Photonic Interconnects for Gigabit Multicomputer Communications

Current and near-term optoelectronic technology can implement photonic interconnects for multicomputer parallel-processing systems with thousands of nodes and teraflop performance.

Jon R. Sauer, Daniel J. Blumenthal, and Aruna V. Ramanan

The increasing complexity and variety of today's scientific computational problems might be handled efficiently by large-scale parallel multiprocessor computing systems.[1] The requirements for next-generation computing include developing systems capable of Teraflop processing rates, Terabit per second communications speeds, and Terabyte storage capacity (sometimes referred to as the 3Ts of supercomputing).

Successful architectures will balance resources between the processing, communications, and storage system components. Important measures of performance for the next generation of parallel processors will be scalability in number of processing nodes and flexibility in how the hardware maps to various problem domains. In this article we will discuss how photonic and optoelectronic technologies could play an important role in future highly scalable and flexible interconnects for multicomputer parallel processing systems.

There are basically three hardware subsystems of a parallel processing system: processors, memories, and interconnection networks. The subsystem designs and their organization into a parallel processing system constitutes the overall architecture. Perhaps one of the most important considerations in successful high-performance systems is the scalability of the architecture to a large number of processors. This measure of performance is becoming increasingly important to manufacturers of parallel computers and will place an increasing demand on the interconnection network.

The Interconnect Bottleneck

Interconnects for computer communications and multiprocessor applications require some different attributes than those offered by telecommunication networks. Computer traffic consists mainly of interprocess communications. This type of communication is dominated by short status reports, brief memory requests, and the resulting often large data transfers interspersed with short acknowledgments in the reverse direction. The length of a message can vary from a single word to a large block of data. Therefore, the computer communications interconnect operates over a wide dynamic range of message lengths. Traffic is irregular in both time and destination by nature, and the latency of the shortest status messages often is critical to system performance. Additionally, tolerance to errors is far less than most telecommunications traffic.

Electronics has been an extremely successful implementation technology for parallel processing systems and their interconnection networks due to the ability to integrate many low power devices onto a small area. However, the level of integration onto a single VLSI chip is ultimately limited due to device pinout and on-chip power limitations. Therefore, in order to reach ever-higher processing capabilities by utilizing larger numbers of processors and memories, parallel processing systems based on multiple chips and even multiple computers will need to be used. The geographical scale of these systems can vary widely; examples are systems with multiple chips interconnected on a single board, multiple board systems, or multiple computer systems which reside in a single room. Multiple computer systems also can exist within a single building or campus, and it is envisioned that systems could extend over metropolitan and even continental scales.

Some of the limitations due to the interconnect may be minimized by architectural features such as the latency-hiding HEP and Tera machines [2, 3]. However, electronics does impose a communications bottleneck since the bandwidth of each link and the physical distance these links can cover are limited by power dissipation and electronic crosstalk. The result is a limitation in practical increases in the clock speed and number of processing nodes.

Alleviation of the Bottleneck

Optoelectronic systems have the potential to alleviate some of the bottlenecks imposed by all-electronic interconnects. This might be possible by combining the extremely high bandwidth and parallelism of optics with the logic and buffering capabilities of electronics to produce higher performance interconnection networks. These systems

JON R. SAUER heads a program in optical interconnects for computer communications.

Daniel J. Blumenthal is currently enrolled in the Ph.D. program in electrical and computer engineering at the University of Colorado at Boulder.

ARUNA V. RAMANAN is currently a Research Assistant at the University of Colorado and is working toward the Ph.D. degree.

1055-6877/92/$03.00 1992© IEEE

could potentially scale to a large number of processing and memory nodes over relatively large distances. Another important point is that optoelectronics is maturing to the point where current and near-term technologies can be used to successfully implement these interconnects.

As an example of how photonics might be applied to this class of interconnects, we use an architecture which is currently under construction at the University of Colorado. [4] This photonic computer interconnection network has the following features:

- It is a multistage, packet-switched system.
- The packet size is a single computer word, so the users can regard the network as an extended backplane, using it with nearly the same flexibility as their own internal backplanes.
- The network latency approaches the lower limit set by the speed of light.
- Both the number of users and geographical size can be quite large.
- The architecture avoids precision timing between switching nodes, optical logic, and high-speed optic or electronic buffers.
- The architecture exploits deflection routing, simple, pipelined electronic logic, to optical fiber bandwidth for transmission, wideband optical switches, optical amplifiers, and developed optoelectronic technology.

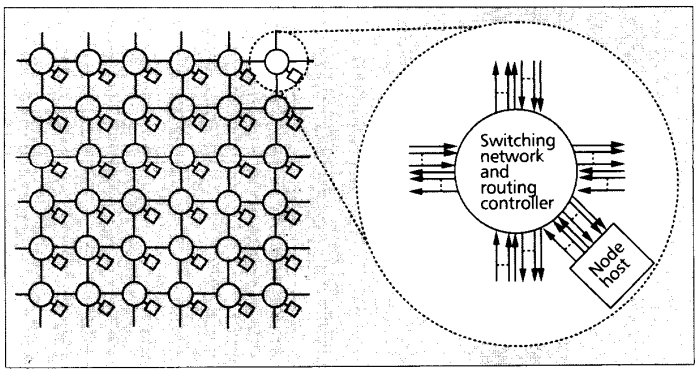
The goals of this article are threefold. First, there will be a brief review of interconnects for parallel processing computers. Second, we will discuss current approaches in electronic interconnects and their limitations. Third, a particular interconnect architecture will be presented as an example of how optoelectronic-based interconnects can overcome these limitations.

In addition to the overall architecture and design philosophy, we discuss network performance measures and an experimental switching node demonstration.

Interconnection Networks

Interconnection networks, used for communications between processing and memory nodes in multiprocessor systems, are often one limiting factor in system performance. The interconnect is responsible for efficient memory access and interprocessor communications. An important performance measure for computer interconnects is latency. *Latency is* defined as the time it takes for a packet or message to reach its destination from the time it is placed on an output queue at the source. The user node *throughput is* the rate at which information is inserted and removed from the network via its network access ports. The network capacity is the sum of the user node throughputs.

Switched interconnects are an important class of interconnect which can provide multiple communication paths between different components of the system. Switch networks can be classified according to the relation between the switching elements and the processing or memory elements. In *direct* connection networks, each switching point has an associated processor, memory, or other user. An example of a direct interconnect is given in Fig. 1, which is an example of an electronic two-dimensional mesh. In *indirect* connection networks, switching points and network users are separated. Crossbar and butterfly switches are well-known classes of indirect switching networks.



■ Figure 1. *An electronic direct-connected interconnection network.*

Multiprocessor interconnects based on direct switching architectures scale better to a large number of processing nodes ($\geq 10^3$) and are more efficient at real-time message passing than indirect switches [5]. Indirect switches introduce large communications latency for large networks, adversely affecting the performance of the multiprocessor system.

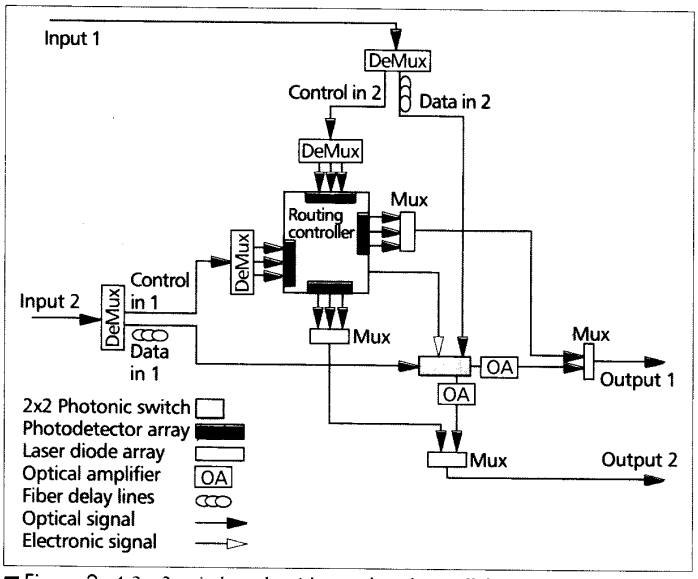
The interconnection network design space is composed of the *topology*, the *routing protocol*, and the *flow control* technique. A particular interconnect architecture represents a point in this design space and is chosen to balance the tradeoff between cost and performance for a particular application.

The interconnection network topology is defined as the connection relation between switching nodes and communications links. Some terms used in this paper to describe network topologies include the following:

- **Symmetric**—A network is symmetric if the graph structure looks identical from the point of view of any node.
- **Degree**—The degree of the network is the number of input or output links at each node.
- **Multistage**—A network is multistage if multiple hops are required in general to provide a path between any two pairs of nodes.
- **Multipath**—In multipath interconnection networks, there exist multiple paths or multiple combinations of nodes and links between source and destination nodes.
- **Wraparound**—In wraparound networks there are continuous paths around the network in at least some direction.

The type of network we discuss in this article has degree 2 and exhibits the characteristics listed above. This network is a wraparound shuffle-exchange network, with properties that will be described later.

Networks which require single-word transfers, with each transfer potentially to a different destination, are efficiently implemented using the packet-switching paradigm. Routing is defined as the method used to choose a preferred path to send a packet from its source to destination, and flow control arbitrates among colliding preferred paths. Routing control may be centralized or distributed. As networks become more widely distributed or incorporate more nodes, centralized control causes latency, throughput, and complexity to degrade. In distributed routing, also referred



■ Figure 2. *A 2 x 2 switch node with wavelength parallel channels.*

to as self-routing, packets carry destination information with them for processing at each routing node. This form of distributed processing reduces the burden on a centralized processor, and increases the throughput at each switching node.

We concentrate on a flow control technique which takes advantage of the strengths of photonic interconnects. *Deflection routing*, where the term routing encompasses both routing and flow control, requires a multipath network for its implementation. This technique can be implemented without static storage for throughgoing packets at each node.

Photonics' Role

For off-chip communications over distances greater than one meter, all-electronic interconnects represent a serious information bottleneck, constricting the performance of a multiprocessing system. We therefore seek an architecture which can duplicate the advantages of the ubiquitous electronic bus for large-scale, extended multicomputer architectures over a wide variety of interconnection distances.

Photonics represents a maturing technology with the potential to help eliminate this information bottleneck when applied correctly to the interconnect architecture. Optics provides a more efficient method of communication than electronics for distances greater than intrachip due to the impedance-match between quantum opto-electronic devices and electromagnetic propagation [6].

Current State of Technology

Photonics includes a variety of technologies which can be used to implement the basic functionality of an interconnect. These include the interface between processor electronics and transmission optics; the transmission, switching, and amplification of optical signals; and minimal processing of optical signals. The state of optoelectronic technology is at the point where it is viable to design interconnects with clear performance advantages using current and near-term technology. The success of optoelectronics will depend on the ability to make optoelectronic integrated circuits (OEICs) which can compete with electronics in terms of functionality, robustness, cost, and performance. A good overview of research and application of OEIC's can be found in [7].

The wide optical bandwidth (≥ 1 THz) and low loss (≤ 0.2 dB/km) of optical fiber allows for high transmission rates using parallel wavelength channels over long distances [8]. The development of broadband photonic switches provides a mechanism to dynamically change the interconnection pattern between multiple optical fibers at rates exceeding 1 GHz [9]. These devices spatially redirect a wide wavelength band of optical power, and so can switch many parallel-by-wavelength data streams without electro-optic conversion and electronic switches.

Losses imposed by photonic switches can be compensated for with wideband (≥ 30 nm @ 1550 nm wavelength) erbium-doped fiber-optic amplifiers [10, 11]. Device technology which allows independent modulation of multiple optical wavelengths is maturing rapidly. Monolithic multiple wavelength laser diode arrays with 140 unique lasers spaced at 0.7 nm channel separation, each with linewidth under 0.01 nm have been reported [12]. Integrated-optic diffraction-gratings will provide for low-cost, passive multiplexing and demultiplexing of parallel wavelength transmission channels over a single fiber [13]. Other key photonic systems components include integrated detector arrays capable of high-speed, high-sensitivity operation in conjunction with pre-amplifier electronics [14].

Photonic Multiprocessor Interconnects

Photonic interconnects can improve the performance of large-scale multiprocessor computers by extending the distance and bandwidth of the word-wide electronic bus beyond that available with electronic interconnects. Communications using word-wide parallel transmission through parallel optical channels is possible due to the wide optical bandwidth of optical fiber, photonic switches, and optical amplifiers.

In a photonic switch, the bandwidth of the switch fabric can be commensurate with that of the optical fiber, much higher than single-channel electronics. In order to maintain this bandwidth throughout the switching process, conversion to electronics for elastic buffering and contention resolution should be avoided.

Because photons have only one speed, that of light, purely optical memory with random access in time is currently not possible. Therefore, a photonic switch must have a flow-through architecture so the data is neither slowed down nor resynchronized by an internal clock.

The question to ask is how can we use the abundant, inexpensive resources in optics to alleviate the burden on the expensive resources. Consider the 2 x 2 switch shown in Fig. 2. We transmit a complete word in parallel over a single fiber using different wavelengths to replace each channel of the electronic word-wide bus. Therefore, each data bit and each control bit is transmitted on a different optical wavelength. The wide optical bandwidth of the fiber, photonic switch, and optical amplifier, are utilized to reduce the overall complexity

of the node and reduce the power requirements.

The cost of going to multiple wavelengths is in the fabrication of laser arrays with multiple sources, with each laser operating at a different stable wavelength. At each node, we must separate control bits from data bits. This function may be done simply and passively by coding the data bits at one group of wavelengths (e.g., about 1.55 μm) and the control at another well-separated group of wavelengths (e.g. 1.3 μm), allowing the use of inexpensive wavelength demultiplexers.

An example of the location of data and control bits as a function of optical wavelength is shown in Fig. 5. The control bits are separated spatially at the control processor using a demultiplexer capable of resolving the individual wavelengths. Additional components which are needed include output wavelength multiplexers to recombine routing and data at the switch outputs, and a multiple wavelength laser array at the output of the routing processor. The routing output processor is relatively simple since only one switch crosspoint is controlled.

A High-performance Photonic Interconnect

In this section we review the high-performance interconnect described in [4, 15]. The interconnect topology is a multistage direct-switching network consisting of photonic switching nodes interconnected by fiber optic links.

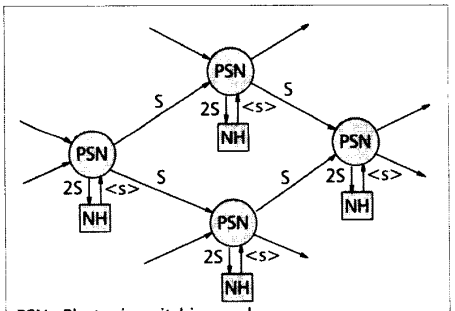
A processing node or other node host is associated with each switching node as shown in Fig. 3. The internal electronic word-size data path for each node host is optically extended to distances much greater than one meter using the wide bandwidth of optical fiber and photonic switches. Static buffering is eliminated by computing routing decisions on the fly without slowing down the optical payload data.

The interconnect architecture is based on a shuffle-exchange topology [16] using the deflection routing and flow-control protocol [17]. This protocol is ideally suited to routing nodes in which storage is difficult (i.e., photonic switching nodes) due to the simplicity and pipelined speed of the routing computation.

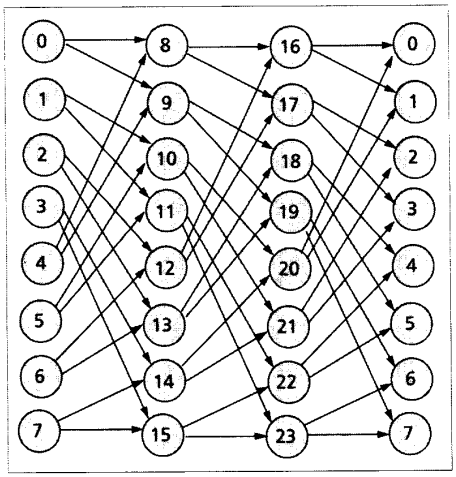
A necessary condition for implementation of deflection routing is to provide a means for deflected packets to reach the destination by an alternate path or through retransmission. The shuffle-exchange topology with wraparound provides the multipath/multihop characteristics required for implementation of the deflection routing protocol. It also minimizes the network depth for the nodes with degree 2. (The following definitions are used in Fig. 3: S—Peak link bandwidth. ⟨s⟩—Average user access bandwidth onto the network.)

Interconnect Topology

The topology is a multi-hop shuffle-exchange network consisting of photonic switching nodes interconnected by unidirectional optical fiber links as shown in Fig. 4. Shuffle-exchange or perfect-shuffle networks are a general class of multistage interconnection networks which can realize an arbitrary permutation from N inputs to N outputs in $O(\log_2 N)$ stages.



■ Figure 3. *A high-performance wide area interconnect.*



■ Figure 4. *A 24-node ShuffleNet interconnect.*

The choice of network topology is based primarily on two related features: concurrency and connectivity. Concurrency, the ability to support many packets simultaneously, is essential to best exploit the potentially large bandwidth available in photonic networks. The network access electronics work at speeds considerably less than a lightwave medium can support, where the mismatch is more than three orders of magnitude. Therefore, the network should be designed so that a given access node need only be concerned with a small fraction of the total network traffic.

A 24-node ShuffleNet, depicted in Fig. 4, consists of $N = kp^k$ nodes arranged in k columns of p^k nodes each, where p is the fan-in and fan-out of each node. In order to keep the node routing logic as simple as possible, as is desired for a lightwave flow-through architecture, we will consider only networks with $p = 2$. The nodes are all sources and sinks of network traffic, each having a bidirectional link to a host, and the network wraps around on itself with the last column connecting the first.

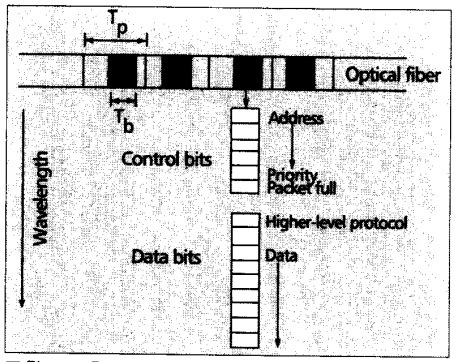
Deflection Routing and Flow Control

Deflection or hot-potato routing was first proposed by Baran [17] in 1963 as a shortest-path control strategy for distributed multipath communications networks. At the time, transmission and memory were both costly resources. Optimal utilization of the link meant storing a backlog of

• • • • •

The choice of network topology is based primarily on two related features: concurrency and connectivity.

The technique used to encode both routing information and data is bit per wavelength (BPW) encoding.



■ Figure 5. *Bit per wavelength encoding (BPW).*

data at the node at the expense of memory and delay. The main idea of the proposed control strategy was to minimize the necessary processing and storage overhead at each node by passing a message (the hot potato) to an alternate route, rather than holding it in storage when the required link is busy. For properly designed networks, the penalty for deflecting data under this protocol can be reduced to acceptable limits.

In the late 1970s, the MIMD multiprocessor HEP supercomputer was produced commercially and utilized hot-potato routing with a priority class assigned to packets. The first published account of this implementation is presented by Smith in [2].

In 1985 Maxemchuk [18] described a regularly connected topology, with deflection routing, termed the Manhattan Street Network. The penalty for deflecting information in this mesh topology network is a path increase of only four hops, yielding a high throughput with bufferless nodes. Another important topology, which has been investigated for implementation of deflection routing, is the shuffle-exchange network [16]. This minimum depth network achieves the smallest maximum length path of a two-connected network, thus reducing the penalty for deflecting. A detailed account of the history of deflection routing can be found in [19].

Optical Packet Structure

The packet format encompasses several generic features which take into account finite switching times, data duration, and other timing issues. Additionally, a certain level of timing tolerance is built in through sparse population of the data in the time domain. Only a small portion of the packet frame is occupied by the data as shown in Fig. 5.

In a synchronous network, T_p is defined by a global clock. It is the duration encompassing a guard band to account for nonzero switching time, jitter uncertainty, synchronization uncertainty, and a duration T_b for each optical packet. Note that the optical data and routing information are transmitted in parallel via multiwavelength channels.

Bit per Wavelength (BPW) Encoded Packets

The technique used to encode both routing information and data is bit per wavelength (BPW) encoding. A single packet is transmitted in multiple wavelength channels in one time slot, at low-duty cycle, such that the network is sparsely populated in the time domain as illustrated in Fig. 5. The number of wavelengths required to encode a packet is $O(\log_2 N) + M$, where N is the number of nodes and M is the number of data bits in a packet.

Each packet contains address, priority, packet full/empty, ordering information, and data as illustrated in Fig. 5. The network operates either synchronously or self-timed with a packet period T_p and actual optical signals occupying a single bit width T_b. The $O(\log_2 N)$ control wavelengths $\{\lambda_c\}$ and M data wavelengths $\{\lambda_d\}$ are transmitted at well-separated wavelength bands to simplify demultiplexing. The practical number of wavelength channels possible on a single fiber over large distances is on the order of 100 due to fiber nonlinearities [20]. Due to the abundant bandwidth of optical fiber, this technique is efficient for transmission of single-word packets. Also with BPW encoding, parallel pipelined routing control processors can operate on all bits simultaneously, thus maximizing processor throughput and using electronics efficiently.

Performance

It is important to define measures which characterize network performance. Performance predictions difficult to obtain analytically can be obtained through computer simulations. At the conclusion of this section we briefly describe typical network simulation results and the performance characteristics they describe.

Performance Metrics

The interconnection network is used by the node hosts to exchange packets of information. The performance of the network can be measured in terms of a number of interrelated metrics. In the context of this article these metrics are:

- **Contention probability**—the probability that packets entering a node simultaneously will prefer the same output port to the network.
- **Latency**—The time between the placement of a packet by the source in the network input buffer until it is delivered to its destination.
- **Capacity or throughput**—The user's throughput is the rate at which packets are injected into and received from the network. The network capacity is the sum of the throughput of all the users.
- **Network utilization**—This is the ratio of the actual network capacity to the maximum possible capacity when resource contention is ignored.

Latency

To the first order, latency can be described by the number of hops. The number of hops in a multihop network is defined as the number of link node pairs a packet must traverse from source to destination. Ignoring the effects of contention, the statistically expected number of hops between two randomly selected nodes in a ShuffleNet with $N = kp^k$ nodes can be solved for analytically. The expected number of hops is proportional to the mean latency and is given, for $k \geq 4$, approximately by [15]:

$$\lim_{k \geq 4} \langle E \rangle = \frac{3(k-1)}{2}$$

The total mean latency can be defined as the sum of the average time a packet spends in the originating node host output queue, the total latency in the links traversed, and the average node processing latency at each hop. Defining an average output queue wait time $\langle Q \rangle$, an average link $\langle L \rangle$ and an average node processing time $\langle P \rangle$ the average

latency can be expressed as

$$\langle D\rangle = \langle Q\rangle + \frac{n}{c}\langle E\rangle\langle L\rangle + (\langle E\rangle + 1)\langle P\rangle$$

where c is the speed of light in a vacuum and n is the effective index of refraction for single mode fiber.

Throughput

The peak bandwidth of the fiber optic link is denoted by the parameter S. The node architecture is constructed such that an incoming port may access the node host without contention via parallel access ports, yielding a bandwidth to the host of $2S$. For an N user network, the user access bandwidth depends on the expected number of hops $\langle E\rangle$ a packet will travel to its final destination and a derating factor (η) to account for the effects of traffic asymmetry in time and address. The total average network throughput $\langle T\rangle$ is then given by

$$\langle T\rangle = \frac{2NS}{\eta\langle E\rangle}$$

For example, a network with 2048 nodes, a packet rate of 0.3 GHz, a data parallelism of 64 channels, and a derating factor of 3, the expected number of hops is 10.5, the average user access is 1.2 Gb/sec, the peak user access is 38 Gb/sec, and the network throughput is 2.5 Tb/sec.

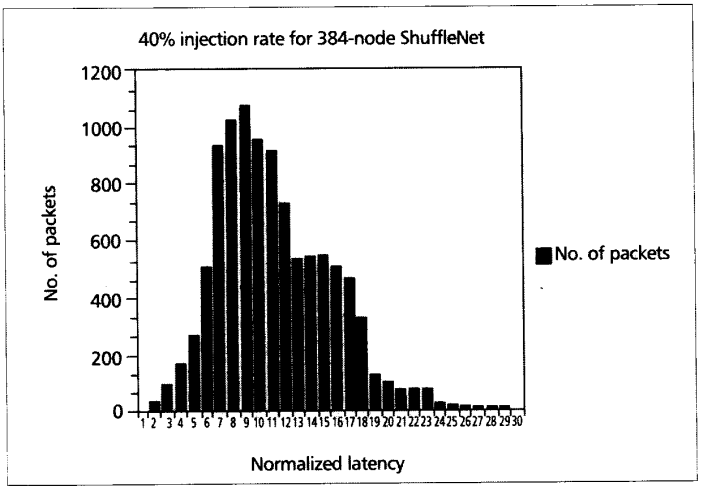
Latency Characteristics

The interconnection network performance is under investigation. We have developed a simulation package for packet routing interconnection networks. The number of nodes, the internode distance, and packet size have been parameterized.

A histogram obtained through computer simulation for a 384-node ShuffleNet network, consisting of six columns is shown in Fig. 6. The histogram is a plot of frequency vs. normalized latency. The latency is normalized by internode transit time, and is the sum of the number of hops between source and destination and a single wait cycle at the initial output queue.

The simulation histogram represents actual random traffic generated in an equivalent ShuffleNet with contention. A random request pattern generates network messages which are uniformly distributed in space and time. The network was moderately loaded with the nodes injecting messages 40 percent of the maximum possible rate.

The histogram of Fig. 6 shows a spread in the tail of the latency distribution. This tail is caused by the increase in latency due to deflections. Simulations show that the average latency for a moderately loaded network varies from about 1.2 to 1.5 times the contention-free latency as the network size grows from 64 to 2048 nodes.



■ Figure 6. *Simulated latency vs. load histogram for N = 384.*

Node Architecture

Each switching node consists of wideband photonic switches, a routing control processor, data-control separators, delay lines, and optical amplifiers. Each input port from the network has access to the node host through a separate channel. In this node architecture, contention occurs only for the output ports to the network. Although the node external degree is 2, the node is effectively a 4 x 4 constructed of 2 x 2 photonic switch elements. This design accommodates both network and node host connections. Construction of a single external degree $p = 2$ node requiring a 4 x 4 switch, fiber delays (d), optical amplifiers (A), data-control separators (indicated by dots), a pipelined control processor, and connections to the node host.

Control of the switch is handled by a parallel pipelined processor which operates on optical control information extracted from each node input and electronic outputs from the node host. Fiber optic delays are required to compensate for the control processor latency such that data is synchronized with control at the switching points. Optical data signal levels are restored by optical amplifiers at the node outputs. The optical control information is regenerated by the control processor and reinserted on the network at the node outputs. The maximum processing rate or throughput is limited by the slowest processing element in the pipeline. Pipelining of the routing requests allows packets to be served as they enter the node without elastic buffering in the data path. The control processor has simultaneous access to requests from both network inputs and both node host output queues. The processor arbitrates access to the network from the node host by disabling one or both of the host output queues when full packets are present at the node inputs.

An Experimental 2 x 2 Photonic Switch

A 2x2 self-routing photonic switch based on the techniques described in this paper has been demonstrated [21]. In this single 2 x 2 switch demonstration, the packet header contains two address bits plus one priority bit for contention resolution, and the packet payload is represented by a single bit. Control information is optically encoded at 830 nm with 3.6 nm channel spacing and data at 1300 nm. A two-bit address field, corresponding to four addresses, is mapped onto one of the two output ports of the switch by a lookup table control. When two packets desire to route to the same output port, the higher-priority packet will be switched to the correct port while the other packet is deflected to the remaining port. In the case where priorities are equal, a fair resolution measure is used. We maintain the switch in its prior state under this condition, promoting fairness for statistically independent packets.

The effective throughput for this switch, based

• • • • •

An important limitation to transmission of many simultaneous wavelengths is nonlinear crosstalk in the optical fiber.

on transmission of one data bit per packet, is T = 2.5 Mb/s. Using the same TTL electronics with laser arrays, 64-bit wide words could be transmitted, increasing the throughput to 160 Mb/s. The node latency $D = 400ns + (225m)/(2x10^8 m/s) = 1.525ms$. These performance parameters will be improved upon in current experiments by increasing the electronic processing rates and decreasing latency with state-of-the-art GaAs-based lookup tables.

Growth and Scaling Issues

Important issues to consider are nondisruptive growth of the network over potentially wide areas, and scalability of the technology used to implement the interconnect. Fault tolerance is also important, since the loss of data in a computer communications network is serious. These issues are addressed in more detail in [22].

Interconnect architectures that tolerate large increases in the numbers of access nodes and geographical separation between nodes are said to "scale" with respect to these quantities. For networks on the order of 1000 nodes and computer words of 64 bits, approximately 85 bits are required, including additional communications overhead such as packet ordering. This capacity can be met by either direct intensity modulation of individual lasers or subcarrier multiplexing of significantly fewer lasers [23].

An important limitation to transmission of many simultaneous wavelengths is nonlinear crosstalk in the optical fiber. With 85 lasers spaced ≥ 20 GHz apart, transmission of approximately 1 mW per laser over distances greater than 10 km results in crosstalk power penalty of less than 0.5 dB [20]. The limitation on internode distance is a complex function of intrabit dispersion, switching time, duty cycle, and other factors.

Synchronous network operation requires either a globally distributed clock or transmission of a clock with the data, reducing sensitivity to thermal fiber variations. The maximum phase error allowable between the two inputs due to slowly varying temperature changes is $\pm\ 0.5T_p$ and can be accommodated for at the control processor using slow feedback techniques. Thermal variations affect processing on a node-to-node scale. The incoming control bits are accumulated over the period

$$T_{avg} = T_p - 2\tau_s - \tau_{jitt} - \tau_{synch} \qquad (4)$$

where τ_s is the switch rise or fall time, τ_{jitt} the electronic jitter, and τ_{synch} the synchronization error. The number of bits within the data bandwidth will limit the throughput or latency of the switch. The source to destination distance is relevant for intrabit dispersion within the data. For example, if we transmit 64 data bits, each 1 ns in duration, and contained within the erbium amplifier bandwidth of 25 nm over 60 km of dispersion flattened fiber (dispersion of 1.5 ps/nm/km), the end bits will separate by approximately 2.25 ns.

Assuming the parameters in equation 3 are $\tau_s = 1ns$, $\tau_{jitt} = 0.5\,ns$, and $\tau_{synch} = 1.5\,ns$, the packet rate is $T_p = 6.25\,ns$. The effective throughput is $T = (64\ bits)/(6.25\ ns) \approx 10$ Gb/sec and the node latency $D = 8.25\,ns$. This rate can be increased by introducing a negative dispersion element at the switch to realign the data bits, doubling the effective data rate while increasing the switch latency by approximately 1 ns.

Thus, as a result of the bit-per-wavelength transmission format with log_2N addressing, built-in timing tolerance, and the ease of incorporating wideband optical amplifiers, this architecture scales well with respect to both numbers of access nodes and geographical size.

Conclusions

In this article we have discussed the use of photonic interconnects in multicomputer parallel processing systems. For electronic interconnect implementation, the primary limitations arise from transmission drive power requirements, limited bandwidth, and the crosstalk-limited length. We provided arguments for how photonic interconnects can relieve these bottlenecks in order to allow systems to scale to larger numbers of nodes without degrading the interconnect performance. These attributes are primarily due to the wide optical bandwidth of optical fiber, optical amplifiers, and photonic switches, and the efficient impedance matching of quantum devices to the optical fiber transmission medium. It was emphasized that current optoelectronic and photonic technologies are at the point where their integration into large-scale multistage interconnect architectures is both feasible and beneficial in order to reduce the complexity and increase the throughput of each switching node.

We introduced, as an example, a network architecture capable of interconnecting thousands of processors with multigigabit average access rate per user, and peak access rates an order of magnitude higher. The network topology is a shuffle-exchange, multihop, multipath, wraparound direct interconnect. This network utilizes self-routing and a deflection flow control technique to simplify and speed the processing. Routing information and data are encoded using a novel optical wavelength word-parallel technique with each bit coded at a different wavelength. The performance measures of latency and throughput were discussed.

We also described an experimental demonstration of these concepts in a 2 x 2 photonic switching node. Through the use of architectural designs such as this interconnect, photonics can be more deeply integrated into computer interconnect architectures to yield systems that are both scalable and have higher performance than those built solely with conventional electronics.

Acknowledgements

The authors wish to thank Dr. H. Jordan for the many helpful discussions. This work is funded by the NSF, AFOSR and ARMY.

References

[1] E. Corcoran, "Calculating Reality," *Scientific American*, pp. 100-09, Jan. 1991.

[2] B. J. Smith, "A Pipelined, Shared Resource MIMD Computer," in *Proceed. of the 1978 Int. Conf. on Parallel Proc.*, pp. 6-8, Bellaire, Mich., 1978.

[3] R. Alverson, *et al.*, "The Tera Computer System," in *Proceed. of 1990 Int. Conf. on Supercomputing*, ACM, pp. 1-6, June 1990.

[4] J. Sauer, D. Blumenthal, and A. Ramanan, "Multigigabit Photonic Interconnects for Multicomputer Communications," Tech. Report 92-05, University of Colorado at Boulder, Optoelectronics Computing Systems Research Center, Feb. 1992.

[5] C. L. Seitz, *VLSI and Parallel Computation*, Ch. 1 (Morgan Kauffman Publishers, 1990).

[6] D. A. B. Miller, "Optics for Low-energy Communication Inside Digital Processors: Quantum Detectors, Sources, and Modulators as Efficient

Impedance Converters," *Optics Lett.*, pp. 146-48, Jan. 1989.
[7] M. Dagenias, *et al.*, "Applications and Challenges of OEIC Technology: A Report on the 1989 Hilton Head Workshop," *IEEE J. Lightwave Tech.*, Vol. 8, pp. 846-62, June 1990.
[8] "Special Issue on Dense Wavelength Division Multiplexing Techniques for High Capacity and Multiple Access Communications Systems," *IEEE J. Sel. Areas Comm.*, Vol. 8, Aug. 1990.
[9] L. Thylen, "Integrated Optics in LiNbO3: Recent Developments in Devices for Telecommunications," *IEEE J. Lightwave Tech.*, pp. 847-61, June 1989.
[10] "Special Issue on Optical Amplifiers," *IEEE J. Lightwave Tech.*, Vol. 9, Feb. 1991.
[11] M. J. F. Digonnet (ed.), "Fiber Laser Sources and Amplifiers," in *Proceed. of the SPIE*, No. 17, Boston, Mass., Sept. 1989.
[12] C. J. Chang-Hasnain, *et al.*, "Monolithic Multiple Wavelength Surface Emitting Laser Arrays," *IEEE J. Lightwave Tech.*, pp. 1655-73, Dec. 1991.
[13] V. N. Morozov, *Optical Processing and Computing*, pp. 169-74. (Academic Press, 1989).
[14] D. L. Rogers, "Integrated Optical Receivers Using MSM Detectors," *IEEE J. Lightwave Tech.*, pp. 1635-38, Dec. 1991.
[15] J. R. Sauer, "A Multi-Gbit/s Optical Interconnect," OE LASE '90, Los Angeles, Calif., Jan. 1990.
[16] M. G. Hluchyj and M. J. Karol, "Shufflenet: An Application of Generalized Perfect Shuffles to Multihop Lightwave Networks," *IEEE J. Lightwave Tech.*, Vol. 9, No. 9, pp. 1386-97, 1991.
[17] P. Baran, "On Distributed Communications Networks," *Trans. Comm. Sys.*, pp. 1-9, 1964.
[18] N. Maxemchuk, "Comparison of Deflection and Store-and-Forward Techniques in the Manhattan Street and Shuffle-exchange Networks," IEEE INFOCOM '89, Ottawa, Ontario, April 1989.
[19] A. Krishna, "Communications with few buffers: Analysis and Design," Coordinated Science Laboratory Technical Report #UILU - Eng-90-2259, University if Ill., Dec. 1990.
[20] A. R. Chraplyvy, "Limitation on Lightwave Communications Imposed by Optical-fiber Nonlinearities," *JLT*, Vol. 8, No. 8, pp. 1548-57, 1990.
[21] D. J. Blumenthal, *et al.*, "Demonstration of a Deflection Routing 2 x 2 Photonic Switch for Computer Interconnects," *IEEE Photon. Tech. Lett.*, Vol. 4, No. 4, pp. 169-73, 1992.
[22] J. R. Sauer, "Multi-Gbit/s Optical Computer Interconnect," (Boston, MA), Proc. of SPIE, OE Fibers '91, Boston, Mass., Sept. 1991.
[23] R. Olshansky, V. A. Lanzisera, and P. M. . Hill, "Subcarrier Multiplexed Lightwave Systems for Broad-band Distribution," *IEEE J. Lightwave Tech.*, Vol. 7, No. 7, pp. 1329-42, 1989.

Biographies

JON R. SAUER was educated in Stanford and Tufts universities, with a Ph.D. from the latter in 1970 in particle physics. He conducted research in particle and accelerator physics at Harvard; Stanford; Fermilab, Indiana; and Argonne until 1980. He then joined Bell Labs. Dr. Sauer briefly worked on the Denelcor HEP MIMD supercomputer before rejoining Bell Labs in 1984 in Denver. Dr. Sauer accepted an appointment as an adjunct research scientist to the Center for Optoelectronic Computing Systems, Boulder, Colorado. He then joined the university in Boulder in 1990 as a professor in the electrical and computer engineering department. He currently heads a program in optical interconnects for computer communications.

DANIEL J. BLUMENTHAL [M] received B.S. and M.S. degrees in electrical engineering from the University of Rochester and Columbia University, respectively. He is currently completing the Ph.D degree program in Electrical Engineering at the University of Colorado. Mr. Blumenthal has co-authored more than 15 papers in the areas of photonic switching and optical networks since 1986.

ARUNA V. RAMANAN received the M.Sc. degree in physics from the University of Delhi, India, in 1973; the M.Phil. degree from Jawarharlal Nehru University, New Delhi in 1981; and the M.S. in electrical engineering from the University of Colorado, Boulder, in 1988. She was with the department of electrical and computer engineering at Kuwait University from 1982 to 1986. Ramanan is currently a Research Assistant at the University of Colorado and is working toward the Ph.D. degree.