Building Ultralow-Latency Interconnection Networks Using Photonic Integration

Ultralow-latency interconnection networks have become a necessity in modern high-performance computing systems. Recent advances in photonic integration technology are paving the way for a disruptive step in the design of these networks. We present SPINet, an optical interconnection network architecture designed for implementation using photonic integration, providing an end-to-end photonic path while completely avoiding optical buffering. SPINet resolves contentions through message dropping, but facilitates message recovery using a novel physical-layer acknowledgment protocol.

Assaf Shacham
Aprius, Inc.

Keren Bergman
Columbia University

Contemporary high-performance computing (HPC) systems use the distributed shared-memory (DSM) paradigm, in which the entire memory is logically shared among all the processors but is physically implemented using memory modules distributed across many computing nodes. This approach simplifies programming, provides portability of software, and exhibits improved scalability over traditional shared-memory systems.[1] Large-scale DSM systems, however, suffer from a fundamental communication problem that significantly affects scalability: increased latency of remote memory accesses.[2] With faster processor speeds, the remote-access latency problem is becoming critically more pronounced, as each memory access consumes a correspondingly larger number of clock cycles.

The latency of a message through a network has three contributing components:

- time of flight t_p, governed by the transmission-line propagation velocity and the physical distance between the nodes, and assumed to be fixed;
- serialization latency t_b in each individual router in the path, affected by packet size b and transmission rate R, as $t_b = b/R$; and
- routing latency t_r in each router, which is the sum of forwarding latency t_f, queuing latency t_q, and t_b.

Published by the IEEE Computer Society.

The total latency can therefore be expressed as

$$D = t_p + t_b + (h - 1)t_r = t_p + h(b/R) + (h - 1)(t_f + t_q). \quad (1)$$

Here, $h \geq 2$ is the number of hops in the message's path, including the source and the destination.[3] As systems scale to large port counts, their diameter, defined as the length of the minimal path connecting the most distant pair of nodes, grows.[4] The resulting hop count (h) that packets experience is larger, contributing to a high latency that can be detrimental to overall system performance.[2]

Equation 1 suggests three methods for reducing message latency: increasing transmission rate R; reducing the network diameter by using high-radix routers; and lowering queuing latency t_q in the routers. Reducing memory access latency by combining these methods is key to improving the performance of future HPC systems.[2]

Interconnection networks with low latency and high bandwidth have therefore become an important component in the design of HPC systems. To construct such networks, designers currently use high-speed serial links as well as high-performance cross-point switching fabrics. However, a performance gap is beginning to emerge between the processors, whose performance scales quickly according to Moore's law, and the interconnecting medium, which fails to advance at the required rate because of fundamental physical limitations.[5] Electrical losses caused by the skin effect distort high-speed signals transmitted across large electronic backplanes. Sophisticated signal-processing techniques, such as pre-emphasis and equalization, can mitigate these effects to some extent, but they add to the overall latency and are expensive both in power and area.[5] An alternative approach is to use short transmission lines and an indirect topology such as a mesh or a torus, based on low-radix routers, but this approach leads to further increases in the overall latency because each packet must traverse a larger number of hops.[4]

Researchers have suggested photonic interconnection networks as a potentially disruptive technology with the capability to overcome these limitations and provide the required performance scaling.[6–8] Optical fibers' enormous bandwidth—approximately 32 THz—facilitates the transmission of multiple data streams on a single fiber at very high data rates using wavelength-division multiplexing (WDM). The low loss in fibers, nearly zero for the distances relevant to interconnection networks, alleviates the need for regeneration and effectively removes the signal transmission limitation. The photonic medium also allows for bidirectional transmission and switching of high-rate data using optical switching elements completely transparent to the modulated data. This property, known as *bit-rate transparency*, makes optical switching very attractive in terms of power consumption: The power consumed by an optical switch is independent of the bandwidth routed through it. The power per unit bandwidth in optical interconnection networks can therefore be dramatically reduced if a high transmission bandwidth is used (for example, by exploiting WDM). Several experimental optical packet-switching systems use semiconductor optical amplifiers (SOAs) as on/off photonic gates, providing a substantial gain over a wide switching band, and subnanosecond switching times.[6,8]

However, photonic technology presents some fundamental design challenges, specifically because of its lack of efficient buffering and processing capabilities. Although researchers are investigating some promising technologies, such as photonic crystals, that may prove useful in constructing photonic memories and logic gates, such technologies have failed to reach commercialization thus far. Optical buffers based on recirculating fiber delay lines have been demonstrated, as have interferometeric optical logic gates, but their dimensions and bulkiness prohibit them from becoming cost-effective solutions.

The main impediment to the construction of photonic interconnection networks lies in the high cost and large footprints associated with using discrete optical elements such as lasers, modulators, switches, and passive optics. Photonic integration, the

Recent advances in photonic integration

The main engine driving the explosive growth of electronics over the last four decades has been the ability to integrate an ever-growing number of transistors onto silicon dies. This very large scale integration (VLSI) led and still is leading to constant improvements in performance, enabling an economy of scale in which new applications emerge while costs continue to drop.

For many years, engineers and researchers have been trying to reproduce this electronic success in the fabrication of photonic integrated circuits (PICs). Photonic integration can be divided into two main technology thrusts, coinciding with two material systems. First, III–V semiconductors—such as indium phosphide (InP) and gallium arsenide (GaAs)—are traditionally used for the manufacturing of photonic devices, especially lasers and photodiodes, because of their good electro-optic properties.[1] However, integrating these materials has traditionally been very difficult, especially achieving the process uniformity and reproducibility required for mass production in such processes as epitaxy, dry etching, and lithography.[2]

Photonic integration in the second material system, silicon, can potentially leverage the enormous experience and knowledge of silicon processing amassed in the electronics industry. Unfortunately, silicon has very poor electro-optic properties resulting from its being an indirect bandgap material. This means that optical gain, a necessary condition for the creation of laser sources, is very hard to generate in silicon,[1] so light sources, for example, must be externally coupled to the silicon chip.

Although research efforts have been going on for decades, the escalating need for low-power, high-bandwidth interconnects in telecommunications and data communications has lately intensified attention on photonics. This sidebar describes several breakthroughs that have occurred in the last two or three years (see Figure A). If photonic integration, both in silicon and InP, matures and acquires the required reproducibility, reliability, and low cost, the opportunities are seemingly limitless.

Micro-ring resonator-based silicon modulators

Modulators are key elements in optical communications systems. They are used to encode a high-speed electronic signal on constant-wave laser light, thus converting it to a stream of light pulses. Silicon micro-ring resonator-based modulators, fabricated by a group of researchers at Cornell University, exhibit good optical properties accompanied by unprecedented low power consumption, small footprint, and modulation rates up to 12.5 Gbps.[3,4] These devices, although facing challenges in thermal stability and manufacturing reliability, are very promising components for future chip-to-chip and intrachip optical communication systems.

Silicon optical delay lines

Optical buffering is a fundamental problem in optical communication systems. There is no optical equivalent for random access memory (RAM), so at this point research efforts focus on optical delay lines, in which pulses of light experience large group delay and are effectively stored. An IBM research group has shown group delay in excess of 500 ps for data rates up to 20 Gbps,[5] effectively storing 10 bits in an area smaller than 0.09 mm^2. The researchers emphasize, however, that fundamental trade-offs exist between the group delay, device size, insertion loss, and operational bandwidth.

Evanescent, electrically pumped AlGaInAs-silicon laser

Generation of light in silicon is very difficult because of its indirect bandgap. Collaborating researchers at the University of California, Santa Barbara, and Intel worked around this problem by bonding AlGaInAs (aluminum gallium indium arsenide, a III–V semiconductor) with silicon to fabricate the world's first electrically pumped AlGaInAs-silicon laser.[6] Light is amplified in the AlGaInAs, but the laser cavity is actually defined in silicon, so that light need not be coupled into the silicon chip. With this method, hundreds of lasers can be fabricated in one bonding step, so it is suitable for high-volume, low-cost integration.

CMOS-compatible photonics

Luxtera of Carlsbad, California, has built a library of optical communication sub-blocks—such as modulators, waveguides, photodetectors, and holographic lenses—all compatible and integratable with standard CMOS processes in silicon on insulator (SOI) technology.[7] This is a first attempt to provide a toolkit of optical components that chip designers can use to implement high-speed, low-power chip-to-chip interconnects. Luxtera's library can also be used to construct optical subsystems such as WDM transceivers in silicon, thus potentially reducing costs.

fabrication of circuits implementing many photonic functions in a single package, is promising as a way to eliminate these final barriers (see the sidebar, "Recent advances in photonic integration"). Because the elements making up the prohibitive cost of optical networks lie mainly in the assembly and packaging of very large systems, and because a significant share of the power consumption rises from coupling losses between individually packaged devices, integration of large parts of the network on a single photonic integrated circuit (PIC) alleviates these drawbacks. Monolithic indium phosphide (InP) PICs containing 50 photonic functions, reported in the scientific literature,[9] are now commercially available. Additionally, silicon-based optical and electro-optical components such as modulators, photodetectors, and waveguides—all compatible with standard CMOS processing techniques—have recently become available, promising an unprecedented potential for low-cost electronic-optical interfacing.[10]

Large-scale photonic integration in InP

Finally, Infinera of Sunnyvale, California, takes a different approach: large-scale photonic integration in InP. The company fabricates PICs implementing complete WDM transmission systems with very high bandwidths.[2] Infinera's first product is a PIC integrating nearly 50 photonic functions: 10 transmission systems (lasers, 10-Gbps modulators, and power- and wavelength-control devices) multiplexed onto a single-output optical fiber through an on-chip waveguide grating. The complementary receiver PIC consists of a wavelength demultiplexer and 10 high-speed photodetectors.[2] These two PICs comprise a 100-Gbps transmission system, which the company soon followed with a 400-Gbps system (10 × 40-Gbps) and a 1.6-Tbps (40 × 40-Gbps) transmitter PIC. The main challenges to this approach are cost reduction and proving the reliability of the PICs in compliance with standards set by the VLSI industry.

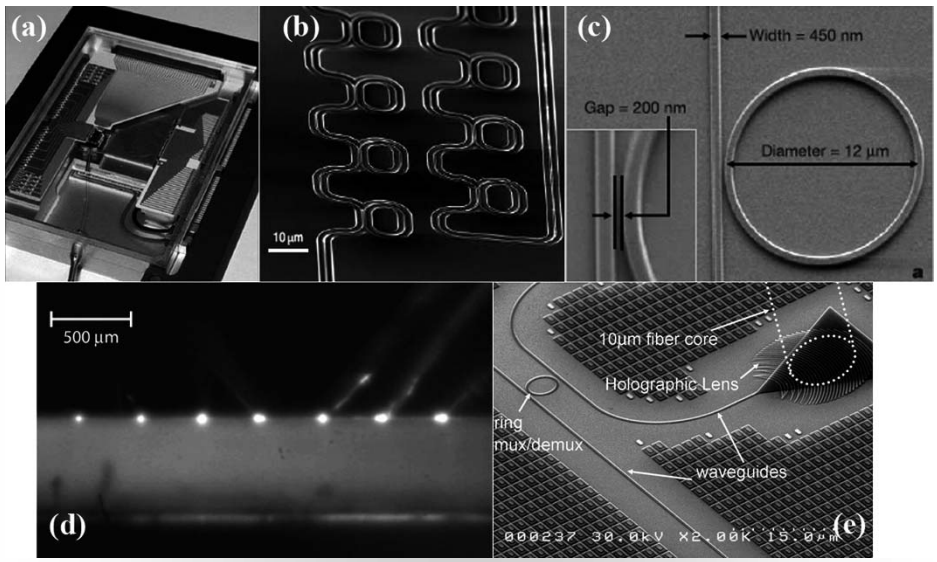

Figure A. 10 × 10-Gbps transmitter PIC (Infinera[2]) (a); optical delay lines in silicon (IBM,[5] image courtesy Yurii Vlasov) (b); microring resonator modulator (Cornell,[3] image courtesy Michal Lipson) (c); hybrid AlGaInAs-silicon laser (Intel/UCSB,[6] image courtesy John Bowers) (d); and photonic CMOS devices (Luxtera[7]): a lens, a filter, and waveguides (e).

References

1. A. Yariv and P. Yeh, *Photonics: Optical Electronics in Modern Communications*, Oxford Univ. Press, 2006.
2. R. Nagarajan et al., "Large-Scale Photonic Integrated Circuits," *IEEE J. Selected Topics Quantum Electronics*, vol. 11, no. 1, Jan.-Feb. 2005, pp. 50-65.
3. Q. Xu et al., "Micrometre-Scale Silicon Electro-Optic Modulator," *Nature*, vol. 435, May 2005, pp. 325-327.
4. Q. Xu et al., "12.5 Gbit/s Carrier-Injection-Based Silicon Micro-Ring Silicon Modulators," *Optics Express*, vol. 15, 2007, pp. 430-436.
5. F. Xia, L. Sekaric, and Y.A. Vlasov, "Ultracompact Optical Buffers on a Silicon Chip," *Nature Photonics*, vol. 1, Jan. 2007, pp. 65-71.
6. A.W. Fang et al., "Electrically Pumped Hybrid AlGaInAs-Silicon Evanescent Laser," *Optics Express*, vol. 14, no. 20, Oct. 2006, pp. 9203-9210.
7. C. Gunn, "CMOS Photonics for High-Speed Interconnects," *IEEE Micro*, vol. 26, no. 2, Mar.-Apr. 2006, pp. 58-66.

When photonic integration is harnessed to construct interconnection networks, however, buffering becomes very difficult. The optical packet occupies a fixed length of a waveguiding medium. This length is the product of the speed of light in the medium and the duration of the packet. A typical 100-ns packet occupies 20 meters of silica fiber or approximately 6 meters of semiconductor waveguide. Consequently, buffering optical packets within a PIC is not currently practical; interconnection networks based on photonic integration should be truly bufferless, using other means of contention resolution.

In this article, we present SPINet (Scalable Photonic Integrated Network), an architecture designed to be implemented using photonic integration technology. Based on an indirect multistage interconnection network (MIN) topology, SPINet exploits WDM to simplify the network design and provide very high bandwidths. SPINet does not employ buffering, instead resolving contention by dropping contending messages. A novel physical-layer acknowledgment protocol provides immediate feedback, notifying the terminals whether their messages are accepted, and facilitates retransmissions when necessary in a manner resembling

collision detection techniques in traditional multiple-access media networks.

Related work

Researchers have proposed several architectures that use photonic switching for the design of HPC interconnection networks. The Rapid architecture suggests the use of a hierarchical structure of passive optical networks, with fixed wavelength assignment and dynamic time-slot preallocation.[7] Although this structure is simple to implement and does not require active optical switching elements, hopping between different networks requires optical-electronic (O/E) and electronic-optical (E/O) conversions, queuing, and reallocation of time slots, all contributing to latency and complexity. Furthermore, the use of passive networks that typically are based on power splitting limits scalability, so this architecture would require large diameters in large systems.

The data vortex is an optical packet-switching MIN, scalable to a very large port count, and composed of simple, SOA-based, bufferless switching nodes.[8] This network resolves contentions by deflecting contending packets to alternative paths that connect switching nodes in the same stage. The packets' addresses are encoded optically, and the payload, encoded on multiple wavelengths to attain very high bandwidths, is maintained in the optical domain from source to destination. Researchers have proven the data vortex scheme feasible on a 12 × 12 network fully implemented in a laboratory setup, and have confirmed error-free transmission at a peak bandwidth of 160 Gbps.[8,11]

The limitations of the data vortex architecture arise from the unpredictable packet path that results from the network's complex topology and its use of deflection routing. The different number of switch element hops taken by each packet has a performance penalty of unpredictable latency and packet reordering. It also places constraints on the physical design of the optical and optoelectronic elements such as the SOAs and the receivers. The buffering of packets on deflection fibers within the network prohibits the integration of a data vortex interconnection network, because the physical space occupied by the packets cannot be reduced.

In the Osmosis project, researchers have constructed an SOA-based 64 × 64 crossbar switching fabric and developed a high-speed scheduling algorithm to maximize its utilization.[6] SOAs function as the cross-points in the switching fabric, and the Osmosis team has demonstrated 40-Gbps peak bandwidth per port. Internal paths are wavelength-multiplexed in the switching fabric to reduce the number of SOAs and the total cost. The researchers suggest scaling Osmosis beyond the 64-port system using a fat-tree topology, performing O/E/O conversion and queuing between stages.

With a running prototype, the Osmosis architecture comes closest to actual implementation and possible commercialization, but it might not be the ideal choice for very-large-scale systems. A scheduled crossbar requires additional communication lines, limits scalability, and is often accompanied by a notable queuing delay. In the suggested fat-tree architecture, packets can experience this delay several times, resulting in a possibly significant latency penalty.

Architecture overview

We have focused on developing an architecture that provides high bandwidth and very low latencies (less than 1 μs, application to application) for a large-scale interconnection network (more than 1,024 ports). Other properties, such as error-free transmission—a raw bit error rate (BER) of 10^{-12}—cost-effectiveness, and power-efficiency are also of interest.

Basic architecture

A SPINet network is a transparent optical MIN, composed of 2 × 2 SOA-based bufferless photonic switching nodes.[12,13] The specific topology can vary between implementations, and switching nodes of higher radices can be used if they are technologically available. The entire network is integrated on a PIC. Its terminals are physically located on the computing nodes of the HPC system, and are connected to the PIC through optical fibers. The terminals are assumed to be synchronized, the network is slotted and synchronous, and

messages have a fixed duration. The minimal slot duration is determined by the round-trip propagation time of the optical signal from the terminals to the PIC. A slot time of 100 ns can, therefore, accommodate a propagation path of nearly 20 meters and a distance of 10 meters between the terminals and the PIC.

The topology we study in this article is an Omega network, an example of a binary banyan topology.[4,14] An $N_T \times N_T$ Omega network consists of $N_S = \log(N_T)$ identical stages. Each stage consists of a perfect-shuffle interconnection followed by $N_T/2$ switching elements, as Figure 1a shows. In the Omega network, each switching node has four allowed states (straight, interchange, upper broadcast, and lower broadcast). In this implementation, we have modified the switching nodes by removing the broadcast states and introducing four new states (upper straight, upper interchange, lower straight, lower interchange). In these four states, the node passes data from only one input port to an output port and drops the data from the other port (see Figure 1b).

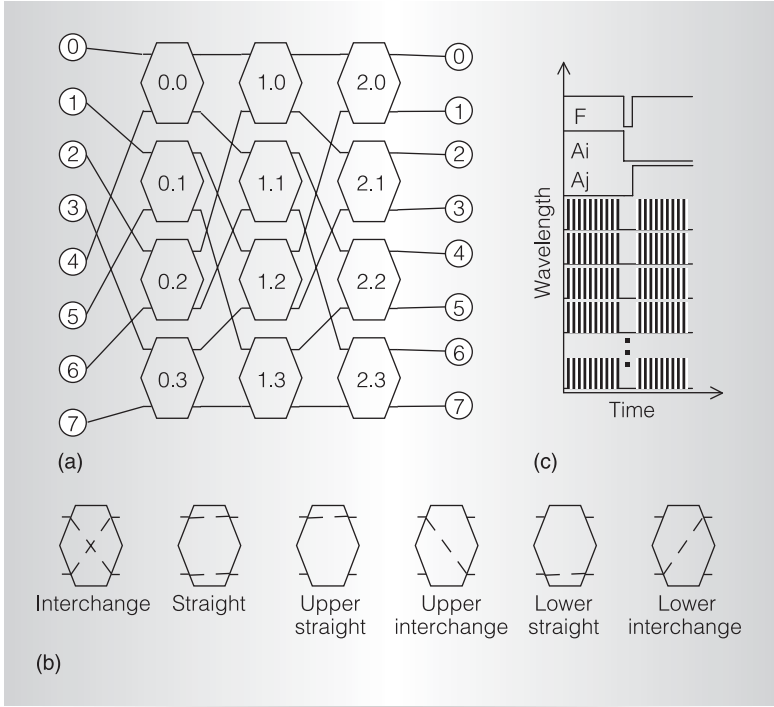

Figure 1. An 8 × 8 Omega network (a); switching nodes' six states (b); and wavelength-parallel messages (c). Header bits and payload are encoded on dedicated wavelengths.[12]

At the beginning of each slot, any terminal may start transmitting a message, without a prior request or grant. The messages propagate in the fibers to the input modules on the PIC and are transparently forwarded to the switching nodes of the first stage. (We discuss the functionality of the input modules in the next section.) At every routing stage when the leading edges of the messages are received from one or both input ports, a routing decision is made, and the messages continue to propagate to their requested output ports. In the case of output-port contention in a switching node, the network drops one of the contending messages. The choice of which message to drop can be random, alternating, or priority-based. Because the propagation delay through every stage is identical, all the leading edges of the transmitted messages reach all the nodes of each stage at the same time.

The nodes' switching states, as determined by the leading edges of the packets, remain constant throughout the duration of the message, so the entire message follows the path acquired by the leading edge. Because the propagation delay through the PIC is very short compared to the duration of the messages, the messages stretch across the PIC, effectively creating *transparent lightpaths* between the inputs and outputs. When the messages reach the output modules, they are transparently forwarded on the output fibers to the appropriate terminals; at the same time, the destination terminal generates an acknowledgment optical pulse and sends it on the previously acquired lightpath in the opposite direction. Because the nodes' switching elements preserve their states and support bidirectional transmission, the source terminal receives the acknowledgment pulse, which serves as confirmation of the message's successful reception.

When the slot time is over, all terminals cease transmission simultaneously, the switching nodes reset their switching states, and the system is ready for a new slot. The slot duration is set to ensure that the acknowledgment pulses are received at the source terminals before the slot ends. Hence,

before the beginning of the next slot, every terminal knows whether its message was accepted; when necessary, it can choose to immediately retransmit the message.

Using ultralow-latency propagation through the PIC, SPINet eliminates the need for central scheduling, instead employing the distributed computing power of the switching nodes to produce an input-output match at every slot. This process of *implicit arbitration* enables scalability to large port counts without burdening a central arbiter with computations of complex maximal matches. Because SPINet uses blocking topologies to reduce hardware complexity, the network's utilization is lower than that of a traditional maximum-matched nonblocking network (as in switching fabrics for high-performance Internet routers). In later sections, we'll investigate techniques that exploit the properties of integrated photonics to increase utilization by adding a small number of stages.

SPINet uses the wavelength domain to facilitate a routing mechanism in the switching nodes that can instantly determine and execute the routing decision upon receiving the leading edges, without any additional information exchange between the switching nodes. The mechanism also maintains a constant switching state for the duration of the messages. The messages are constructed in a wavelength-parallel manner, similar to that used in the data vortex architecture,[8] trading off a part of the enormous bandwidth of optical fibers to simplify the switching-node design. As Figure 1c shows, the routing header and the message payload are encoded on separate wavelengths and are received concurrently by the nodes. The header consists of a frame bit that denotes the message's existence and a few address bits. Each header bit is encoded on a dedicated wavelength and remains constant throughout the message duration. When a binary network is used, a single address bit is required at every stage, and therefore the number of wavelengths required for address encoding is the number of routing stages in the network, or $\log_2$ of the number of ports. The switching nodes' routing decisions are based solely on the information extracted from the optical header, as encoded by the source. The switching nodes neither exchange additional information nor add any to the packet. The payload is encoded on multiple wavelengths at the input terminal, which segments it and modulates each segment on a different wavelength, using the rest of the switching band. A guard time is allocated before payload transmission, accommodating the SOA switching time, clock recovery in the payload receivers, and synchronization inaccuracies.

The wavelength-parallel structure also facilitates a low-cost solution for E/O conversion: the replacement of prohibitively expensive high-speed serial modulators and receivers with a set of lower-speed devices to attain the same bandwidth. These devices can be, for example, low-cost, directly modulated 10-Gbps vertical-cavity, surface-emitting lasers (VCSELs), or even on-chip silicon transceivers.[10] The achievable low-cost bandwidth and the network's inherent wide switching band and bit-rate transparency suggest a bandwidth-time trade-off: For a given required peak bandwidth, an $S\times$ speedup is used in the wavelength domain, whereas the packet rate (the load offered to the network) is reduced by a $1/S$ factor. For example, for a port designed to provide a peak bandwidth of 40 Gbps, eight wavelengths are modulated at 10 Gbps each, and a new packet is generated once every two slots on average. This $2\times$ speedup lowers the load offered to the network, leading to reduced queuing latency. We can therefore describe the peak bandwidth as

$$B_{\text{peak}} = \frac{R \times N_{\text{payload}} \times \eta}{S}, \quad (2)$$

where R is the serial data rate (for example, 10 Gbps), N_{payload} is the number of payload wavelengths used, and η is the slot efficiency parameter, defined as

$$\eta = \frac{T_{\text{SLOT}} - T_{\text{G}}}{T_{\text{SLOT}}} \quad (3)$$

where T_{SLOT} is the slot duration, and T_{G} is the guard time.

Using path diversity to increase utilization

Self-routing in banyan networks, such as the Omega, obviously limits throughput, owing to the network's internal blocking.

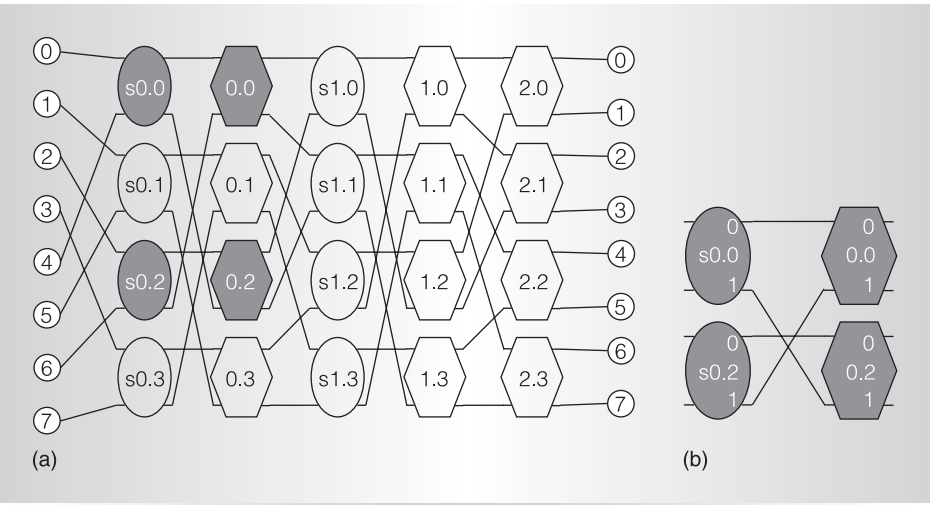

Figure 2. An 8 × 8 Enhanced Omega network (a) and detail showing a pair of buddies (b). Hexagons indicate routing nodes, and ellipses indicate deflecting nodes. Nodes 0.0 and 0.2 are buddies. Nodes s0.0 and s0.2 are a part of the scattering stage.[12]

On the other hand, because of their $O(N^2)$ complexity, strictly nonblocking networks aren't practical for large systems.[3,14] Rearrangeably nonblocking networks require central management, contradicting the self-routing paradigm, and also typically use higher-radix switching nodes.[4] We therefore use the binary banyan structure and attempt to mitigate its internal blocking by increasing the path diversity. Different messages can travel through different paths while preserving the simplicity of the switching nodes and the minimal latency routing through the PIC. This article suggests and investigates two techniques of using path diversity to mitigate contentions and increase utilization.

To enable these techniques, we introduce a new type of switching node: the *deflecting node*. Deflecting nodes are completely identical to the routing nodes introduced earlier, except for the manner in which they deal with contending messages. Whereas a routing node drops one of the contending messages, a deflecting node deflects it to the undesired port. Deflecting nodes never lose messages, but to avoid routing errors certain restrictions must be observed on the placement of the deflection nodes in the network.

Enhanced Omega network. In the Enhanced Omega (EOM) topology (Figure 2), a stage of deflecting nodes, termed the *scattering stage*, is placed before each routing stage except the last one.[12] Each pair of deflecting nodes in the scattering stage is connected to a pair of routing nodes called *buddies*. According to the banyan buddy property, any node in a banyan network has a buddy node that is connected to the same nodes in the following stage.[14] The deflecting nodes examine the same address bit as the subsequent routing nodes, identify messages that would contend, and attempt to scatter them to different routing nodes. The buddy property guarantees that the scattering operation causes no routing errors. Scattering stages cannot be inserted before the last routing stage, because each of these routing nodes is connected to two output modules and has no buddy. Therefore, construction of an EOM network requires the addition of $N_S - 1$ stages to the original N_S-stage Omega network.

Distribution network and path adjustments. A second technique to increase utilization makes use of the ultralow propagation latency of the optical signals and acknowledgment pulses to make path adjustments before the payload transmission begins.[15] To support this mechanism, a distribution network of deflecting nodes, with an Omega topology, is inserted before

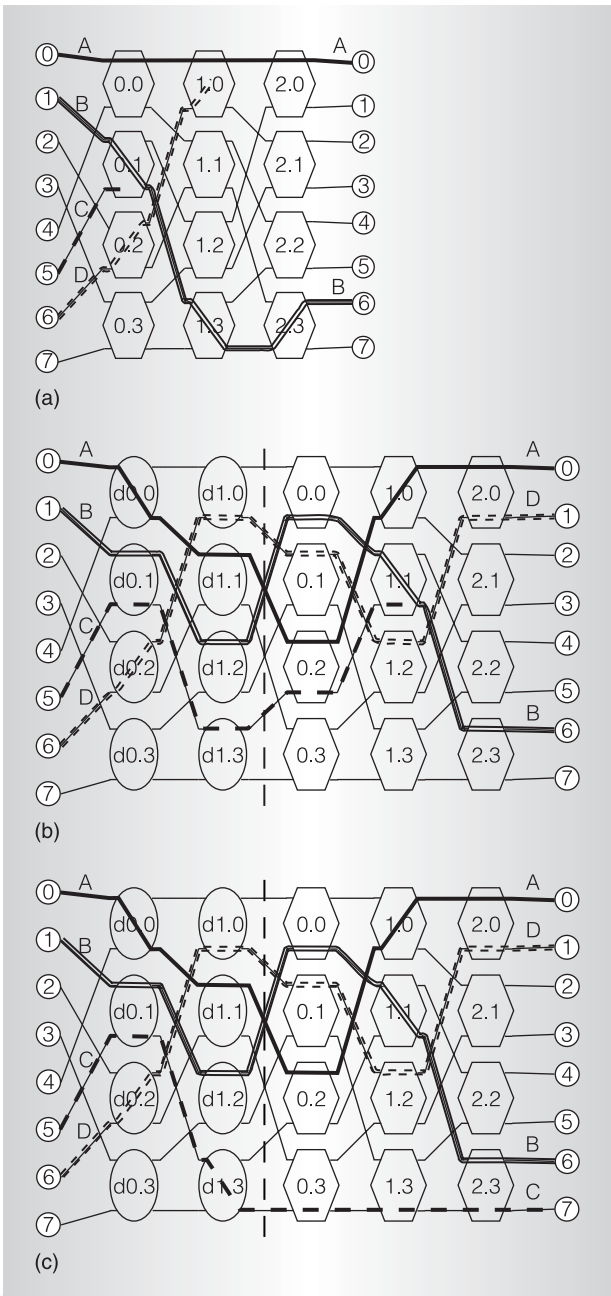

Figure 3. Example of the benefits of the distribution network: The Omega network (a) drops messages C and D. In Omega with a distribution network (b), message D is received, and C is dropped. With path adjustments (c), the network adjusts message C's path, and it is now successfully received.[15]

the routing network. A distribution address, encoded on reserved wavelengths, is attached by the input modules to incoming messages when they are injected into the PIC. The messages are routed through the distribution network and then to their destinations by the routing network. The advantages of the distribution network are twofold: First, by encoding a random distribution address, the load on the routing network can be balanced regardless of the real traffic pattern.[4] Second, the physical-layer acknowledgment protocol can be exploited to make path adjustments by changing the distribution address if the network drops the message on the first attempt. These path adjustments can be made in several iterations within the same time slot, during the guard time that precedes the payload transmission. Each iteration takes as long as the acknowledgment reception latency—that is, the sum of the round-trip time across the PIC and the response time of the output modules in generating the acknowledgment pulses ($T_{PIC\text{-}RT} + T_{ACK}$). The guard time is increased accordingly, affecting the slot efficiency parameter (η) according to Equation 3. The number of iterations must therefore be optimized against the improved utilization gained from multiple iterations. Figure 3 shows an example of the path adjustment process.

Performance analysis

Although researchers have studied the performance of banyan networks extensively using analytical models as well as simulations,[14,16] the performance analysis of SPINet presents a different problem space. Banyan networks are often analyzed as packet-switching networks, where each packet progresses one stage per time slot. The ultralow latency of the PIC, enabling the fast acknowledgment protocol and path adjustments within the same time slot, is a new technical possibility requiring a different analysis. We conducted a set of synthetic-workload simulations to study various aspects of SPINet's performance: We compared the EOM to the Omega network, studied the effect of path adjustments, and applied nonuniform traffic patterns to the network.

We conducted our simulations with infinite input queues at every source under the condition that every rejected message

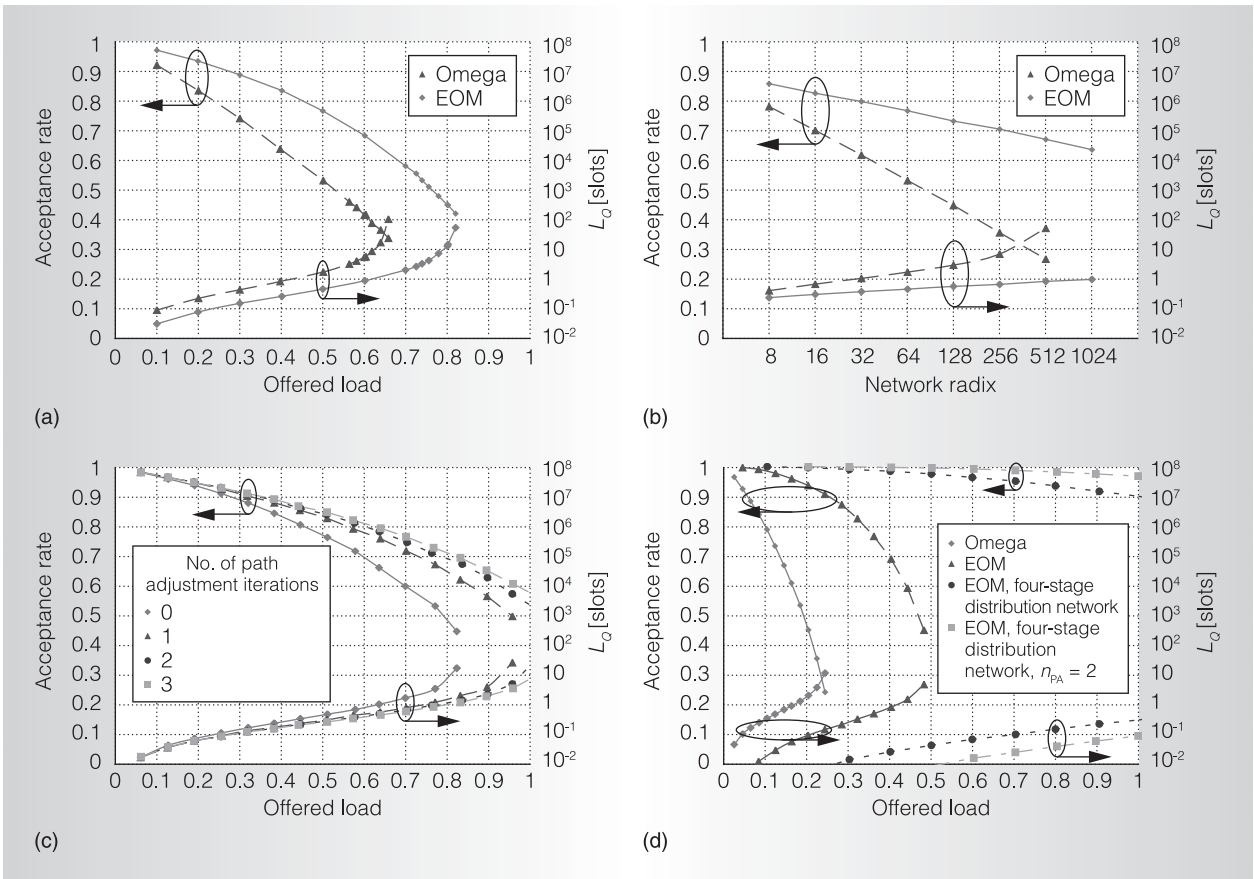

Figure 4. Performance of Omega and EOM topologies under various loads and operation modes: Omega and EOM versus offered load (r) under uniform Bernoulli traffic, 64-port network (a); Omega and EOM versus network radix, under uniform Bernoulli traffic, $r = 0.25$ (b); 64-port EOM with different numbers of path adjustment iterations (n_{PA}) (c); and bit-reversal traffic, 64-port network (d).

reattempts transmission until it transmits successfully. For every simulation, we measured two key metrics: the mean queuing latency (L_Q) in time slots and the acceptance rate—the probability that a given injection attempt is successful. Because every rejection leads to an additional injection attempt, the acceptance rate can be viewed as a measure of the system's efficiency. For each measurement, we took ensemble averages over 10 batches of 6,000 messages.

In the following simulations, we used a wavelength speedup factor (S) of 2, and computed the offered load (r) as $r = R_{data}/B_{peak}$, where R_{data} is the tributary data rate and B_{peak} is the port peak bandwidth. When not stated otherwise, the simulations involved 64-port networks, and distribution networks were four stages long.

Design parameters

Since banyan networks have a well-known analytical model for uniform Bernoulli traffic,[14] we use the Omega as a reference to assess the EOM's performance gain under this traffic profile. Initially, we compared the performance of the Omega and EOM topologies under uniform Bernoulli traffic for varying loads (Figure 4a) and varying switch radices (Figure 4b). Because the naïve Omega implementation saturates at a load of $r \approx 0.65$ (when $S = 2$), we conducted the second simulation (Figure 4b) under a load of $r = 0.50$.

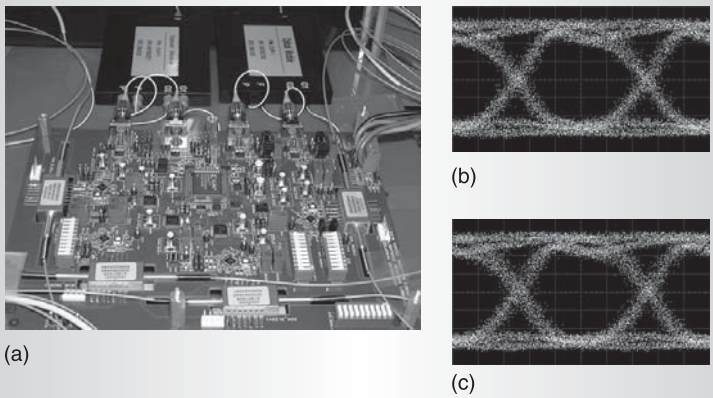

Figure 5. Experimental 2 × 2 switching node (a), and eye diagrams, at 10 Gbps, of input (b) and output (c) optical signals.

As mentioned earlier, path adjustment iterations contribute to utilization but reduce the slot efficiency parameter (η). It is therefore of interest to study their contribution to find an optimization point. We simulated the performance of an EOM network with a four-stage distribution network under uniform Bernoulli traffic, with a varying number of path adjustment iterations (n_{PA}); Figure 4c shows the results.

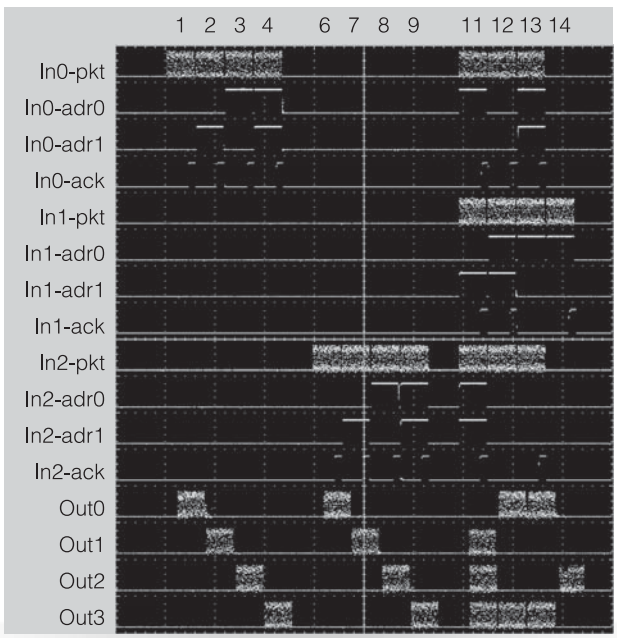

Figure 6. Optical waveforms of the signals at the network's input and output ports, demonstrating correct routing and contention resolution between optical packets.

The performance gain is noticeable but seems to have a diminishing return, which becomes very small for $n_{PA} > 2$.

Nonuniform traffic

Uniform traffic is considered a favorable traffic pattern for indirect networks such as the Omega.[4] Unfortunately, real traffic in HPC interconnection networks is not uniform. As an example of more realistic, nonuniform traffic, we consider the *locality of reference* scenario.[16] This scenario models a case in which every processor has a "favorite" memory module. The performance in this scenario strongly depends on the exact locality pattern applied to the network. Whereas a banyan network can route sorted traffic without any contentions (as in Batcher-banyan networks),[14] bit-reversal traffic, conversely, is known to be an adversarial traffic pattern to banyan networks.[4] Figure 4d illustrates the performance of four SPINet configurations (Omega, EOM, EOM plus a distribution network with $n_{PA} = 0$, and EOM plus a distribution network with $n_{PA} = 2$) under bit-reversal traffic. The improvements of the EOM, the distribution network, and the path adjustments are noticeable. Because the bit-reversal pattern, as any other permutation pattern, has no output port contentions, it significantly outperforms uniform traffic once the distribution network mitigates its internal contention effect.

Experimental demonstration

Although fabrication of a fully integrated network requires resources beyond the reach of a research laboratory, we can demonstrate the feasibility of SPINet's underlying concepts using discrete optic and optoelectronic elements. We have reported the fabrication of an SOA-based, nonblocking 2 × 2 switching node on a printed circuit board, and have confirmed error-free transmission (BER $< 10^{-12}$) of 160-Gbps peak bandwidth (16 wavelength × 10 Gbps) on all input and output paths.[13] Figure 5 is a photograph of a 2 × 2 switching node with eye diagrams of the 10-Gbps optical signal at the input and the output.

We assembled a 4 × 4 demonstration network composed of four routing nodes organized in two stages to test and demon-

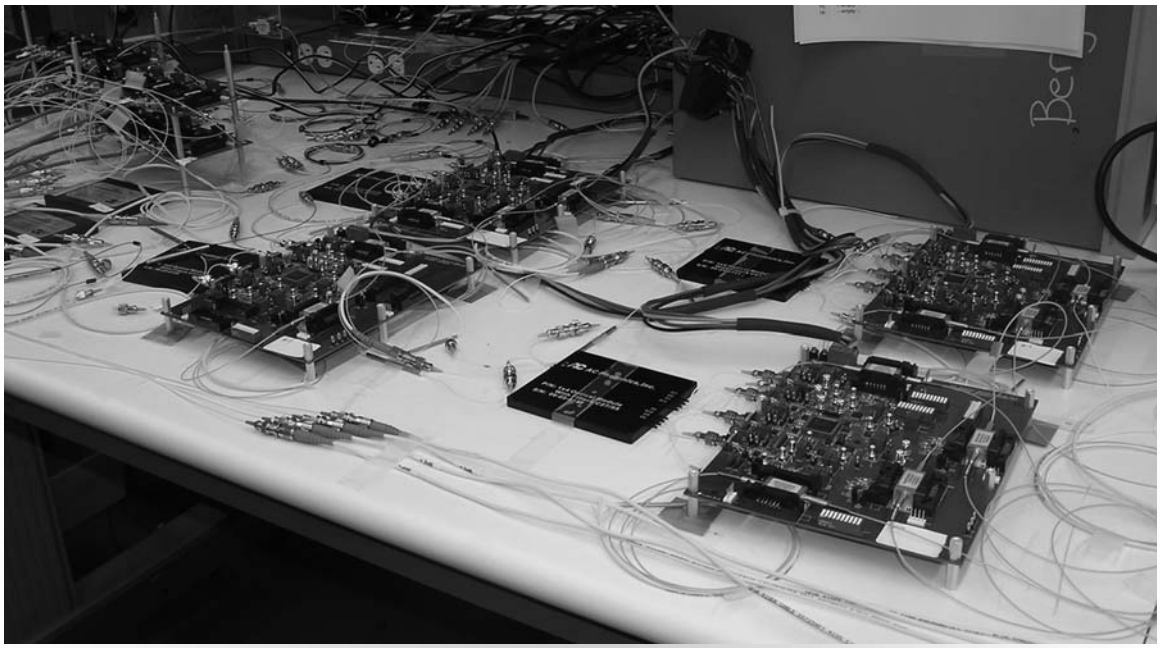

Figure 7. Photograph of the four-node experimental setup along with the packet generation circuitry.

strate SPINet's concepts: decoding of optical headers, correct transmission and switching of messages, and routing of acknowledgment pulses in the opposite direction. As an experimental setup to test the network, 192-ns packets, carrying control and payload wavelengths, were generated independently for three input ports and driven into the network. The packets were addressed to all four output ports, where a communications signal analyzer (CSA) analyzed them.

Figure 6 is a waveform plot as received by the CSA: The three inputs are labeled I0, I1, and I2, and for each input the optical waveforms of a payload wavelength (pkt), a first-stage address (address MSB, labeled adr0), and a second-stage address (address LSB, labeled adr1) are shown as driven into the network. Additionally, the figure shows the acknowledgment pulses (ack) for each input port and the ejecting packets for each output ports (Out0, Out1, Out2, and Out3).

The test pattern is 16 slots long, with slots numbered 0 to 15. In slots 1 through 4, and 6 through 9, optical packets are transmitted from input ports I0 and I2, respectively, to each output port, demonstrating the functionality of all output ports.

The following four slots demonstrate interaction between packets and contention resolution:

- Slot 11 shows three packets routed to their destinations without contentions (I0→Out2, I1→Out1, and I2→Out3), and ack pulses received appropriately.
- Slot 12 demonstrates an output port contention between messages I0→Out0 and I2→Out0. One message is dropped (I2→Out0), and I2 does not receive an ack pulse.
- Slot 13 demonstrates an internal contention between messages I0→Out3 and I1→Out2. Message I1→Out2 is dropped, and I1 does not receive an ack pulse. The dropped message from the previous slot (I2→Out0) is now retransmitted successfully.
- Slot 14 shows the dropped message from slot 13 (I1→Out2) being retransmitted and received successfully. I1 receives an ack pulse.

Figure 7 is a photograph of the assembled four-node network along with the input packet generation system.

Implementation considerations

Now we consider the actual implementation of an integrated 64 × 64 network, similar to the one simulated in this article, along with some design and timing parameter calculations.

Area, power, and physical-layer latency

An EOM network with $N_T = 64$ terminals and a four-stage distribution network will require $N_S = (2\log_2 N_T - 1) + 4 = 15$ stages, of $N_T/2 = 32$ switching nodes each, or a total of 480 switching nodes. Williams et al. reported an integrated SOA-based 2 × 2 switching element with an area of 1 mm^2 and consuming an average power of 15 mW.[17] Because the switching element is the largest and most power-consuming element in the switching node, a reasonable estimate for the switching node area is 1 mm × 2 mm, including the interstage waveguides. A reasonable estimate for the average total power of the 64-port PIC would be 10 W. The input and output modules contain laser sources and some electronic circuitry for the generation of distribution addresses and acknowledgment pulses. These modules' added area and power dissipation are not large compared to the network size, because only 64 input and output modules are required. The whole network can, therefore, fit on a 4 cm × 4 cm PIC.

A routing node's latency is governed by the latency of the passive optics, photodetectors, and electronic logic gates, along with the switching time of the SOAs. The combined latency for an integrated device can be approximated at less than 0.3 ns, taking into account the simplicity of the logic circuit and current trends in device performance. The total message latency is, therefore, less than 4.5 ns for the 15-stage network, and we thus estimate the round-trip latency at 9 ns.

Performance

According to the results in Figure 4c, when operated with a speedup of $S = 2$ at $r = 0.8$, with two path adjustment iterations ($n_{PA} = 2$), the network attains an acceptance rate of 0.7 and a mean queuing latency $L_Q = 1.0$ slot. For 100-ns slots ($T_{SLOT} = 100$ ns) and a guard time (T_G) of 6 ns, assuming the acknowledgment reception latency ($T_{PIC} + T_{ACK}$) is 9 ns, we compute the slot efficiency parameter as

$$\eta = \frac{T_{SLOT} - T_G - n_{PA}(T_{PIC} + T_{ack})}{T_{SLOT}} = 0.76 \quad (4)$$

Using wavelength-parallel messages (16 wavelength × 10 Gbps, as previously demonstrated experimentally[13]), we can also calculate the average bandwidth per port (see Equation 2) and the aggregate bandwidth:

$$B_{PORT} = \frac{R \cdot N_{PAYLOAD} \cdot \eta}{S} \cdot r = 48.6\text{Gbps}$$

$$B_{AGGREGATE} = B_{PORT} \cdot N_T = 3.1 \text{ terabits/second}$$

To scale to larger systems, a designer can choose either to construct a PIC with a larger port count, if the required integration level is available, or use the multistage topology suggested by Luijten et al.[6] The latency effects of each one of these options can be simulated and straightforwardly computed using Equation 1. For the discussed single-stage 64-port system, where t_b is the slot time (100 ns), and maximum fiber length (l_f) is 16 m, the mean message latency (D) is

$$t_p = l_f/c + T_{PIC} = 89 \text{ ns}$$

$$t_r = t_q + t_b = (L_Q \times T_S) + t_b = 200 \text{ ns}$$

$$D = t_p + t_b + t_r = 389 \text{ ns}$$

where t_r is the routing latency, including the queuing latency incurred by retransmissions, and c is the speed of light in silica (2×10^8 m/s).

Implementation challenges

Current technology (as of 2006) still hasn't reached the level of photonic integration required to fully construct SPINet

networks. The main obstacle lies in efficient hybrid integration of InP SOAs, silica waveguides, and silicon electronics to construct multiple wideband switching nodes. However, recent advances in monolithic,[9] as well as silicon-based,[10] integration suggest that the enabling technologies are currently existent or close to materialization. Designers could perform partial integration of sections of the network (for example, each stage on a different PIC), as an intermediate step toward a complete network integration.

Two technology trends are intersecting these days: First, computing systems are hitting power consumption limits both in terms of cooling technologies and operating costs, thus requiring major technology shifts to address the power challenges. Second, photonic integration technologies, both in silicon and InP, are becoming commercially available with unprecedented integrability and reliability. The convergence of these two separate trends enables the construction of interconnection networks based on photonic integration technology, offering dramatically reduced power consumption owing to the low loss in optical transmission and to bit-rate transparency. The SPINet interconnection network architecture for future HPC systems exploits the vast potential of photonic integration to offer low latency, high bandwidth, and low power consumption. The work reported here demonstrates the performance potential. Future work will include a more detailed experimental demonstration, development of integrated network prototypes, and improvements to the network control plane, such as asynchronous transmission and priority-based routing. MICRO

Acknowledgments

We thank Howard Wang for his contribution to the assembly of the experimental setup. This work was supported in part by the National Science Foundation under grant CCF 05-23771 and by the US Department of Defense under subcontract B-12-664.

References

1. J. Protic, M. Tomasevic, and V. Milutinovic, ''Distributed Shared Memory: Concepts and Systems,'' *IEEE Parallel & Distributed Technology*, vol. 4, no. 2, Summer 1996, pp. 63-79.
2. D. Dai and D.K. Panda, ''How Can We Design Better Networks for DSM Systems?'' *Proc. 2nd Int'l Workshop Parallel Computer Routing and Comm.* (PCRCW 97) LNCS 1417, Springer 1998, pp. 171-184.
3. J.P.G. Sterbenz and J.D. Touch, *High-Speed Networking: A Systematic Approach to High-Bandwidth Low-Latency Communication*, Wiley & Sons, 2001.
4. W.J. Dally and B. Towles, *Principles and Practices of Interconnection Networks*, Morgan Kaufmann, 2004.
5. D.A.B. Miller, ''Rationale and Challenges for Optical Interconnects to Electronic Chips,'' *Proc. IEEE*, vol. 88, no. 6, June 2000, pp. 728-748.
6. R. Luijten et al., ''Viable Opto-Electronic HPC Interconnect Fabrics,'' *Proc. ACM/IEEE Supercomputing Conf.* (SC 05), IEEE Press, 2005, p. 18.
7. A.K. Kodi and A. Louri, ''Design of a High-Speed Optical Interconnect for Scalable Shared-Memory Multiprocessors,'' *IEEE Micro*, vol. 25, no. 1, Jan.-Feb. 2005, pp. 41-49.
8. A. Shacham et al., ''A Fully Implemented 12×12 Data Vortex Optical Packet Switching Interconnection Network,'' *J. Lightwave Technology*, vol. 23, no. 10, Oct. 2005, pp. 3066-3075.
9. R. Nagarajan et al., ''Large-Scale Photonic Integrated Circuits,'' *IEEE J. Selected Topics Quantum Electronics*, vol. 11, no. 1, Jan.-Feb. 2005, pp. 50-65.
10. C. Gunn, ''CMOS Photonics for High-Speed Interconnects,'' *IEEE Micro*, vol. 26, no. 2, Mar.-Apr. 2006, pp. 58-66.
11. B.A. Small, T. Kato, and K. Bergman, ''Dynamic Power Considerations in a Complete 12×12 Optical Packet Switching Fabric,'' *IEEE Photonic Technology Letters*, vol. 17, no. 11, Nov. 2005, pp. 2472-2474.
12. A. Shacham, B.G. Lee, and K. Bergman, ''A Scalable, Self-Routed, Terabit Capacity, Photonic Interconnection Network,'' *Proc. 13th Ann. IEEE Symp. High-Performance*

Interconnects (HOTI 05), IEEE CS Press, 2005, pp. 147-150.

13. A. Shacham, B.G. Lee, and K. Bergman, ''A Wideband, Non-Blocking, 2×2 Switching Node for a SPINet Network,'' *IEEE Photonic Technology Letters*, vol. 17, no. 12, Dec. 2005, pp. 2742-2744.
14. A. Pattavina, *Switching Theory—Architecture and Performance in Broadband ATM Networks*, Wiley & Sons, 1998.
15. A. Shacham and K. Bergman, ''Utilizing Path Diversity in Optical Packet Switched Interconnection Networks,'' *Proc. Optical Fiber Comm. Conf.* (OFC 2006), CD-ROM, IEEE Press, 2006.
16. D.S. Meliksetian and C.Y.R. Chen, ''A Markov-Modulated Bernoulli Process Approximation for the Analysis of Banyan Networks,'' *Proc. ACM SIGMETRICS Conf. Measurement and Modeling of Computer Systems* (SIGMETRICS 93), ACM Press, 1993, pp. 183-194.
17. K.A. Williams et al., ''Integrated Optical 2×2 Switch for Wavelength Multiplexed Interconnects,'' *IEEE J. Selected Topics Quantum Electronics*, vol. 11, no. 1, Jan.-Feb. 2005, pp. 78-85.

Assaf Shacham completed the work described in this article at Columbia University, where his doctoral dissertation was titled ''Architectures of Optical Interconnection Networks for High Performance Computing.'' He is now employed by Aprius, where he is engaged in developing high-performance computer interconnects. Shacham has published more than 20 papers in peer-reviewed journals and conferences. He has a PhD and an MS in electrical engineering from Columbia, and a BSc in computer engineering from the Technion, Israel Institute of Technology.

Keren Bergman is a professor of electrical engineering at Columbia University, where she heads the Lightwave Research Laboratory. Her research includes optical interconnection networks, photonic packet switching, and nanophotonic networks on chips. Bergman has a BS from Bucknell University and an MS and a PhD from the Massachusetts Institute of Technology, all in electrical engineering. She is a senior member of the IEEE and a Fellow of the Optical Society of America (OSA). She serves as Associate Editor for *IEEE Photonic Technology Letters* and the OSA *Journal of Optical Networking*.

Direct questions and comments about this article to Assaf Shacham, Columbia University, Department of Electrical Engineering, 500 W. 120th St., 1300 Mudd, New York, NY 10027; assaf@computer.org.