The Data Vortex Optical Packet Switched Interconnection Network

Odile Liboiron-Ladouceur, *Member, IEEE*, Assaf Shacham, *Member, IEEE*, Benjamin A. Small, *Member, IEEE*, Benjamin G. Lee, *Student Member, IEEE*, Howard Wang, *Student Member, IEEE*, Caroline P. Lai, *Student Member, IEEE*, Aleksandr Biberman, *Student Member, IEEE*, and Keren Bergman, *Senior Member, IEEE*

Invited Paper

***Abstract*—A complete review of the data vortex optical packet switched (OPS) interconnection network architecture is presented. The distributed multistage network topology is based on a banyan structure and incorporates a deflection routing scheme ideally suited for implementation with optical components. An implemented 12-port system prototype employs broadband semiconductor optical amplifier switching nodes and is capable of successfully routing multichannel wavelength-division multiplexing packets while maintaining practically error-free signal integrity (BER $< 10^{-12}$) with median latencies of 110 ns. Packet contentions are resolved without the use of optical buffers via a distributed deflection routing control scheme. The entire payload path in the optical domain exhibits a capacity of nearly 1 Tb/s. Further experimental measurements investigate the OPS interconnection network's flexibility and robustness in terms of optical power dynamic range and network timing. Subsequent experimental investigations support the physical layer scalability of the implemented architecture and serve to substantiate the merits of the data vortex OPS network architectural paradigm. Finally, modified design considerations that aim to increase the network throughput and device-level performance are presented.**

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

I. Introduction

NEARLY all contemporary large-scale high-performance information systems, including supercomputers, high-capacity data storage, and telecommunications core routers, require high-bandwidth low-latency interconnection networks. In these systems, performance is highly dependent upon the efficiency of vast information exchanges between sometimes thousands of clients (e.g., processors, memory, network hosts).[1] It is therefore critical for the interconnect infrastructure to support high-bandwidth low-latency communications that are highly scalable, thus transparently facilitating overall system performance [1]. With the immense growth in data traffic and required computation capacities, conventional electronic interconnection networks utilized for these applications are reaching capacity limits in their ability to meet the increased demands of the surrounding clients.

While it is well established that optical data encoding and the utilization of fiber-optic and photonic media possess the potential to provide orders of magnitude more bandwidth at near speed-of-light transmission latencies [2], critical optical technology shortcomings must be addressed. The key challenge for optical interconnection networks is to fully leverage the immense bandwidth of the fiber-optic components through techniques such as wavelength-division multiplexing (WDM) while avoiding the inadequacies of photonic technologies, particularly the absence of robust optical buffers and registers.

The data vortex architecture was specifically designed as a packet switched interconnection network for optical implementation; the topology supports large port counts scalable to thousands of communicating terminals. In order to accommodate scalability and to address the problematic absence of reliable dynamic photonic buffers, the conventional butterfly network topology is modified to contain integrated deflection routing pathways. The primary consideration in the architecture's organization is enabling the optical packets to approach speed-of-light time-of-flight latencies. The switching nodes are therefore designed to be as simple as possible and to contain very little routing logic; in this way, the packets do not sit in buffers while routing decisions are made. These design considerations naturally result in a modular architecture, which is scalable and self-similar in such a way that a small system implementation can predict the performance of significantly larger networks.

The data vortex design differs markedly from the conventional approaches to optical packet switched (OPS) networks, in which electronic architectures are often simply mapped into the optical media in a manner that can fail to capitalize on the unique properties of optical transmission. In the data vortex architecture, when processing and routing decisions are required, high-speed digital electronic circuitry is employed in a way that complements the semiconductor optical amplifier (SOA) wide-band photonic switching elements. This allows the high-bandwidth optical payload, which is encoded on multiple wavelengths in

Manuscript received August 15, 2007; revised November 9, 2007. Published August 29, 2008 (projected). 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-12-664.

O. Liboiron-Ladouceur is with McGill University, Montreal, PQ, Canada.

A. Shacham is with Aprius Inc., Sunnyvale, CA 94085 USA.

B. A. Small, B. G. Lee, H. Wang, C. P. Lai, A. Biberman, and K. Bergman are with the Electrical Engineering Department, Columbia University, New York, NY 10027 USA.

Digital Object Identifier 10.1109/JLT.2007.913739

[1]TOP500 List for June 2006, http:www.top500.org/lists/2006/06

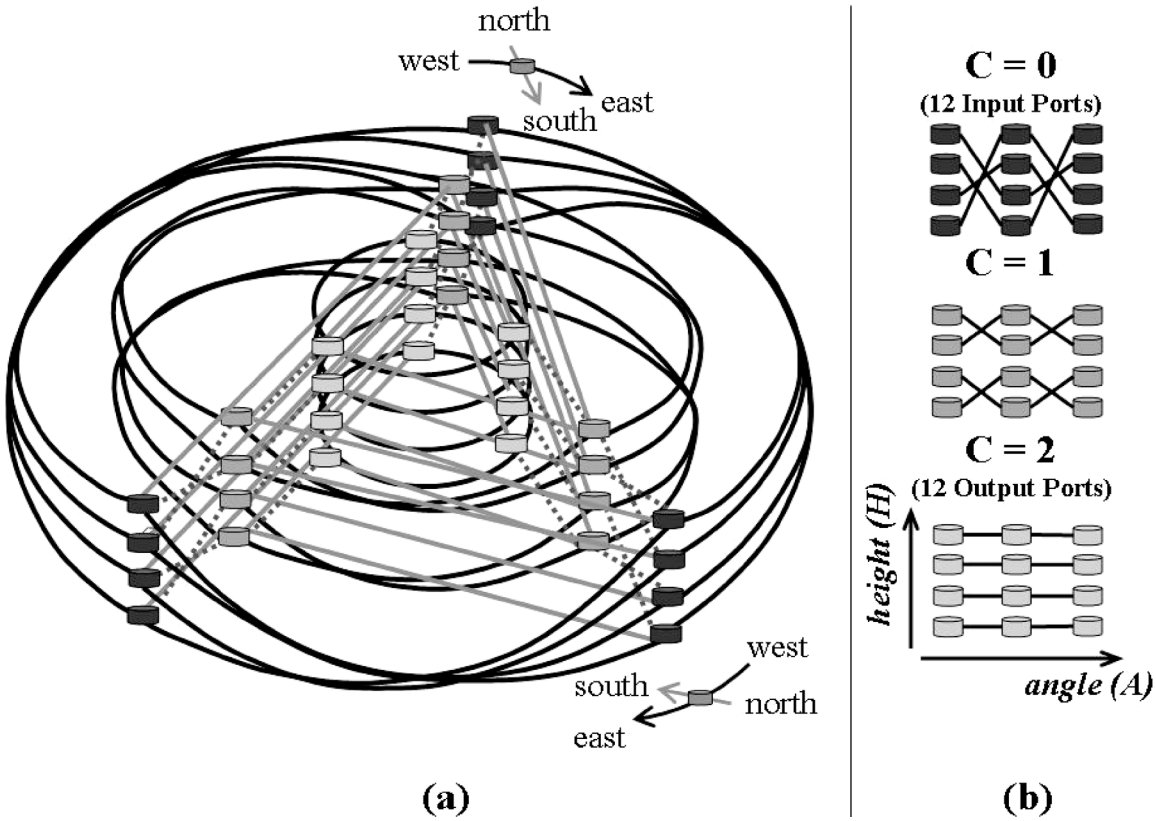

Fig. 1. (a) Illustration of a 12×12 data vortex topology with 36 interconnected $(\mathrm{C} = 3, \mathrm{H} = 4, \mathrm{A} = 3)$ and distributed 2×2 nodes (cylinders). Straight lines are ingression fibers, curved lines are deflection fibers, and dotted lines are electronic deflection signal control cables. (b) The banyan-like crossing pattern shows the deflection path connectivity for each cylinder.

order to maximize transmission capacity, to transparently traverse the network.

The data vortex topology [3] was first described in [4] and first investigated in [5]–[7], and its architecture was further analyzed in [8]–[11]. A 12-port data vortex prototype was implemented and its routing performance investigated [12]. The scalability of the physical layer was analyzed and demonstrated in [13] and [14], and further experimental studies of the optical dynamic range and packet format flexibility were performed [15], [16]. Sources of signal degradation in the data vortex were investigated in [17] and [18], and data resynchronization and recovery was achieved using a source synchronous embedded clock in [19]. Extensible and transparent packet injection modules and optical packet buffers for the data vortex were presented in [20]. Finally, alternative data vortex architecture implementations and performance optimization were explored in [21]–[23].

In this paper, we present a comprehensive discussion of the data vortex interconnection network and provide a complete review of the architectural investigations and experimental research. Section II reviews the topology, including a discussion of the deflection routing, network scalability, and node structure. In Section III, the implementation of a fully interconnected 12-port data vortex is presented along with a proposed synchronization approach for recovering short packets. The interoperability of an injection control module with the data vortex is demonstrated as well. Section IV describes the characterization of the physical layer and its quantitative impact on the system scalability. Section V presents both alternative architectural designs and performance optimization considerations at the device level for next generations of the data vortex interconnection network.

II. Data Vortex Interconnection Network

The data vortex topology [Fig. 1(a)] integrates internalized virtual buffering with banyan-style bitwise routing specifically designed for implementation with fiber-optic components. The structure can be visualized as a set of concentric cylinders or routing stages, which are cyclic subgroups that allow for deflections without loss of routing progress. Moreover, the hierarchical multiple-stage structure is easily scalable to larger network sizes while uniformly maintaining fundamental architectural concepts [5].

A. Topology

The data vortex topology is composed entirely of 2×2 switching elements (also called nodes) arranged in a fully connected, directed graph with terminal symmetry but not complete vertex symmetry. The single-packet routing nodes are wholly distributed and require no centralized arbitration. The topology is divided into C hierarchies or cylinders, which are analogous to the stages in a conventional banyan network (e.g., butterfly). The architecture also incorporates deflection routing, which is implemented at every node; deflection signal paths are placed only between different cylinders. Each

cylinder (or stage) contains A nodes around its circumference and $H = 2^{C-1}$ nodes down its length. The topology contains a total of $A \times C \times H$ switching elements, or nodes, with $A \times H$ possible input terminal nodes and an equivalent number of possible output terminal nodes. The position of each node is conventionally given by the triplet (c, h, a), where $0 \le c \le C-1, 0 \le h \le H-1, 0 \le a \le A-1$.

The 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 are of the same length throughout the entire system, as are the deflection fibers. The deflection fibers' height crossing patterns [Fig. 1(b)] direct packets through different height levels at each hop to enable banyan routing (e.g., butterfly, omega) to a desired height and assist in balancing the load throughout the system, mitigating local congestion [5]–[7], [12].

Incoming packets are injected into the nodes of the outermost cylinder and propagate within the system in a synchronous, time-slotted fashion. The conventional nomenclature illustrates packets routing to progressively higher numbered cylinders as moving inward toward the network outputs. During each timeslot, 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 the cth most significant bit in the node's height coordinate (h). If the bits match, the packet ingresses to node $(c+1, h, a+1)$ through the node's south output. Otherwise, it is routed eastward within the same cylinder to node $(c, G_c(h), a+1)$, where $G_c(h)$ defines a transformation which expresses the above-mentioned height crossing patterns (for cylinder c) [6], [7]. Thus, packets progress to a higher cylinder only when the cth address bit matches, preserving the $c-1$ most significant bits. In this distributed scheme, a packet is routed to its destination height by decoding its address in a bitwise banyan manner. Moreover, all paths between nodes progress one angle dimension forward and either continue around the same cylinder while moving to a different height or ingress to the next hierarchal cylinder at the same height. Deflection signals (Fig. 2), discussed further in Section II-B, only connect nodes on adjacent cylinders with the same angular dimension; i.e., from $(c+1, h, a)$ to a node at position $(c, G_{c+1}(h)), a)$.

The paths within a cylinder differ depending upon the level c of the cylinder. The crossing or sorting pattern (i.e., the connections between height values defined by $G_c(h)$) of the outermost cylinder $(c = 0)$ must guarantee that all paths cross from the upper half of the cylinder to the lower half of the cylinder; thus, the graph of the topology remains fully connected and the bitwise addressing scheme functions properly. Inner cylinders must also be divided into $2c$ fully connected (i.e., Hamiltonian) and distinct subgraphs, depending upon the cylinder. Only the final level or cylinder $(c = C-1)$ may contain connections between nodes of the same height. The cylindrical crossing must ensure that destinations can be addressed in a binary tree-like configuration, similar to other binary banyan networks.

Addressing within the data vortex architecture is entirely distributed and bitwise, similar to other banyan architectures: as a packet progresses inward, each successive bit of the binary address is matched to the destination. Each cylinder tests only one bit (except for the innermost one); half of the height values permit ingression for 1 values and half for 0 values, arranged in a banyan binary tree configuration. Within a given cylinder c,

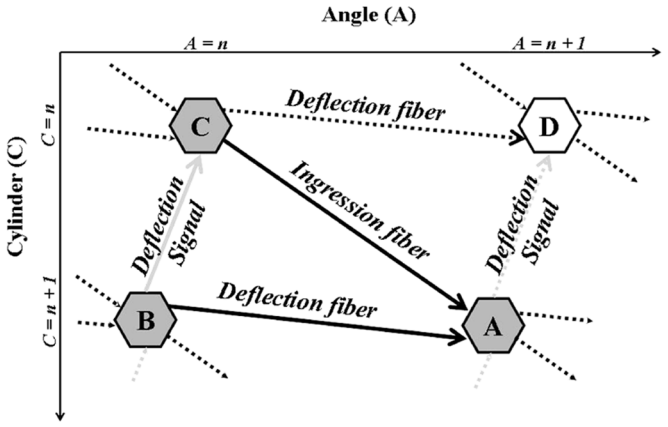

Fig. 2. Schematic representation of the deflection triangle. In order to avoid packet collision at node A, the electrical deflection signal (gray lanes) sent by node B to node C will force the packet at node C to be deflected to node D (black lanes).

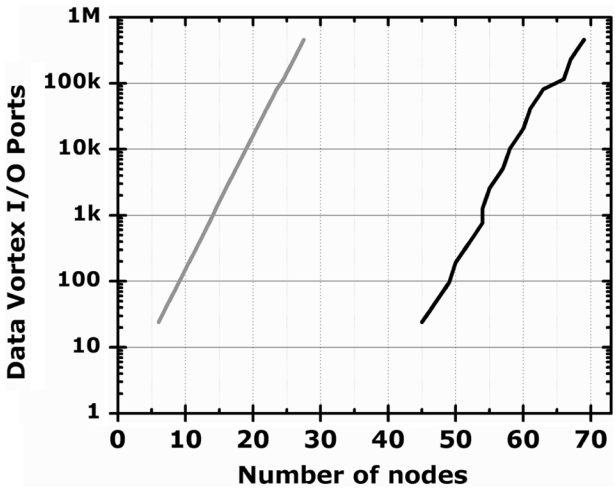

Fig. 3. Plot of the number of encountered nodes M as a function of the number of I/O ports N. The left curve represents the median number of hops and the right depicts the 99.999th percentile [13].

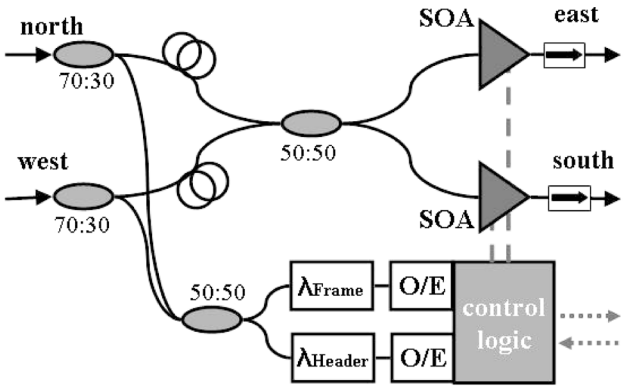

Fig. 4. (a) Switching node structure of the data vortex constructed with couplers (ellipses), filters (λ), isolators (boxed arrows), and a PCB that integrates the control logic components, receivers (O/E), and SOAs.

nodes at all angles at a particular height (i.e., $(c, h, \cdot)$) match the same $(c+1)$th significant bit value, while the paths guarantee preservation of the c most significant address bits. Thus, with each ingression to a successive cylinder, progressively more precision is guaranteed in the destination address. Finally, on the last cylinder $c = \mathrm{C} - 1$, each node in the angular dimension is assigned a least significant value in the destination address so that the packets circulate within that cylinder until a match is found for the last $\sim \log_2$ (A) bits (so-called angle-resolution addressing) [12].

As will be discussed in Section II-D, each switching node is bufferless and is designed to check exactly one bit of the destination address, in addition to the packet frame. When the selected address bit matches the value assigned to the node due to its position within the cylinder, the packet is allowed to ingress into the next cylinder on the ingression fiber unless a deflection signal is received. When the selected address bit does not match, or when a deflection signal is received, the packet is routed within the same cylinder on the deflection fiber and the node sends a deflection signal indicating that the next node will soon be busy with the deflected packet. Therefore, every switching element always has an available deflection fiber (east) and ingression fiber (south) used for routing matches.

While it may seem wasteful to have twice as many optical paths as necessary, having a guaranteed deflection fiber pathway allows for an extraordinarily simple routing logic that can be executed extremely quickly [12], [24]. The distributed bitwise addressing scheme also helps to ensure that routing decisions do not dominate network latency. No buffers are used, so the network latency can be reduced to approach optical time-of-flight.

B. Deflection Routing

Although packets can originate from either of the two input ports (north or west), recall that the node design is capable of routing only one packet at a time. This fundamental constraint yields a simplistic construction of the node but requires an architectural implementation of internal blocking or deflection. Deflection within the data vortex thus differs from conventional deflection routing topologies that allow for deflection routing at the completion of each discrete node hop. Instead, the data vortex deflection implementation prevents two packets from simultaneously entering the same switching node, and thus colliding, by controlling or blocking one of the two nodes connected upstream of the input ports.

Therefore, a packet deflection in cylinder c occurs for only two reasons: 1) a packet is deflected in the adjacent cylinder $c+1$, or 2) the packet in cylinder c cannot be injected due to a mismatch between its destination address and the node's height value. A packet that would otherwise ingress into that node on cylinder $c+1$ from cylinder c (when address bits match) can thus be required to remain in cylinder c when the deflection signal indicates that the desired node is busy. This deflection structure results in "backpressure" from the inner cylinders $(c = C-1, C-2, \ldots)$ to the outer cylinders $(c = 0, 1, \ldots)$. A deflected packet must traverse two additional node hops before the address matches again, as a consequence of the crossing patterns.

The deflection signaling structure incorporates the output and input terminals such that output nodes can receive busy signals from the output queuing subsystem, and the input nodes can transmit similar busy signals to the input interface. Thus, packets that attempt input at the first cylinder $c = 0$ may receive a signal that indicates that the desired input node is busy; the packet must then be queued to reattempt injection or be discarded. The deflection signal relationship can be represented geometrically as a triangle. In fact, this triangular deflection unit is the fundamental building block of the entire data vortex topology. Only the arrangement of each leg differs from cylinder to cylinder, in accordance with the specific crossing patterns used. The deflection signal's crossing pattern must be the same as that of the deflection fiber since connections between cylinders do not undergo height translation.

In order to maintain correct timing for deflection signaling, particular latency conditions must be satisfied. Since implementations of this architecture avoid buffering within the switching elements, latencies are caused entirely by the optical and electrical (for control signals only) paths' times-of-flight. To maintain accurate deflection signaling, deflection signals must be transmitted sufficiently early such that the node receiving the deflection signal can direct its packet appropriately. The timing of cylinder $c+1$ containing the node initiating the deflection must therefore precede the timing of cylinder c containing the node receiving the deflection signal. Thus, the ingression fibers must be shorter than the deflection fibers by an amount equal to the processing and transmission time of the deflection signal [24]. Consequently, the timing cycles of the inner cylinders precede those of the outer ones.

When the aforementioned timing condition is met for the data vortex implementation, global clocking for every switching element is not required. If packets are only injected at timeslots that correspond to the deflection fiber latency, packets will maintain this slotted arrival schedule at every position within the hierarchal topology. Again, recall that no buffers or storage devices are utilized; hence when physical time-of-flight requirements are met, they hold for all packets at each node.

C. Scalability of the Topology

The data vortex topology exhibits a flexible modular architecture that is scalable to large numbers of input and output ports $(N \times N)$. To increase the size of the interconnection network, the number of ports can be augmented by increasing the cylinders C, angles A, or height H parameters. Per the topology discussion in Section II-A, the number of ports N is defined as the product of the number of angles A and heights H

$$N = H \times A. \tag{1}$$

The cylinders represent the stages of the data vortex. The cylinder number C is defined by H and corresponds to one plus the base-two logarithm of H, as each routing node consist of two inputs (west and north) and two outputs (south and east)

$$C = \log_2(H) + 1. \tag{2}$$

Following the above two equations, a topology can be appropriately designed. The modular architecture enables several possible topologies for the same number of I/O ports. It has been shown that for a given number of ports, a shorter height and greater number of angles provide lower latencies as compared to taller (large H) and narrower (small A) topologies [10].

In optical multistage interconnection networks, an important parameter is the number of routing nodes a packet will traverse before reaching its destination. For the data vortex utilizing 2×2 switching elements, the number of nodes M scales logarithmically with the number of ports N [25], [26]

$$M \sim \log_2 N. \tag{3}$$

Packets will propagate through M cascaded nodes; generally, M increases with the network size but M is also tightly coupled to the network load. Various scenarios have been simulated to illustrate the effect of the topology and network load on M [9], [11]. The number of cascaded nodes directly affects the overall network latency. Due to the deflection characteristic of the network, the number of cascaded nodes M has a non-Gaussian statistical distribution arising from the contention resolution scheme [5], [7]. Simulations of the data vortex as a large-scale switching fabric have shown that for a heavily loaded 10 k $\times$ 10 k port data vortex implementation, 99.999% of the injected packets propagate through fewer than 58 internal switching nodes with a median hop count of 19 (Fig. 3) [13].

D. Packet Self-Routing

The modular architecture of the data vortex and its impressive scalability are made possible by the simplicity of the routing node structure (Fig. 4). The nodes are evenly distributed across the data vortex topology in a manner that facilitates contention resolution while minimizing latency and maximizing throughput. Packets propagate from one of the two inputs of the node (north or west) to one of the two outputs (east or south). Two SOAs are used to select one of the two outputs based on the control information encoded along with the payload data within the optical packet (see Section II-E).

The two SOA gates are enabled by laser drivers that are controlled by the electronic decision circuitry internal to each node, producing the routing decisions on a per-packet basis without the need of a central scheduler. To produce the routing decision, the nodes in the first cylinder $(c = 0)$ detect the presence of Header 0, the nodes in the second cylinder $(c = 1)$ detect Header 1, and so forth. The frame information is used to validate the presence of a packet in the routing node. If the header bit does not match the user-programmed value for the node, or if the interconnected node in an inner cylinder is busy, the packet is deflected to the next angle in the same cylinder. Here, current is supplied to the east SOA only. If the header bit matches that of the node and no deflection signal is received, then the packet ingresses to the angle in the next cylinder (or stage), as current is supplied to the south SOA only. 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 the innermost cylinder at its destination height, it is routed to the correct output port based on the nodes' angle parameters a. Allocated bits of the header address are typically used for the angle information, but modified decoding logic can be used at the innermost cylinder to reduce the number of channels needed for the address.

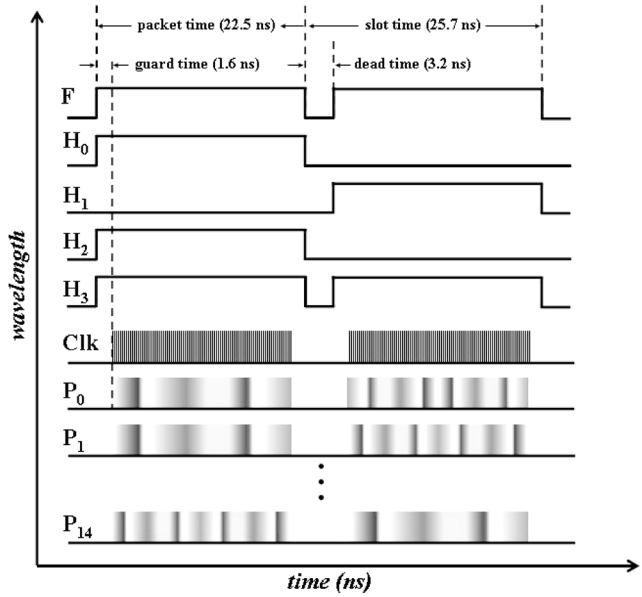

Fig. 5. Timing diagram of the multiwavelength packets with control wavelengths, a clock wavelength, and multiple payload wavelengths.

E. Packet Format

The data vortex architecture utilizes a high-bandwidth multiple-wavelength packet format (Fig. 5). The header address is encoded in the frequency domain through WDM wavelength allocation in order to minimize latency. The implementation of the header address at a lower bit-rate than the payload also facilitates the electronic decoding and interpretation of the routing information. A source-synchronous clock can be encoded on a dedicated wavelength to sample the short packets at the destination node [19]. The remainder of the C-band can be filled with high-bit-rate payload data, since the data vortex is transparent to the optical bit-rate, resulting in very high aggregate bandwidths.

The packet timing is precisely defined to maintain contention resolution and network synchronization. The overall slot time corresponds to the time-of-flight between two adjacent nodes such that one packet is contained in only one node at a time. In addition to the packet time, the timeslot includes the dead time and guard time. The dead time distinguishes two temporally adjacent packets and accommodates slight process variations in real-world devices, while the guard time is determined by the finite time required to enable the SOAs.

III. Data Vortex Implementation

A 12-port fully connected data vortex system prototype has been implemented using 36 switching nodes integrated onto six

TABLE I
DATA VORTEX IMPLEMENTATION WAVELENGTH ASSIGNMENT

Type	Function	Wavelength (nm)
Frame	Validator for c=0,1	1555.55
Header 0	Encoded Address, c=0	1535.00
Header 1	Encoded Address, c=1	1533.50
Header 2	Encoded Address, c=2	1550.90
Header 3	Encoded Address, c=2	1531.12
Clock	Used at destination	1553.33
Payload 0 to 15	> 160 Gb/s	1536.6 to 1559.8

printed circuit boards (PCBs) and interconnected by passive optics modules, single-mode fiber, and electronic cables. The design of this system along with improvements for data recovery and packet injection is discussed below.

A. Design of a Functional 12-Port Data Vortex

In this section, we discuss the design specifications used to demonstrate a fully functional optical packet switched interconnection network. Note that only off-the-shelf components were used in the experimental implementation of the data vortex.

In the case of the 12-port data vortex $(H = 4, A = 3)$, a 4-bit header encoded on four header wavelengths is required to address each system output. For a 10 000-port system, 14 headers would be required to route packets individually from each input port to each output port. While the packet is transparently routed through the node, the header address is decoded at each node by filtering a frame wavelength and one cylinder-specific header wavelength. Upon entering the node from either input port, 30% of the input packet's power is extracted for control signal decoding. The control wavelengths are filtered with 100-GHz optical bandpass filters and directed to photodetectors. The receiver data path is designed to be dc-coupled to accommodate bursty packet arrival. The remainder of the packet power (70%) is routed through a fixed length optical delay line and a 50:50 coupler, which combines the two input ports and splits the optical packet to the two SOAs. The SOAs are followed by isolators in order to mitigate counter-propagating amplified spontaneous emission (ASE) noise.

The entire node's measured latency is 15.8 ns [24]. A packet slot time of 25.7 ns is determined by the total time-of-flight latency between two nodes within the same cylinder and the deflection-timing requirement. Each packet is 22.5 ns long and includes a guard time of 1.6 ns inserted at both the beginning and the end of the packet, yielding a net payload window of 19.3 ns (Fig. 5). A guard time is required for the SOA switching transition time, which is approximately 1 ns. The control information required to route the packets consists of one frame and a 4-bit header. Sixteen payload channels are modulated at 10 Gb/s for 160 Gb/s of aggregated bandwidth. Packet payload bandwidths approaching terabit/second can be straightforwardly achieved by increasing the channel count and modulated data rates (e.g., 25 channels modulated at 40 Gb/s). The selected address field wavelength channels are shown in Table I. In the innermost cylinder $(c = 2)$, a decoding logic uses both header 2 and header 3 to determine the output port address.

In the experimental test bed illustrated in Fig. 6, the payload wavelengths are generated by distributed feedback lasers (DFBs) modulated with a 2^9-1 pseudorandom bit sequence at 10 Gb/s and decorrelated with a length of optical fiber by approximately 450 ps/nm. An SOA segments the continuous data stream into packets. Prior to injection in the data vortex, the header and frame information is independently modulated and coupled to the multiwavelength packet to meet the required packet structure (Fig. 5). The timing of all the signals are closely calibrated using the data-timing generator (DTG). At the output port, the packets are preamplified with an erbium-doped fiber amplifier, filtered for a specific payload channel, and converted to an electrical signal through a dc-coupled receiver module. The error rate of the data contained within the individual payload wavelengths of each packet is measured using a bit-error-rate (BER) tester (BERT). The BERT, which enables extensive analysis of the performance of the data vortex network, is externally gated and synchronized with the pulse pattern generator (PPG).

To validate the design, routing experiments were performed to demonstrate the correct addressing of packets through the system and to verify the functionality of its contention resolution scheme [3]. All 12 output ports are addressed and contentions are resolved between switching nodes according to the data vortex internal deflection routing mechanism. The system is capable of routing packets with 160 Gb/s (10 Gb/s $\times$ 16 WDM channels) payloads from any one of the 12 input ports to any one of the 12 output ports. The average (and median) latency for the system is 110 ns, corresponding to five node hops [27].

The emulation of realistic network traffic utilizing a supercomputing interconnection network traffic workload was also demonstrated [28]. The evaluation workload uses processor-memory accesses from an application in a SPLASH-2 parallel computing benchmark suite [29]. The methodology captures the behavior of shared memory parallel execution, providing message traffic typically found in a similarly configured multiprocessor system. The simulation results show that all packets are routed correctly; furthermore, appropriate deflections and address decoding are also observed.

B. Resynchronization and Recovery

Data traffic in interconnection networks such as the data vortex often consists of short and bursty message exchanges. One of the key challenges is resynchronization and recovery of data at the destination node without the use of conventional phase-locked loop designs. In the data vortex, a clock synchronous to the data payload can be embedded in the packet and used as the timing reference at the destination node [19]. The embedded clock avoids the complexity of low-skew clock distribution through large-scale synchronous interconnection networks and simplifies the message recovery circuitry as compared to asynchronous networks.

This approach has been demonstrated in the implemented data vortex network by recovering WDM messages routed through five switching nodes. The messages are entirely recovered and processed at the destination node using an embedded clock signal. The clock-to-data skew, defined as the relative timing between the payload and the embedded clock, must remain within the setup-time and hold-time requirements of the deserializer. However, a main contributor to clock-to-data

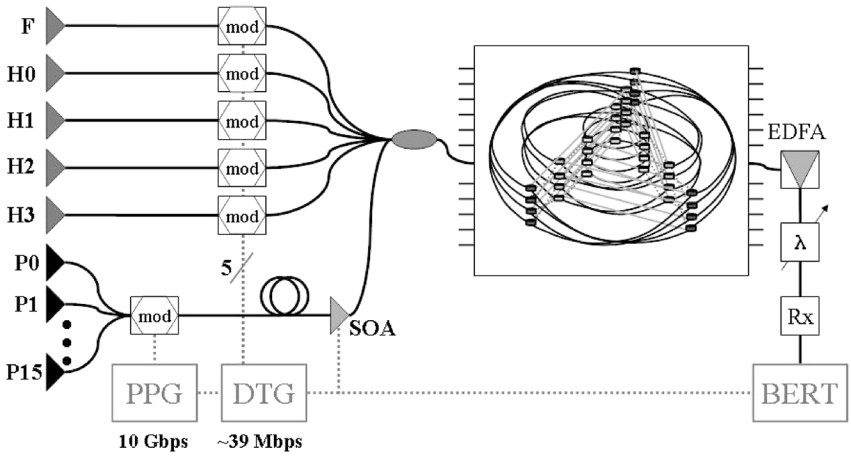

Fig. 6. Schematic of the data vortex test bed, including $LiNbO_3$ modulators (mod), a gating SOA, a pulse pattern generator (PPG), a data-timing generator (DTG), a burst-mode receiver (Rx), and a tunable bandpass filter (λ)[12].

skew is group velocity dispersion (GVD). A sampling window, corresponding to over 500 m of interconnected fiber, was measured in [19]. A very large-scale OPS interconnection network such as a 10 k $\times$ 10 k data vortex would incorporate only about 200 m of interconnected fiber.

C. Node Extension for Packet Injection

In an OPS network with an internal contention resolution scheme such as in the data vortex, packet injection must be managed. By extending the routing node, an injection control module (ICM) has been experimentally demonstrated and shown to mediate packet injection into OPS routers such as the implemented 12×12 data vortex network [20]. The basis of the ICM is a reprogrammable nonblocking 2×2 wide-band switching node that controls a feedback fiber delay line (FDL). The module is further composed of SOAs and fast opto-electronic and electronic circuitry that facilitate the dynamic decoding of control signals from the router to manage injection on a packet-by-packet basis (Fig. 7). The module thus operates at the packet timescale and can delay a packet until the router input port is available for injection. The physical complexity of the ICM is independent of the number of packet delay cycles.

At each timeslot, the frame bit is extracted from the incoming packet and an electronic busy signal may or may not be received from the router. Based on these inputs, a routing decision is made. In the case where the busy signal is absent, the packet is injected into the router (specifically the outer cylinder nodes of the implemented data vortex). If the busy signal is present, the packet is buffered on the FDL to reattempt injection during the following slot. At the beginning of the next timeslot, another packet may be received at the input port and thus the control circuit must make a routing decision for both packets. If the busy signal is no longer active, indicating that injection is permitted, one packet is injected while the other is delayed. If the router port is still busy, one packet is dropped. The successful implementation of the FDL-based ICM demonstrates that the photonic packet injection module is useful as a functional subsystem for OPS systems such as the data vortex.

IV. Data Vortex Physical Layer Analysis

Although the topology may scale, it does not necessarily follow that the optical physical layer also scales in terms of maintaining end-to-end signal integrity without requiring regeneration. Various effects in the optical domain may limit the true scalability of the data vortex. In this section, we investigate the effects of the physical layer on the optical packets.

A. Dynamic Range

First, the dynamic range of optical power levels over which the OPS network remains functional is a critical consideration to the physical layer scalability and system robustness. When signal power levels are too low, they are buried in the optical noise floor, yielding poor error-rate performance. At the other extreme, the SOAs may become saturated, leading to enhanced channel crosstalk. In order to determine the system's robustness to these changes, the optical power of the multiple-wavelength packet is varied uniformly over a wide range of values by inserting a booster SOA before the gating SOA shown in Fig. 6 and by adjusting the gain of these SOAs. The optimal signal power for each payload wavelength is found to be approximately -15 dBm, corresponding to a packet power of approximately -1 dBm for the 16 payload wavelengths. The payload is combined with the five routing wavelengths, each at a power of -13 dBm. At this level, the total SOA input power is about -7 dBm, since around 6 dB of passive optical losses precede an SOA in each node [24], producing a signal power less than the SOAs' input saturation power of -5 dBm. Payload powers greater than -13 dBm should be avoided, as they will bring the device into the saturation regime, yielding poor BER performance (Fig. 8). For 16 payload wavelengths, the dynamic range at a 10^{-12} (10^{-9}) BER threshold is determined to be 6.7 ± 0.3 dB (8.2 ± 0.5 dB), depending on the wavelength of interest (Fig. 8) [16].

B. Packet Format Flexibility

Depending on the application and on the type of network behavior, it is important that the network support various packet formats. Two experiments supporting the flexibility of

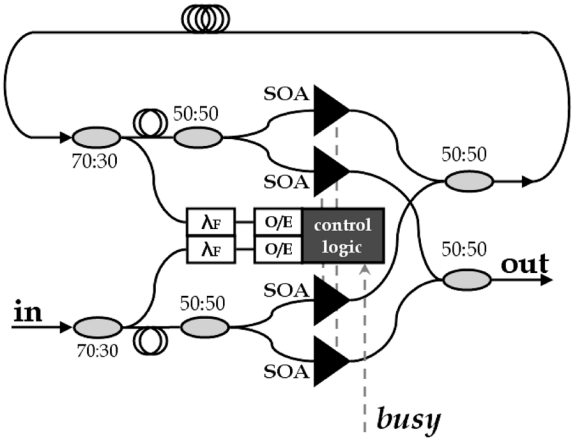

Fig. 7. The injection control module is composed of SOAs, optical couplers, filters, p-i-n photodetectors, optical fiber, and an electronic control circuit [20].

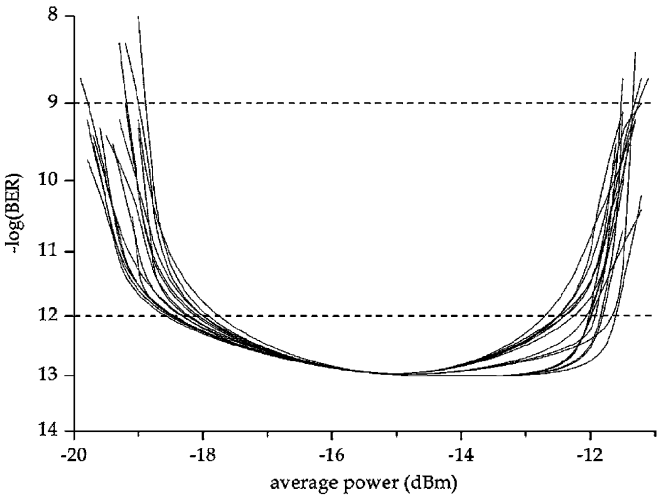

Fig. 8. Measured BER versus received power for the 16 payload wavelengths illustrating the dynamic power range at BERs of 10^{-9} and 10^{-12} [16].

the data vortex packet format have been presented [15], [16]. These confirm that the system can simultaneously route packets containing variable-sized payloads by altering the number of payload wavelengths in the packet and the time duration of the payload data stream. The first experimental demonstration utilizes a variable number of payload wavelengths, while the second utilizes a variable payload time duration while varying the relative location within the timeslot. Specifically, packets with payloads composed of 4, 8, 12, or 16 wavelengths, and time durations of 9.6 and 19.3 ns, are simultaneously routed through a five-node path in the network. Therefore, in a single 25.7-ns timeslot, the data capacity can range from 48 to 384 bytes, corresponding to a net bandwidth in the range of 15–120 Gb/s. The 6-dB variation in power between the packets of 4 and 16 payload wavelengths falls within the previously discussed 6.7-dB dynamic range. As a result, BERs of 10^{-12} or better were verified for all possible packet formats (4, 8, 12, or 16 wavelengths and 9.6- or 19.3-ns durations). The demonstrated flexibility very clearly illustrates the transparent nature of the implemented architecture and switching node design.

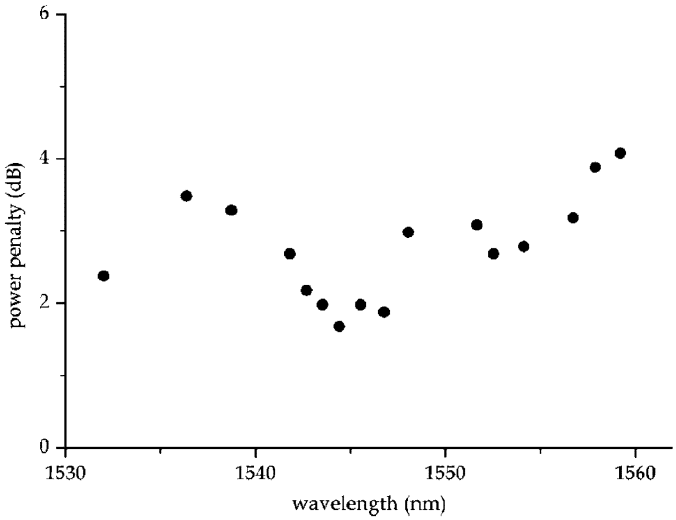

Fig. 9. Power penalty of each of the 16 payload wavelengths for a five-node path through the network [16].

C. Routing Node Power Penalty

In many OPS networks, the predominant switching element used is the commercially available SOA, which has a fast switching time, high extinction ratio, broad gain bandwidth, and high potential for integration. The primary function of the SOA is to efficiently and transparently route a broadband packet as well as to compensate for small switching node losses. The SOA, however, introduces ASE noise along with the amplified signal, causing degradation of the optical signal-to-noise ratio (OSNR). In [18], the eye diagram quality factor Q of a typical packet payload channel ($\sim$1547 nm) is shown to degrade from 21.2 dB before injection to 19.0 dB after propagation through a five-node path in the network, which indicates an average decline of approximately 0.4 dB per node. Measurements have also indicated an average OSNR degradation of 2.7 dB per node for a five-node path [16]. Although the SOA's noise is not expected to increase linearly for a large number of nodes, the precise scaling is difficult to quantify in a concise manner [31]. The total power penalty at the receiver for a five-node path is found to vary between 1.7 and 4.1 dB, measured at a BER of 10^{-9}, for all 16 payload wavelengths; the average power penalty is 2.6 dB (Fig. 9).

D. Node Cascadability

The node cascadability ultimately determines the physical layer size scalability of an SOA-based optical switching network. Node cascadability is influenced by the launched input power of the packets, the power per channel, the number of channels, and the bandwidth across which the channels are distributed. The scalability of a packet-switched optical interconnection network using SOA switching elements has been previously investigated [13]. Using a recirculating loop test-bed environment, the SOA switching nodes were constructed in accordance with the data vortex network architecture. Here, it was shown that BERs lower than 10^{-9} can be maintained through 58 node hops (sufficient for a 10 k $\times$ 10 k port interconnection network) for eight wavelength channels spanning 24.2 nm of

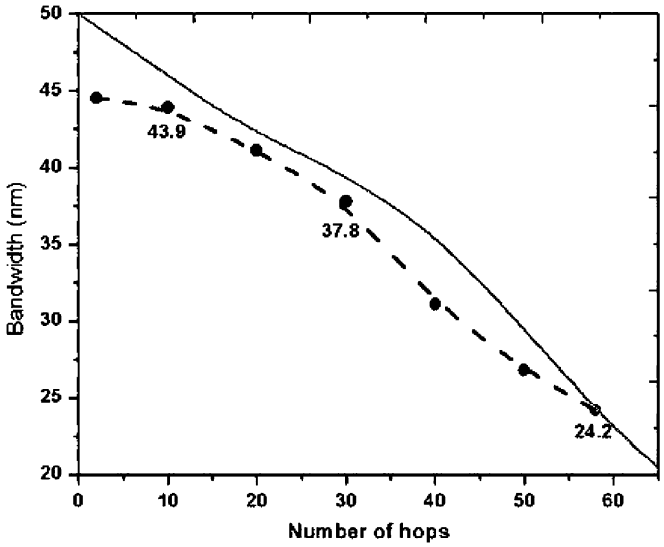

Fig. 10. Functional bandwidth for eight payload channels versus number of hops. The solid gray line shows simulation results [13].

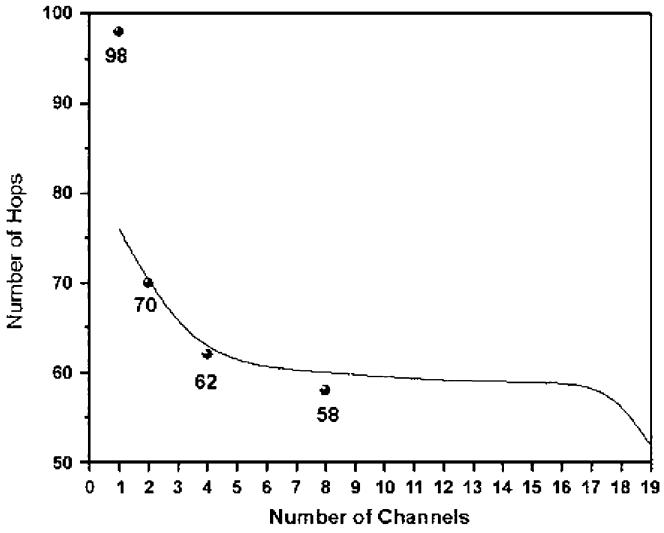

Fig. 11. Experimental number of hops obtainable with BERs $< 10^{-9}$ for one, two, four, and eight channels. The line shows simulation results [13].

the C-band at 10 Gb/s per channel. With the packet payload residing in a single 10-Gb/s channel, 98 hops can be achieved while maintaining the same BER. The measured results were confirmed by a phenomenological model, which matched the empirical results to within approximately 0.2 dB (Figs. 10 and 11).

E. Gain Profile Effect

The switching nodes are intended to be as optically transparent as possible, such that the incoming and outgoing packet powers at a given node are identical. Because the optical path through the switching node contains passive optical couplers and other sources of insertion loss, the SOA gain must be set to compensate for these losses. Due to the wavelength dependence of the passive and active optical components within the node, each successive switching node introduces a small amount of wavelength variation across a packet's power spectrum. At each node, the gain/loss varies by less than 1.0 dB across the wavelengths of interest (about 1530–1560 nm), and a packet propagating through a five-node path in the network experiences less than 4.9 dB of net gain/loss variation [16].

F. Polarization Gain Dependence

In order to determine appropriate physical layer constraints, it is important to understand the impact of polarization-dependent gain (PDG), which results from the larger amplification of the transverse electric (TE) mode over the transverse magnetic (TM) mode in bulk active materials. Although polarization dependence in these devices is typically small (generally less than 1 dB), it becomes significant in multistage optical interconnection networks. Again using a recirculating loop test bed with polarization controllers (PCs), it has been shown that the maximum number of cascaded nodes varies by as much as 20 elements for SOA-based designs with PDG of less than 0.35 dB. This corresponds to a 100-fold decrease in the number of interconnected ports of an optical interconnection network such as the data vortex [14]. It hence becomes evident that, for larger network sizes, PDG compensation techniques are necessary to minimize the dramatic shift in performance.

G. Timing Accuracy Effect

The data vortex network topology relies heavily on passive, unclocked, node-to-node timing constraints for the distributed self-routing of packets. First, due to the lack of convenient dynamic buffering [31], packet timeslots are preserved by the design of the routing path latencies (e.g., fiber lengths within and between nodes). Secondly, although the individual nodes do not require a clock signal, the electronic deflection signals sent between nodes must be timed precisely. In order to test the timing constraints of the implemented network, experiments were performed to characterize the effect of time-of-flight inaccuracies on propagated packets [17]. One observation is that self-routed packets grow slightly shorter as they propagate through the network. This is expected and is the result of the finite SOA rise and fall times, which have been measured to be approximately 0.9 ns each [24]. However, at each node hop, the packet headers are truncated by only 0.4 ns on average. This can be attributed to the gradual slopes of the transition edges and the high sensitivity of the low-speed detectors within each node, which can trigger a routing decision on the slightest rise (or fall) of incident optical power. With faster SOA switching elements, the packet truncation may be further reduced. In the current implementation, the packet payloads are constructed with guard times at the leading and trailing edges (1.6 ns each) to accommodate the finite transition times. The measured timing margin is sufficient for eight node hops ($3.2 \div 0.4$ ns). In general, as the data vortex scales to a larger number of input and output ports (N $\times$ N), the number of routing nodes M required for each packet scales logarithmically (3) [6], [7]. Therefore, the packet guard time T_G required for a system based on the data vortex architecture is on the order of

$$T_G \geq \tau \log_2 N \tag{4}$$

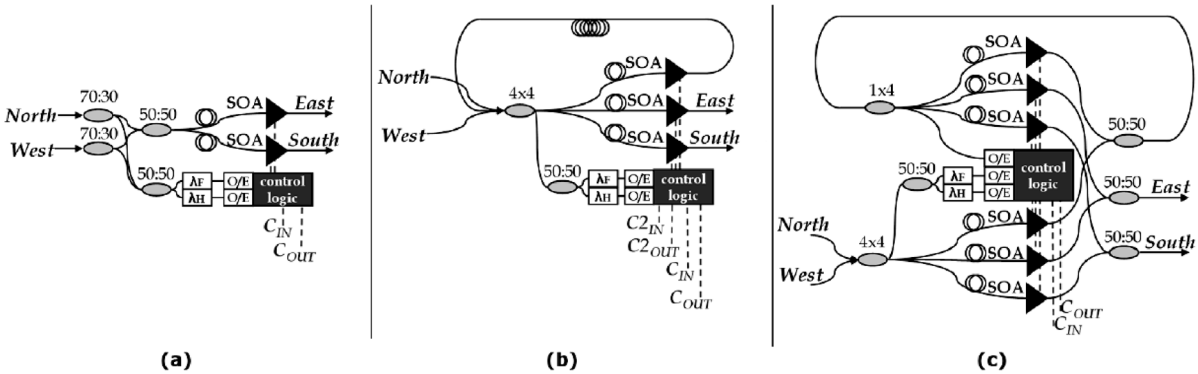

Fig. 12. (a) Original switching node; (b) blocking switching node; and (c) nonblocking switching node [11].

where τ is the average truncation at each switching node. This result is very encouraging, particularly since improvements in switching speeds yield direct reductions in τ.

The degradation in the SOA's optical response was ameliorated by modifying the standard SOA package [22]. The D-SOA was created by bonding a 10.7 Gb/s current driver die to the SOA's active region in a temperature-controlled hybrid integration platform. The new device was enclosed within a modified 28-pin butterfly package with two high-frequency sub-SubMiniature version B (SMB) input connectors. The connectors serve to preserve the signal integrity of the differential digital logic signal that enables the D-SOA, which is produced as the packet is routed through the node. Additionally, the current driver has an integrated compensation network consisting of a series-damping resistor and a shunt RC optimized for the bond wire's 0.4 nH inductance. The current driver can inject up to 100 mA into the SOA, corresponding to a gain of 6 dB. In order to maintain a fast transition time, a small dc bias current is delivered to the D-SOA to provide a carrier density slightly below threshold. The D-SOA's 20%–80% transition time was measured as 500 ps, compared to 900 ps using a commercial SOA. The signal quality was also markedly improved, exhibiting minimal ripple compared to SOAs with standard packaging techniques [22].

V. Next-Generation Data Vortex

The creation and experimental verification of a small-scale data vortex system is an important step in the realization of a full-fledged high-performance computing systems (HPCS) interconnect based on the topology. The successful operation of this experimental system along with the optimistic results of the physical layer characterization suggest that a larger data vortex switching fabric can be built and utilized as an interconnection network in a HPCS.

Detailed performance studies have been pursued in simulation to characterize the utility of various modifications to the architecture, control, network dimensions, and injection/ejection polices of a standard data vortex architecture [10]. In addition to system-scale considerations, the implementation of the next-generation data vortex system will benefit greatly from improvements at the device level. Below, four areas of improvement and a different node structure approach are highlighted for future design considerations and implementations of the data vortex OPS network.

A. Internal Buffering

A time-domain method for contention resolution for the data vortex architecture has been proposed involving the insertion of FDL-based recirculating buffers into the switching nodes [11]. Methodologies for contention resolution remain a central focus for the data vortex, which uses a combination of time- and space-domain contention resolution to implement virtual buffering. On the switching-node scale, contentions are traditionally resolved in the space-domain; packets are deflected to an undesired port when the requested port is unavailable. This space-domain technique translates to time-domain resolution, as the contending packets reach their final destination at different times. In this way, the packets are deflected in internal paths until their destination becomes available.

Unfortunately, deflection routing can have detrimental effects on the latency of the network. Thus, an alternate method for resolving contentions in the data vortex switching nodes has been explored by inserting a FDL recirculating buffer into each switching node. In the first approach [Fig. 12(b)], a node may handle only a single packet per timeslot. This approach, referred to as the blocking switching node, utilizes a 1×3 SOA-based switch to route the packet from any of its three inputs (west, north, or FDL) to one of its three outputs (east, south, or FDL). To guarantee that only a single packet is received during each timeslot, several control cables are added to enable the transmission of intrastage deflection signals. This approach introduces additional blocking to the network. Furthermore, each FDL traversal requires two timeslots. To overcome these shortcomings, a second approach is suggested that uses a 2×3 SOA-based switch. The second approach [Fig. 12(c)], referred to as the nonblocking switching node, can manage two simultaneous incoming packets: one packet from either the west or north input ports and one packet from the FDL. Although this node is internally blocking between the west and north ports, the FDL does not block nor can it be blocked by any of the inputs. Moreover, this approach allows the FDL traversal to be set to one timeslot, reducing the latency penalty incurred by blocking of the output port.

Using synthetic Bernoulli uniform random traffic, the performance of the data vortex with the alternate time-domain con-

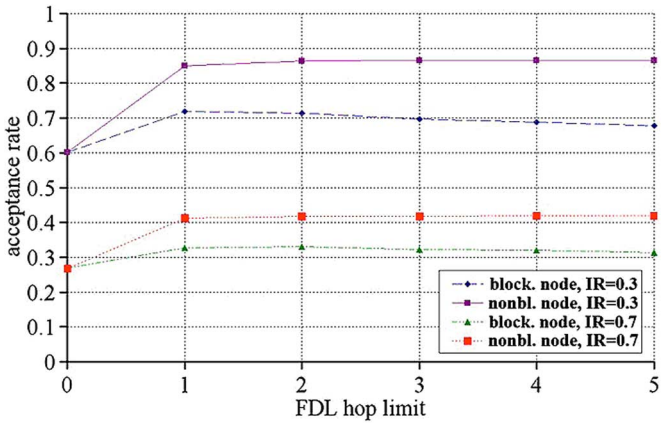

Fig. 13. Simulation results showing the effect of the FDL hop limit on acceptance rate at injection rates of 0.3 and 0.7 [21].

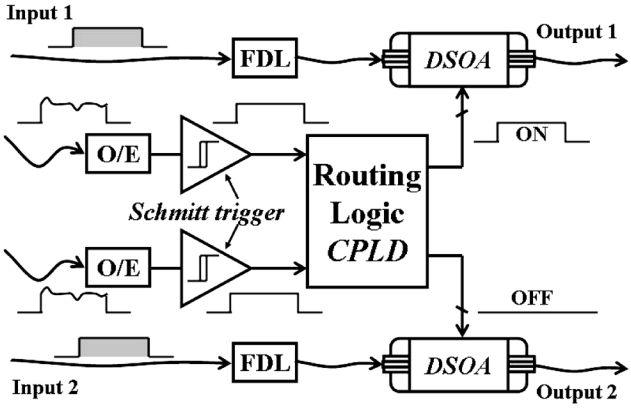

Fig. 14. Schematic of the multistage-optimized switching node with Schmitt trigger comparators in the routing logic and fast-switching D-SOAs, optical receivers (O/E), and fiber delay lines [22].

tention resolution technique has been simulated with a C++ program, and the results were compared to the original data vortex virtual buffering mechanism. Each of the new switching node design approaches (blocking and nonblocking) was evaluated with respect to the following metrics: acceptance rate and throughput, mean latency, and latency distribution.

Simulations demonstrate that the acceptance rate monotonically decreases with the injection rate and that the FDL-based configurations improve the acceptance rate only at medium-to-high injection rates, while asymmetric injection (injecting on only a fraction of the angles) improves the acceptance rate at low injection rates (Fig. 13). Additionally, for a data vortex realization (with given H, C, and A parameters), simulated throughput results indicate that the network utilizing the nonblocking node yields the highest throughput saturation value.

The latency of the data vortex architecture is also an important metric. Simulations show that a network incorporating FDL-based buffers has lower subsaturation latency and performs better than an original data vortex network with symmetric and asymmetric injections. More performance metrics were simulated in [21] and demonstrated the advantages of using FDL-based recirculating buffers as an alternate technique for contention resolution.

B. Nondeterministic Latency

Due to the nondeterministic nature of the paths traversed by packets propagating through the data vortex network, packet reordering is a common occurrence. Thus, the application space of the original data vortex interconnection network is limited by this topological characteristic. Furthermore, nondeterministic routing through the network gives rise to the possibility of unbounded latency distributions; these not only are detrimental to latency sensitive applications, such as memory accesses, but also contribute to physical-layer-induced limitations on network scalability. Investigations have explored the utility of variations on the traditional data vortex network in an attempt to alleviate this shortcoming [11].

C. Optimization of the Node

In addition to considering the node routing efficiency, self-routed networks rely on error-free routing through the internal switching nodes. Bit errors in payload data may be tolerable due to data encoding and forward error-correction techniques employed at the source and destination devices. However, any errors manifested as the result of misroutes at a node will have fatal effects on the operation of the network (i.e., packet loss and collisions with other packets). As packets propagate through a cascade of switching nodes, ASE noise and nonlinear effects such as cross-gain modulation can create erratic behavior in the optical signal. Further, slight changes in incident optical power at the photodetectors in each node will be interpreted as glitches in the electrical routing signals gating the SOAs and may lead to packet truncation and/or collision.

In order to increase the noise immunity of our switching nodes, a dual thresholding scheme has been implemented via the use of a Schmitt trigger comparator circuit following the header receivers [22]. A prototype evaluation switching node (Fig. 14) has been implemented with the aforementioned Schmitt trigger comparator circuitry, the D-SOA devices mentioned in Section IV-G, and the control logic by means of complex programmable logic devices [22]. The addition of a hysteric response increases the noise margins of the electronic decision logic by maintaining exaggerated transition thresholds. Thus, low swing glitches in the routing signal due to noise at the switching node receivers are eliminated. This improvement maintains correct routing functionality in the presence of glitches, yielding increased switching node robustness.

D. Faster Switching Node

The switching time and latency of the current node structure are limited by the electrical signal processing time. To increase the node throughput and minimize the packet truncation (Section IV-G), an all-optical node structure approach can potentially minimize the processing time of the current node structure [32]. The logic performed by the switching node would be performed in the optical domain. In fact, an all-optical self-routing switching node for the data vortex architecture has been implemented [33]. The node uses two Mach–Zehnder interferometers (MZIs) integrated with SOAs to perform the routing logic on the extracted header signal and the control signal from the inner cylinder node. A 1-dB power penalty (BER $< 10^{-9}$)

was measured partly due to the additional ASE of the MZI-SOA structure, compared to 0.4 dB for the current node structure (Section IV-C). Hence, more work is required to enable the application of all-optical node structures to large OPS network sizes.

E. Packet Structure

Although improvements to the aforementioned characteristics of the SOA can be implemented, modifications to the packet format may also mitigate the limitations imposed by imperfect optical switching devices. First, the routing wavelengths can be optimally assigned following the characterization of the spectral gain profile of the chosen SOAs. This will serve to mitigate the wavelength-dependent gain mismatches introduced by the devices to different channels in a multiple-wavelength packet. Specifically, channels experiencing less gain through the SOA devices will be relegated to contain higher order address information, since destination tag routing employed by the network architecture will utilize these channels first. Secondly, gain-flattening filters can be implemented to further compensate for the apparent gain imbalance.

VI. Conclusion

A novel paradigm for high-capacity low-latency OPS interconnection networks based on the data vortex architecture has been enunciated. The architectural design capitalizes on the immense bandwidth provided by optical signals encoded and transmitted over contemporary fiber-optic components and avoids common pitfalls and shortcomings of photonic technologies. A comprehensive discussion of the data vortex topology, switching node design, packet routing, and control has been provided. Further, an experimental system implementation, which shows the feasibility of large-scale data vortex interconnection networks, has been discussed. A complete characterization of the implemented network's physical layer has provided additional performance information, especially highlighting the cascadability of the individual switching nodes. Finally, design considerations for the topology and components of the next-generation data vortex have been addressed.

References

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

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

[3] C. Reed, "Multiple level minimum logic network," U.S. Patent 5 996 020, Nov. 30, 1999.

[4] N. F. Maxemchuk, "Regular and mesh topologies in local and metropolitan area networks," *AT&T Tech. J.*, vol. 64, no. 7, pp. 1659–1686, Sep. 1985.

[5] Q. Yang, K. Bergman, G. D. Hughes, and F. G. Johnson, "WDM packet routing for high-capacity data networks," *J. Lightw. Technol.*, vol. 19, pp. 1420–1426, Oct. 2001.

[6] Q. Yang and K. Bergman, "Traffic control and WDM routing in the data vortex packet switch," *IEEE Photon. Technol. Lett.*, vol. 14, pp. 236–238, Feb. 2002.

[7] Q. Yang and K. Bergman, "Performance of the data vortex switch architecture under nonuniform and bursty traffic," *J. Lightw. Technol.*, vol. 20, pp. 1242–1247, Aug. 2002.

[8] 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.

[9] C. Hawkins, B. A. Small, D. S. Wills, and K. Bergman, "The data vortex, an all optical path multicomputer interconnection network," *IEEE Trans. Parallel Distrib. Syst.*, vol. 18, pp. 409–420, Mar. 2007.

[10] C. Hawkins and D. S. Wills, "Impact of number of angles on the performance of the data vortex optical interconnection network," *J. Lightw. Technol.*, vol. 24, pp. 3288–3294, Sep. 2006.

[11] A. Shacham and K. Bergman, "On contention resolution in the data vortex optical interconnection network," *J. Opt. Netw.*, vol. 6, no. 6, pp. 777–788, Jun. 2007.

[12] 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, pp. 3066–3075, Oct. 2005.

[13] O. Liboiron-Ladouceur, B. A. Small, and K. Bergman, "Physical layer scalability of a WDM optical packet interconnection network," *J. Lightw. Technol.*, vol. 24, pp. 262–270, Jan. 2006.

[14] O. Liboiron-Ladouceur, K. Bergman, M. Boroditsky, and M. Brodsky, "Polarization-dependent gain in SOA-based optical multistage interconnection networks," *J. Lightw. Technol.*, vol. 24, pp. 3959–3967, Nov. 2006.

[15] B. A. Small, B. G. Lee, and K. Bergman, "Flexibility of optical packet format in a complete 12×12 data vortex network," *IEEE Photon. Technol. Lett.*, vol. 18, pp. 1693–1695, Aug. 2006.

[16] B. A. Small, T. Kato, and K. Bergman, "Dynamic power consideration in a complete 12×12 optical packet switching fabric," *IEEE Photon. Technol. Lett.*, vol. 17, pp. 2472–2474, Nov. 2005.

[17] B. A. Small and K. Bergman, "Slot timing consideration in optical packet switching networks," *IEEE Photon. Technol. Lett.*, vol. 17, pp. 2478–2480, Nov. 2005.

[18] B. G. Lee, B. A. Small, and K. Bergman, "Signal degradation through a 12×12 optical packet switching network," in *Proc. Eur. Conf. Opt. Commun.*, Sep. 2006, paper We3.P.131.

[19] O. Liboiron-Ladouceur, C. Gray, D. C. Keezer, and K. Bergman, "Bit-parallel message exchange and data recovery in optical packet switched interconnection networks," *IEEE Photon. Technol. Lett.*, vol. 18, pp. 770–781, Mar. 15, 2006.

[20] A. Shacham, B. A. Small, and K. Bergman, "A wideband photonic packet injection control module for optical packet switching routers," *IEEE Photon. Technol. Lett.*, vol. 17, pp. 2778–2780, Dec. 2005.

[21] A. Shacham and K. Bergman, "Optimizing the performance of a data vortex interconnection network," *J. Opt. Netw.*, vol. 6, no. 4, pp. 369–374, Apr. 2007.

[22] O. Liboiron-Ladouceur and K. Bergman, "Optimized switching node for optical multistage interconnection networks," *IEEE Photon. Technol. Lett.*, vol. 19, pp. 1658–1660, Oct. 15, 2007.

[23] Q. Yang, "Improved performance using shortcut path routing within data vortex switch network," *Electron. Lett.*, vol. 41, no. 22, pp. 1253–1254, Oct. 2005.

[24] B. A. Small, A. Shacham, and K. Bergman, "Ultra-low latency optical packet switching node," *IEEE Photon. Technol. Lett.*, vol. 17, pp. 1564–1566, Jul. 2005.

[25] Y. Pan, C. Qiao, and Y. Yang, "Optical multistage interconnection networks: New challenges and approaches," *IEEE Commun. Mag.*, vol. 37, pp. 50–56, Feb. 1999.

[26] H. Ahmadi and W. E. Denzel, "A Survey of modern high-performance switching techniques," *IEEE J. Sel. Areas Commun.*, vol. 7, pp. 1091–1103, Sep. 1989.

[27] 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," in *Proc. Opt. Fiber Commun. Conf.*, Mar. 2005, paper OWK1.

[28] 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 subsystems," *J. Opt. Netw.*, vol. 3, no. 11, pp. 802–809, Nov. 2004.

[29] J. M. Arnold, "The SPLASH-2 software environment," *J. Supercomput.*, vol. 9, no. 3, pp. 277–290, Sep. 1995.

[30] H. A. Haus, "The noise figure of optical amplifiers," *IEEE Photon. Technol. Lett.*, vol. 10, pp. 1602–1604, Nov. 1998.

[31] R. S. Tucker, K. Pei-Chen, and C. J. Chang-Hasnon, "Slow-light optical buffers: Capabilities and fundamental limitations," *J. Lightw. Technol.*, vol. 23, pp. 4046–4066, Dec. 2005.

[32] C. K. Yow, Y. J. Chai, R. V. Reading-Picopoulos, I. H. White, C. G. Leburn, A. McWilliam, A. A. Lagatsky, C. T. A. Brown, W. Sibbett, G. Maxwell, and R. McDougall, "Experimental demonstration of femtosecond switching of a fully packaged all-optical switch," in *Proc. Opt. Fiber Commun. Conf.*, Mar. 2005, paper OThE4.

[33] H.-D. Jung, I. T. Monroy, A. M. J. Koone, and E. Tangdiongga, "All-optical data vortex node using an MZI-SOA switch array," *IEEE Photon. Technol. Lett.*, vol. 19, pp. 1777–1779, Nov. 15, 2007.

Odile Liboiron-Ladouceur (M'95) was born in Montréal, QC, Canada. She received the B. Eng degree in electrical engineering from McGill University, Montréal, in 1999 and the M.S. and Ph.D. degrees from Columbia University, New York, in 2003 and 2007, respectively, all in electrical engineering.

Her thesis work was related to the data plane and physical layer analysis of optical interconnection networks. She is currently a Postdoctoral Researcher with the Photonics Systems group, McGill University, where she works on ultra-high-data-rate optical transport networks. She is the author or coauthor of 30 papers in peer-reviewed journal and conferences.

Dr. Liboiron-Ladouceur received a postdoctoral fellowship from the Natural Sciences and Engineering Research Council of Canada.

Assaf Shacham (S'03–M'07) received the B.Sc. degree *(cum laude)* in computer engineering from The Technion—Israel Institute of Technology, Haifa, in 2002 and the M.S. and Ph.D. degrees in electrical engineering from Columbia University, New York, in 2004 and 2007, respectively.

He is now with Aprius Inc, Sunnyvale, CA, where he is engaged in the development of high-performance computer interconnects. In the past he was with Intel Inc. and Charlotte's Web Networks, both in Israel, as an IC Design Engineer. He has authored or coauthored more than 25 papers in peer-reviewed journal and conferences. His research interests are computer architecture, high-performance optical computer interconnects, interconnection networks, and networks-on-chip.

Benjamin A. Small (S'98–M'06) received the B.S. (with honors) and M.S. degrees in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, in 2001 and 2002, respectively, and the M.Phil. and Ph.D. (with distinction) degrees in electrical engineering from Columbia University, New York, in 2005.

He is currently a Postdoctoral Research Scientist with the Department of Electrical Engineering, Columbia University. His interests include optoelectronic device physics and modeling as well as optical packet switching interconnection network traffic analysis and system-level behavior.

Benjamin G. Lee (S'04) received the B.S. degree in electrical engineering from Oklahoma State University, Stillwater, in 2004 and the M.S. degree in electrical engineering from Columbia University, New York, in 2008, where he is currently pursuing the Ph.D. degree in electrical engineering.

His research interests include silicon photonic devices, integrated optical switches, and semiconductor optical amplifiers in packet-switching applications.

Howard Wang (S'03) received the B.S. and M.S. degrees in electrical engineering from Columbia University, New York, in 2006 and 2008, respectively, where he is currently pursuing the Ph.D. degree in electrical engineering.

Caroline P. Lai (S'07) received the B.A.Sc. degree (with honors) in electrical engineering from the University of Toronto, Toronto, ON, Canada, in 2006 and the M.Sc. degree in electrical engineering from Columbia University, New York, in 2008, where she is currently pursuing the Ph.D. degree in electrical engineering.

Aleksandr Biberman (S'05) received the B.S. degree (with honors) in electrical and computer and systems engineering from Rensselaer Polytechnic Institute, Troy, NY, in 2006, and the M.S. degree in electrical engineering from Columbia University, New York, in 2008, where he is currently pursuing the Ph.D. degree in electrical engineering.

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 currently a Professor of electrical engineering at Columbia University, New York, where she also directs the Lightwave Research Laboratory. Her current research programs involve optical interconnection networks for advanced computing systems, photonic packet switching, and nanophotonic networks-on-chip. She is the Editor-in-Chief of the *Journal of Optical Networking*

Dr. Bergman is a Fellow of the Optical Society of America. She is currently an Associate Editor of IEEE PHOTONICS TECHNOLOGY LETTERS.