An Experimental Validation of a Wavelength-Striped, Packet Switched, Optical Interconnection Network

Assaf Shacham, *Member, IEEE*, and Keren Bergman, *Fellow, IEEE, Fellow, OSA*

***Abstract*—We experimentally validate a complete optical packet switched interconnection network, implementing the SPINet architecture. The scalable photonic integrated network (SPINet) architecture capitalizes on wavelength division multiplexing (WDM) to provide very large transmission bandwidths, simplify network design, and reduce the network's power dissipation. Contention resolution is performed in the optical domain, and a novel physical layer acknowledgement protocol is employed to mitigate the associated latency and performance penalties. Moreover, the SPINet architecture is specifically designed to enable on-chip integration by not using any kind of optical delay lines. Experiments presented include a complete functionality verification, error-free routing of 80 Gb/s wavelength-striped optical packets (8 wavelengths each modulated at 10 Gb/s) with a bit-error rate (BER) better than 10^{-12}, and novel performance-enhancement techniques such as path adjustments and load balancing.**

***Index Terms*—Multistage interconnection networks, multiprocessor interconnection, optical interconnections, photonic switching systems.**

I. Introduction

OVER THE LAST four decades, progress in high-performance computing (HPC) systems has been dominated by remarkable advances in semiconductor technologies, namely improved fabrication technologies, circuit design techniques, and processor microarchitectures. These advances, manifested in Moore's law [1], have led to the extremely high performance presented by today's CMOS-based microprocessors. Large scale distributed systems (e.g., HPC), while benefiting from this progress, face a severe problem, exacerbated with every generation: the interconnection network, whose performance is crucial to the overall system performance, does not keep up with Moore's law. The physical laws governing the propagation of signals in electrical transmission lines are beginning to limit the performance of communication systems at high data rates (several Gb/s and higher). Dielectric losses and losses caused by the skin effect limit the transmission distance, requiring large power and dedicated circuits to overcome inter-symbol interference (ISI) [2]. The resulting systems are expensive, power hungry, and suffer from a large latency that impacts the overall system performance [3].

Three critical parameters determine the performance of an HPC interconnection network: latency, bandwidth, and power dissipation. The **latency** is dominated by the propagation velocity across the transmission lines (*time of flight*) and by the queueing latency, which is inevitable for systems with large port counts. As systems scale in port counts and clock speeds, the latency grows in absolute terms and even more so in terms of processor clock speeds [3]. Many HPC systems are currently using latency hiding and masking techniques to overcome this problem [4], but these techniques incur performance and power costs and cannot be used in every application.

Modern processors further challenge network designers by demanding ever-growing off-chip **bandwidth** requirements (e.g., 512 Gb/s in the IBM Cell Broadband Engine processor [5]). Providing this bandwidth for remote memory accesses and for interprocessor communication become extremely challenging and power-consuming.

Finally, the **power** consumed by electronic interconnects in HPC systems to meet the bandwidth and latency requirements is becoming the most critical design constraint. The power expended in HPC systems on computation and communications grows very quickly with the data rates [6], and the associated cost and heat dissipation problems have become limiting factors in the design and deployment of many such systems [7]. The power dissipated by the interconnection network is a large fraction of the total system power. As systems become spatially larger, a larger amount of power is required to overcome the losses in transmission lines. This is done either by periodic regeneration or by sophisticated signal processing techniques such as pre-emphasis and equalization [8]. Additional concerns such as cabling density, bending radii, and cooling airflow also present important and growing challenges to HPC interconnection networks designers.

Optical transmission technologies have the potential to mitigate or even eliminate most of these problems. The bandwidth of optical fibers, nearly 32 THz [9], can be utilized through wavelength division multiplexing (WDM) to carry very high data rates, exceeding 10 Tb/s [10]. The low loss in optical fibers alleviates the need for regeneration and sophisticated signal processing techniques. Bending radii and spatial volumes issues are also alleviated when optical fibers are used [11]. These reasons have led to an increasing trend towards using multimode fibers as point-to-point links in local area networks, HPC, and server systems [11].

Point-to-point optical links, however, provide only a partial relief to the power consumption and bandwidth problems. In order to properly address the power and latency challenges, optical switching must be employed taking advantage of *bit rate transparency* by offering end-to-end photonic paths across the

Manuscript received November 14, 2007. Current version published April 17, 2009. This work was supported in part by the National Science Foundation under Grant CCF-0523771 and in part by the U.S. Department of Defense under Subcontract B-12-664.

A. Shacham is with Aprius, Inc., Sunnyvale, CA 94085 USA (e-mail: assaf@ee.columbia.edu).

K. Bergman is with the Department of Electrical Engineering, Columbia University, New York, NY 10027 USA (e-mail: bergman@ee.columbia.edu).

Digital Object Identifier 10.1109/JLT.2008.928541

0733-8724/$25.00 © 2009 IEEE

network. Bit rate transparency is a property of optical switches representing the fact that the power consumption of such switches is independent of the bit rate of the routed data [12]. An optical switching gate, for example, is switched only at the packet rate, and the power expended to control it is independent of the amount of data routed through it. Optical interconnection networks of various structures and topologies have therefore been suggested as solutions for HPC interconnection networks [13]–[17].

Although the advantages of using optical technology to construct interconnection networks are widely accepted, some major challenges have to be addressed when designing an optical interconnection network. Two architectural challenges rise from inherent limitations of the optical medium: the lack of efficient buffering technologies and the limited processing capabilities of optics. Electronic interconnection networks rely heavily on memory elements such as registers and random access memory (RAM) to perform essential functions such as contention resolution and data storage during the processing of control information. Packet processing is also easily performed by electronic processing resources abundant in silicon. An optical interconnection network, conversely, has no such luxuries. Contentions must be resolved without (or with minimal) buffering, and processing must be simple and be done as quickly as possible since data cannot be delayed for an arbitrarily long time. Optical switching devices also present challenges that are negligible or nonexistent inside electronic switching devices, such as noise, optical nonlinearities, and signal degradation. All these limitations and factors must be carefully considered and addressed.

SPINet (Scalable Photonic Integrated Network) [18], [19] is an optical interconnection network architecture designed with these considerations in mind. It uses wavelength-striped packets, where several wavelengths are simultaneously modulated and transmitted to attain a very large transmission bandwidth. SPINet also trades a portion of the abundant bandwidth to encode routing control information on dedicated wavelengths in a format which simplifies the routing logic and accelerates address processing. The network's power consumption is reduced by relying extensively on bit rate transparency, performing electronic computation at low speeds, and by turning-off parts of the network when they do not route useful data. Contentions are resolved by dropping contending messages in the network, but a novel *physical layer acknowledgement* protocol, based on the bidirectionality property of optical waveguides and switches, mitigates the latency penalty associated with dropping packets.

The last few years have seen remarkable advances in the field of photonic integration. Photonic integrated circuits (PIC) with unprecedented number of optical functions are fabricated in III-V semiconductors and are commercially available [20], [21]. Semiconductor optical amplifiers (SOA) have also been integrated, as amplification and switching devices, along with other optical elements onto PICs in InP [22] and in bonded AlGaInAs-silicon [23]. Silicon is also finally emerging as a bona-fide photonic material with photonic elements such as modulators, waveguides, and SiGe receivers, available as library cells in standard CMOS processes [24]. Once photonic integration technology reaches the appropriate level of maturity, SPINet is specifically designed to take advantage of these advances, as it does not require any optical buffering in delay lines. SPINet is based on a network of waveguides, optical couplers and SOAs which have all been demonstrated in PICs. An entire network is, therefore, amenable to integration on a single PIC.

The SPINet architecture was previously presented in several publications, focusing mainly of the development of the architecture and detailed studies of its performance [18], [19]. A software model of the network was developed and simulations have shown that a 64-port network, with and aggregate bandwidth of 3.1 Tb/s and an average latency of 390 ns can be integrated on a PIC dissipating a power estimated at 10 W [19]. These results should be compared to the typical latency of contemporary low-latency HPC interconnect technologies (e.g., InfiniBand) at 2 μs [25].

The network's performance was studied under uniform, nonuniform, and bursty traffic patterns and advanced means of load balancing and contention resolution were developed and modeled [19], [26]. The system management scheme developed in [19] shows that the physical layer acknowledgement mechanism allows us to retransmit dropped packets and effectively convert packet loss to latency-penalties, which are minimized by the ultralow latency of the medium. The 390-ns result quoted here includes these retransmissions, so the packet loss rate (PLR) in that system is essentially 0. For more information about the raw PLR of the system as a function of the applied load, the reader is referred to [19].

To experimentally validate the network concepts, we first constructed an experimental 2 × 2 switching node [27], comprised of optical, electro-optic, and electronic elements. This paper, an extension of [28], provides a cohesive experimental validation of the SPINet network concepts. We report the construction of a fully assembled 4 × 4 network of such switching nodes, and a set of experiments demonstrating critical physical layer and network layer concepts: correct routing functionality, contention resolution, data integrity in the presence of acknowledgement pulses, and path adjustments.

This paper is organized as follows: Section II is dedicated to a brief overview of the SPINet architecture. In Section III, we present the implemented demonstration network followed by a description of the experimental setup in Section IV. Three experiments are detailed in Section V: full addressing verification, data integrity validation and demonstration of in-slot path adjustments. Finally, a summary and conclusions are provided in Section VI.

II. Architecture Overview

A SPINet network is a transparent optical multistage interconnection network (MIN), comprised of 2 × 2 bufferless photonic switching nodes. Each switching node is, in fact, a nonblocking 2 × 2 optical switch which can assume any one of the 6 switching states shown in Fig. 1(a). The switch is implemented using four SOA gates. SOAs are chosen for their wide switching band, sub-ns switching time, low power consumption and integrability [22], [29]. Alternative switching techniques, such as

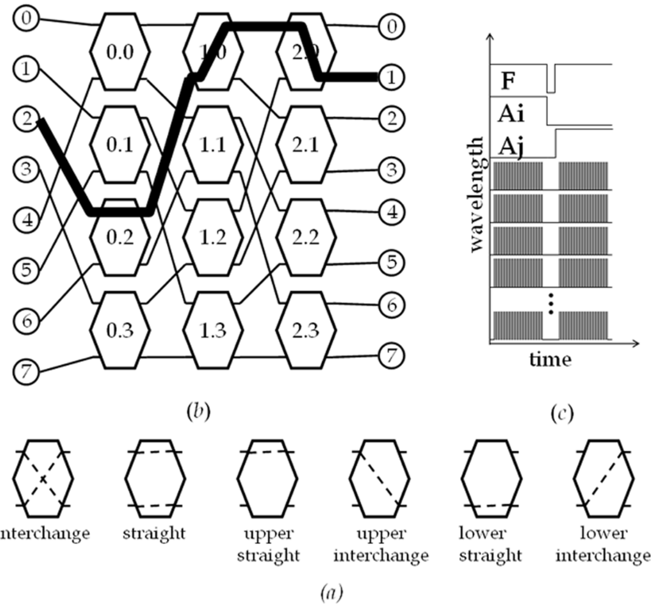

Fig. 1. (a) Switching node has six possible switching states. (b) 8 × 8 omega network with a lightpath across it. (c) Wavelength-striped packet structure, comprised of low-speed control wavelengths (*Frame* and *Address*) and high-speed payload wavelengths.

silicon photonic switches [30], can be considered once they are mature enough to provide these advantages.

In its envisioned implementation, the entire network is integrated on a PIC, its terminals are physically located on the compute nodes of the HPC system, and are connected to the PIC by optical fibers. The terminals are assumed to be synchronized, the network is slotted and synchronous, and messages have a fixed duration. The minimal slot duration is determined by the roundtrip propagation time of the optical signal from the terminals to the PIC. A slot time of 100 ns can, therefore, accommodate nearly 20-m propagation path and a distance of 10 meters between the terminals and the PIC.

The network can be of any indirect binary Banyan topology, where 2 × 2 switching nodes are used [31]. In this implementation we choose the Omega topology, because of its simple stage interconnection pattern. An $N_T \times N_T$ Omega network consists of $N_S = \log_2 N_T$ identical stages. Each stage consists of a perfect shuffle interconnection followed by $N_T/2$ switching elements, as shown in Fig. 1(b), where $N_T = 8$. The switching nodes are non-blocking, so that data from both input ports may be received and routed simultaneously.

At the beginning of each slot, any terminal may start transmitting a message, without a prior request or grant. The messages propagate in the fibers to the input modules on the PIC, and are transparently forwarded to the switching nodes of the first stage. At every routing stage when the leading edges of the messages are received from one or both input ports, a routing decision is made and the messages continue to propagate to their requested output ports. In the case of output port contention in a switching node, one of the contending messages is dropped. Since the propagation delay through every stage is identical, all the leading edges of the transmitted messages reach all the nodes of each stage at the same time.

The switching states of the nodes, as determined by leading edges, remain constant throughout the duration of the message, so the entire message follows the path acquired by its leading edge. Since the propagation delay through the PIC is very short compared to the duration of the messages, the messages stretch across the PIC, effectively creating transparent *lightpaths* between the inputs and outputs [see Fig. 1(b)]. When the messages reach the output modules, they are transparently forwarded on the output fibers to the appropriate terminals, while at the same time an *acknowledgement optical pulse* is generated and sent on the previously acquired lightpath in the opposite direction. Since the switching elements in the node preserve their states and support bidirectional transmission, the acknowledgement pulse is received in the source terminal and the successful reception of the message is confirmed. The input- and output- network interface modules reside on the PIC, with the switching nodes, so the fast exchange of ack pulses is enabled. This point is especially important to facilitate path-adjustments within the same slot, as will be discussed later.

When the slot time is over, all terminals cease transmission simultaneously, the switching nodes reset their switching states, and the system is ready for a new slot. The slot duration is set to assure that the acknowledgement pulses are received at the source terminals before the slot ends. Hence, before the beginning of the following slot, every terminal knows whether its message was accepted and therefore can choose to immediately retransmit the message when necessary.

It should be noted that during the slots when switching nodes do not handle packets, they consume nearly zero power. The low-speed photodetectors are still biased to detect incoming control signals, requiring some current, but the electronic control logic does not switch, and the SOA switches are not turned on. The nodes are therefore effectively off when packets are not routed through them. Further, the SOAs are operated in the small gain regime (amplification is used only to compensate for stage losses), driven with low current (approx. 50 mA). As we switch over a wide band, drifts in temperature do not impact the performance significantly, so thermal control is not required for the stable operation of the interconnection network. These features can offer significant power savings, especially when compared to high-speed electronic signaling standards (e.g., PCI Express, Infiniband). These standards require the system to constantly transmit high-speed idle patterns to maintain interface synchronization, while dissipating large amount of power (e.g., 25–36 W in Mellanox InfiniScale III InfiniBand switch) [32]. In [19], we estimate the power consumption of a 64-port SPINet PIC, with an aggregate bandwidth of 3.1 Tb/s to be 10 W.

The routing and addressing mechanism relies on using the abundant bandwidth offered via WDM, to construct the messages. The message header is encoded on the messages such that it can be instantly decoded in the switching node and the routing decision can be executed upon receiving the leading edges of the message. This is achieved by encoding the control information (e.g., address) on dedicated wavelengths, one bit per wavelength [Fig. 1(c)]. The control wavelengths remain constant throughout the duration of the message to maintain a constant switching state for the entire time the message traverses the switching node. The message payload is segmented, and the individual segments are simultaneously encoded on several other wavelengths, at a higher data rate (e.g., 10 Gb/s per wavelength), to offer a large transmission bandwidth by employing parallel optical transceivers in the interfaces. A guard time is allocated before payload transmission, accommodating for the SOA switching time, for clock recovery in the payload receivers, and for synchronization inaccuracies.

This technique, termed *wavelength-striping*, where a message stretches among several wavelengths which are routed together as one unit is used in several proposed short-reach applications, similar to the one discussed in this paper (see, for example, [16], [33]).

The message header, encoded on the control wavelengths, is comprised of a *frame* bit that denotes the existence of the message and several *address* bits. When a binary network is used, a single address bit is required at every stage, and therefore the number of wavelengths required for address encoding equals the number of routing stages in the network, or $\log_2 N_T$, where N_T is the number of network terminals. The message payload and the header are received concurrently by the nodes. The routing decision in the switching nodes is based solely on the information extracted from the optical header, as encoded by the source. No additional information is exchanged between switching nodes or added to the optical messages by them. The payload is neither decoded nor changed and is routed transparently using the SOA gates.

Wavelength striping contributes to reduced power consumption in the network and the interfaces: the SOA switches consume a constant amount of power independent of the number of wavelengths routed through them, so the power per unit bandwidth can be decreased [29]. Furthermore, several moderate-speed transceivers consume less power than an equivalent-bandwidth ultra-high-speed transceiver [11].

The wavelength-striped packet structure is also compatible with the parallel optical transmission adopted in low-cost local area optical networks. Parallel transmission on separate fibers is currently commercially used [34], and coarse WDM (CWDM) transceivers with 4 wavelengths per fiber are also available [35]. As technology improves, solutions for multiplexing a larger number of wavelengths on a single fiber are found, offering low-cost alternatives to high-speed serial transmission [11], [34]. New advances in silicon photonics also suggest that a large number of moderate-speed silicon modulators can be integrated and multiplexed on a single CMOS-compatible silicon chip [24], [36]. We argue wavelength striping is a cost effective and power efficient way to aggregate traffic from multiple nodes onto a high-bandwidth transmission since it provides a natural interface to the lower data rate parallel electronic links and avoids the need to serialization or de-serialization. Thus, despite the seemingly large component count, wavelength-striping is actually a cost-effective technique for short-reach optical interconnections.

Another issue that has to be addressed is the wide transmission band. Narrowband transmission is a desired property from dispersion management and component design standpoint. By modulating the control wavelengths in the single-wavelength per bit fashion, we occupy a wider band, which creates some challenges to the component design. This approach is advantageous to the SPINet architecture network performance because it enables the instantaneous extraction and detection of the relevant address bit in the switching nodes. A narrowband approach, using a *serial* header will require complex circuits such as clock and data recovery (CDR) as well as header-regeneration circuits to reside in each switching node (as is the case in label-switching networks), rendering the entire system unfeasible for integration. We address the challenges posed by the wide transmission band by using wideband SOAs as switching elements, and carefully assigning the modulation wavelengths such that they all fit in the SOA switching bands. The 30-nm band of contemporary SOAs [29] can fit as many as 37 WDM channels of the ITU grid and 75 DWDM channels, such that a 10-stage system is within reason.

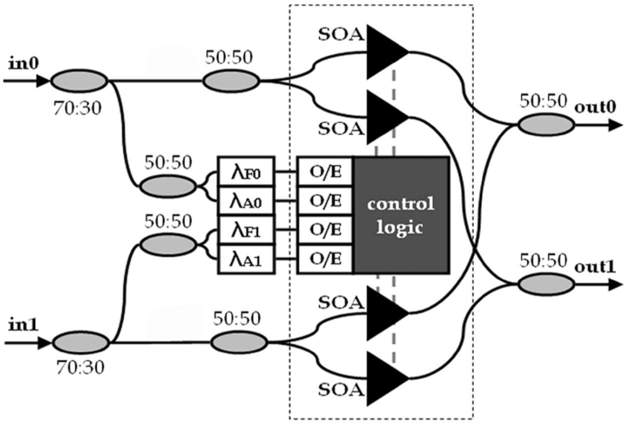

Fig. 2. Switching node is based on a 2 × 2 SOA switch supplemented with couplers (ovals), filters (λ), photodetectors (O/E), and control electronics.

Blocking topologies, such as Omega, are chosen to implement the network because they require less hardware ($2N_T \log_2 N_T$ switching gates) to map N_T input ports to N_T output ports. A non-blocking topology will require N_T^2 switching gates to perform the same task. Several techniques have been suggested to mitigate the blocking properties of the network by increasing its path diversity. Adding stages to the original network can offer more paths between each input-output pair (i.e., increase the *path diversity*), thus facilitating contention avoidance [18], or load balancing and path adjustments [19], [26].

Load balancing, for example, is achieved by a distribution network that precedes the routing network. In the distribution network, the messages are routed according to a distribution address, encoded in a manner similar to the routing address. When a random distribution address is encoded traffic patterns which are adversarial to the Omega topology, such as bit-reversal or other permutation patterns, are distributed uniformly among input ports thus appearing to the routing network as uniform traffic, which is easier to handle [3]. Additionally, when a message is dropped and the acknowledgement pulse is not received, the distribution address can be changed within the same slot, and a different path may be found. This technique is termed *path adjustments*.

These techniques (load balancing and path adjustments) were evaluated using a cycle-accurate simulation model [19], and have been shown to increase the network throughput by more than 20% for a 64-port network, under uniform traffic. The performance gain under non-uniform traffic patterns was also evaluated and found to be even larger, confirming that the load balancing techniques are effective in counter-acting the effects of non-uniform traffic. For additional details on using path diversity and other techniques to improve performance, the reader is referred to [19]. In Section V we present an experimental

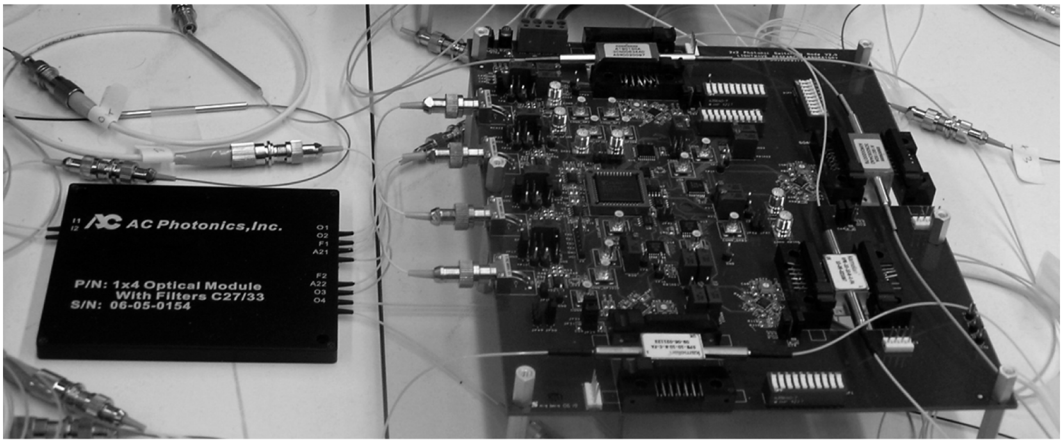

Fig. 3. Photograph of the 2 × 2 switching node.

demonstration of the full network functionality including load balancing and path adjustments.

III. Implemented Network

The SPINet network architecture is designed to take advantage of photonic integration technologies, so that the entire MIN can be integrated on a single PIC. Its concepts can nonetheless be demonstrated using macro-scale photonic elements. To test the network functionality and its ability to route optical packets error-free while correctly routing acknowledgement pulses in the opposite direction, we have constructed an experimental demonstration network. The demonstration network is comprised of six 2×2 switching nodes, organized in 3 stages. Stage 0 is used as a distribution stage, for load balancing and path adjustments, and stages 1 and 2 are a routing network, connecting four input ports and four output ports. In this section we describe the implementation details of the switching nodes and the network.

A. Switching Node

The implementation of the switching node was reported in detail in [27]. In this subsection we provide a brief summary of that report.

The optical switching element in the switching node is a 2×2 switch based on four SOA gates organized as a gate matrix (Fig. 2). The switch is supplemented by optical couplers, wavelength filters, photodetectors, and control electronics. The switching node is capable of processing optical addresses and executing a routing decision while the wavelength-striped data is maintained in the optical domain.

When the leading edges of the messages enter the node, 30% of their power is tapped off, while the rest of the signal propagates in a parallel fiber path. The tapped power is then split by a 3 dB coupler and directed into wavelength filters to extract the *frame* bit ($\lambda_F = 1555.75$ nm, denoting message existence) and the *address* bit, denoting the requested output port. The address filter's wavelength value depends on the stage in the network. The four bits (*frame* and *address* from each message) are detected using 115 MHz *p-i-n* photodetectors and are driven into the control logic, implemented using a Xilinx complex programmable logic device (CPLD). The CPLD processes the header bits according to a pre-programmed truth-table to reach a routing decision. It then turns on the appropriate SOAs, using current drivers. The SOAs are driven with current to provide 8.5 dB gain that compensates for the coupling and connector losses.[1] At the end of the message, when the optically-encoded header bits turn off, the SOAs are switched off accordingly.

The total latency of the node has been measured to be 38.4 ns. The optical coupler and filter network take 19.6 ns and the latency of the electronic path is 18.8 ns. The latencies result mainly from the fact that separately packaged elements are used, so the signal spends most of its time in pigtail fibers, electronic traces, packages, and debug integrated circuits. We estimate that the node latency can be reduced by an order of magnitude in an integrated implementation.

A photograph of the switching node appears in Fig. 3.

B. Six-Node Network

The switching nodes are organized as a three-stage Omega network, with two nodes in each stage. Stage 0 is used as a distribution stage, where a 1-bit distribution address is used to encode the selected path between the two paths existing from each input port to each output port. The wavelength used to encode the distribution address is $\lambda_{A0} = 1535.04$ nm. Stages 1 and 2 are used as the routing network between the four input ports and the four output ports, so a 2-bit routing address is required. The routing address is encoded on $\lambda_{A1} = 1533.47$ nm and $\lambda_{A2} = 1550.92$ nm, which are decoded in stage 1 and stage 2, respectively.

The stages are directly connected without intermediate fibers, so the 38.4 ns latency of each stage is equal to the node latency. The total latency of the three-stage network is therefore 115.2 ns. Optical circulators are connected at the network output ports to inject the optical acknowledgement pulses and to eject the light reflected from the interfaces. Circulators are also placed at the input ports to extract the pulses. The total latency of the network, including the circulators, is 117.6 ns.

A photograph of the implemented six-node network appears in Fig. 4. While this network is clearly expensive and complex to implement, it is reasonable to expect that once photonic integration technology is mature enough to benefit from an economy

[1] Due to limited availability some of the SOAs in the output stage, are from a different manufacturer and can only provide 5.5 dB gain under these operating conditions. The nodes in stage 2, therefore, exhibit approximately 3 dB loss.

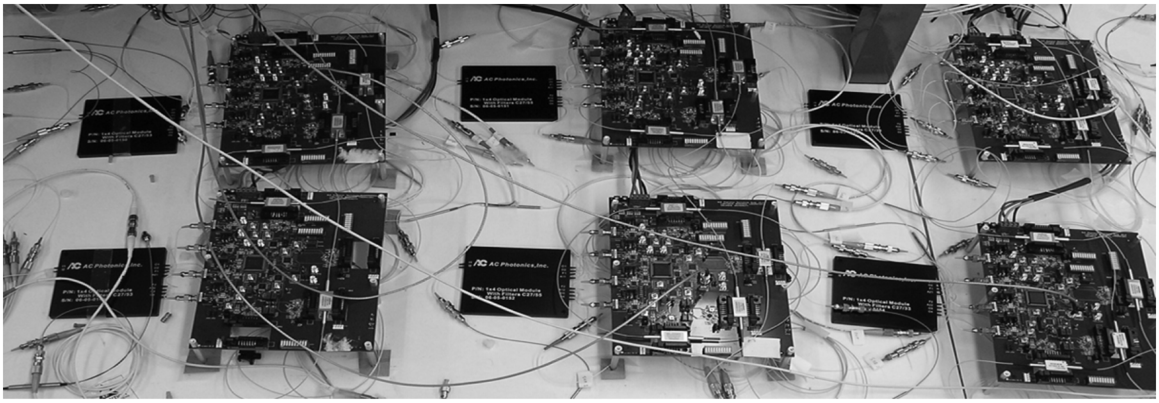

Fig. 4. Photograph of the 3-stage, 6-node, 4 × 4 implemented demonstration network.

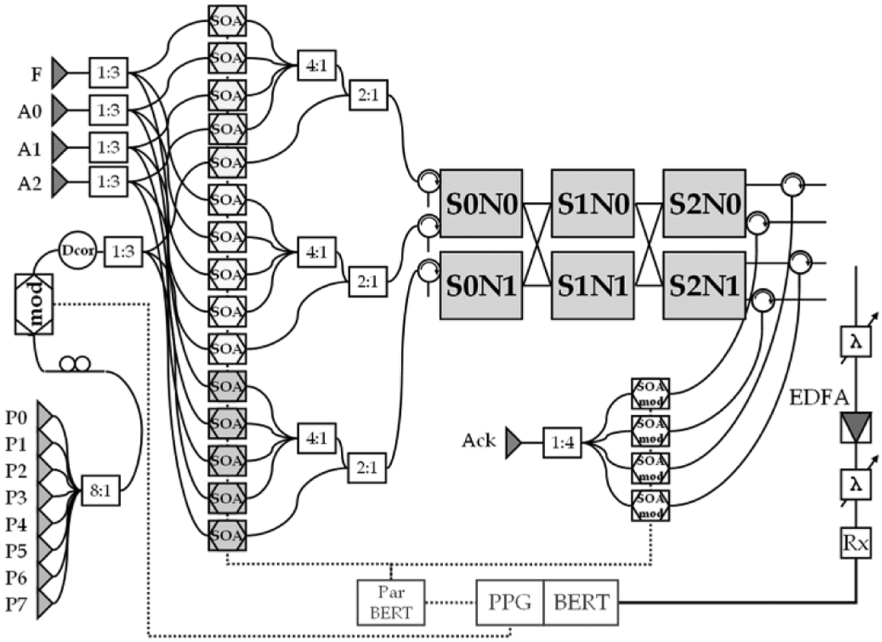

Fig. 5. Experimental setup is comprised of DFB lasers (triangles), SOA modulators, optical couplers (rectangles), and additional equipment, as annotated. The switching nodes are annotated with their respective stage and height indices.

of scale, an integrated version of such a network can be fabricated rather cost-effectively. PIC-based systems are already being mass-produced and commercially deployed [37].

IV. Experimental Setup

An experimental setup (Fig. 5) is assembled to construct wavelength-striped optical packets, inject them into the network and analyze them when they reach the output ports. The experimental setup generates optical packets for 3 independent output ports by separately generating the payload and control signals and coupling them together before injection. The packet analysis instrumentation can be connected to any of the four output ports of the network.

Eight CW-DFB lasers with wavelengths ranging from 1539.8 nm to 1559.8 nm are multiplexed onto a single fiber using a passive 8:1 optical coupler, and are then modulated by a $\mathrm{LiNbO_3}$ modulator. The modulator is driven with a $2^7 - 1$ pseudo random bit sequence PRBS at 10 Gb/s by a pulse pattern generator (PPG). The modulated wavelengths are decorrelated with a 24-km length of optical fiber (SMF-28) by approximately 420 ps/nm and then split by a passive 3:1 coupler into 3 modulated wavelength-striped data streams, which are then gated and amplified using SOAs to form discrete optical packets. The DFB lasers' wavelengths are chosen such that a rigorous testing of the network may be performed. The spacing within 2 pairs of wavelengths is only 0.8 nm, and two payload wavelengths are spaced by 0.8 nm from λ_F and λ_{A2}, respectively, to test the possible crosstalk between control and payload wavelengths.

To generate the control wavelengths, 4 CW-DFB lasers, at the appropriate wavelengths ($\lambda_F, \lambda_{A0}, \lambda_{A1}$, and λ_{A2}), are split using 3:1 couplers and are then modulated with the appropriate control information by an Agilent ParBERT using SOAs as modulators.[2] The $12 (= 4 \times 3)$ modulated control streams are grouped using three 4:1 couplers to three optical fibers, each carrying 4 modulated control wavelengths (*Frame, Address0, Address1, Address2*). The control wavelengths are multiplexed

[2] SOAs are used for convenience and to conserve lab-space. Directly modulated lasers can be used as well.

with the wavelength-striped packets. Each of the three active input ports is therefore injected with SPINet packets comprised of a 4-wavelength header and an 8-wavelength payload. The acknowledgement pulses are generated in the same way: a CW-DFB ($\lambda_{\text{ack}} = 1547.90$ nm in the first experiment and $\lambda_{\text{ack}} = 1556.53$ nm in the second experiment) laser is split and then modulated by SOAs to create 4 independent *ack* streams.

All the packet generation SOAs are controlled by an Agilent ParBERT programmed with patterns designed specifically for each experiment. The acknowledgement patterns are programmed such that an *ack* pulse is transmitted when a packet is expected in the respective output port according to the input patterns.

The packet analysis system includes a 10 Gb/s DC-coupled *p-i-n* receiver which follows an erbium doped fiber amplifier (EDFA). The EDFA is preceded by a tunable filter to select the desired payload wavelength and reject other wavelengths and SOA amplified spontaneous emission (ASE). Another tunable filter is placed between the EDFA and the receiver to reject the EDFA ASE. The 10 Gb/s receiver is connected to a BER tester (BERT), which is synchronized with the PPG and gated for packet analysis by the ParBERT. An optical spectrum analyzer and a sampling oscilloscope are used at the output to inspect the optical signals.

Three experiments are performed using the experimental setup to verify the correct functionality of the network. In the first experiment, full addressing of all output ports is demonstrated and several contention scenarios are presented. In the second experiment, the bit error rate (BER) of the data traversing the network is measured, in the presence of acknowledgement pulses. Finally, in the third experiment, contention resolution using path adjustments is demonstrated.

V. Experiments and Results

Using the experimental demonstration network and the packet generation and analysis setup, we conduct three experiments validating the SPINet network concepts: decoding of optical addresses, correct routing of messages and *ack* pulses, error free data routing, and path adjustments. The experiments are detailed in this section.

A. Full Addressing

The goal of the first experiment, originally reported in [28], is to demonstrate the correct routing functionality from the input ports to all output ports as well as to verify that all ports are active and addressable. Packets are injected into 3 input ports (*input0*, *input1*, and *input2*). The 16-slot pattern is programmed such that initially (slots 0–3) all output ports are addressed from *input0*, then (slots 5–8) all output ports are addressed from *input2*. Finally, (slots 10–13) several cases of contentions and packet-dropping are shown when packets are injected simultaneously from three input ports. Whenever a packet is received at any output port, a 16-ns ack pulse is injected, using an optical circulator, into the output fiber in the opposite direction. In this experiment, the packets are 250-ns long, spaced by a dead-time of 16.6 ns, forming 266.6-ns long slots.

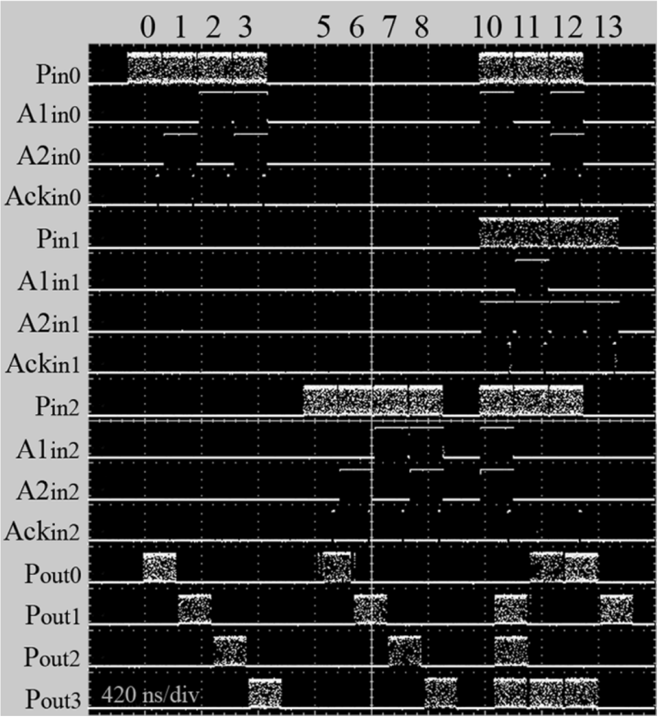

Fig. 6. Optical waveforms of the input and output signals in experiment *A*: *payload*, *address1*, *address2*, and *ack* for each input port, and the packets' payload for each output port [26].

The optical waveforms of the packets, as appearing at the inputs and the outputs, and the *ack* pulses, as received at the inputs, are shown in Fig. 6, verifying correct routing functionality in both directions.

To assist in the interpretation of the contention scenarios in Fig. 6, the following examples are given: in slot 11, 3 packets are injected: *input0→output0*, *input1→output3*, and *input2→output0*. The contention on *output0* is resolved by dropping the message from *input2* as can be seen by the fact that acks are received only in *input0* and *input1*. The packet from *input2* to *output0* is retransmitted successfully and acknowledged in slot 12.

Slot 12 presents a case of an internal contention caused by the blocking topology of the network. Two packets (*input0→output3* and *input2→output0*) are routed successfully and the third packet (*input1→output1*) is dropped following a contention, although *output1* is not busy. This packet loss is a result of internal blocking in the network, and the missing *ack* pulse in *input1* facilitates the retransmission in slot 13. Path adjustments, experimentally shown in Subsection V.C below, can enable a successful transmission within the same slot and thus increase the utilization of the network.

B. Data Integrity

Many optical effects common to SOAs may cause signal degradation that can lead to errors in data transmission. These effects include noise, gain compression and other non-linear effects that can result from interaction between the optical signals and ASE and among the optical signals (for example, data streams and *ack* pulses) [38]. It is, therefore, very important to verify error-free routing of the optical packets through the 3-stage network, while *ack* pulses are traveling in the opposite direction.

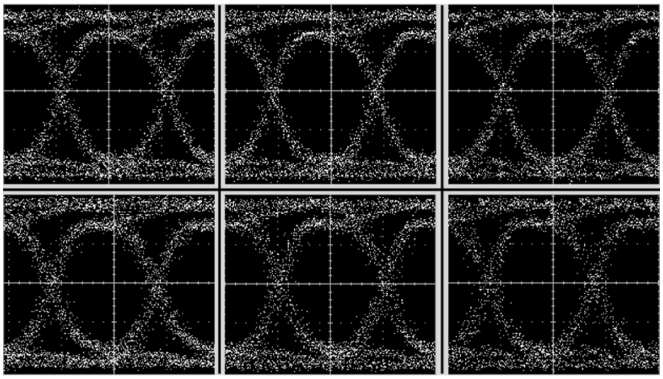

Fig. 7. Electronic eye diagrams in experiment B, at 10 Gb/s for three representative wavelengths at the network input (top) and at the network output (bottom). The wavelengths shown are 1539.8 nm (left), 1551.72 nm (center), and 1559.8 nm (right). The vertical scale is 100 mV/div.

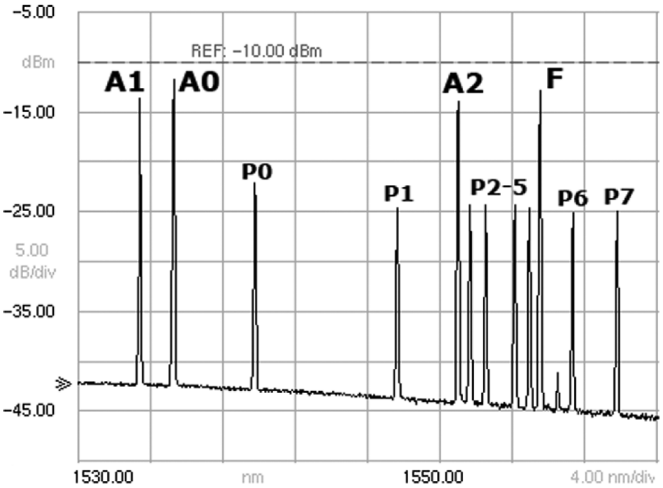

Fig. 8. Spectrum of the optical packets as they emerge from *output3* in experiment *B*. *Frame*, *address*, and *payload* wavelengths are annotated.

The Anritsu BERT is connected to *output3*, and a pattern is programmed such that all three input ports and all four output ports are active. *Ack* pulses are injected into *output3* whenever a packet is received. The Bert is operated in *gated* mode, where an external *gate* signal controls when actual BER measurements are performed. The gate signal is generated by the PPG, instructing the BERT to only measure BER when a packet is expected in the network output. Because the pattern is periodic, this technique enables BER measurement on packetized traffic. Additionally, the gating method and the short duration of the packets dictate that the PRBS pattern used is a short pattern, to allow for initial BERT-lock. As described above, we use PRBS $(2^7 - 1)$.

Error free transmission (BER of 10^{-12} or better) is confirmed on all eight wavelengths. Eye diagrams of three representative wavelengths (1539.8 nm, 1551.72 nm, and 1559.8 nm) at the networks input ports and at output ports are given in Fig. 7. The output spectrum is given in Fig. 8.

This experiment confirms the feasibility of multi-stage error-free transmission of wavelength-striped packets and *ack* pulses. The transmission power, as is evident in Fig. 8, must be kept below the SOA input saturation level so that gain compression and inter-channel crosstalk are avoided. The input power, therefore, was -7 dBm (peak) for the control-wavelengths. Fig. 8 shows average power at the outputs. Taking the packets duty cycle into account, the overall insertion loss is about 1 dB. The launch power of the data wavelengths is about -15 dBm. Because the switching nodes preserve the optical power in the packets and do not exhibit a substantial insertion loss, the low launch power does not pose a problem. Recent studies have shown that wavelengths-striped packets can be routed error-free through tens of stages in similar SOA-based networks [39]. These results, combined with the fact that 10 stages, for example, are required to connect a 1024×1024 network, suggest that the SPINet architecture's scalability is not limited by signal integrity considerations, but rather by integration and packaging constraints.

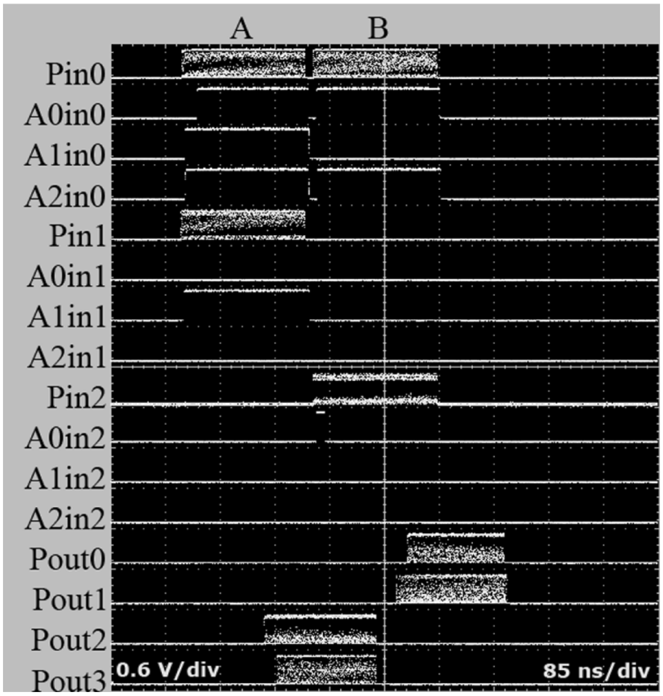

Fig. 9. Optical waveforms of the input and output signals in experiment *C*: *payload*, *address0* (distribution address), and *address1* and *address2* (routing address), for each input port, and the packets' payload for each output port.

C. Path Adjustments

As described in Section II, the insertion of a distribution network provides additional paths connecting each input-output pair. In this experiment, we demonstrate the *path adjustment* technique which exploits this path diversity [26]. When an acknowledgement pulse is not received at the input-module when it is expected, following a message transmission, the input-module understands that the message was dropped. The Message transmission can, therefore, be restarted with a new distribution address such that a different path, which may be open, is taken. Since the entire network, including the input- and output-modules, is assumed to be integrated on a single PIC, the roundtrip latency is not expected to be large. The overhead time expended on path adjustments should, therefore, be balanced against the increased utilization. A detailed discussion of the performance optimization of path adjustments is given in [19].

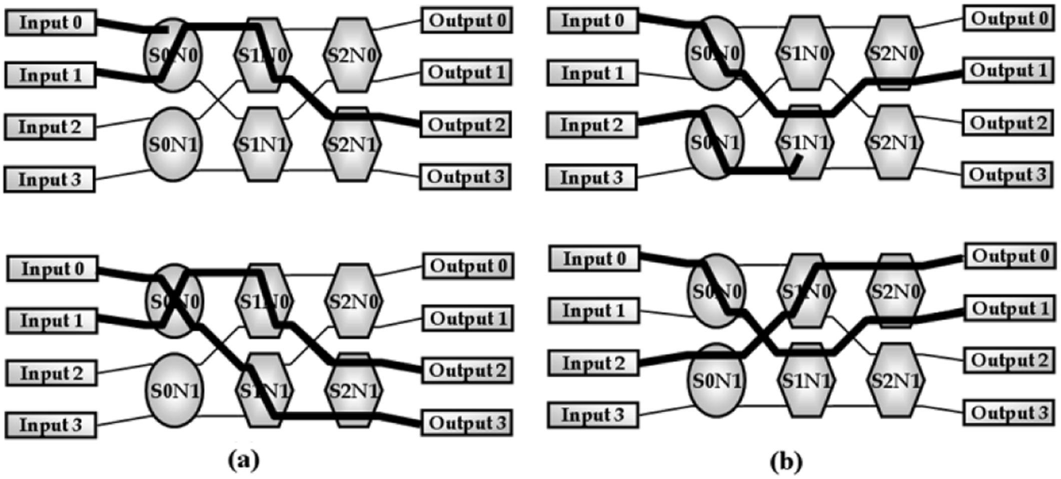

Fig. 10. Explanatory diagram of the path adjustment experiment (*C*). In the first slot (a), *input0* is blocked and changes its distribution address. In the second slot (b), *input2* is blocked and forced to adjust its path.

To support path adjustments, the node implementation has to be slightly modified. The original implementation of the switching node, using a simple truth-table, relies on the fact that all the leading edges are received approximately at the same time (all within the guard-band that precedes the payload transmission). To enable path adjustments, the system must now guarantee that paths that are being adjusted will not interfere with paths that have already been successfully established. To support this we add a 2-bit memory to each switching node to save state information. Using the saved state information, it can now be ensured that when a message is routed the state of the switching node cannot be changed until its transmission is completed. If a new message tries to change the switching node's state while adjusting its path, it is blocked. In cases where the new message's requested output port conforms with the current node's state, it will be routed successfully. The correctness of this scheme was validated using computer simulations [19], [26]. Since the electronic control of the network is implemented as a programmable CPLD, this change of functionality is easy to implement.

The pattern in this experiment is programmed to generate two scenario demonstrating path adjustments. The slot duration in this experiment is 206 ns and it is divided to a 194-ns long packet and a 12-ns long dead-time.

The experiment waveforms, as seen on the oscilloscope, are shown in Fig. 9. The reader is invited to follow the waveforms using the explanation below and the explanatory diagrams in Fig. 10.

In slot *A* (Fig. 10(a)), one packet (*input1→output2*) is routed successfully while the other packet (*input0→output3*) is blocked. *Input0* then changes its distribution address (*A0in0*) from 0 to 1, and finds an open path to *output3*. In slot *B* (Fig. 10(b)), the packet (*input2→output0*) is blocked by the packet (*input0→output1*). Again, by changing the distribution address (*A0in2*) from 1 to 0, *Input2* finds an open path to *output0*. In Fig. 9, it is evident that the adjusted packets reach the output ports after the packets that were initially successfully routed and, therefore, can only use a shorter transmission period until the end of the slot. Without the path adjustment, however, the messages would be blocked. As mentioned earlier, the time allocated for path adjustments is a factor that has to be optimized against the added utilization gained.

VI. Conclusion

Optical interconnection networks offer many advantages that address the communication bottleneck in HPC systems. The large bandwidth, low latency, and potential reduction in power dissipation make optical technology attractive for employment in future systems. The SPINet optical interconnection network architecture leverages on the unique properties of optical switching to provide a low-latency, high-bandwidth, low-power interconnection solution while addressing the unavoidable design challenges. Its high performance was proven by simulation-based studies under various scenarios and traffic patterns [19].

In this paper, we have reviewed the architecture and provided a compelling experimental demonstration of its feasibility. Encoding and decoding of optical addresses, photonic routing of wavelength-striped data in the optical domain, and several techniques of contention resolution have been shown. Data integrity is proven by routing the optical packets with a BER of 10^{-12} or better on all eight wavelengths while acknowledgement pulses are propagating in the opposite direction. Studies show that the scalability to larger networks is limited by the progress of photonic integration technology, which has shown some very promising advances over the last few years.

As electronic high performance interconnection networks approach their fundamental limits, photonic interconnection networks provide solutions to many imperative constraints. The SPINet architecture is an example of such a complete, scalable, experimentally verified solution. It is reasonable to expect that as photonic integration technology progresses, the implementation complexity and cost will be reduced to the level where this type of architecture is made commercially viable.

Acknowledgment

The authors would like to thank H. Wang for his help in the assembly of the experimental setup and A. Biberman for his careful proof-reading of the manuscript.

REFERENCES

[1] G. E. Moore, "Cramming more components onto integrated circuits," *Electronics*, vol. 38, no. 8, pp. 114–117, Apr. 1965.

[2] D. A. B. Miller, "Rationale and challenges for optical interconnects to electronic chips," *Proc. IEEE*, vol. 88, no. 6, pp. 728–749, Jun. 2000.

[3] W. J. Dally and B. Towles, *Principles and Practices of Interconnection Networks*. San Francisco, CA: Morgan Kaufmann, 2004.

[4] J. L. Hennessy and D. A. Patterson, *Computer Architecture: A Quantitative Approach*, 3rd ed. San Francisco, CA: Morgan Kaufmann, 2002.

[5] M. Kistler, M. Perrone, and F. Petrini, "Cell multiprocessor communication network: Built for speed," *IEEE Micro*, vol. 26, no. 3, pp. 10–23, May/Jun. 2006.

[6] T. Mudge, "Power: A first-class architectural design constraint," *IEEE Computer*, vol. 34, no. 4, pp. 52–58, 2001.

[7] L. A. Barroso, J. Dean, and U. Hölzle, "Web search for a planet: The Google cluster architecture," *IEEE Micro*, vol. 23, no. 2, pp. 22–28, Mar./Apr. 2003.

[8] J. Liu and X. Lin, "Equalization in high-speed communication systems," *IEEE Circuits Syst. Mag.*, vol. 4, no. 2, pp. 4–17, 2004.

[9] G. P. Agrawal, *Fiber-Optic Communication Systems*. New York: Wiley, 2002.

[10] K. Fukuchi, T. Kasamatsu, M. Morie, R. Ohhira, T. Ito, K. Sekiya, D. Ogasahara, and T. Ono, "10.92-Tb/s (273 × 40-Gb/s) triple-band/ultradense WDM optical-repeatered transmission experiment," presented at the Optical Fiber Communications Conf. (OFC), Mar. 2001, paper PDP-24.

[11] A. F. Benner, M. Ignatowski, J. A. Kash, D. M. Kuchta, and M. B. Ritter, "Exploitation of optical interconnects in future server architectures," *IBM J. Res. Dev.*, vol. 49, no. 4/5, pp. 755–775, Jul. 2005.

[12] R. Ramaswami and K. N. Sivarajan, *Optical Networks: A Practical Perspective*, 2nd ed. San Francisco, CA: Morgan Kaufmann, 2002.

[13] R. Luijten, C. Minkenberg, B. R. Hemenway, M. Sauer, and R. Grzybowski, "Viable opto-electronic HPC interconnect fabrics," in *Proc. ACM/IEEE (SC—05) Conf. Supercomp.*, Nov. 2005, p. 18.

[14] A. K. Kodi and A. Louri, "Design of a high-speed optical interconnect for scalable shared-memory multiprocessors," *IEEE Micro*, vol. 25, no. 1, pp. 41–49, Jan./Feb. 2005.

[15] I. H. White, K. A. Williams, R. V. Penty, T. Lin, A. Wonfor, E. T. Aw, M. Glick, M. Dales, and D. McAuley, "Control architecture for high capacity multistage photonic switch circuits," *J. Optical Networking*, vol. 6, no. 2, pp. 180–188, Jan. 2007.

[16] A. Shacham, B. A. Small, O. Liboiron-Ladouceur, and K. Bergman, "A fully implemented 12 × 12 data vortex optical packet switching interconnection network," *J. Lightw. Technol.*, vol. 23, no. 10, pp. 3066–3075, Oct. 2005.

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

[18] A. Shacham, B. G. Lee, and K. Bergman, "A scalable, self-routed, terabit capacity, photonic interconnection network," in *Proc. Hot Interconnects: IEEE 13th Ann. Symp. High Performance Interconnects*, Aug. 2005, pp. 147–150.

[19] A. Shacham and K. Bergman, "Building ultralow latency interconnection networks using photonic integration," *IEEE Micro*, vol. 27, no. 4, pp. 6–20, Jul./Aug. 2007.

[20] R. Nagarajan *et al.*, "Large-scale photonic integrated circuits," *IEEE J. Sel. Topics Quantum Electron.*, vol. 11, no. 1, pp. 50–65, Jan./Feb. 2005.

[21] R. Nagarajan *et al.*, "Large-scale photonic integrated circuits for long-haul transmission and switching," *J. Opt. Netw.*, vol. 6, no. 2, pp. 102–111, Feb. 2007.

[22] R. Nagarajan *et al.*, "Monolithic, 10 and 40 channel inp receiver photonic integrated circuits with on-chip amplification," in *Proc. Opt. Fiber Commun. Conf. (OFC)*, Mar. 2007, paper PDP32.

[23] H. Park, A. W. Fang, O. Cohen, R. Jones, M. J. Paniccia, and J. E. Bowers, "A hybrid AlGaInAs-silicon evanescent amplifier," *IEEE Photon. Technol. Lett.*, vol. 19, no. 4, pp. 230–232, Feb. 15, 2007.

[24] C. Gunn, "CMOS photonics for high-speed interconnects," *IEEE Micro*, vol. 26, no. 2, pp. 58–66, Mar./Apr. 2006.

[25] S. Sur, M. J. Koop, L. Chai, and D. K. Panda, "Performance analysis and evaluation of Mellanox ConnectX InfiniBand architecture with multi-core platforms," presented at the Hot Interconnects: IEEE 15th Ann. Symp. High Performance Interconnects, Aug. 2007.

[26] A. Shacham and K. Bergman, "Utilizing path diversity in optical packet switched interconnection networks," presented at the Opt. Fiber Commun. Conf. (OFC), Mar. 2006, paper OTuN5.

[27] A. Shacham, B. G. Lee, and K. Bergman, "A wideband, non-blocking, 2 × 2 switching node for a SPINet network," *IEEE Photon. Technol. Lett.*, vol. 17, no. 12, pp. 2742–2744, Dec. 2005.

[28] A. Shacham, H. Wang, and K. Bergman, "Experimental demonstration of a complete SPINet optical packet switched interconnection network," presented at the Opt. Fiber Commun. Conf. (OFC), Mar. 2007, paper OThF7.

[29] I. Armstrong, I. Andonovic, and A. E. Kelly, "Semiconductor optical amplifiers: Performance and applications in optical packet switching," *J. Opt. Netw.*, vol. 3, no. 12, pp. 882–897, Dec. 2004.

[30] Q. Xu, B. Schmidt, S. Pradhan, and M. Lipson, "Micrometre-scale silicon electro-optic modulator," *Nature*, vol. 435, pp. 325–327, May 19, 2005.

[31] A. Pattavina, *Switching Theory—Architecture and Performance in Broadband ATM Networks*. New York: Wiley, 1998.

[32] Mellanox InfiniScale III Switch Overview [Online]. Available: http://www.mellanox.com/products/switch silicon.php

[33] T. Lin, K. A. Williams, R. V. Penty, I. H. White, and M. Glick, "Capacity scaling in a multihost wavelength-striped SOA-based switch fabric," *J. Lightw. Technol.*, vol. 25, no. 3, pp. 655–663, Mar. 2007.

[34] L. A. B. Windover *et al.*, "Parallel-optical interconnects >100 Gb/s," *J. Lightw. Technol.*, vol. 22, no. 9, pp. 2055–2073, Sep. 2004.

[35] B. E. Lemoff, M. E. Ali, G. Panotopoulos, G. M. Flower, B. Madhavan, A. F. J. Levi, and D. W. Dolfi, "MAUI: Enabling fiber-to-the-processor with parallel multiwavelength optical interconnects," *J. Lightw. Technol.*, vol. 22, no. 9, pp. 2043–2054, Sep. 2004.

[36] A. Narasimha, B. Analui, Y. Liang, T. Sleboda, and C. Gunn, "A fully integrated 4 × 10 Gb/s DWDM optoelectronic transceiver in a standard 0.13 μm CMOS SOI," presented at the Int. Solid State Circuits Conf., Feb. 2007, paper 2.1.

[37] Infinera DTN Product Brief [Online]. Available: http://www.infinera.com/products/dtn.html

[38] M. J. Connelly, *Semiconductor Optical Amplifiers*. Boston, MA: Kluwer Academic, 2002.

[39] O. Liboiron-Ladouceur, B. A. Small, and K. Bergman, "Physical layer scalability of WDM optical packet interconnection networks," *J. Lightw. Technol.*, vol. 24, no. 1, pp. 262–270, Jan. 2006.

Assaf Shacham (S'03–M'07) received the B.Sc. degree (*cum laude*) in computer engineering from the Technion, Israel Institute of Technology, Haifa, Israel, in 2002, and the M.S. and Ph.D. degrees, both in electrical engineering, from Columbia University, New York, in 2004 and 2007, respectively. His doctoral research work focused on architectures of optical interconnection networks for high performance computing.

He has published more than 25 papers in peer-reviewed journals and conferences. He is now employed by Aprius Inc., Sunnyvale, CA, where he is engaged in developing high-performance computer interconnects.

Keren Bergman (S'87–M'93–SM'07) received the B.S. degree from Bucknell University, Lewisburg, PA, in 1988, and the M.S. and Ph.D. degrees from the Massachusetts Institute of Technology, Cambridge, in 1991 and 1994, respectively, all in electrical engineering.

She is a Professor of electrical engineering at Columbia University where she also directs the Lightwave Research Laboratory. She leads multiple research projects in optical packet switched networks, distributed grid computing over optical networks, photonic interconnection networks, nanophotonic networks-on-chip, and the applications of optical networking in high-performance computing systems.

Dr. Bergman is a Fellow of OSA. She is currently Associate Editor for IEEE PHOTONIC TECHNOLOGY LETTERS and the Editor-in-Chief for the *OSA Journal of Optical Networking*.