2$. Due to the gain saturation effect in the RSOA, the optical power P_{TO3} will then be $\sim (Psat - 3)$ dBm, with *Psat* the output saturation power of the RSOA. Clearly, the signal generated by the O/E converter in the TD will be $V_{TO3} = V_{TO1}/2$. Simply setting *Vth* between V_{TO1} and $V_{TO1/2}$ makes it possible for *line-card_1* to recognize that the token for output *1* is not available and that it must retry with a random backup time to gain the grant for token request B and prevent, at the same time, resource starvation [16].

To generalize, all the token requests for output N occurring during the transmission of packet A' will be denied. However, multiple token requests for different outputs can be satisfied simultaneously since there is a RSOA for each output. Note that, if two requests arrive at approximately the same time at the RSOA (details appear in Section III), both the requestors receive approximately *Psat/2* reflected power and hence the detectors at neither node triggers, which corresponds to a situation that neither token request has been granted (this is unlikely to happen in asynchronous architectures). This condition is still sufficient to guarantee mutual exclusion of the data plane output. The AO-TOKEN technique does not require a centralized electrical control plane and the acquisition of the token is handled all-optically. Note that due to the gain saturation effect in the RSOA, the token request B also causes a voltage drop in the O/E converter output of the TD of *line-card_2*. This notifies the transmitting node that at least one other node tried to transmit to the current output. This information can improve the fairness of the protocol while reducing the overhead of token requests for multiple packets destined to the same output port. Detailed network performance analysis of the AO-TOKEN interconnect architecture is beyond the scope of this paper, but can be found in [16].

III. Experimental Demonstration

A. Experiment Setup

The testbed used for the proof-of-concept experimental demonstration of the AO-TOKEN technique is illustrated in Fig. 3. Two polarization-diversity TXs are connected to input ports 1 and 2 of a 200 GHz-spacing 8 × 8 AWGR with uniform insertion loss of 8 dB and cyclic frequency characteristic (ULCF [21]). Polarization Controllers (PCs) at the AWGR inputs align the signal polarization with the PBS at the AWGR output. Alternatively, all Polarization Maintaining (PM) components may be used. Each TX includes an external-cavity Tunable Laser (TL) with line-width $\approx$ 2 MHz, a PBS and a Polarization Beam Combiner (PBC) to multiplex the data and token path in the polarization domain. The token arm of the TX includes a 10 GHz Mach Zehnder (MZ) modulator. Two 10 GHz MZs are also used in the data arm as the data modulator and gate. The gate is controlled by an FPGA (running at 155 MHz), and it stays open unless the token request is not granted. Note that in an actual system no optical gate would be necessary since packets would be stored in the line-cards' TX buffers and transmitted only upon the token acquisition. An Erbium-Doped Fiber Amplifier (EDFA) with 21 dBm output power (not shown for the sake of brevity) is placed right after

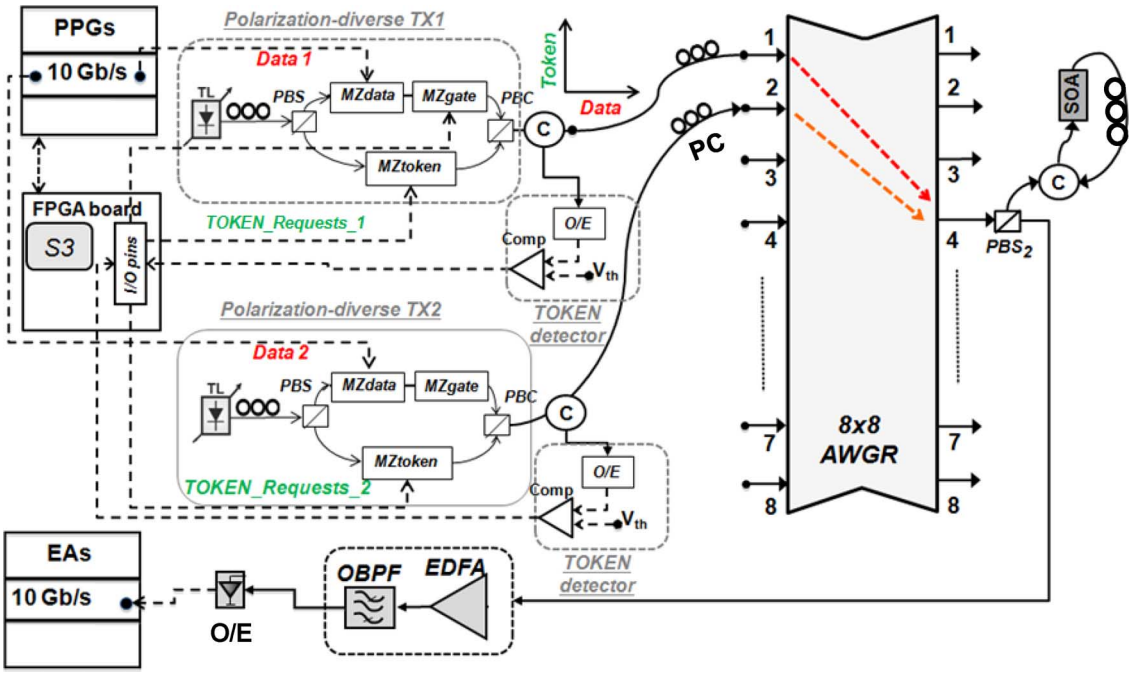

Fig. 3. Experimental setup: PBS—polarization beam splitter; PC—polarization controller; PBC- polarization beam combiner; MZ—Mach Zehnder modulator; AWGR—arrayed waveguide grating router; PPG—pulse pattern generator; C—optical circulator, SOA—semiconductor optical amplifier; EDFA—erbium-doped fiber amplifier; OBPF—optical band-pass filter.

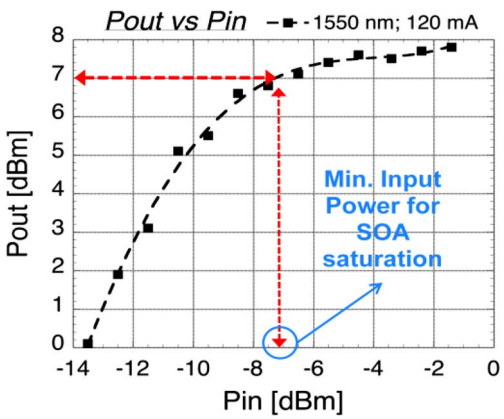

Fig. 4. Measured Pout vs Pin characteristic of the SOA used in the experiment setup of Fig. 3.

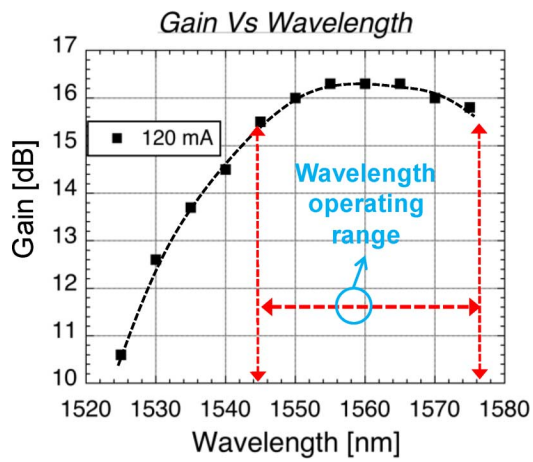

Fig. 5. Measured Gain vs Wavelength characteristic of the SOA used in the experiment setup of Fig. 3.

each TL to compensate for the loss of the MZ in the token path and for the MZ acting as a gate in the data path. Notably, these two MZs would not be necessary in an actual implementation. The output power requirement for the tunable lasers in the future implementations of the proposed LIONS in absence of EDFA and MZ modulator gate would be $+7$ dBm. This is calculated considering -6 dBm saturation SOA input power and the following insertion loss values: 0.5 dB for circulators, PBS and PBC; 8 dB insertion loss for AWGR; 3 dB power splitting ratio at the PBS. The 6 dB insertion loss for MZ in the token arm was not included in this calculation since fast tunable lasers with blanking capability will be utilized.

The FPGA also generates the token requests coming from inputs 1 and 2 (see Fig. 6(c) and (d)) to recreate a similar condition to that explained in the Fig. 2(b) timing diagram, while a sequence of 10 Gb/s 406.9 ns-long packets is generated with a pattern generator with each packet containing a portion of $2^{31} - 1$ PRBS (see Fig. 6(a)). Note that for demonstration purposes, TX2 at input 2 generates only the contending token requests.

A PBS (PBS2) is placed at AWGR output 4 to separate the token from the data. The PBS extracts the token requests, which enter in a RSOA implemented here with an optical circulator and an SOA, as shown in Fig. 3. The SOA used here is an Alcatel 1901 with small signal gain G of 25 dB, saturation output power P_{sat} of $+6$ dBm, polarization dependent gain PDG of 1 dB and 3-dB bandwidth $BW_{3\text{ dB}}$ of 90 nm (see Figs. 4 and 5). The PC at the SOA output maximizes the optical power going back through the PBS_2 and AWGR, and reaching the TD. The second PBS_2 output connects to an O/E converter for BER mea-

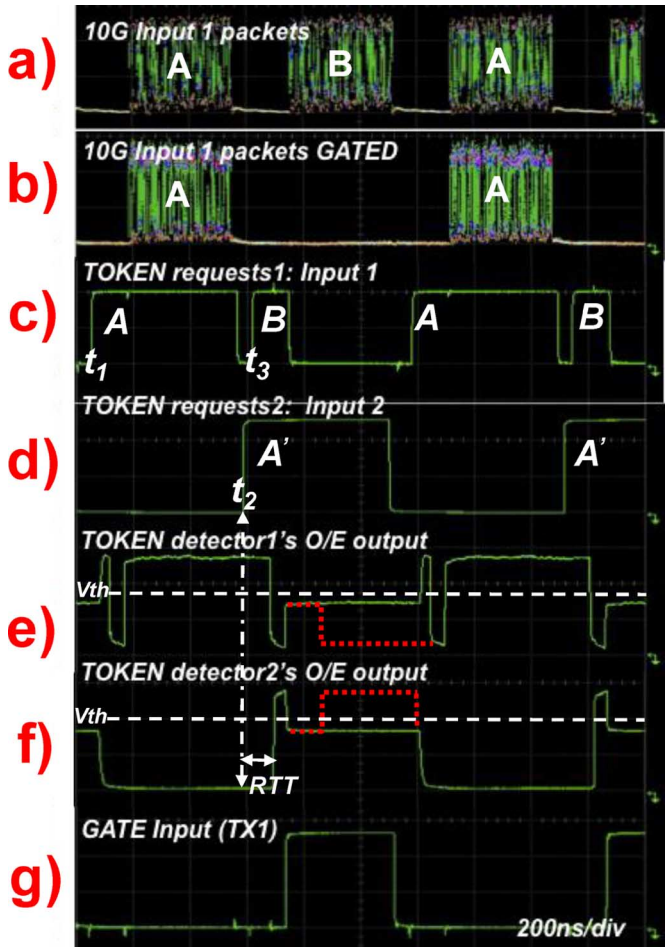

Fig. 6. Measured traces that experimentally demonstrate the working principle explained in Fig. 2(b). (a) Packets generated by TX1 (b) Packets generated by TX1 after the optical gate. B packets are blocked because the corresponding token requests are not granted. (c) Token requests generated by TX1. (d) Contending token requests generated by TX2. (e) Output of the O/E converter in TD1. (f) Output of the O/E converter in TD2. (g) Gate signal generated by FPGA and applied to the optical MZ gate in TX1. By opening the loopback between the TD and FPGA to measure the PDs outputs, the FPGA does not get the trigger signal to stop the token request B. As a result, the trace in Fig. 6(e) does not go to "0" and the trace in Fig. 6(f) do not return to the high value. A dashed red line in trace (e) and (f) shows how the real traces would look like with the feedback loop closed. RTT: round trip time for the token requests.

surements on the data path. The TD was implemented here with a 10 GHz DC-coupled photo-receiver and a 3.84 Gb/s voltage comparator.

B. Experiment Results and Discussion

Fig. 6 shows the measured traces for the packets at AWGR input 1, token requests at AWGR inputs 1 and 2, TD1's and TD2's O/Es outputs, and gate1 output for (a)–(g) respectively. Fig. 6(c) shows a delay of $\approx$ 160 ns between the time that a token request is generated and the time that the relative packet is transmitted. This latency, caused by the fiber pigtails of the bulky components used in this experiment, is equal to the Roundtrip Time (RTT) necessary for each token request to reach the SOA, being reflected back, extracted with an optical circulator and detected by the TD. A delay of 25 FPGA clock cycles (equal to the RTT discussed above) is used to determine whether or not the token is available. At time instant t_1, the FPGA generates the token request A for output 4, and starts a counter. If, at the 25th clock cycle, the FPGA senses a transition at one of its input pin connected to the TD1 output, it means that the token for output 4 is available (O/E output in TD1 is above threshold and comparator output goes from high to low—negative logic) and therefore packet A can be transmitted. In this case the FPGA leaves the gate open (no gate signal is generated). A different situation happens at time instant t_3, when the TX at input 1 generates token request B for output 4. In fact, at time instant t_2 TX2 at input 2 has already generated the token request A' and found the token for output 4 was available. When token request B reaches the SOA, it has been already saturated by the token request A'. It is due to the SOA gain saturation that the optical power arriving at the input of TD1 at time instant $t_3 + RTT$ is not sufficient to generate an electrical signal above threshold. Since the FPGA does not sense the expected transition at the TD1 output, it determines that the token is not available and generates a gate signal (see Fig. 6(g)) to prevent the transmission of packet B (see Fig. 6(b)), while simultaneously terminating the non-granted token request B (see Fig. 6(c)). Note that the TD1 O/E converter output shown in Fig. 6(e) does not go to zero when B does (see Fig. 6(c)). This is due to the fact that the feedback loop between TD1 and the FPGA has been opened to measure the O/E converter output trace.

As explained in the previous section, the token request B also causes a voltage drop in the O/E converter output of the TD2 (also due to the gain saturation effect in the SOA). This is shown in Fig. 6(f).

Fig. 8 shows the BER measurements for the experiment described above. The BER curve with red dots refers to the TX signal after the MZ gate driven by the FPGA (see Fig. 6(b)). Only packets A are granted, while transmission of packets B is always denied. Error-free operation is achieved, and the power penalty of the gated signal compared to the back-to-back measurement (no gating involved) is ≈ 0.5 dB. This is due to the OSNR degradation caused by the insertion loss of the gate ($\approx$ 7 dB). These results demonstrate that the token technique works properly, granting the transmission of packets only upon successful token acquisition. The BER curve with blue diamonds (referring to the switched packets at AWGR output 4, after the PBS_2) shows negligible power penalty, compared to the red curve. This result demonstrates that the coherent crosstalk caused by the token requests on the corresponding packets under transmission is not critical. In fact, the polarization extinction ratio of the PBS_2 ($\approx$ 30 dB) and the power values used in the experiment (-12 dBm and -6 dBm are the power values at the PBS_2 outputs for data and token signal respectively) guarantee a signal to coherent crosstalk ratio $\approx$ 25 dB [22].

Since the AO-TOKEN technique exploits the saturation effect in SOAs, it is subject to the gain wavelength dependence in the SOAs. Under the assumption that the V_{th} in the TD is kept constant, as in this experiment, the gain wavelength dependence can limit the wavelength operating range of the AO-TOKEN technique. By moving TX1 and PBS_2 to AWGR input 1 (8) and output 1 (8), while keeping TX2 connected to AWGR input 2, it is possible to test the AO-TOKEN technique at the lower

8x8 AWGR table

1536.3	1537.9	1539.5	1541.1	1542.7	1544.3	1545.9	1547.5
1537.9	1539.5	1541.1	1542.7	1544.3	1545.9	1547.5	1549.1
1539.5	1541.1	1542.7	1544.3	1545.9	1547.5	1549.1	1550.7
1541.1	1542.7	1544.3	1545.9	1547.5	1549.1	1550.7	1552.3
1542.7	1544.3	1545.9	1547.5	1549.1	1550.7	1552.3	1553.9
1544.3	1545.9	1547.5	1549.1	1550.7	1552.3	1553.9	1555.5
1545.9	1547.5	1549.1	1550.7	1552.3	1553.9	1555.5	1557.1
1547.5	1549.1	1550.7	1552.3	1553.9	1555.5	1557.1	1558.7

Fig. 7. 8×8 AWGR wavelength routing table. Dashed circles indicate the wavelength values used for TX2. Continuos circles indicate the wavelength values used for TX1. Colors are associated with BER curves in Fig. 8.

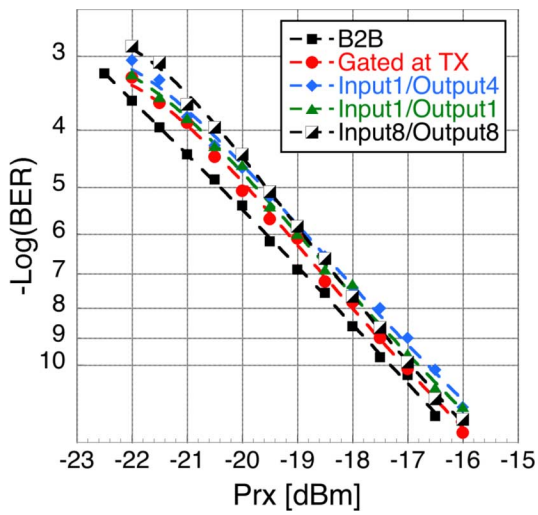

Fig. 8. BER measurements. Filled black squares: back-to-back; red dots: TX1 signal after the MZ gate; other curves: packets from TX1 switched at different AWGR outputs.

(or higher) wavelength values in the AWGR routing table (see Fig. 7).

BER measurements represented by the curves with green triangles and black and white squares demonstrate that the AO-TOKEN technique works correctly at the lower and higher wavelength of the AWGR used in this experiment (resulting in a useable wavelength range of 22.5 nm). 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., ≈ 35 nm (see Fig. 5).

There is a lower bound to the minimum time interval between two or more token requests arriving at the input of the RSOA; below this limit, none of the token requests will be granted. In this experiment, the speed of the FPGA interfacing with the

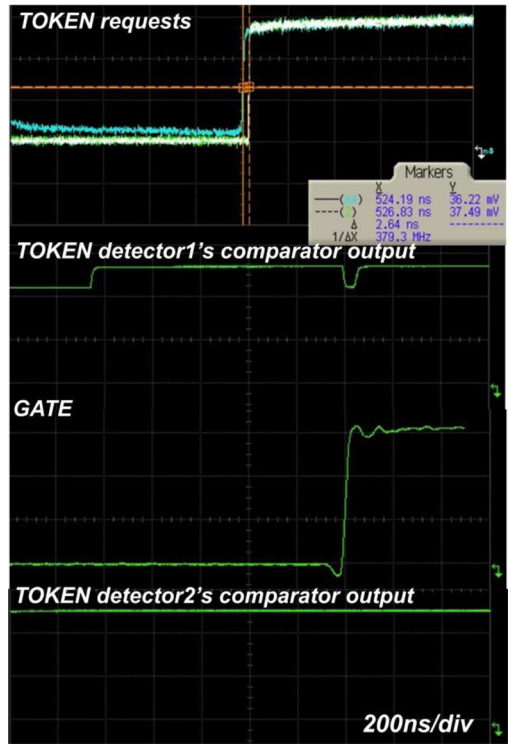

Fig. 9. Measured traces showing that when the time interval between two token requests arriving at the input of the SOA is less than one FPGA clock cycle, none of the token requests is granted.

token detector determines this lower bound, as shown in Fig. 9, where the oscilloscope trace at the top shows two token requests arriving at the SOA input with only a 2.64 ns difference in the arrival times. The TD1's comparator is sufficiently fast to catch the output of the O/E converter going above $V_{\rm th}$ for the short amount of time. However, the FPGA running at 155 MHz (6.4 ns clock cycle) cannot recognize the transition at the TD1's comparator output. Then, the FPGA generates the gate signal, denying the packet transmission. At the same time it is possible to observe that the TD2's comparator output never goes below $V_{\rm th}$, as expected. With a faster FPGA and TD, the ultimate limit to the time interval between two token requests will be given by the SOA gain response time, which is in the order of a few hundreds of picoseconds (ps).

The probability of an unresolved contention will be affected by a variety of factors (traffic pattern, distance to AWGR, SOA gain recovery, port count, etc). Of course, the switching performance is affected by the probability of unresolved contention and some traffic patterns will result in no unresolved contentions (e.g., in uniform random traffic, the contention probability is independent from the port count), while others such as hot spot will result in a much higher probability of unresolved contentions. The switching performance is affected by the number of nodes contending for a given output (a result of port count and traffic pattern), and the minimum duration of a token request. The asynchronous nature of our architecture, together with the random back-up time used in the retransmission strategy [16] should limit the effect of these unresolved contentions.

TABLE I
AVERAGE OPTICAL POWER VALUES FOR THE TOKEN AND DATA PATH IN THE EXPERIMENT OF FIG. 3

	PBS_1 out	AWGR In1	PBS_2 out	SOA out	TD1 in
Data	+12 dBm	-4 dBm	-12 dBm		
Token	+12 dBm	+3 dBm	-6 dBm	+6 dBm	-6 dBm

C. Polarization Crosstalk

As discussed above, coherent crosstalk caused by the token plane on the data plane is not critical for the AO-TOKEN technique exploiting polarization diversity. However, another type of crosstalk has to be considered, i.e., the incoherent crosstalk caused by possible token requests happening during the transmission of a packet. Even though these requests will be denied and terminated immediately after the token requests responses are detected, a crosstalk will be produced by these token requests on the packet under transmission due to the finite polarization extinction ratio of the PBS_2 at the AWGR output. The power of this incoherent crosstalk, together with its frequency distribution, will depend on the arrival times of the token requests, which in turn depend upon the distribution of the traffic injected into the network. Fortunately, the work in [23] shows that when the incoherent crosstalk is ≤ -10 dB, the penalty introduced is negligible. Given the power values used in this experiment (-12 dBm and -6 dBm for data and token signals respectively, as shown in Table I) and the polarization extinction ratio at PBS_2 output (30 dB), it is not difficult to see that even in the worst case (7 inputs sending token requests simultaneously to an in use output), the crosstalk level will be significantly below the -10 dB value mentioned previously. Considering the optical power values, the polarization extinction ratio mentioned above, and a -10 dB maximum crosstalk level, the signal to incoherent crosstalk ratio at the PBS_2 data port is 24 dB. Under this scenario, the maximum number of ports should be $10^{[(24\ \mathrm{dB}-10\ \mathrm{dB})/10]} = 25$; however, if we neglect the insertion loss of the MZ gate (which would not be necessary in an actual implementation), the signal to incoherent crosstalk ratio would be 30 dB, increasing the maximum port count to ≈ 100.

D. SOA Bandwidth, AWGR and Tunable Lasers

As the number of nodes increases, the AWGR channel spacing needs to decrease because of the limited wavelength operating range, as discussed above and shown in Fig. 5. If we consider the wavelength operating range of our technique, ($\sim$ 35 nm as shown in Fig. 5), 25 GHz spacing (0.2 nm) [24] is more than enough for 128 ports. A 25 GHz-spacing AWGR expects to incur less than 1 dB excess loss compared to the 200 GHz counterpart due to the low propagation losses ($\sim$ 0.1 dB/cm) in silica waveguides. In fact, the insertion loss difference between the 200 GHz-spacing AWGR in Fig. 3 and the 50 GHz-spacing of Fig. 10 is smaller than 1 dB.

As demonstrated in [25], the tunable lasers are capable of tuning to the desired target wavelengths with a resolution below 0.02 nm and a switching time of few tens of nanoseconds [25], [26]. However, for port count higher than 128, the AWGR and TLD requirements might become prohibitive.

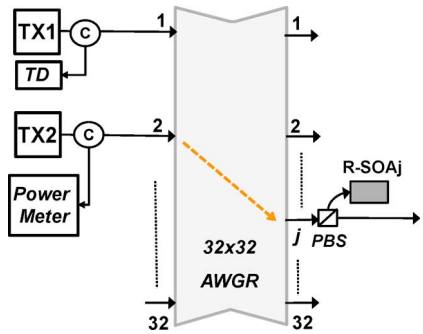

Fig. 10. Experimental setup to measure the noise contribution given by each RSOA. When TX2 is ON, the power meter measures an average optical power equal to -3 dBm (signal power). When TX2 is OFF, the power meter measures an average optical power equal to -33 dBm.

E. RSOA Optical Noise

Since the RSOAs are always ON, the optical noise produced by each RSOA is fed back into the 128 receivers. Although the ASE noise power fed back is filtered and thus relatively low, the total noise power added to the received token becomes not negligible when the switch port count becomes high.

We performed an experimental evaluation, as shown in Fig. 10, based on the SOA used in the experiment. We measured the noise power contribution that the SOA gives at the AWGR input ports of a 50 GHz spacing 32×32 AWGR. This value is equal to -33 dBm, while the power of the reflected CW signals is -3 dBm. This means a ratio of 30 dB. If we add all the contributions given by 128 ports (in reality would be 127), each token detector would see a noise level equal to -33 dBm$+10 \times \log(127) = -11.96$ dBm, which is 8.96 dB below the signal level. We conclude that for 128 ports, the RSOA-noise problem is not the limiting factor.

F. Four Wave Mixing

Each RSOA operates in saturation, thus FWM will occur. The FWM products, on the AWGR wavelength grid, are practically acting as crosstalk and fed back into the receivers, where they add to the token responses before to be fed into the TDs. Similar to the noise power, for a large number of port count, the overall crosstalk could cause errors at the threshold decision, and thus error in calculating contentions.

The RSOAs are based on short length SOAs that achieves gain saturation at low power between various optical channels, but results in very low FWM efficiency between the competing optical channels.

As shown in [27], the FWM efficiency in SOAs rolls off rapidly beyond $\sim$10 GHz frequency spacing, and the FWM between the optical channels can be negligibly small while the SOA can still achieve cross gain saturation.

We verified this by simulation (with VPI Photonics software) and experiment, evaluating the normalized output power of the FWM conjugate signal with the output power of the probe signal [27]. This ratio depends also on the detuning between the pump and probe signals [27], which in our application have the same power, being both represented by a token request. Assuming an AWGR with 25 GHz spacing, the detuning values to be considered here is then 0.2 nm and higher (0.4, 0.6, 0.8, etc. . .).

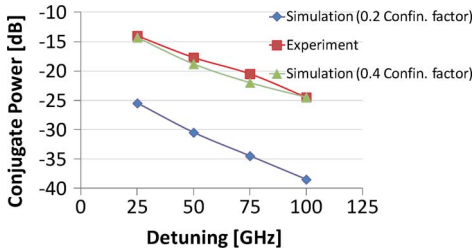

Fig. 11. Normalized FWM conjugate power as function of the channel spacing between two token requests. Squares: experiment results; Triangles: VPI simulation results matching the experiment; Diamonds: simulation results with SOA design optimized for reducing non-linear effects.

Fig. 11 includes an experimental validation of crosstalk level due to FWM (red squares). We used two co-polarized CW signals with the same power level as in the experiment, changing the spacing from 25 to 100 GHz. Note that the SOA used in this experiment to implement the RSOA functionality is designed for non-linear processing (with a high confinement factor ≈ 0.4 and a device length of 1.2 mm). This is confirmed using VPI Photonics software by adjusting the SOA parameters to closely match the simulation (green triangle) with experimental results. By changing the SOA parameters (reducing the confinement factor to 0.2), it is possible to significantly reduce the FWM conjugate power (see blue diamonds), obtaining values similar to what reported in [27]. Then, even in the worst case where $\mathrm{N}-1$ nodes send token requests simultaneously to the same node, the total FWM conjugate power that will interfere with the main wavelength (worst case will be for the token request at the middle wavelength) at the receiver will be -26 dB below the main wavelength. In all the other cases, the fast roll-off of the conjugate FWM power, together with the constraints given by the AWGR routing table, guarantee a crosstalk level even lower than -26 dB. Hence, this level of crosstalk [22] guarantees that FWM is not a limiting factor for the system scalability.

IV. Conclusion

This paper proposes and experimentally demonstrates an all-optical token technique for contention resolution in an AWGR-based optical switch. The technique eliminates the need of any centralized electrical control plane, making use of a fully-distributed token-based contention resolution scheme which exploits the saturation effect in SOAs and the AWGR wavelength routing. The proposed asynchronous switch architecture can either work as packet or fast circuit switch. A full network level performance analysis of the proposed switch architecture is currently under investigation [16].

As mentioned in Section III.B, the wavelength operating range of the proposed technique is limited by the wavelength-dependent gain of the SOA. Considering the 1-dB bandwidth of the SOA used in this experiment, i.e., $\approx$ 35 nm, and an AWGR channel spacing of 25 GHz (0.2 nm), the number of connected nodes could reach 128. Analysis of two types of polarization crosstalk is presented for the proposed scheme. As discussed previously, only the incoherent crosstalk can limit the scalability. However, the crosstalk problem can be completely eliminated by replacing the polarization diversity scheme with a space diversity scheme (which would employ an additional AWGR for the token plane). Using space diversity would not only eliminate the crosstalk between token and data plane, but would also remove the need for any PM components. Moreover, the polarization domain could be used in the data plane to double the total capacity of the switch. In fact, since the proposed architecture does not make use of optical wavelength converters [28], which are notoriously agnostic to polarization multiplexing transmission schemes, the AO-TOKEN architecture can be used with any coherent transmission scheme and modulation format.

The scalability analysis given in Section III.C, III.D, III.E and III.F shows that the ultimate limit to the scalability of the AO-TOKEN architecture is set by the RSOA wavelength operating range and AWGR technology [24].

As explained in [7], it is possible to exploit k wavelengths per AWGR output port to strongly reduce the contention probability and improve the overall networking performance of the AO-TOKEN interconnect architecture [16]. In order to implement this, it would be necessary to use k RSOAs at each AWGR output port. The k RSOAs would connect to each AWGR output port through a $1:k$ optical demux, as explained in [7]. Note that, since the proposed architecture is asynchronous, and the first request arriving at the SOA input wins, it is not possible to handle priorities based on round-robin arbitration schemes used in other types of switches. In the case that two or more requests arrive at the same time, a switch with classic arbitration scheme would assign the contented resource based on certain priorities and a round-robin algorithm. In our scheme, the requests will be first all rejected and then different priorities could be handled at the higher layer in the backoff time algorithm (see [16] for more details) that is used to resend those requests. Those packets/nodes of higher priority could have a smaller backoff time than those of lower priority. We are currently investigating this aspect. Moreover, as explained in [7], it is possible to exploit k wavelengths per AWGR output port to strongly reduce the contention probability and improve the overall networking performance of the AO-TOKEN interconnect architecture [16]. This contention probability reduction, only possible thanks to wavelength routing offered by AWGR, can make the priority aspect less critical.

References

[1] T. Benson *et al.*, "Understanding data center traffic characteristics," in *ACM SIGCOMM Computer Commun. Rev.*, 2010, vol. 40, no. 1, pp. 92–99.
[2] 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.
[3] N. Farrington *et al.*, "Helios: A hybrid electrical/optical switch architecture for modular data centers," *J. SIGCOMM Comput. Commun. Rev.*, vol. 40, no. 4, pp. 339–350, 2010.
[4] L. A. Barroso and U. Hölzle, "The data center as a computer: An introduction to the design of warehouse-scale machines," *Synthesis Lectures on Computer Architecture*, vol. 4, no. 1, pp. 1–108, 2009.
[5] F. Abel *et al.*, "Design issues in next-generation merchant switch fabrics," *IEEE-ACM Transactions on Networking*, vol. 15, no. 6, pp. 1603–1615, Dec. 2007.
[6] R. Hemenway *et al.*, "Optical-packet-switched interconnect for supercomputer applications," *J. Opt. Netw.*, 2004.
[7] X. Ye *et al.*, "DOS—A scalable optical switch for data centers," in *Proc. ACM/IEEE Symposium on Architectures for Networking and Communications Systems (ANCS)*, 2010, pp. 1–12.
[8] O. Liboiron-Ladouceur *et al.*, "The data vortex optical packet switched interconnection network," *J. Lightw. Technol.*, vol. 26, no. 13, Jul. 2008.

[9] I.-S. Joe and O. Solgaard, "Scalable optical switches with large port count based on a waveguide grating router and passive couplers," *IEEE Photon. Technol. Lett.*, vol. 20, pp. 508–510, 2008.
[10] Y. Yong-Kee *et al.*, "High-speed optical switch fabrics with large port count," *Opt. Exp.*, vol. 17, no. 13, pp. 10990–10997, 2009.
[11] R. Proietti *et al.*, "All-optical physical layer NACK in AWGR-based optical interconnects," *IEEE Photon. Technol. Lett.*, 24, no. 5, pp. 410–412, 2012.
[12] X. Ye *et al.*, "Buffering and flow control in optical switches for high performance computing," *IEEE/OSA J. Opt. Commun. Netw.*, vol. 3, no. 8, pp. A59–A72, 2011.
[13] R. Proietti *et al.*, "Performance of AWGR-based optical interconnects with contention resolution based on all-optical NACKs," in *Proc. OFC*, 2012, Paper OTu1B.3.
[14] N. Calabretta, "FPGA-based label processor for low latency and large port count optical packet switches," *J. Lightw. Technol.*, vol. 30, no. 19, pp. 3173–3181, 2012.
[15] R. Proietti *et al.*, "All-optical token technique for contention resolution in AWGR-based optical interconnects," in *Proc. OSA Tech. Dig.*, Paper CM2A.7.
[16] R. Proietti *et al.*, "Scalable and distributed contention resolution in AWGR-based data center switches using RSOA-based optical mutual exclusion," *IEEE J. Sel. Topics Quantum Electron.*, to be published.
[17] M. J. Connelly, *Semiconductor Optical Amplifiers*. New York: Springer, 2002.
[18] E. Dijxstra, "Solution of a problem in concurrent programming control," *Commun. ACM*, vol. 8, no. 9, p. 569, 1965.
[19] C. K. Chan, K. L. Sherman, and M. Zirngibl, "A fast 100-channel wavelength-tunable transmitter for optical packet switching," *IEEE Photon. Technol. Lett.*, vol. 13, no. 7, pp. 729–731, Jul. 2001.
[20] H. C. Shim, "Reflective semiconductor optical amplifier," U.S. 8149503, Apr. 3, 2012.
[21] K. Okamoto *et al.*, "32×32 arrayed-waveguide grating multiplexer with uniform loss and cyclic frequency characteristics," *Electron. Lett.*, vol. 33, no. 22, pp. 1865–1866, 1997.
[22] H. Kim and S. Chandrasekhar, "Dependence of coherent crosstalk penalty on the OSNR of the signal," in *Proc. OFC*, 2000.
[23] L. Buckman, L. Chen, and K. Lau, "Crosstalk penalty in all-optical distributed switching networks," *IEEE Photon. Technol. Lett.*, vol. 9, no. 2, pp. 250–252, 1997.
[24] Y. Hida *et al.*, "400-channel arrayed-waveguide grating with 25 GHz spacing using 1.5%-Δ waveguides on 6-inch Si wafer," *Electron. Lett.*, vol. 37, no. 9, pp. 576–577, 2001.
[25] Y. Runxiang *et al.*, "Rapid high-precision in situ wavelength calibration for tunable lasers using an athermal AWG and a PD array," *IEEE Photon. Technol. Lett.*, vol. 24, no. 1, pp. 70–72, 2012.
[26] L. A. Coldren *et al.*, "Tunable semiconductor lasers: A tutorial," *J. Lightw. Technol.*, vol. 22, no. 1, pp. 193–202, 2004.
[27] A. Mecozzi *et al.*, "Four-wave mixing in traveling-wave semiconductor amplifiers," *IEEE J. Quantum Electron.*, vol. 31, no. 4, pp. 689–699, 1995.
[28] T. Durhuus *et al.*, "All-optical wavelength conversion by semiconductor optical amplifiers," *J. Lightw. Technol.*, vol. 14, no. 6, pp. 942–954, 1996.

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

He is 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.

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

He is currently a Postdoctoral Researcher with the Next Generation Networking Systems Laboratory, University of California, Davis, where he works on 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 received the B.Eng. degree in electrical engineering from Peking University, Beijing, China, in 2007. Currently he is working towards the Ph.D. degree at the Department of Electrical and Computer Engineering, University of California, Davis.

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

Christopher J. Nitta received the Ph.D. degree in computer science from University of California, Davis, in 2011.

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

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

He is a Professor of Electrical and Computer Engineering at University of California, Davis. His current research encompasses 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 ACM and received the NSF CAREER award.

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 currently serves as a Professor of electrical engineering at 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, leading technical efforts in optical networking research and systems integration. He participated 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 (OSA), and is a recipient of the DARPA Award for Sustained Excellence (1997), the Bellcore CEO Award (1998), the Outstanding Mid-Career Research Award (UC Davis, 2004), and the Outstanding Senior Research Award (UC Davis, 2011).