A Fully Implemented 12 × 12 Data Vortex Optical Packet Switching Interconnection Network

Assaf Shacham, *Student Member, IEEE*, Benjamin A. Small, *Student Member, IEEE*, *Student Member, OSA*, Odile Liboiron-Ladouceur, *Student Member, IEEE*, and Keren Bergman, *Member, IEEE*, *Fellow, OSA*

***Abstract*—A fully functional optical packet switching (OPS) interconnection network based on the data vortex architecture is presented. The photonic switching fabric uniquely capitalizes on the enormous bandwidth advantage of wavelength division multiplexing (WDM) wavelength parallelism while delivering minimal packet transit latency. Utilizing semiconductor optical amplifier (SOA)-based switching nodes and conventional fiber-optic technology, the 12-port system exhibits a capacity of nearly 1 Tb/s. Optical packets containing an eight-wavelength WDM payload with 10 Gb/s per wavelength are routed successfully to all 12 ports while maintaining a bit error rate (BER) of 10^{-12} or better. Median port-to-port latencies of 110 ns are achieved with a distributed deflection routing network that resolves packet contention on-the-fly without the use of optical buffers and maintains the entire payload path in the optical domain.**

***Index Terms*—Interconnection networks (multiprocessor), optical interconnections, packet switching, photonic switching systems, wavelength division multiplexing.**

I. Introduction

OPTICAL packet interconnection networks have been suggested as possible solutions to the interchip communications bottleneck in high-performance computing systems (HPCS). As communications between processors and memory elements is rapidly becoming the main challenge in the design of next-generation HPCS, electronic interconnection networks may not be able to address the latency, scalability, and throughput requirements [1], [2].

Contemporary processors are capable of working at gigafloating-point operations per second (GFLOPS), and high-speed memory elements can be written and read at data rates of hundreds of gigabits per second. It is well recognized that the performance bottleneck is shifting towards the data exchange medium between processors and memory elements in multiprocessor systems [3]. Increasing the pin count of electronic integrated circuits is becoming more demanding, while growing data rates lead to increased power consumption of communication chips and greater difficulty in transmitting high-speed electronic signals over distances of tens of meters. These factors clearly render future electronic interconnection networks complex to design and expensive to manufacture. Fiber-optic technology may be a preferable transmission medium for multiprocessor HPCS interconnects [4]–[11].

Semiconductor optical amplifiers (SOAs) offer substantial gain, sub-nanosecond switching time, low latency, and relatively uniform gain across the International Telecommunications Union (ITU) C-band. They have therefore been utilized as switching gates in optical packet switching (OPS) networks [4], [7]–[13]. It has been shown that wavelength division multiplexing (WDM) optical packets can be transmitted through many SOAs while maintaining sufficient signal integrity [14], [15].

A suitable optical interconnection network is able to provide the high bandwidth and ultralow latency required for a multiprocessor HPCS [2], [4], [5]. The data vortex architecture has been presented (see, for example, [11], [16]) as an architecture for OPS systems that provides ultrahigh bandwidth and nearly optical time-of-flight limited latency. Furthermore, topological simulations have shown that the data vortex can scale to thousands of ports while its latency is proportional to $O(\log N)$, where N is the number of ports [16], [17]. This high degree of scalability is possible because the architecture is composed of discrete and independent simple OPS nodes that are capable of routing synchronous time-slotted wavelength-parallel WDM packets. Another notable feature of the data vortex architecture is its ability to resolve packet contentions on-the-fly without utilizing optical buffers [11], [16]–[18].

Multiple-wavelength OPS networks have the potential for providing very high transmission bandwidths. Although wavelength-parallel structures can be problematic in long-haul telecommunications applications due to chromatic dispersion and other effects, HPCS interconnection networks are typically confined to length scales shorter than 100 m. At these propagation distances, chromatic dispersion and other fiber nonlinearities are insignificant [2], [19]. Further, in a self-contained OPS network, packet synchronization can be achieved by controlling the timing of packet sources electronically, thus avoiding complex optical processing. This approach also greatly simplifies switching node design and allows for the implementation of a distributed packet routing architecture.

Previous literature has presented the data vortex architecture with basic performance analysis [16]–[18]. Other literature has discussed individual switching node design, implementation, and functionality [20], [21], and subsystems containing eight switching nodes [22], [23] have been demonstrated. The scalability of multinode systems has also been investigated from a physical perspective [14], [15].

Manuscript received December 1, 2004; revised July 11, 2005. This work was supported in part by the National Science Foundation under Grant ECS-0322813 and by the U.S. Department of Defense under subcontract B-12-644.

The authors are with the Department of Electrical Engineering, Columbia University, New York, NY 10027 USA (e-mail: assaf@ee.columbia.edu).

Digital Object Identifier 10.1109/JLT.2005.856242

This paper presents a fully implemented data vortex OPS interconnection network, as introduced in [11]. The implemented system is comprised of 36 SOA-based switching nodes. It demonstrates complete packet routing functionality from 12 input ports to 12 output ports for data packets containing eight payload wavelengths modulated at 10 Gb/s each, with median latencies of approximately 115 ns, while maintaining a bit error rate (BER) of 10^{-12} or better.

The remainder of this paper is organized as follows. Section II provides an overview of the data vortex architecture. Section III presents the design and implementation specifics of the 36-node data vortex system. The test bed designed to test and demonstrate the system capabilities is described in Section IV. Section V discusses three experiments that demonstrate the correct routing functionality of the network, the contention resolution through deflection routing, and the integrity of the routed payload data. Finally, conclusions are discussed in Section VI.

II. Architecture Overview

The data vortex is a distributed deflection routing interconnection network architecture designed to fully exploit the properties of fiber-optic technology in order to achieve ultrahigh bandwidth, low latency, and a high degree of scalability. The data vortex topology is comprised of simple 2×2 single-packet optical switching nodes often visualized as a set of concentric cylinders (Fig. 1) [16]. A data vortex switching fabric is defined by three topological parameters: C, the number of cylinders; H, the number of nodes along a cylinder height (i.e., the cylinder height); and A, the number of nodes along a cylinder circumference (i.e., the angle). Hence, every switching node is identified by the triplet (c, h, a), $0 \leq c < C$, $0 \leq h < H$, $0 \leq a < A$, denoting its location within the system. The number of cylinders scales with the height parameter as $C = \log_2 H + 1$, and the number of nodes per cylinder is $A \times H$. Since packets are injected into the nodes of the outermost cylinder and are ejected from the nodes of the innermost cylinder, a data vortex of $N_t = A \times H$ input and output ports will have $N = A \times H \times C = A \times H \times (\log_2 H + 1)$ nodes and the size of the switch N scales with the number of ports as $N_t \log_2 N_t$.

Switching nodes are interconnected using a set of ingression fibers, which connect nodes of the same height in adjacent cylinders, and deflection fibers, which connect nodes of different heights within the same cylinder. The ingression fibers must be of the same length throughout the entire system, as must be the deflection fibers. The deflection fibers' height crossing patterns (Fig. 1) direct packets through different height levels at each hop to enable banyan routing (e.g., butterfly, shufflenet) to a desired height, and assist in balancing the load throughout the system, mitigating local congestion [16], [17].

Incoming packets are injected into the nodes of the outermost cylinder and propagate within the system in a synchronous time-slotted fashion. During each time slot, each node either processes a single packet or remains inactive. As a packet enters node (c, h, a), the cth bit of the packet header is compared to cth most significant bit in the node's height coordinate (h).

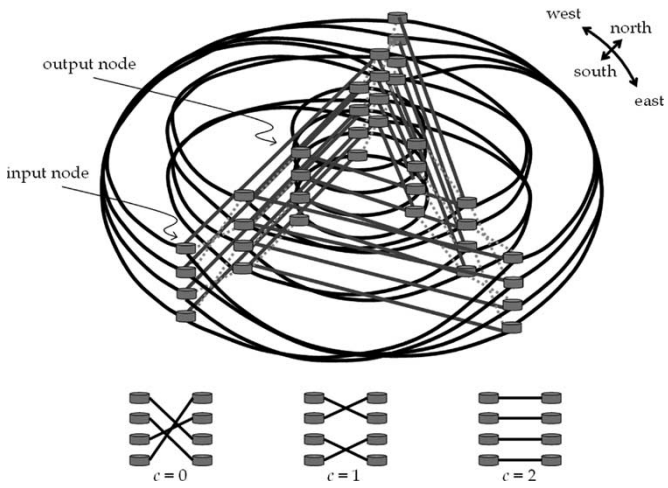

Fig. 1. Diagram illustrating the data vortex topology. A data vortex of $C = 3$, $H = 4$, and $A = 3$ (top), with height crossing patterns of the three cylinders (bottom). Curved lines are deflection fibers, straight lines are ingression fibers, and dotted lines are electronic deflection signal control cables.

If the bits are equal, the packet ingresses to node $(c + 1, h, a + 1)$. Otherwise, it is routed within the same cylinder to node $(c, G_C(h), a + 1)$, where $G_C(h)$ is a transformation that expresses the above-mentioned height crossing patterns [17]. Thus, packets progress to a higher cylinder only when the cth address bit matches in a manner that preserves the $c - 1$ most significant bits. In this distributed scheme, a packet is routed by decoding its address in a bitwise banyan manner to its destination height.

Once a packet reaches its destination height, in the innermost cylinder, it is routed to the desired system output switching node according to the node's angle parameter (a). The angular routing can be implemented in two ways: 1) encoding the angle part of the header by assigning a single bit per angle and using the same single-bit address matching algorithm as used in the height-resolution switching nodes, or 2) using a more compressed angle encoding that allows for a shorter header but requires a more complicated decoding logic for the inner cylinder nodes. In either case, if the angle of the node matches the angle encoded in the packet header, it is routed to the output port; otherwise, it is forwarded to the subsequent node in the same height and cylinder and undergoes the same angle matching until it reaches its destination node.

Contentions are resolved in the data vortex by way of deflection signaling between adjacent nodes. Because each switching node has two input ports but is designed to process only one packet per time slot, this mechanism prevents two packets from reaching a node at the same slot. Whereas conventional deflection routing is implemented internally for each output port within individual nodes, the data vortex approach resolves contentions between switching nodes, thus requiring simple signaling between adjacent nodes. The deflection rule gives priority to packets traversing a cylinder over packets trying to ingress into that cylinder [16], [18], [21]. The deflection signals are transmitted over electronic cables that connect adjacent nodes. Specifically, node A in cylinder c has two inputs: node B in the same cylinder and node C in cylinder $c - 1$ (Fig. 2). Whenever node B receives a packet that should be routed to node A, the

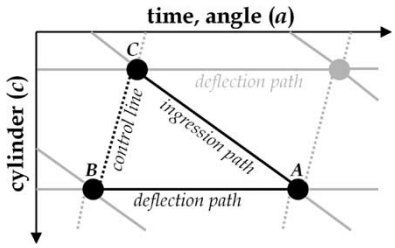

Fig. 2. Schematic representation of the deflection triangle. In order to avoid packet collision at node A, the electronic deflection signal, transmitted by node B on the control line, must reach switching node C in time for C to route a packet to the deflection path instead of the ingression path. This structure is repeated throughout the network between every pair of adjacent cylinders regardless of the angle coordinate.

packet is forwarded to node A on the deflection fiber, and node B sends an electronic deflection signal to node C. If, at the same timeslot, node C receives another packet that should also be routed to node A, the received electronic deflection signal prevents this packet from being forwarded to A on the ingression fiber. Thus, the packet is deflected to another node, and node C generates an appropriate deflection signal to ensure that this packet does not collide at the next hierarchical deflection triangle, which is congruent to the triangle ABC. This triangular node arrangement is repeated throughout the data vortex architecture so that every switching node has exactly one switching node available for deflection [16], [18], [21]. Deflected packets statistically make two extra hops, on average, but are eventually routed to their destinations through different paths [16], [17]. Thus, the contentions are resolved without buffers at the price of a slight increase in the routing latency, which is comparable to the latency penalty introduced by store-and-forward routing. The high degree of path diversity presented in the data vortex reduces the latency penalty that is inherent to deflection routing [3].

The deflection mechanism places some timing constraints on the system design, as the deflection signal must be received in time to be used in the routing decision process. That is, using the previous example (Fig. 2), node C must receive both the deflection signal and the packet concurrently. Therefore, node B must receive and forward its packet before node C does so that B has time to transmit the deflection signal to C. This requires that the deflection path from B to A be longer than the ingression path from C to A such that packets maintain synchronization at every switching node, regardless of the packets' origin. More generally, the deflection fiber latency must be equal to the sum of the latencies of the deflection signaling cable, the switching node processing time, and the ingression fiber. This triangular timing requirement can be met by correctly designing the deflection paths' optical fibers to have a particular length, which is the same throughout the system; the ingression paths' fibers are also uniform in length, as are the electronic deflection signal control lines. The three lengths found in the triangular deflection structure (Fig. 2) are repeated throughout the network, regardless of network size and regardless of the network dimensions (C, H, A) [16], [18]. Since all switching nodes are identical in their processing latency, the triangular timing requirement is easy to maintain

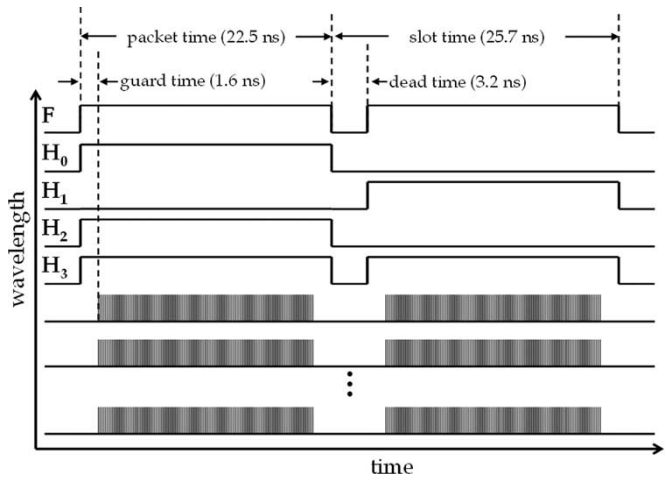

Fig. 3. Timing diagram of the wavelength-parallel packets used in the system. The packets consist of control wavelengths (F, H_0, H_1, H_2, H_3) that remain constant throughout the duration of a packet in addition to multiple payload wavelengths modulated at a high rate. The packet timing parameters are indicated: slot time, packet time, guard time, and dead time.

by choosing all the ingression fibers and control cables to be of the shortest possible length that can support the physical layout of the system, and setting the deflection fibers' lengths accordingly.

The multiple-wavelength packet structure allows for simple processing and high bandwidth (Fig. 3). A wavelength-parallel packet contains a limited number of wavelengths that are allocated for control bits, which encode the header address and framing bits. These control bits, encoded on one wavelength each, are the only information used during the routing process and remain fixed over the duration of the packet. At each routing node, only two filters and two simple low-speed receivers are required to decode the relevant control information [20], [21]. The remainder of the transmission band is exploited for the packet payload, which is segmented and encoded on multiple wavelengths at higher data rates, thus offering an ultrahigh transmission bandwidth. The multiple-wavelength packet payload is detected only at the final destination, where the multiple wavelengths are demultiplexed and received electronically.

The switching node's required functionality is limited to the routing of a packet to one of its two output ports based on the processing of two control bits and a deflection signal. The simple logic operation required in the node simplifies the address resolution processing requirements and allows for the implementation of simple low-latency nodes [20], [21]. The packets are transparently routed as a single structure to one of the two output ports as determined by the routing decision logic, preserving all control and payload information. No header information is lost or added, and no wavelength conversion is required. The node's east port is connected to a deflection fiber, and the node's south port is connected to an ingression fiber (Fig. 4).

This architecture lends itself to implementation with commercially available electronic and fiber-optic technologies. The node-based design is scalable, allowing for the successful construction of a complete switching fabric.

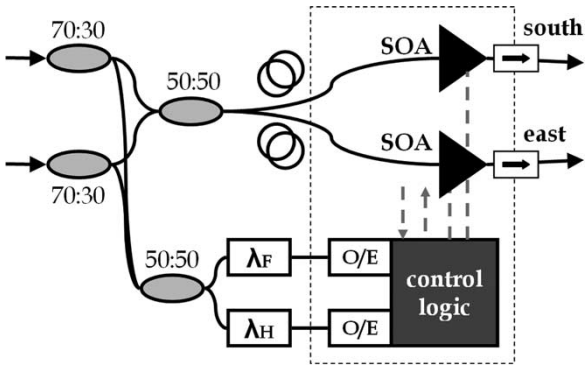

Fig. 4. Schematic diagram of the switching node constructed using optoelectronic and electronic devices that are integrated onto a PCB (within the dashed line) and passive optical components: couplers (ellipses), filters (λ_F, λ_H), and isolators (boxed arrows) [21].

III. Design and Implementation

The presented system is a fully implemented data vortex OPS interconnection network. It consists of three cylinders $(C = 3)$ with a height dimension (H) of 4 and an angle dimension (A) of 3. These dimensions yield a system port count of $N_t = A \times H = 12$ input ports and 12 output ports, and a node count of $N = A \times H \times C = 36$. Five wavelengths are encoded with the control information required to route packets in the system: a packet presence bit ("frame") allows for packet detection in the nodes and is encoded on C27 (1555.75 nm). The 4-bit header, required to map the 12 system outputs, is encoded on four header wavelengths: H_0 (C53, 1535.04 nm) and H_1 (C55, 1533.47 nm) are used for height resolution; H_2 (C33, 1550.92 nm) and H_3 (C58, 1531.12 nm) use 2-bit encoding for angle resolution in the innermost cylinder $(c = 2)$ rather than the single bit matching used for height resolution in other cylinders. In this 2-bit code, angle $a = 0$ is encoded as [xx01], $a = 1$ is encoded as [xx10], and $a = 2$ is encoded as [xx11]; the code [xx00] is not used. This paper notates addresses in binary as $[H_0H_1H_2H_3]$. The 12 system output ports are therefore addressed as follows: [0001], [0010], [0011], [0101], [0110], [0111], [1001], [1010], [1011], [1101], [1110], [1111].

The system is comprised of 36 switching nodes [20], [21]. The nodes are integrated onto six printed circuit boards (PCBs) and passive optics modules, and are organized in racks to allow for efficient interconnection by short fibers, according to the considerations mentioned in Section II. Each switching node contains two optical input ports and two optical output ports, accessible with optical fiber pigtails. The nodes use control information encoded in an incoming packet, and a received deflection signal, to make the routing decision and to generate an output deflection signal.

Each switching node is divided into two physical modules [20], [21] (Fig. 4). The passive optics module consists of four fiber-optic couplers and two wavelength filters, integrated with optical fibers into a 3 in $\times$ $4\frac{1}{2}$ in $\times$ $\frac{1}{2}$ in plastic box. Upon entering the node from either input port, 30% of the input packet's power is extracted for control signal decoding. Two relevant control wavelengths, frame or headers, are filtered with 100-GHz optical passband filters and directed to photodetectors. The wavelengths filtered correspond to the exact header bits used in the particular node, according to its c coordinate. The rest of the packet is routed through a fixed length optical delay line and a 50:50 coupler, which couples the two input ports and directs them to the SOAs. The SOAs are followed by isolators, used to mitigate the back-propagation of amplified spontaneous emission (ASE) noise from downstream nodes. The coupling and connector losses of the packet path and the control wavelength detection path in the node are 5.1 dB and 9.6 dB, respectively.

A 3 in $\times$ 6 in standard PCB integrates the remaining components: two low-speed optical receivers that are used to detect the isolated control wavelengths, electronic decision circuitry, and two SOA gates that execute the routing decision and compensate for coupling losses (Fig. 5). The electronic decision circuit is implemented with high-speed positive coupled emitter logic (PECL) logic gates in order to achieve minimal latency. Both of the detected signals are matched against a preprogrammed user-controlled value, and the outcomes are used, along with a received deflection signal, to determine which output port is activated. If both matches are successful, and if a deflection signal is not received, the packet is routed to the south port. If one of the matches is not successful or if a deflection signal is received, the packet is routed to the east port and a deflection signal is transmitted to the appropriate node [20], [21].

Optical receivers detect the information encoded on the previously isolated control wavelengths using p-i-n photodetectors designed with a minimum average power sensitivity of -26 dBm at 155 Mb/s. The low frequency operation is sufficient for the system's relatively low packet rate. Because of the bursty nature of the control signals, determined by the packet arrival rate, the receiver data path is designed to be dc coupled in order to eliminate the common-mode drift. A transimpedance amplifier (TIA) and a limiting amplifier follow the p-i-n. The limiting amplifier rectifies the differential signal to PECL levels, as required by the electronic circuitry. The offset correction feedback circuitry of the limiting amplifier assumes an even duty cycle for the received signal in order determine its decision threshold and is therefore bypassed. Instead, a fixed threshold voltage is set externally, depending upon the control information input power.

The routing decision is executed by SOA gates, powered by laser drivers that are controlled by the electronic circuitry. The SOAs are commercially available devices with a noise figure of 6.5 dB and an unsaturated input power of approximately 0 dBm at an operating current of 50 mA. Based on the routing decision, only one of the SOAs is driven, providing the gain required to compensate for the coupling losses (5.1 dB, as mentioned above) [20], [21]. The second SOA, which receives no current, blocks the packet from leaking into the unused output port with a switching ratio of more than 50 dB. The SOAs are switched with rise and fall times of nearly 1 ns. The exact drive current of the SOAs is set based upon the experimental losses of each node.

Several other electronic devices such as tunable delay integrated circuits, connectors, and potentiometers are integrated onto the node board to allow close control of the processing latency, the SOA gain, and the receiver threshold voltages (Fig. 5). The switching node total latency is 15.8 ns, 4.3 ns of

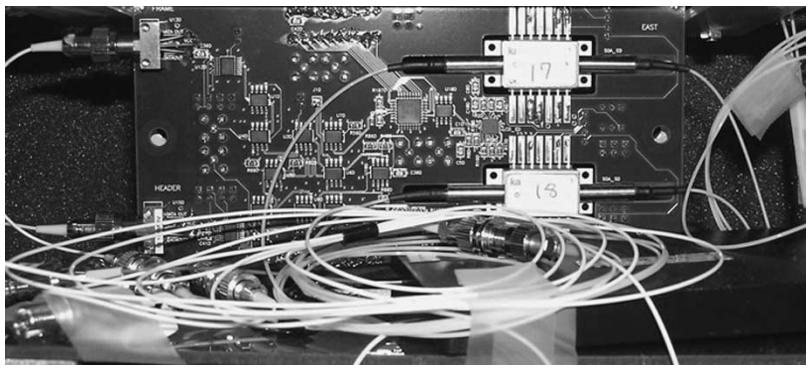

Fig. 5. Photograph of the switching node. Two photodetectors are on the left (for the frame and header wavelengths). The control electronics is comprised of high-speed discrete logic gates. The SOAs are near the right edge of the PCB. Passive optical components are contained within the plastic box on the bottom, which is connected using optical fibers to the photodetectors and SOAs.

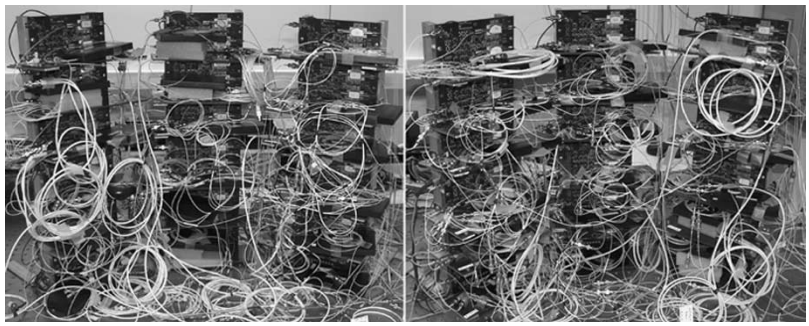

Fig. 6. Photograph of the implemented 12-port data vortex network. 36 switching nodes, implemented on PCBs and passive optics modules, are interconnected with optical fibers and electronic control cables to form a fully functional data vortex switching fabric. The nodes are divided into six PCB towers with six nodes in each tower. The towers are then divided into two groups according to the nodes' height coordinates. The input nodes (outer cylinder) are the lowermost and uppermost nodes; output nodes (inner cylinder) are located in the middle levels. Optical fibers are seen to connect nodes in accordance with the topological schematic shown in Fig. 1.

which is spent in the passive optics and the remaining 11.5 ns is dedicated to detection, processing, and the SOA rise time. The net electronic processing latency is 5.5 ns [20], [21].

In cylinders $c = 0$ and $c = 1$, the frame wavelength and a single header wavelength (H_c) are extracted. The frame is matched against a value of logical "1" to verify the existence of a packet, and the header is matched against a value equal to the cth most significant bit of the node's height. However, in cylinder $c = 2$, two header wavelengths H_2 and H_3 are extracted and matched against a 2-bit code that represents the node angle, as described in Section II.

For convenience and controllability, the nodes are integrated in groups of six onto 20 in $\times$ 6 in PCBs. The six printed circuit boards along with the passive optics boxes are arranged in six racks of six nodes each (Fig. 6). The six nodes integrated on each board all belong to the same angle, two adjacent height levels, and all three cylinders. The rationale behind the grouping is the minimization of the lengths of the control cables that connect nodes of the same angle and different height levels. In order to minimize the ingression fibers' lengths and thus further relax the triangular timing constraint, as discussed in Section II, the racks are organized in two sets of 18 nodes each, all belonging to the same two adjacent height levels. Only deflection fibers connect nodes of different halves (nodes at $h = 0,1$, and nodes at $h = 2,3$) and link one set to the other (Fig. 7).

The progression fibers are 70 cm long (latency of 3.4 ns) and the electronic deflection cables are 15 cm long (1.0 ns). Since the net processing time of the nodes is measured to be 5.5 ns, in order to comply with the triangular timing constraint and to allow deflection signals to reach the nodes in time, the deflection fibers' lengths must be 200 cm (latency of 9.9 ns). Finally, a key parameter in the system is its slot time, which must be equal to an eastward node hop, the time it takes a packet to propagate through a switching node and the subsequent deflection fiber. The latency of the switching node is 15.8 ns, so the slot time is chosen to be 25.7 ns [20], [21].

IV. Test Bed

In order to test and demonstrate the full functionality of the implemented 12 $\times$ 12 data vortex switching network, optical packet generation and analysis subsystems are assembled from conventional fiber-optic components (Fig. 8). The payload wavelengths are generated by eight discrete distributed feedback (DFB) laser sources and are modulated with a 10 Gb/s pseudo-random bit sequence (PRBS) of length 2^7 - 1 by a single $LiNbO_3$ modulator. The eight wavelengths span 23.2 nm in the C-band, from 1536.6 nm (C51) to 1559.8 nm (C22), and contain pairs of wavelength with 0.8-nm WDM spacing (Fig. 9). The payload wavelengths are then decorrelated with a 24-km length of optical fiber by approximately

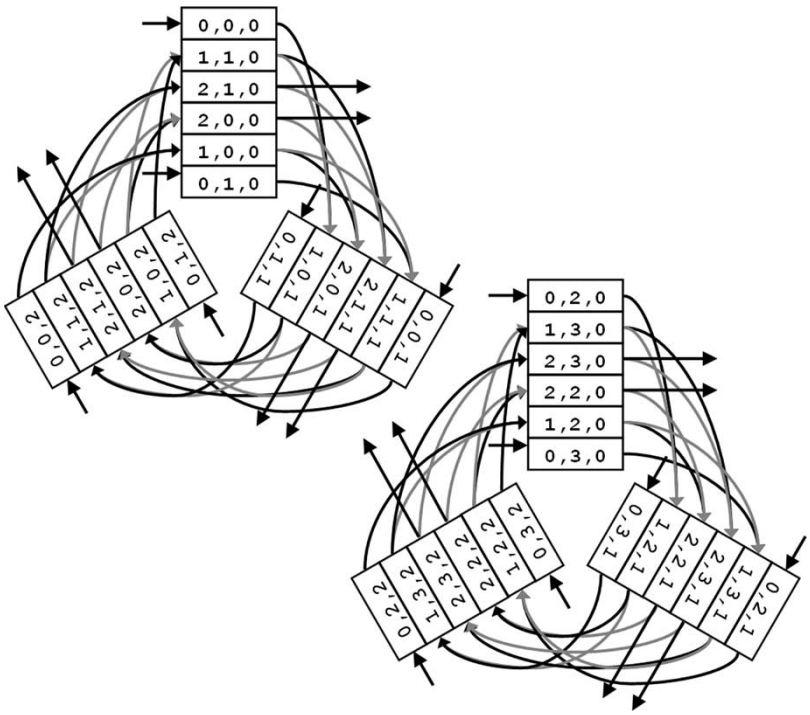

Fig. 7. Diagram of the 36-node network organized as two groups. Progression fibers are curved black arrows and deflection fibers are curved gray arrows. The deflection fibers that connect one group to another are omitted for clarity. The system input and output ports are represented by straight black arrows. The node labels denote the system coordinates (c, h, a).

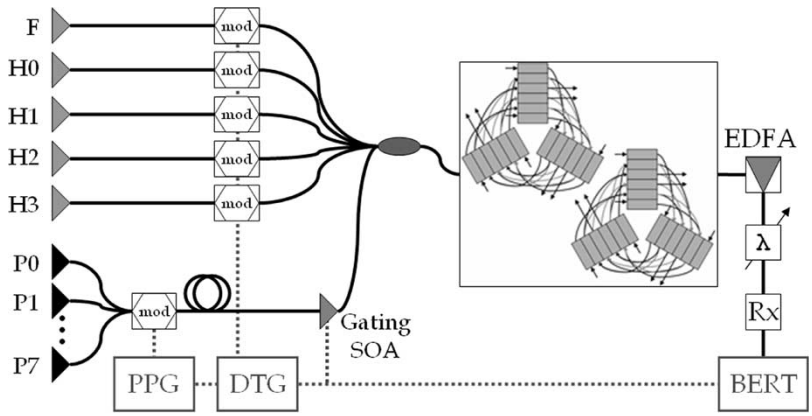

Fig. 8. Schematic diagram of the testbed. $LiNbO_3$ modulators (mod), SOAs, pulse pattern generator (PPG), DTG, p-i-n-TIA receiver (Rx), and tunable filter (λ).

450 ps/nm (more than 3 bits between the closest payload wavelengths). After the continuous data streams emerge from the decorrelator, they are segmented into packets by a gating SOA that is controlled by the data timing generator (DTG). Five additional laser sources at the control wavelengths are simultaneously modulated with the frame and header signals, also driven by the DTG.

The packetized payload and control wavelengths are then multiplexed together and aligned to form complete packets. Each packet's length is 22.5 ns, including a guard time of 1.6 ns that is inserted at both the beginning and the end of the packet, yielding a net payload duration of 19.3 ns. The guard time is necessary to allow for the SOA switching rise and fall times. The packets are spaced by a dead time of 3.2 ns to accommodate timing discrepancies in the switching nodes. Thus, the total packet slot time in the system is 25.7 ns, as discussed in Section III (Fig. 3).

The power levels are determined to ensure error-free detection of the control signals and to provide a sufficient optical signal-to-noise ratio (OSNR) for the payload wavelengths. An upper limit on the total packet power is set by the switching SOAs' input saturation power, which is approximately 0 dBm for the devices used in the system. The SOAs must be kept in the linear regime to avoid crosstalk between adjacent wavelengths due to cross gain modulation (XGM). In order to meet the sensitivity requirement of the nodes' receivers, the control wavelengths' peak power is approximately -10 dBm at the data vortex inputs (surpassing the -26 dBm receiver sensitivity, including the 9.5-dB loss in the nodes' passive optics). An error-free detection of the control signals in the

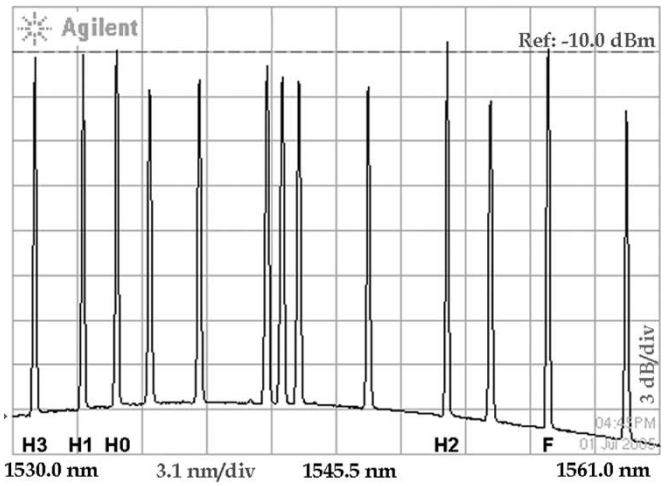

Fig. 9. Spectrum of the packets at the network output. The 13 signals include five control wavelengths, annotated frame (F) and headers (H0 through H3), and eight payload wavelengths. The average packet power levels shown include different duty factors for payload wavelengths and for different control wavelengths, which result from the specific packet sequence.

switching nodes is very important because any noise in the node's photodetectors is manifested as noise in the SOAs' gating signals, which could lead to payload data errors. The payload wavelengths are transmitted with an average packet power of about -13 dBm at the inputs so the maximum total packet power, as seen by the SOAs, is approximately -1 dBm. These power levels are maintained from node to node because the SOAs compensate for the nodes' internal losses, allowing for propagation through multiple successive switching nodes.

Upon emerging from the system, the payload wavelengths are preamplified with an erbium-doped fiber amplifier (EDFA), filtered to select a particular wavelength, and directed to a dc-coupled 10.7 Gb/s p-i-n TIA receiver module, which is capable of receiving packetized data. A BER tester (BERT) is used to measure the error rate of the data contained within the individual payload wavelengths in the packets. The BERT receives packets periodically, so it must also receive an external gating signal synchronized with the input packet generator in order to measure the cumulative BER of the packetized payload data. The gating ensures that the BERT measures only the BER within packets and ignores the dead time between packets. A communication signal analyzer is also used to measure the signal waveforms and obtain eye diagrams.

V. Experimental Results

Three experiments demonstrate the correct addressing functionality of the system, its contention resolution via deflection routing, and the signal integrity of the data routed through it. In the first experiment, packets are sent to all 12 system output ports from a single input port, their correct delivery verified, and their latencies measured. In the second experiment, two packets, which are addressed to the same system output, are injected simultaneously; the contention is resolved internally; and both packets reach the destination port at different times, as predicted. The third experiment measures the BER of the eight payload wavelengths after being routed through two paths comprised of three and seven switching nodes.

In order to clarify the data vortex's packet routing mechanism, an example of a packet's path through the system follows. The packet is injected at system input #1 at node ($c = 0$, $h = 0$, $a = 0$) and is addressed to system output [1111] at node ($c = 2$, $h = 3$, $a = 2$). At the input node (0,0,0), the frame and H_0 are decoded; because the packet's header $H_0 = 1$ and because the node's most significant height bit is "0" ($h = 0 = 00_2$), the values do not match and the packet is routed east to node (0,2,1). At this node, the most significant height bit is equal to "1" ($h = 2 = 10_2$) and matches H_0; therefore, the packet is routed south and ingresses to node (1,2,2). Here, H_1 (which is "1" for this packet) is compared to the node's second most significant height bit, which is "0"; the packet is therefore routed east to node (1,3,0). At this node, the second most significant height bit is "1"; the successful match results in an ingression (south) to node (2,3,1). In cylinder $c = 2$, the innermost cylinder, H_2 and H_3 are extracted for angular resolution routing. Because this packet has $H_2H_3 =$ "11", in node (2,3,1), where the angle value is $a = 1$ (encoded as "10"), the values of the packet's header and the node angle do not match, so the packet is routed east. Finally, at node (2,3,2), the angle value ($a = 2 = 11_2$) matches the packet header and the packet is routed south and exits the network. The packet's path in the system can be notated in shorthand as $(0,0,0)_{\mathrm{E}} \rightarrow (0,2,1)_{\mathrm{S}} \rightarrow (1,2,2)_{\mathrm{E}} \rightarrow (1,3,0)_{\mathrm{S}} \rightarrow (2,3,1)_{\mathrm{E}} \rightarrow (2,3,2)_{\mathrm{S}}$.

A. Routing Experiment

A stream of 12 packets, each addressed to one of the 12 system output ports, is injected into system input #1 at node (0,0,0). After being routed through different paths within the system, the packets reach their destination ports with latencies varying from three hops (60 ns) to seven hops (160 ns). It should be mentioned that every packet takes exactly three south hops, which are 6.4 ns shorter than the east hops, as discussed in Section II. Fig. 10 illustrates the wavelength-parallel control signals of the injected packets and the resulting ejected payloads. For example, the fifth injected packet, addressed with [1001], makes one east hop in the outermost cylinder ($c = 0$) before ingressing to cylinder $c = 1$, from which it continues to ingress to cylinder $c = 2$ and exits immediately at angle $a = 0 : (0,0,0)_{\mathrm{E}} \rightarrow (0,2,1)_{\mathrm{S}} \rightarrow (1,2,2)_{\mathrm{S}} \rightarrow (2,2,0)_{\mathrm{S}}$. Because the number of routing hops required by each packet varies, it is possible for two packets in the sequence (Fig. 10) to exit different nodes at the same time.

B. Contention Resolution Experiment

The unique contention resolution mechanism inherent to the data vortex architecture is demonstrated in this experiment. Two packets with the same destination address are simultaneously injected at two different input ports. The system input ports selected are system input #4 at node (0,1,0) and system input #7 at node (0,2,0). The latency from each of the selected system inputs to the destination system output at node (2,3,0), addressed [1101], is four hops. In the absence of the other packet, each

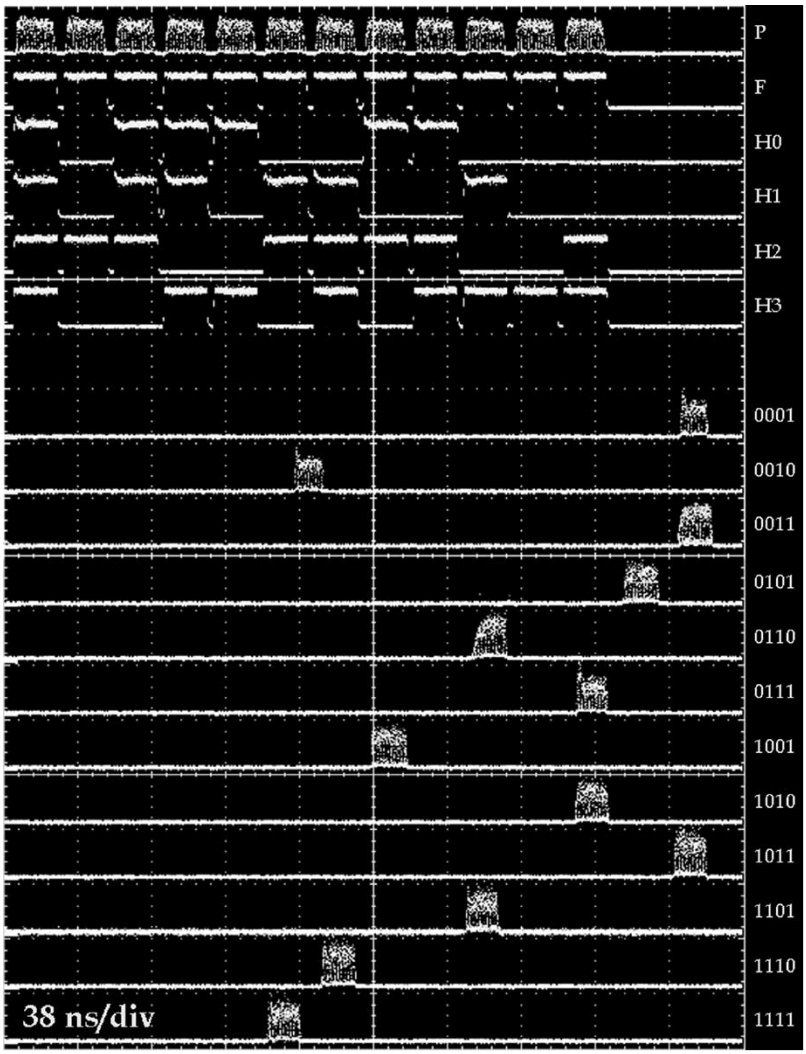

Fig. 10. Waveforms of input and output signals for the routing verification experiment. From the top down: packet payload (P), frame (F), and headers (H0 through H3), followed by the payload at the 12 output ports, annotated by their respective header encodings. The latencies of the packets vary between 60 ns (three hops) and 160 ns (seven hops).

packet would take a four-hop path and emerge out of the system output after 82 ns: $(0,1,0)_{\mathrm{E}} \rightarrow (0,3,1)_{\mathrm{S}} \rightarrow (1,3,2)_{\mathrm{S}} \rightarrow (2,3,0)_{\mathrm{S}}$ from system input #4, and $(0,2,0)_{\mathrm{S}} \rightarrow (1,2,1)_{\mathrm{E}} \rightarrow (1,3,2)_{\mathrm{S}} \rightarrow (2,3,0)_{\mathrm{S}}$ from system input #7. The deflection routing mechanism resolves the contention internally, causing the packet from system input #4 to be deflected with a three hop penalty: $(0,1,0)_{\mathrm{E}} \rightarrow (0,3,1)_{\mathrm{E}} \rightarrow (0,0,2)_{\mathrm{E}} \rightarrow (0,2,0)_{\mathrm{S}} \rightarrow (1,2,1)_{\mathrm{E}} \rightarrow (1,3,2)_{\mathrm{S}} \rightarrow (2,3,0)_{\mathrm{S}}$. In Fig. 11, the first packet appears at the output port after the originally predicted latency of 82 ns, whereas the other packet emerges after 158 ns, which corresponds to seven hops. In this situation, the packets compete for node (1,3,2), but the first packet, which comes from (1,2,1), has priority and deflects the packet from node (0,3,1) to node (0,0,2).

C. Payload Integrity Experiment

The integrity of the packet payload while being routed through the network is demonstrated by two routing paths of three and seven node hops, which represent the shortest and longest paths, as discussed above. A stream of packets addressed to output port [1111] at node (2,3,2) is injected into the system input #10 at node (0,3,0) and therefore passes through three nodes: $(0,3,0)_{\mathrm{S}} \rightarrow (1,3,1)_{\mathrm{S}} \rightarrow (2,3,2)_{\mathrm{S}}$. The BER of these packets is measured to be better than 10^{-12} on all eight payload wavelengths. Similarly, another stream of packet address to output port [1111] is injected into the system input #3 at node (0,0,2). The packets in this stream require seven node hops to reach the output: $(0,0,2)_{\mathrm{E}} \rightarrow (0,2,0)_{\mathrm{S}} \rightarrow (1,2,1)_{\mathrm{E}} \rightarrow (1,3,2)_{\mathrm{S}} \rightarrow (2,3,0)_{\mathrm{E}} \rightarrow (2,3,1)_{\mathrm{E}} \rightarrow (2,3,2)_{\mathrm{S}}$. The BER of these packets is also measured to be better than 10^{-12} on all eight payload wavelengths. Fig. 12 shows eye diagrams of the 10 Gb/s data from three representative payload wavelengths that cover the packets' wavelength span.

This demonstration of error-free routing confirms that multiple-wavelength packets can be routed through multiple-hop paths in the system while maintaining optical signal quality. A more detailed discussion of the effects of packet propagation through similar SOA-based photonic switching networks is given in [15].

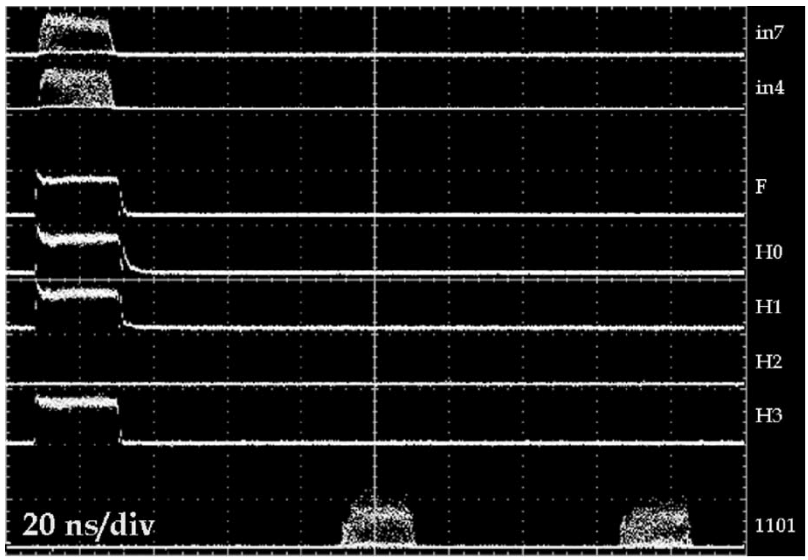

Fig. 11. Waveforms of input and output signals for the contention resolution and deflection signal verification experiment: the injected packets at the input ports with their payload and control signals (frame and header). The optical signal of both payloads at the output port verifies that the contention is resolved.

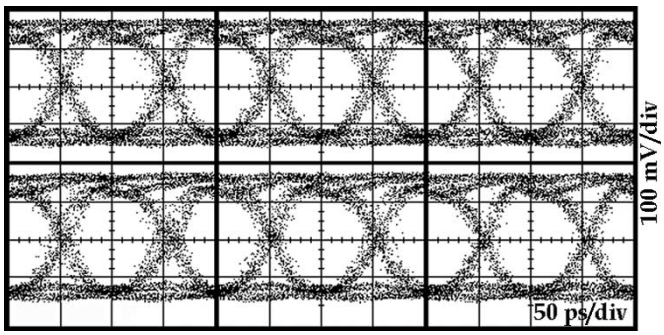

Fig. 12. Input (top) and output (bottom) electronic eye diagrams of three of the packet's payload signals, each modulated at 10 Gb/s, as seen after seven node hops. The wavelengths shown are 1538.98 nm (C48, left), 1543.73 nm (C42, center), and 1559.79 nm (C22, right). The vertical scale represents a voltage swing of 300 mV for a receiver single-ended port and 600 mV for the received differential signal.

VI. Conclusion

A fully functional 12-port optical packet interconnection network based on the data vortex architecture is demonstrated. The system performs address decoding, wavelength division multiplexing (WDM) packet routing, and local contention resolution. The design of the system and three experiments that demonstrate different aspects of its functionality are described in detail. All 12 output ports are addressed, and correct packet routing is verified with latencies near 100 ns. Contentions are resolved between switching nodes according to the data vortex internal deflection routing mechanism, and an 80 Gb/s multiple-wavelength payload is routed with a BER of 10^{-12} or better.

The successful operation of this system suggests that a larger data vortex switching fabric can be built and used as an interconnection network in a HPCS. The immense bandwidth afforded by utilization of WDM data encoding within every packet and the ultralow latencies measured in the conducted experiments demonstrate the clear advantages of the network over electronic alternatives.

Acknowledgment

The authors would like to thank J. P. Mack for his contribution to the assembly of the system.

References

[1] NRC, *The Future of Supercomputing: An Interim Report*. Washington, DC: National Academies Press, 2003.

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

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

[4] G. I. Papadimitriou, C. Papazoglou, and A. S. Pomportsis, "Optical switching: Switch fabrics, techniques, and architectures," *J. Lightw. Technol.*, vol. 21, no. 2, pp. 384–405, Feb. 2003.

[5] S. Yao, S. Dixit, and B. Mukherjee, "Advances in photonic packet switching: An overview," *IEEE Commun. Mag.*, vol. 38, no. 2, pp. 84–94, Feb. 2000.

[6] R. S. Tucker and W. D. Zhong, "Photonic packet switching: An overview," *IEICE Trans. Electron.*, vol. E82-C, no. 2, pp. 202–212, Feb. 1999.

[7] F. Masetti, D. Chiaroni, R. Dragnea, R. Robotham, and D. Zriny, "High-speed high-capacity packet-switching fabric: A key system for required flexibility and capacity," *J. Opt. Netw.*, vol. 2, no. 7, pp. 255–265, Jul. 2003.

[8] S. Araki, Y. Suemura, N. Henmi, Y. Maeno, A. Tajima, and S. Takahashi, "Highly scalable optoelectronic packet-switching fabric based on wavelength-division and space-division optical switch architecture for use in the photonic core node," *J. Opt. Netw.*, vol. 2, no. 7, pp. 213–228, Jul. 2003.

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

[10] T. Lin, K. A. Williams, R. V. Penty, I. H. White, M. Glick, and D. McAuley, "Self-configuring intelligent control for short reach 100 Gb/s optical packet routing," presented at the Optical Fiber Communication Conf., Anaheim, CA, 2005, Paper OWK5.

[11] B. A. Small, O. Liboiron-Ladouceur, A. Shacham, J. P. Mack, and K. Bergman, "Demonstration of a complete 12-port terabit capacity optical packet switching fabric," presented at the Optical Fiber Communication Conf., Anaheim, CA, 2005, Paper OWK1.

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

[13] K. A. Williams, G. F. Roberts, T. Lin, R. V. Penty, I. H. White, M. Glick, and D. McAuley, "Integrated optical 2×2 switch for wavelength multiplexed interconnects," *IEEE J. Sel. Topics Quantum Electron.*, vol. 11, no. 1, pp. 78–85, Jan./Feb. 2005.

[14] W. Lu, O. Liboiron-Ladouceur, B. A. Small, and K. Bergman, "Cascading switching nodes in data vortex optical packet interconnection network," *Electron. Lett.*, vol. 40, no. 14, pp. 895–896, Jul. 2004.
[15] O. Liboiron-Ladouceur, W. Lu, B. A. Small, and K. Bergman, "Physical layer scalability demonstration of a WDM packet interconnection network," in *Proc. 17th Annu. Meeting IEEE Laser and Electro-Optics Society (LEOS)*, Rio Grande, PR, 2004, pp. 567–568, Paper WM3.
[16] Q. Yang, K. Bergman, G. D. Hughes, and F. G. Johnson, "WDM packet routing for high-capacity data networks," *J. Lightw. Technol.*, vol. 19, no. 10, pp. 1420–1426, Oct. 2001.
[17] Q. Yang and K. Bergman, "Performances of the data vortex switch architecture under nonuniform and bursty traffic," *J. Lightw. Technol.*, vol. 20, no. 8, pp. 1242–1247, Aug. 2002.
[18] ——, "Traffic control and WDM routing in the data vortex packet switch," *IEEE Photon. Technol. Lett.*, vol. 14, no. 2, pp. 236–238, Feb. 2002.
[19] G. P. Agrawal, *Fiber-Optic Communication Systems*, 3rd ed. New York: Wiley, 2002.
[20] A. Shacham, B. A. Small, O. Liboiron-Ladouceur, J. P. Mack, and K. Bergman, "An ultra-low latency routing node for optical packet interconnection networks," in *Proc. 17th Annu. Meeting IEEE Laser and Electro-Optics Society (LEOS)*, Rio Grande, PR, 2004, pp. 565–566, Paper WM2.
[21] B. A. Small, A. Shacham, and K. Bergman, "Ultra-low latency optical packet switching node," *IEEE Photon. Technol. Lett.*, vol. 17, no. 7, pp. 1564–1566, Jul. 2005.
[22] W. Lu, B. A. Small, J. P. Mack, L. Leng, and K. Bergman, "Optical packet routing and virtual buffering in an eight-node data vortex switching fabric," *IEEE Photon. Technol. Lett.*, vol. 16, no. 8, pp. 1981–1983, Aug. 2004.
[23] B. A. Small, A. Shacham, K. Bergman, K. Athikulwongse, C. Hawkins, and D. S. Wills, "Emulation of realistic network traffic patterns on an eight-node data vortex interconnection network subsystem," *J. Opt. Netw.*, vol. 3, no. 11, pp. 802–809, Nov. 2004.

Assaf Shacham (S'03) was born in Israel in 1976. He received the B.Sc. (*cum laude*) degree in computer engineering from The Technion, Israel Institute of Technology, Haifa, Israel, in 2002, and the M.S. degree in electrical engineering from Columbia University, New York, in 2004. He is currently working towards the Ph.D. degree in electrical engineering at Columbia University.

From 1999 to 2001, he worked for Intel Inc., Haifa, Israel, as a Circuit Designer in the Mobile Products Group. He then joined Charlotte's Web Networks in 2001 and spent two years working as a Logic Design Engineer in the core router hardware group. His interests include architecture and design aspects of optical packet switching interconnection networks.

Benjamin A. Small (S'98) received the B.S. (with honors) and M.S. degrees in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA, in 2001 and 2002, respectively, and the M.Phil. degree in electrical engineering from Columbia University, New York, in 2005. He is currently working towards the Ph.D. degree in electrical engineering at Columbia University.

His interests include optoelectronic device physics and modeling as well as optical packet switching interconnection network traffic analysis and system-level behavior.

Odile Liboiron-Ladouceur (S'95) was born in Montréal, QC, Canada, in 1976. She received the B.Eng. degree in electrical engineering from McGill University, Montréal, QC, Canada, in 1999, and the M.S. degree in electrical engineering from Columbia University, New York, in 2003. She is currently working towards the Ph.D. degree in electrical engineering at Columbia University.

From 1999 to 2000, she worked for Teradyne Inc., as an Applications Engineer in the mass storage business unit. She then joined Texas Instruments in 2000 and spent two years working in the fiber-optic business unit as a Design Engineer. Her interests include optical interconnection physical layer analysis and clock and data recovery solutions for optical packet switching interconnection networks.

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

She is an Associate Professor of Electrical Engineering at Columbia University, New York, NY. Her research interests include ultrafast optical signal processing, optical packet switching, wavelength division multiplexed (WDM) networks, and optical interconnection networks.

Prof. Bergman is a Fellow of the Optical Society of America (OSA). She currently serves as an Associate Editor for IEEE PHOTONIC TECHNOLOGY LETTERS and for the OSA *Journal of Optical Networking*.