Hierarchical optical ring interconnection (HORN): scalable interconnection network for multiprocessors and multicomputers

Ahmed Louri and Rajdeep Gupta

A new interconnection network for massively parallel computing is introduced. This network is called a hierarchal optical ring interconnection (HORN). The HORN consists of a single-hop, scalable, constant-degree, strictly nonblocking, fault-tolerant interconnection topology that uses wavelength-division multiple access to provide better utilization of the terahertz bandwidth offered by optics. The proposed optical network integrates the attractive features of hierarchical ring interconnections, e.g., a simple node interface, a constant node degree, better support for the locality of reference, and fault tolerance, with the advantages of optics. The HORN topology is presented, its architectural properties are analyzed, and an optical design methodology for it is described. Furthermore, a brief feasibility study of the HORN is conducted. The study shows that the topology is highly amenable to optical implementation with commercially available optical elements.

Key words: Parallel processing, single-hop nonblocking optical interconnection network, wavelength-division multiple access, scalable interconnection network.

1. Introduction

Parallel-processing systems are a proposed solution to the increasing demands for processing power and computation speeds. These systems can consist of thousands of processing elements (PE's) interconnected by means of an interconnection network, such as in massively parallel processing. Because of the large number of PE's contained in these systems, the interconnection network usually determines performance and cost. Such a network must have low interconnection complexity (such as a low node degree, thus a low cost and ease of implementation), a relatively small diameter for such a large number of PE's, a high degree of scalability and expandability, and most importantly, efficient support for both local and remote communications. Recent studies[1,2] have shown that efficient implementation of local communications (spatial locality) is a fundamental requirement for interconnection networks because PE's engage in data transfers more frequently with nearby neighbors than with more distant PE's.

The authors are with the Department of Electrical and Computer Engineering, The University of Arizona, Tuscon, Arizona 85721.

Received 22 December 1995; revision received 20 May 1996.
0003-6935/97/020430-13$10.00/0

It is proving to be very difficult for flat interconnection networks to satisfy the above requirements, especially scalability to a large number of PE's, while still maintaining a small diameter and low cost. Recently there has been strong interest in hierarchical interconnection networks[3,4] that can provide a high degree of scalability while still maintaining a low network latency. The rationale behind hierarchical networks is based on the locality of reference found in the communication profiles of many parallel-processing applications. Therefore, it is desirable to have cluster-based interconnection networks, in which a cluster comprises a relatively small number of PE's. The intracluster level should efficiently support local communication, whereas global communication will take place at the intercluster level.

An additional advantage of hierarchical networks is modularity, but as the number of PE's increases and the performance of each PE increases the demand for higher communication bandwidths and higher interconnect densities also increases. There are, however, some serious technical challenges to making these systems a reality.

A possible solution to the realization of interconnection networks for large parallel processors and massively parallel processors is the use of optical technology.[1,5–11] Optics provides many features such as parallelism, large bandwidth, low power requirements, reduced cross talk, and better isolation

than semiconductor electronics can provide. For exploiting the terahertz bandwidth of optics for large parallel processors, wavelength-division multiple access (WDMA) techniques that enable multiple multiaccess channels to be realized on a single physical channel can be used.

In a WDMA network, the optical spectrum is divided into many different logical channels, each channel corresponding to a different wavelength. These channels can be carried simultaneously on a small number of physical channels, e.g., optical fibers. Additionally, each network node is typically equipped with a small number of transmitters and receivers (transceivers), some of these being dynamically tunable to different wavelengths. For a single-hop packet transmission to occur, one of the transmitters of the sending node and one of the receivers of the destination node must be tuned to the same wavelength for the duration of the packets' transmission.

Several WDMA-based network architectures have been introduced in the literature recently.[12–17] This list is by no means complete but it gives us a broad outlook on types of WDMA networks. Some of these architectures are not size scalable to large numbers,[12,13,16] whereas other architectures are multihop,[13,14] in which a packet may not remain completely in the optical domain between source and destination. This vascillation incurs a major delay at each intermediate node as a result of optical–electrical conversions and processing of the packet for routing and retransmission. Some of these networks[14,15] require tunable transmitters and receivers at each PE, which is very costly at this time. Other networks[15–17] suffer from splitting losses incurred from star couplers. Finally, no distinction is made between local and remote communications in any of these networks, which has significant performance implications.

The above considerations have led us to look into optical hierarchical networks to circumvent the disadvantages of the current WDMA-based networks referred to above. To this end we present a novel interconnection topology known as the hierarchical optical ring interconnection (HORN). The HORN is based on a ring-of-buses hierarchical paradigm and consists of a single-hop, scalable, nonblocking topology. The cost savings are accomplished through the low node degree, while maintaining scalability and achieving excellent performance through the use of WDMA and single-hop techniques. Packets are sent from the source node on a distinct wavelength and arrive at the destination node with the same wavelength. No wavelength reconfiguration is required for changes in traffic. A distinction is made between local and remote communications in that both are implemented independently of one another. Through wavelength reuse we are able to implement both local and remote communications efficiently. PE's consist of a single, (slow) tunable transmitter and a small set of nontunable receivers; consequently tunability is not required at both ends. A connection between any two nodes does not require PE's to forward packets, and no optical-to-electrical (O–E) or electrical-to-optical (E–O) converters are required during routing, hence a single-hop architecture. Finally, fault tolerance is enforced through the use of dual rings.

2. Topology of the Hierarchical Optical Ring Interconnection

In this section we define the structure of the HORN, including the wavelength assignment used, message routing, diameter, link complexity, and fault tolerance. An example of a multiple-access protocol is also given.

A. Definition of the Hierarchical Optical Ring Interconnection

It has been shown that a PE engages in data transfer more frequently with nearby neighbors (local communication) than with more distant nodes (remote communications).[2,18] Therefore, the interconnection topology must be designed so that it can efficiently support local data transfers (spatial locality). This emphasis has led us to consider a hierarchical interconnection network topology in which the lower-level network supports local communications very efficiently. We have chosen the snowflake topology[19] because it is well suited for this type of communication.

The HORN is an optical interconnect approach that achieves the architectural objectives of snowflakes while also providing significant performance improvements in elements such as unity diameter, fault tolerance, nonblocking capability, and scalability through the use of WDMA and wavelength reuse. The dual rings of the HORN are used strictly for routing and for fault tolerance. WDMA is used to achieve multiple logical channels without requiring multiple physical links.

Figure 1 shows a diagram of a three-level HORN in which all PE's are located in the first hierarchical level $H(1)$. The PE's in Fig. 1 are identified by the black-filled circles, and the switching nodes are identified by the gray-filled circles. Switching nodes are located at $H(i)$, where $2 \leq i \leq 3$, and are used for routing purposes. The notation HORN(i) is used to characterize the HORN, where i represents the number of hierarchies. Figure 1 shows a diagram of a HORN (3); all groups are labeled by use of $H(n, g)$ notation, such that $H(n, g)$ identifies individual groups of the HORN, where n refers to the hierarchy and g refers to a group at hierarchy n. The dual rings of the HORN can be seen in Fig. 1.

Two types of communication are possible with a HORN: local and remote. In both cases a packet undergoes O–E conversion only at the source and destination, and no further O–E conversion is required during routing. Local communication takes place when both the source and destination PE's are in the same hierarchical group $H(1, g)$, where, from Fig. 1, $1 \leq g \leq 18$. By contrast, remote communication takes place when the source and destination PE's are in different hierarchical groups. We have

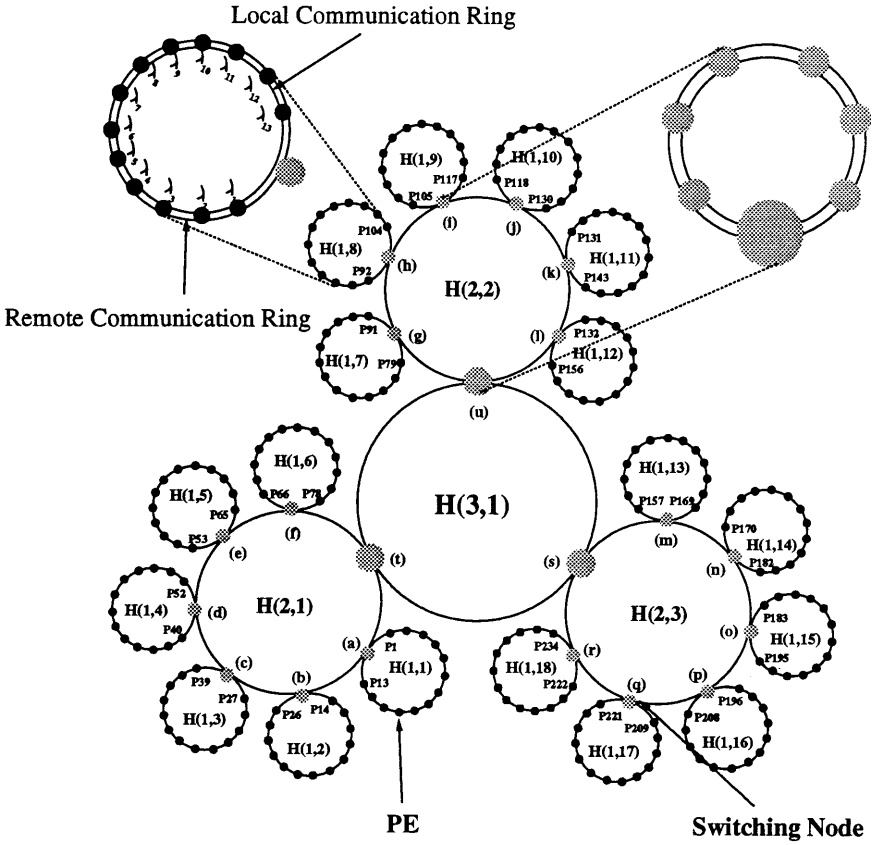

Fig. 1. Diagram of a HORN(4, 3). PE's are indicated by the black-filled circles and switching nodes are indicated by the gray-filled circles. All hierarchical groups are labeled with the notation $H(n, g)$, where n identifies the hierarchy and g identifies a unique group at hierarchy n.

separated local and remote communications from one another to provide a more efficient implementation for both types of communication. Local communication employs the inner ring and remote communication employs the outer ring of a HORN, as shown in Fig. 1. Switching nodes, therefore, are not used for local communication but are used in remote communication. Efficient implementation of local and remote communications is accomplished through the novel wavelength assignment, which is discussed in Subsection 2.B.

B. Optical Wavelength Assignment for the Hierarchical Optical Ring Interconnection

Assigning a unique wavelength to all PE's would be an ideal solution for packet routing because it would make routing a trivial task. However, there is a limited number of available wavelengths, restricting the interconnection size.[15,16] The number of wavelengths available determines the number of logical channels that is supported by a single line of the interconnection. Although this number may be large when considered from an information-capacity point of view, it may not be large enough to support the number of PE's needed for a massively parallel architecture. One method of overcoming this limitation is to reuse wavelengths. We reuse wavelengths in a HORN by allowing those used in local communication to be reused in remote communication.

The number of wavelengths employed for local communication has to equal the maximum number of PE's located in the rings of the first hierarchy of the HORN (N_1):

$$N_1 = \mathrm{Max}[|H(1, i)| \ \forall \ i]. \tag{1}$$

Figure 1 shows an example of the wavelength assignment of the $H(1, 8)$ group. The wavelengths indicated next to each PE correspond to the wavelength that each PE receives. This same wavelength assignment applies to all rings located in the $H(1)$ hierarchy.

An ideal wavelength assignment for remote communication would be one such that a unique wavelength is assigned to all distinct rings of the HORN. Figure 2 shows an example of such a wavelength assignment for the setup of Fig. 1. Notice that the wavelength assignment for local communication is also shown for completeness. Remote communication takes place with a source PE sending data packets on the assigned wavelength to the destination-PE ring. For example, PE's wanting to send to group $H(1, 12)$ do so by sending on λ_{12}, with all PE's in this ring consuming the packet.

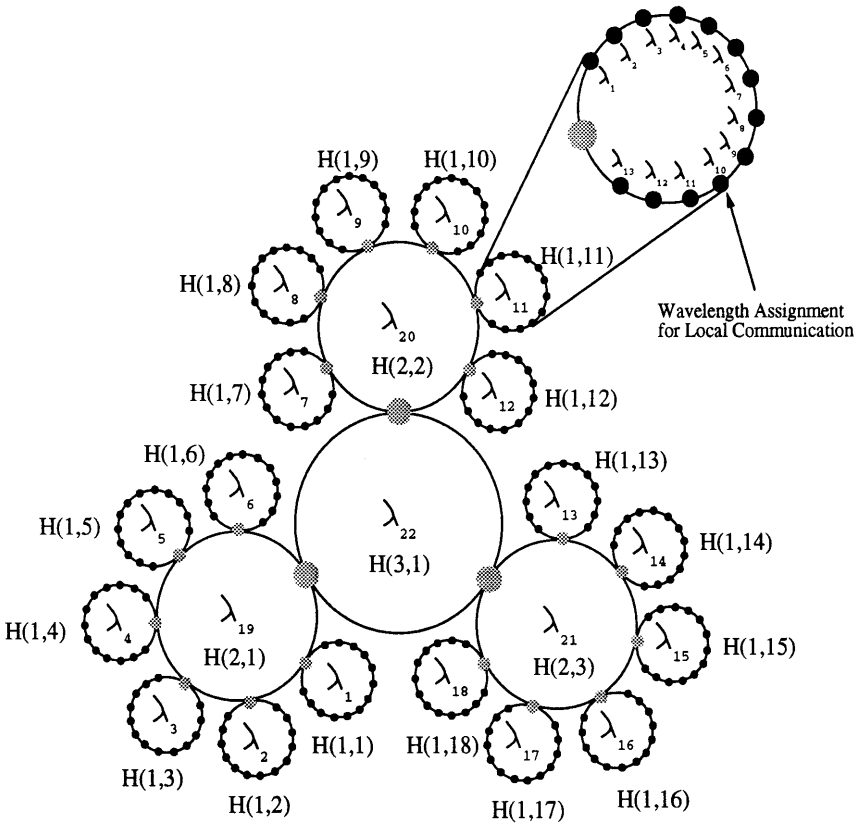

Fig. 2. Wavelength assignment in which there are 22 wavelengths available for remote communications. All rings of all hierarchies are assigned a unique wavelength. The local wavelength assignment for the $H(1, 11)$ ring is shown with all other $H(1)$ rings assigned in the same manner.

PE's, therefore, receive packets sent on the wavelength assigned to their ring as well as on the wavelengths assigned to higher-level rings. For example, PE's located in $H(1, 1)$ receive packets sent on λ_1, λ_{19}, and λ_{22} (Fig. 2). Consequently, multicast and broadcast capabilities are very naturally handled in the HORN. PE's wanting to multicast to groups $H(1, 7)$, $H(1, 8)$, $H(1, 9)$, $H(1, 10)$, $H(1, 11)$, and $H(1, 12)$ do so by sending on λ_{20}, and PE's wanting to broadcast to all PE's do so by sending on λ_{22}.

Through this wavelength assignment we basically have physically separated local and remote communications from one another. The primary reason for doing this in the HORN was to allow for simultaneous local and remote communications.

All PE's consist of a number of fixed, tuned receivers that correspond directly to the number of wavelengths they receive on. If we assume the wavelength assignment shown in Fig. 2, each PE is assigned one wavelength for local communication and h wavelengths for remote communication. Consequently, $h + 1$ receivers are required for each PE. The number of wavelengths on which each PE receives is a small subset of the total number of available wavelengths in the HORN. For the wavelength assignment of Fig. 2 it is equal to 18%.

C. Message Routing in the Hierarchical Optical Ring Interconnection

The design of an interconnection network must permit efficient routing. PE's must be able, at any point in time, to establish a route to an intended destination. The interconnections need each of the PE's to communicate with the intended destinations. That this communication be established is an essential parameter in the design of the interconnection. In a HORN, PE's communicate with destination nodes through either local or remote communication.

One of the novel features of the HORN routing protocol is the separation of local and remote communications. By the physical separation of the two protocols, the routing paths do not coincide with one another, as shown in Fig. 1.

Local communication takes place when both the source and destination PE's are in the same hierarchical group: $H(1, a)_{\text{source}} = H(1, b)_{\text{destination}}$, where $a = b$. The source PE tunes its transmitter to the preassigned wavelength of the destination PE and transmits. The destination PE subsequently consumes the packet. Moreover, a simple WDMA concept is employed, and a diameter of 1 is achieved for local communication.

Remote communication takes place when the source and destination PE's are not in the same hierarchical group: $H(1, a)_{\text{source}} \neq H(1, b)_{\text{destination}}$,

Table 1. Routing Table for the Example AOTF Configuration Shown in Fig. 1

Routing from Switching Node	Wavelengths	Routing to Switching Node	Routing to Group
(a)	λ_2–λ_{18}, λ_{20}, λ_{21}	(t)	
	λ_1		$H(1, 1)$
	λ_{19}, λ_{22}	(t)	$H(1, 1)$
(b)	λ_1, λ_3–λ_{18}, λ_{20}, λ_{21}	(a)	
	λ_2		$H(1, 2)$
	λ_{19}, λ_{22}	(a)	$H(1, 2)$
(c)	λ_1–λ_2, λ_4–λ_{18}, λ_{20}, λ_{21}	(b)	
	λ_3		$H(1, 3)$
	λ_{19}, λ_{22}	(b)	$H(1, 3)$
(d)	λ_1–λ_3, λ_5–λ_{18}, λ_{20}, λ_{21}	(c)	
	λ_4		$H(1, 4)$
	λ_{19}, λ_{22}	(c)	$H(1, 4)$
(e)	λ_1–λ_4, λ_6–λ_{18}, λ_{20}, λ_{21}	(d)	
	λ_5		$H(1, 5)$
	λ_{19}, λ_{22}	(d)	$H(1, 5)$
(f)	λ_1–λ_5, λ_7–λ_{18}, λ_{20}, λ_{21}	(e)	
	λ_6		$H(1, 6)$
	λ_{19}, λ_{22}	(e)	$H(1, 6)$
(g)	λ_1–λ_6, λ_8–λ_{19}, λ_{21}	(u)	
	λ_7		$H(1, 7)$
	λ_{20}, λ_{22}	(u)	$H(1, 7)$
(h)	λ_1–λ_7, λ_9–λ_{19}, λ_{21}	(g)	
	λ_8		$H(1, 8)$
	λ_{20}, λ_{22}	(g)	$H(1, 8)$
(i)	λ_1–λ_8, λ_{10}–λ_{19}, λ_{21}	(h)	
	λ_9		$H(1, 9)$
	λ_{20}, λ_{22}	(h)	$H(1, 9)$
(j)	λ_1–λ_9, λ_{11}–λ_{19}, λ_{21}	(i)	
	λ_{10}		$H(1, 10)$
	λ_{20}, λ_{22}	(i)	$H(1, 10)$
(k)	λ_1–λ_{10}, λ_{12}–λ_{19}, λ_{21}	(j)	
	λ_{11}		$H(1, 11)$
	λ_{20}, λ_{22}	(j)	$H(1, 11)$
(l)	λ_1–λ_{11}, λ_{13}–λ_{19}, λ_{21}	(k)	
	λ_{12}		$H(1, 12)$
	λ_{20}, λ_{22}	(k)	$H(1, 12)$
(m)	λ_1–λ_{12}, λ_{14}–λ_{20}	(s)	
	λ_{13}		$H(1, 13)$
	λ_{21}, λ_{22}	(s)	$H(1, 13)$
(n)	λ_1–λ_{13}, λ_{15}–λ_{20}	(m)	
	λ_{14}		$H(1, 14)$
	λ_{21}, λ_{22}	(m)	$H(1, 14)$
(o)	λ_1–λ_{14}, λ_{16}–λ_{20}	(n)	
	λ_{15}		$H(1, 15)$
	λ_{21}, λ_{22}	(n)	$H(1, 15)$
(p)	λ_1–λ_{15}, λ_{17}–λ_{20}	(o)	
	λ_{16}		$H(1, 16)$
	λ_{21}, λ_{22}	(o)	$H(1, 16)$
(q)	λ_1–λ_{16}, λ_{18}–λ_{20}	(p)	
	λ_{17}		$H(1, 17)$
	λ_{21}, λ_{22}	(p)	$H(1, 17)$
(r)	λ_1–λ_{17}, λ_{19}–λ_{20}	(q)	
	λ_{18}		$H(1, 18)$
	λ_{21}, λ_{22}	(q)	$H(1, 18)$
(s)	λ_1–λ_{12}, λ_{19}, λ_{20}	(u)	
	λ_{13}–λ_{18}, λ_{21}	(r)	
	λ_{22}	(r) and (u)	
(t)	λ_7–λ_{18}, λ_{20}, λ_{21}	(s)	
	λ_1–λ_6, λ_{19}	(f)	
	λ_{22}	(f) and (s)	
(u)	λ_1–λ_6, λ_{13}–λ_{18}, λ_{19}, λ_{21}	(t)	
	λ_7–λ_{12}, λ_{20}	(l)	
	λ_{22}	(l) and (t)	

where $a \neq b$. The key to the routing used in remote communication is the use of acousto-optic tunable filters (AOTF's),[20,21] located in the switching nodes, that are able to route on individual wavelengths. AOTF's, therefore, can be thought of as optical switches. (A detailed description of components is provided in Section 4.) We achieve the configuration of the AOTF to a given routing algorithm by sending an appropriate acoustic wave. Once the AOTF is configured, optical packets experience no delay other than the propagation delay through the acousto-optic cell. Thus, AOTF's are able to operate as transparent optical switches. A more detailed description of AOTF's is given in Subsection 4.B.

Table 1 lists an example configuration of the AOTF's for the HORN shown in Fig. 1, with the assumption of the wavelength assignment of Fig. 2. Table 1 consists of three main columns, with the first (fourth) column identifying all AOTF's located in the switching nodes, the second (fifth) column identifying the wavelengths that need to be routed, and the third (sixth) column identifying where the wavelengths are routed by listing the hierarchical group, switching node, or some combination of the two. AOTF's do not require packets to go through O–E and E–O converters during routing. No intermediate processing of data packets during routing takes place, resulting in a hierarchical interconnection network with a unity diameter.

Figure 3 shows how a source PE located in ring $H(1, 7)$ sends to a destination PE located in ring $H(1, 6)$, under the assumption that the values from Table 1 were used to configure the AOTF's. Packets always travel in a counterclockwise direction in any given ring. The source PE initiates the transfer by sending on the wavelength assigned to the destination PE, λ_6. Subsequently, the packet travels through switching nodes (g), (u), (t), and (f), indicated

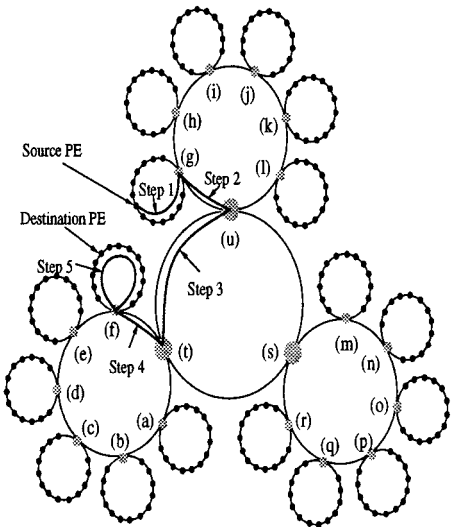

Fig. 3. Steps employed to route a packet from the source PE to the destination PE. Packets always travel in a counterclockwise direction.

as steps 1 through 4 in Fig. 3. Finally, the AOTF of switching node (f) sends the packet to all the PE's of group $H(1, 6)$ (shown as step 5), completing the routing. The routing protocol in the HORN is robust and can survive link failures. Remnants of signals that have already propagated around a closed loop die off, as is further discussed in Section 4.

D. Fault Tolerance of the Hierarchical Optical Ring Interconnection

The self-healing properties of dual rings[22] can be used to achieve fault tolerance in the HORN. Figure 4(a) shows normal operation, with only the primary ring being utilized for communications. Figure 4(b) shows what happens when a link in a hierarchical group breaks. The ring wraps in on itself, with the primary and backup rings now both being utilized for communications. Consequently, the HORN is capable of withstanding single-link failures without any loss of communication channels. The failure of two or more links in a single hierarchical group, however, does isolate parts of the group from other parts. For example, Fig. 4(c) shows what happens when two links in a hierarchical group break; two independent rings are produced. Because parts of the interconnect are physically isolated from one another, communications between the two isolated parts can no longer occur.

E. Multiple-Access Protocols

The HORN requires a multiple-access protocol to prevent packets of the same wavelength from colliding with one another. Examples of multiple-access protocols include carrier-sense multiple access (CSMA), carrier-sense multiple access with collision detection (CSMA/CD), ALOHA, slotted ALOHA, time-division multiple access (TDMA), and arbitration.[23] (The ALOHA protocals, so named because of the Hawaiian greeting, were developed at the University of Hawaii and are a contention-resolution system for networks.) Variations of the HORN are manifested through changes in multiple-access protocols. In this paper, and because of page limitations, we discuss only the HORN TDMA as an example of a multiple-access protocol.

TDMA with WDMA is a very powerful means of sharing the enormous bandwidth (terahertz range) of optics. Time slots are assigned to each of the individual wavelengths, providing two-dimensional sharing of the terahertz bandwidth among multiple users. A fixed time slot is assigned to each PE such that PE's send only during their preassigned time slot. It is a cyclic process in which PE's wait for materialization of the time slot for transmission after previous materializations of the time slots are completed.

Two time-slot-assignment protocols, which are completely independent of one another, are employed in the HORN to provide for a more efficient implementation of TDMA. PE's can choose to use either one or both. One time-slot protocol is utilized for local communication, whereas the other protocol is utilized for remote communication. Therefore, it is possible for a PE to send information using remote communication and to send local communication at

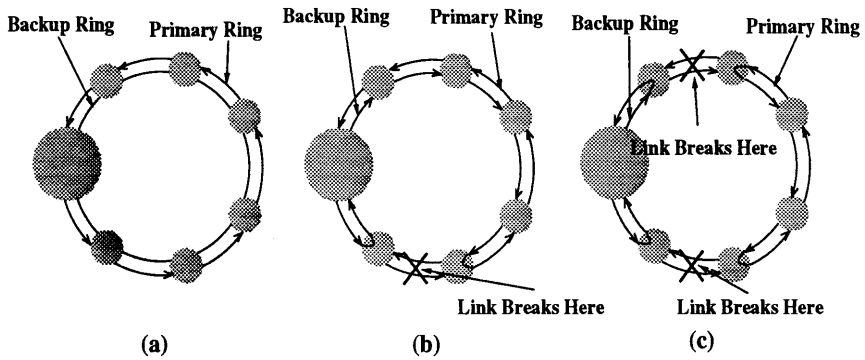

Fig. 4. Fault-tolerance aspects shown for the rings of the upper hierarchies of the HORN: (a) Normal operation of the ring during which only the primary ring is being utilized. (b) Operation when a link in a ring breaks. The ring wraps in on itself, with the primary and remote rings now both utilized. (c) Operation when two links in a ring breaks such that two separate rings are produced.

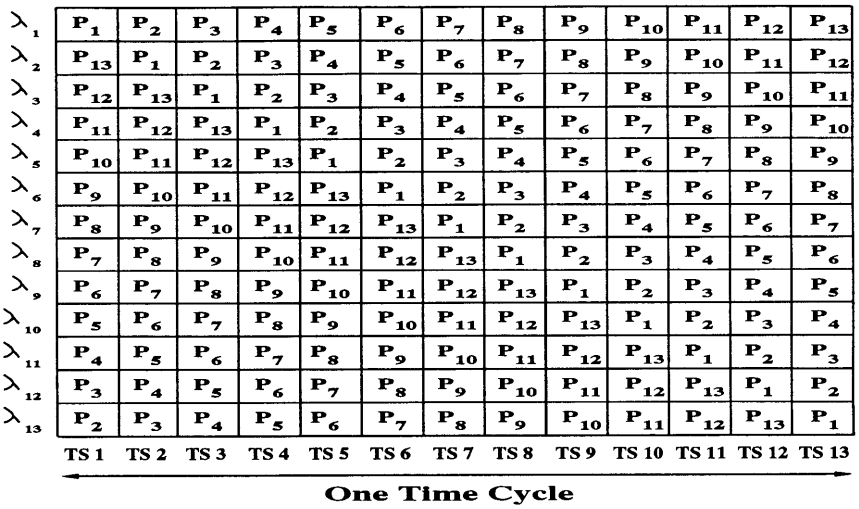

Fig. 5. Time-slot allocation for local communication. The vertical axis shows the wavelengths and the horizontal axis shows the time slots. One time cycle is shown. During one cycle each PE can send on all wavelengths. TS, time slot; P, PE.

the same time if both time slots are active. This is the advantage of having two independent time-slot protocols over one.

Figure 5 shows the time-slot-assignment protocol employed for local communication for the $H(1, 1)$ ring of Fig. 1. The vertical axis shows the wavelengths employed for local communication, and the horizontal axis shows the time scale. The time-slot assignment is constructed to allow each PE a chance to send on each of the available wavelengths for local communication in one cycle. This scheme was originally proposed by Dowd *et al.*[15] for the flat hierarchical architecture interconnection network. The number of communication channels required for the interconnection is independent of the number of PE's. Each PE has a chance to send on all wavelength channels in one cycle. This time-slot assignment guarantees a strictly nonblocking configuration because it is inherent in the TDMA protocol.

Figure 6 shows the time-slot-assignment protocol employed for remote communication. The vertical axis once again lists the wavelengths available for remote communication, and the horizontal axis shows the time scale. It is exactly the same protocol as was employed for local communication but incorporates all PE's of the HORN interconnection. N identifies the number of a PE, and k identifies the number of wavelengths used for remote communication. For example, for the HORN interconnection of Fig. 1, N would be equal to 234 and k would equal 22. Each PE, once again, has a chance to send on all wavelengths in one cycle, hence a strictly nonblocking configuration. Both Figs. 5 and 6 show one time cycle, with other cycles occurring in a round-robin cyclic process as time slots

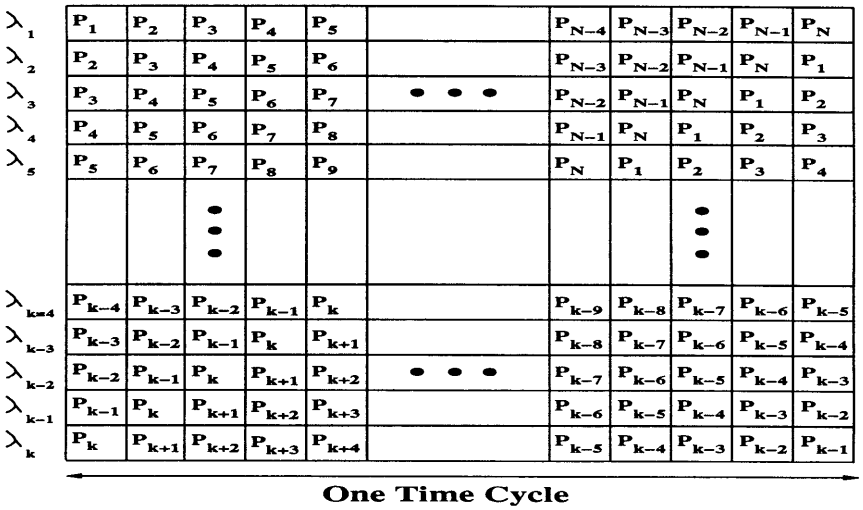

Fig. 6. Time-slot allocation for remote communication. The vertical axis shows the wavelengths and the horizontal axis shows the time slots. One time cycle is shown. During one cycle each PE can send on all wavelengths.

progress. One way to allow PE's to send more packets in a given time cycle is to increase the number of wavelengths. For example, if we again look at Fig. 6 and increase the number of wavelengths from k to $2k$, each PE would now able to send to $2k$ PE's in one cycle instead of sending to k PE's.

3. Scalability Issues of the Hierarchical Optical Ring Interconnection

In this section we discuss size-scalability, cost-scalability, and optical-scalability issues for the HORN interconnection architecture. Communication is also analyzed for some common basic patterns of interprocessor communications that are frequently used in a variety of parallel algorithms. Size scalability refers to the property that the size of the network (e.g., the number of PE's) can be increased with minor or no changes made to the existing configuration. Also, the increase in system size is expected to result in a proportional increase in performance. Cost scalability is measured in terms of the hardware required. For a system to be considered cost scalable, the number of physical hardware components should not grow faster than $O(N^2)$. The optical scalability refers to power-loss and dynamic-range calculations of a single ring. For the HORN interconnection, power-loss and dynamic-range calculations are simplified because packets are regenerated as they progress up and down the hierarchies. These calculations for the HORN, therefore, degenerate to those of a single ring. Complexity scalability refers to cases in which performance does not keep up with the complexity of the interconnection as the number of processing nodes increases. These issues are detailed below.

A. Size Scalability

The overall HORN interconnection structure resulting from the composition of groups $H(n, g)$ can be enlarged modularly to construct $H(n + 1, g)$ groups according to the following two approaches. The first is to make $H(n, g)$ a hierarchy of level n, with $H(n + 1, g)$ becoming a new structure created by the addition to $H(n, g)$ of the number of groups needed to build the next level of the hierarchy. This construction enlarges the hierarchical interconnection in a uniform and regular manner, with no changes made to existing $H(n, g)$ groups. When new hierarchies are formed, AOTF's have to be reconfigured to reflect changes in the HORN topology. The second approach, to be used when incremental growth is desired, is to add new PE's to existing $H(1)$ groups. The maximum number of PE's that can exist in an $H(1)$ group is limited to the number of wavelengths (α) available for maintaining local communication. Reconfiguration of the AOTF's, however, is not required for incremental growth because the HORN topology at the macro level does not change. For both expansions, neither the number of links nor the node degree of existing PE's or switching nodes changes.

Table 2. Cost Scalability Expressions for Components of the HORN[a]

Component	HORN
Transmitters (tunable)	N
Fixed, tuned receivers	$N * h + 1$
Tap	$2N$
AOTF	$\sum_{i=2}^{h} \lvert H(i,j)\rvert$
Fiber links	$2 \sum_{i=1}^{h} \lvert H(i,j)\rvert$
Passive couplers	$\sum_{i=1}^{h} \lvert H(i,j)\rvert$

[a] N is the number of PE's; h identifies the number of hierarchies; $|(H(i, j)|$ refers to the number of PE's or switching nodes at the jth group at hierarchy i.

The increase in size is reflected by a proportional increase in communication channels. The number of communication channels (Λ) available for local communication is equal to

$$\Lambda = N. \tag{2}$$

This relation is a direct consequence of the requirement of assigning a unique wavelength to all PE's in the $H(1)$ rings, as shown in Fig. 2. Figure 2 also shows the number of communication channels available for local communications for the $H(1, 11)$ ring. A local-communication channel (i.e., a unique λ) is available to all PE's at all times, as shown in Fig. 5. It can be seen from Fig. 5 that an increase in the number of PE's in the HORN results in a proportional increase in communication channels.

B. Cost Scalability

For a system to be considered cost scalable, the cost should be less than $O(N^2)$. By this measure, a full crossbar is not considered cost scalable. Beyond that, a system may scale better or worse than another regarding cost. This subsection evaluates the cost complexity of the HORN, which is related to the number of transmitters, receivers, AOTF's, passive couplers, taps, and fiber links required as the number of PE's increases. Table 2 lists these physical components, with N specifying the number of PE's and $|H(i, j)|$ specifying the total number of nodes, whether they are PE's or switching nodes, of the jth group at hierarchy i. The first three rows of Table 2 describe the overall PE complexity, whereas the remaining rows represent the switching-node complexity.

A mathematical expression was derived by Dowd *et al.*[15] that can be used in calculating the number of nodes at a given hierarchy of a given hierarchical interconnection. This expression can be used in calculating the number of switching nodes S in the HORN:

$$S = N\left\{\frac{1}{|H(1, 1)|} + \frac{1}{|H(1, 1)\|H(2, 1)|} + \cdots + \frac{1}{N}\right\}. \tag{3}$$

The number of switching nodes is necessarily lower than $[2N/|H(1, 1)]$, and the number of fiber links is

necessarily lower than $\{4N + 8[2N/|H(1, 1)]\}$ for an arbitrarily large number of hierarchical levels. The numbers 4 and 8 correspond to the node degree of the processing and switching nodes, respectively. By looking at Table 2 and looking at the limit expressions for the switching nodes and fiber links, we were able to conclude that the HORN can be classified as being of $O(N)$ in terms of cost complexity. This level is due to the simple node interfaces and interring connections that hierarchical rings provide. Cost expressions for the HORN are rather small when compared with other conventional networks.[19]

C. Optical Scalability of the Hierarchical Optical Ring Interconnection

Two important parameters for optical scalability are the power loss and dynamic range of the received signals. The above two parameters can limit the size of an interconnection network.[5] In the HORN, however, both calculations degenerate to those of a single ring, as discussed in Section 3. Power loss in the HORN is defined as any loss associated with a ring, such as coupling losses from the fiber to a node, coupling losses from a node to a fiber, connector-insertion losses, and fiber-attenuation losses. The dynamic range for a receiver is defined as the maximum received power to the minimum received power.[22,24] A receiver in the HORN receives various signals from different processing nodes on the same network, on grounds that are strictly dependent on the location of the source and destination nodes. This situation is important because receivers only receive signals within a narrow dynamic range. More quantitative discussions on power-loss and dynamic-range calculations for the HORN are given below.

Let us assume for a ring in the HORN that the power coupling from the bus to a node is x $(0 < x < 1)$ and α is the coupling loss of the tap (further discussed in Section 4). Coupling losses in the tap for N nodes can be defined as

$$10 \log_{10}\left(\frac{P_{\text{in}}}{P_{\text{out}}}\right) = \alpha N, \tag{4}$$

where P_{in} is the power transmitted by the source PE and P_{out} is the power received by the destination PE. Solving for $P_{\text{out}}/P_{\text{in}}$ yields $10^{-\alpha N/10}$. Therefore, the ratio of the output power to the input power, with the assumption of two nodes to a ring, equals $(1 - x)10^{-\alpha/10}$, where the extra term $1 - x$ accounts for the coupling loss from the bus to a node. For N nodes the ratio of the output power to the input power is equal to

$$\begin{aligned}\eta_{\text{ring}} &= x^2(1 - x)^{N-2}10^{-\alpha N/10} \\ &= 10(N - 2) \log_{10}(1 - x) \\ &\quad + 20 \log_{10} x - \alpha N[\text{decibels}].\end{aligned} \tag{5}$$

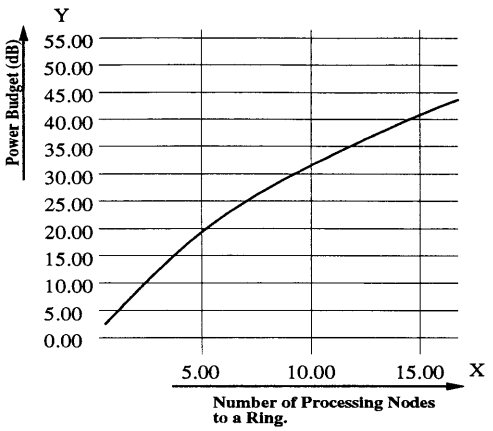

Fig. 7. Graph of the minimum power budget required for different numbers of processing nodes in a ring.

in decibels. If we take the derivative with respect to x and maximize η_{ring} we get

$$x_{\text{optimum}} \approx \frac{2}{N}. \tag{6}$$

Combining Eq. (5) and expression (6) yields

$$\begin{aligned}\eta_{\text{ring,optimum}} &\approx \frac{1}{e^2}\left(\frac{2}{N}\right)^2 10^{-\alpha N/10} \\ &= -[2.6 + 6 \log_2 N + \alpha N][\text{decibels}].\end{aligned} \tag{7}$$

The power budget for the HORN can subsequently be calculated by use of $\eta_{\text{ring,optimum}}$. The power budget for the HORN must ensure that enough power will reach the receiver for reliable performance during the entire system lifetime. The power budget must also incorporate all losses in the system. Therefore, the power budget in the HORN is related to $\eta_{\text{ring,optimum}}$ by the following equation:

$$\begin{aligned}&\text{Power Budget} = P_{tx}[\text{decibels}] - P_{\min}[\text{decibels}] \\ &> [2.6 + 6 \log_2 N + \alpha N][\text{decibels}] = \eta_{\text{ring,optimum}},\end{aligned} \tag{8}$$

where P_{tx} specifies the transmission power and $P_{\min}$ specifies the minimum required receiving power. Figure 7 shows a graph of minimum power-budget values (in decibels) for different values of N, if we assume α is equal to 1 dB.

In the HORN the maximum received signal occurs when the source and destination PE's are located counterclockwise of each other, respectively. The maximum received power is

$$P_{\max} = P_0 x^2 10^{-2\alpha/10}, \tag{9}$$

where P_0 is the transmission power. If, on the other hand, the source and destination PE's are located clockwise in relation to each other, the minimum received power is

$$P_{\min} = P_0 x^2 (1 - x)^{N-2} 10^{-N\alpha/10}. \tag{10}$$

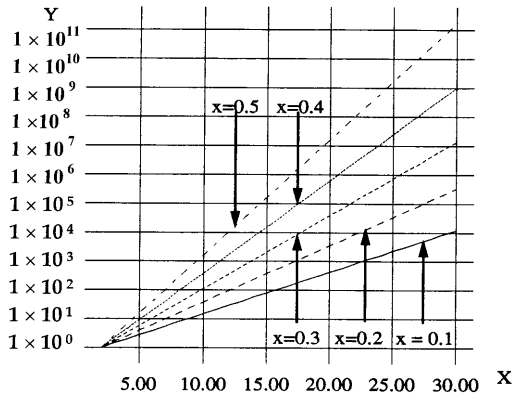

Fig. 8. Graph of the dynamic range plotted versus the processing nodes for different values of the power-coupling loss from the bus to a node (x).

The dynamic range DR is then $P_{\max}/P_{\min}$:

$$\mathrm{DR} = \frac{10^{(N-2)\alpha/10}}{(1-x)^{N-2}} = (N-2)[-10\log_{10}(1-x) + \alpha][\text{decibels}]. \quad (11)$$

Figure 8 shows plots for different numbers of processing nodes in a ring at different values of x.

To demonstrate the feasibility of the HORN implementation we have calculated the total power losses for our proposed system, as shown in Fig. 1. We assume a 1-dB loss occurs from the insertion of the laser signal into the fiber and a 1-dB loss occurs at the detector. Furthermore, the fiber is assumed to be at most 1 m in length at a mean operating wavelength of 960 nm. At this wavelength the fiber has an attenuation of 3.5 dB/km. Thus the fiber loss for the system is 0.0035 dB. Assuming that there are 16 PE's in a ring further equates to a power loss of approximately 43 dB (Fig. 7). Therefore, the total power losses from the input laser diode to the output photodiode is calculated to be approximately 45 dB. For a laser diode (NEC, Model NDL7513P1),[25] operating at 110 mW and a InGaAs photodiode (NEC, Model NDL5461P/P1) that can receive at 10 μW, a power budget of approximately 80 dB is attained. Consequently, this is well within our power-loss requirements for even 16 nodes to a ring.

D. Communication Delay

Communication delay is defined as the time required to complete a given communication operation. There are some common patterns of interprocessor communications that are frequently employed as building blocks in a variety of parallel algorithms.[26] Proper implementation of these basic communication operations is key to the efficient execution of the parallel algorithms that use them. Some of these basic communication operations are as follows: one-to-all broadcasting, all-to-all broadcasting, single-node accumulation, and one-to-all personalized communication. In this subsection, we discuss these operations with respect to the HORN.

A one-to-all broadcast takes place when a single PE sends identical data to all other PE's. Initially, only the source PE has the data of size X that need to be broadcast. After the broadcast there are $N-1$ copies of the initial data, one copy residing at each PE. In the HORN a single communication channel is all that is required for one-to-all broadcasting (e.g., λ_{22} for the wavelength assignment of Fig. 2). The HORN TDMA accomplishes one-to-all broadcasting with the source PE acquiring the wavelength assigned to the highest-level ring. The delay required for a one-to-all broadcast in the HORN TDMA is

$$T(\text{one-to-all})_{\text{TDMA}} \le T_{\text{C}}, \quad (12)$$

where T_C is the duration of a TDMA time cycle,

$$T_C = N * \Delta s. \quad (13)$$

Here, N corresponds to the number of PE's in the system and Δs is the time duration of a single TDMA time slot. The less than or equal to sign in relation (12) reflects the fact that the time slot required to obtain the broadcast channel can occur anywhere in T_C. Therefore, the average delay is equal to

$$T(\text{one-to-all})_{\text{TDMA}} = \frac{T_{\text{C}}}{2}. \quad (14)$$

All-to-all broadcasting (multinode broadcast) is a generalization of a one-to-all broadcast in which all PE's simultaneously initiate a broadcast. A PE dispatches the same packet to every other PE, but different PE's may broadcast different packets. Single-node accumulation is when a single PE accumulates packets from every other PE such that the packets accumulated can be composed of different information. This operation is dual to one-to-all broadcasting. One-to-all personalized communication (or single-node scatter) is when a single PE sends a unique packet to every other PE, and it occurs when a PE wants to send a personalized message to each PE. Consequently, for these types of communications, multiple PE's must obtain a single communication channel (e.g., all-to-all broadcasting and single-node accumulation) or a single PE must obtain multiple communication channels (e.g., one-to-all personalized communication). Thus, the HORN requires N communication channels to satisfy each of these operations. This requirement is unlike a one-to-all broadcast in which a source PE requires a single communication channel. The communication delay, therefore, for the HORN TDMA is

$$T_{\text{TDMA}} = T_C. \quad (15)$$

The equal sign of Eq. (15) reflects the fact that one entire time cycle T_C is required to obtain all the communication channels, unlike for relation (12).

One-to-all broadcasting, all-to-all broadcasting, single-node accumulation, and one-to-all personalized communication all map very efficiently into the HORN architecture. Broadcasts in the HORN require a single communication channel, as shown in

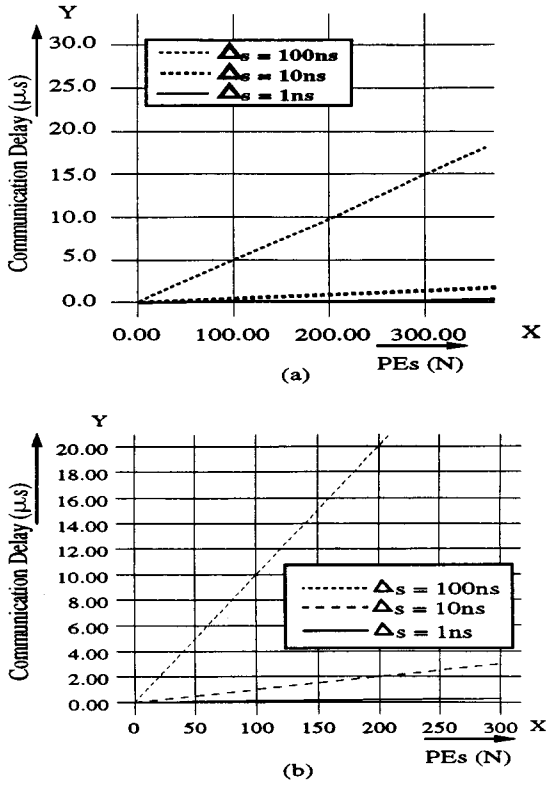

Fig. 9. Graphs for one-to-all broadcasting, all-to-all broadcasting, single-node accumulation, and one-to-all personalized communication. (a) Graph of the communication delay plotted versus N (number of processing nodes) for a variety of Δs (the time duration of a single time slot) in a one-to-all broadcast for the HORN TDMA. (b) Graph of the communication delay plotted versus N for a variety of Δs in an all-to-all broadcast, a single-node accumulation, and a one-to-all personalized communication for the HORN TDMA.

Subsection 2.B. Consequently, a one-to-all broadcast requires one communication channel, whereas the other operations require N communication channels. This is in fact the minimum number of communication channels required to satisfy these respective communication operations.[26]

However, these operations do not occur in isolation from one another, and multiple-access delay (flow-control delay) must also be taken into account. Equations (14) and (15) give an indication of the multiple-access delay for the respective communication operations for both the HORN TDMA and HORN arbitration. Plots given in Figs. 9(a) and 9(b) show the impact of the variation of multiple-access parameters on communication delay.

Figure 9(a) shows the graph of communication delay versus the number of PE's N in a one-to-all broadcast for the HORN TDMA with a variation in Δs. As the time slots (Δs) get smaller, the communication delay decreases and the capacity of the interconnection increases. Figure 9(b) shows the graph of communication delay versus the number of PE's N in an all-to-all broadcast, a single-node accumulation, and a one-to-all personalized communication. An entire T_C is required to fulfill these communication operations, as is reflected by the equal sign in Eq. (15). Three cases of varying Δs are considered. A reduction in Δs leads to a reduction in communication delay.

We accomplish improving the system performance (lowering communication delay) by choosing system parameters (multiple-access parameters) properly. Small time slots Δs lead to a reduction in communication delay. Lowering these system parameters results in an improvement in performance (reduction of communication delay) as a result of the increase in spatial reuse of the communication channels.

4. Optical Implementation of the Hierarchical Optical Ring Interconnection

Figures 10(a) and (b) show block diagrams of the composition of the processing and switching nodes of the HORN, respectively. The processing node [Fig. 10(a)] consists of an erbium-doped fiber amplifier (EDFA), tap (passive coupler), receiver, transmitter, and a second, nontap, passive coupler. EDFA's, represented by dotted lines in Fig. 10, are optional components for the processing nodes. EDFA's are required only if the power-budget calculations of a ring are not met, as discussed in Subsection 3.C. The tap is used to splice the signal from the ring to the receiver, where it can be detected depending on the wavelength of the signal and the wavelength that the receiver is tuned to. The use of EDFA's and taps for a ring was originally proposed in Ref. 27. The switching node shown in Fig. 10(b) consists of a passive coupler, EDFA, and AOTF. AOTF's are optical switches that can route on individual wavelengths.

The top ring in Fig. 10(a) is used for local communication, and the bottom ring is used for remote communication. All rings of the HORN are

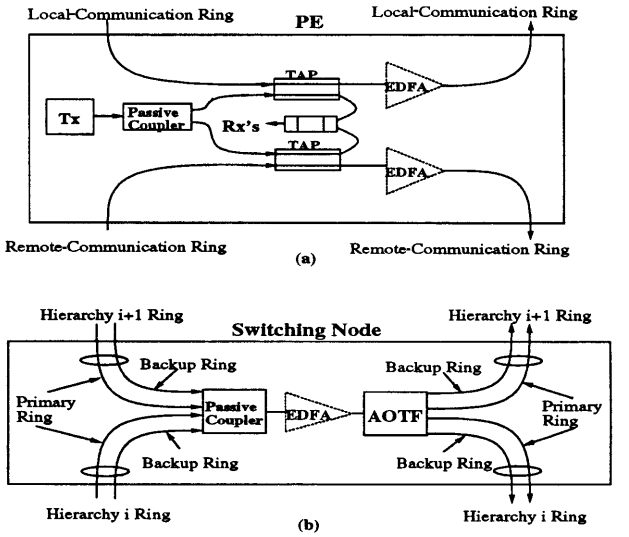

Fig. 10. Electro-optical components of the HORN: (a) Composition of a PE. The label Rx's represents an array of fixed, tuned receivers, and the label Tx represents a single tunable transmitter. EDFA's are shown within dotted lines because they need to be used only if the power budget calculations of a single ring are not met. (b) Composition of a switching node.

implemented by use of a simple buslike topology. We have chosen to use a buslike topology that employs taps instead of star couplers because of the splitting losses of star couplers. If there are a lot of hierarchies in the HORN, a packet will need to traverse all the way up in the hierarchy and then all the way down in the hierarchy in the worst case, which can have significant splitting losses. EDFA's cannot be used to amplify split signals produced from star couplers.[5] The Rx's shown in Fig. 10 are an array of fixed, tuned receivers used for receiving packets from both local and remote transmissions. Consequently, setting up the receivers in this manner allows the HORN to receive simultaneously from both local and remote communications. The transmitter (Tx) is connected to a passive coupler that is able to switch between the local and remote rings, depending on the communication required. This passive coupler can basically be implemented by an optical switch that is set depending on whether local or remote communication is desired by the PE.

The AOTF shown in Fig. 10(b) is used to route on individual wavelengths. A packet is either routed to the hierarchy $i + 1$ ring or the hierarchy i ring. The AOTF is an electronically tuned optical filter that operates on the principle of acousto-optic diffraction. One salient reason for using an AOTF is its electronic control, which requires no optical processing. All AOTF's are initially configured, i.e., see Table 1, and no further reconfigurations are required. Other features that make AOTF's ideal for interconnection networks are their electronic tuning with a fast scan rate and wide tuning range without secondary passbands, allowing it to route on multiple wavelengths.[20,21] Moreover, the combined capabilities of a wide tuning range and relatively large throughput of acousto-optic tunable filters make them favorable for the HORN interconnection. Because of page limitations we defer a detailed description of the optical implementation of the proposed HORN architecture to a subsequent paper.

5. Conclusion

We have put forward in this paper an optical hierarchal interconnection topology, the HORN, that is scalable and has a diameter of 1. Routing is efficiently implemented for both local and remote communications by virtue of wavelength reusability. Furthermore, fault tolerance and nonblocking routing are other characteristics that have been shown for the HORN. We have conducted a detailed scalability analysis for the HORN interconnection to show the potentiality of its use for multiprocessors and massively parallel systems. Finally, we have presented an optical design methodology for the proposed network and shown that the architecture is highly amenable to optical implementation. The physical components required are tunable transmitters, fixed tuned receivers, EDFA's, and passive couplers. We have shown the feasibility of these components as it relates to the HORN interconnection. Consequently, simple and cost-efficient optical implementation of the proposed network with existing optical hardware is possible.

References

1. J. W. Goodman, F. J. Leonberger, S. Y. Kung, and R. A. Athale, "Optical interconnections for VLSI systems," Proc. IEEE **72,** 850–866 (1984).
2. S. P. Dandamudi, and D. L. Eager, "Hierarchical interconnection networks for Multicomputer Systems," IEEE Trans. Comput. **39,** 786–797 (1990).
3. M. Holliday and M. Stumm, "Performance evaluation of hierarchical ring-based shared memory multiprocessors," IEEE Trans. Comput. **43,** 52–67 (1989).
4. Z. G. Vranesic, M. Stumm, D. M. Lewis, and R. White, "Hector: a hierarchically structured shared-memory multiprocessor," IEEE Comput. **28,** 72–79 (1991).
5. B. E. A. Saleh and M. C. Teich, *Fundamentals of photonics*, (Wiley-Interscience, New York, 1991).
6. M. R. Feldman, C. C. Guest, T. J. Drabik, and S. C. Esner, "Comparison between electrical and free-space optical interconnects for fine-grain processor arrays based on connection density capabilities," Appl. Opt. **28,** 3820–3829 (1989).
7. F. Kiamilev, P. Marchand, A. V. Krishnamoorthy, S. C. Esener, and S. H. Lee, "Performance comparison between optoelectronic and VLSI multistage interconnection networks," J. Lightwave Technol. **9,** 1665–1674 (1991).
8. A. Louri and H. Sung, "An optical multi-mesh hypercube: a scalable optical interconnection network for massively parallel computing," J. Lightwave Technol. **12,** 704–716 (1994).
9. A. D. McAulay, *Optical Computer Architectures* (Wiley Interscience, New York, 1991).
10. A. Guha, J. Bristow, C. Sullivan, and A. Husain, "Optical interconnections for massively parallel architectures," Appl. Opt. **29,** 1077–1093 (1990).
11. M. J. Murdocca, *A Digital Design Methodology for Optical Computing* (MIT Press, Cambridge, Mass., 1990).
12. A. E. Willner, C. J. Chang-Hasnain, and J. E. Leight "2-D WDM optical interconnections using multiple-wavelength VCSEL's for simultaneous and reconfigurable communication among many planes," IEEE Photon. Technol. Lett. **5,** 838–841 (1993).
13. M. I. Irshid and M. Kavehrad, "A fully transparent fiber-optic ring architecture for WDM networks," J. Lightwave Technol. **10,** 101–108 (1992).
14. Y. Li, A. W. Lohmann, and S. B. Rao, "Free-space optical mesh-connected bus networks using wavelength-division multiple access," Appl. Opt. **32,** 6425–6437 (1993).
15. P. W. Dowd, K. Bogineni, K. A. Aly, and J. A. Perreult, "Hierarchical scalable photonic architectures for high-performance processor interconnection," IEEE Trans. Comput. **42,** 1105–1120 (1993).
16. L. G. Kazovsky and P. T. Poggiolini, "STARNET: a multigigabit-per-second optical LAN utilizing a passive WDM star," J. Lightwave Technol. **11,** 1009–1026 (1993).
17. J. Bannister, M. Gerla, and M. Kovacevie, "An all-optical multifiber tree network," J. Lightwave Technol. **11,** 997–1008 (1993).
18. G. Bell, "Ultracomputers: a teraflop before its time," Commun. ACM **35,** 27–47 (1992).
19. D. A. Reed and H. D. Schwetman, "Cost-performance bounds for multicomputer networks," IEEE Trans. Comput. **32,** 83–95 (1983).
20. R. J. Berinato, "Acousto-optic tapped delay-line filter," Appl. Opt. **32,** 5797–5809 (1993).

21. R. B. Jenkins and B. D. Clymer, "Acousto-optic comparison switch for optical switching networks with analog addressing techniques," Appl. Opt. **31,** 5453–5463 (1992).
22. J. P. Powers, *An Introduction to Fiber Optic Systems* (Aksen, Homewood, Ill., 1993).
23. G. M. Lundy, "Analyzing a CSMA/CD protocol through a systems of communicating machines specification," IEEE Trans. Commun. **41,** 447–450 (1993).
24. G. P. Agrawal, *Fiber-Optic Communication Systems* (Wiley-Interscience, New York, 1992).
25. "Electron components: optical semiconductor devices," in *A Data Sheet Pack for Optical Semiconductor Devices,"* (NEC, May 1995), pp. 1–200.
26. V. Kumar, A. Grama, A. Gupta, and G. Karypis, *Introduction to Parallel Computing* (Benjamin/Cummings, Redwood City, Calif., 1994).