Scalable optical hypercube-based interconnection network for massively parallel computing

Ahmed Louri and Hongki Sung

Two important parameters of a network for massively parallel computers are scalability and modularity. Scalability has two aspects: size and time (or generation). Size scalability refers to the property that the size of the network can be increased with nominal effect on the existing configuration. Also, the increase in size is expected to result in a linear increase in performance. Time scalability implies that the communication capabilities of a network should be large enough to support the evolution of processing elements through generations. A modular network enables the construction of a large network out of many smaller ones. The lack of these two important parameters has limited the use of certain types of interconnection networks in the area of massively parallel computers. We present a new modular optical interconnection network, called an optical multimesh hypercube (OMMH), which is both size and time scalable. The OMMH combines positive features of both the hypercube (small diameter, high connectivity, symmetry, simple routing, and fault tolerance) and the torus (constant node degree and size scalability) networks. Also presented is a three-dimensional optical implementation of the OMMH network. A basic building block of the OMMH network is a hypercube module that is constructed with free-space optics to provide compact and high-density localized hypercube connections. The OMMH network is then constructed by the connection of such basic building blocks with multiwavelength optical fibers to realize torus connections. The proposed implementation methodology is intended to exploit the advantages of both space-invariant free-space and multiwavelength fiber-based optical interconnect technologies. The analysis of the proposed implementation shows that such a network is optically feasible in terms of the physical size and the optical power budget.

Key words: Optical interconnects, scalability, parallel processing, space-invariant interconnects, multiwavelength multiplexing.

1. Introduction

The quest for Tflops (10^{12} floating-point operations per second) supercomputers combined with the launching of the High Performance Computing and Communication initiative is putting major emphasis on exploiting massive parallelism with greater than 1000 processing elements (PE's) networked to form massively parallel computers (ultracomputers).[1,2] A key element, and a deciding factor in terms of performance and cost of these computers, is the interconnection network.[3] The interconnection network for massively parallel computers must not only be adequate in terms of communication bandwidth, latency, and connectivity, it must also be modular and scalable.

Scalability of a network consists of two aspects: size scalability and generation scalability (or time scalability).[2] Size scalability refers to the property that the size of the network (e.g., the number of communicating nodes) can be increased with a nominal change in the existing configuration. Also, the increase in system size is expected to result in an increase in performance comparable to the increasing size. A generation-scalable network could be implemented in a new technology, and the interconnection bandwidth of the network should grow at the same rate as processing speed and memory. Without increasing interconnection bandwidth, we cannot fully exploit the increased speed of evolutionary PE's. Modular networks enable the construction of a large network out of smaller networks.

Numerous topologies have been explored for parallel computers.[4–7] However, the lack of size scalability and modularity of some of these networks has limited their use in massively parallel computing systems despite their many other advantages. For example, one of the most popular networks for paral-

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

Received 18 October 1993; revised manuscript received 28 April 1994.

0003-6935/94/327588-11$06.00/0.

lel computers is the binary n cube, also called a hypercube. The attractiveness of the hypercube topology is its small diameter, which is the maximum number of links (or hops) a message has to travel to reach its final destination between any two nodes. A binary n-cube network has 2^n nodes, and the diameter is n. Each node is numbered in such a way that there is only one binary digit difference between any node and its n neighbors (node degree) that are directly connected to it. This property greatly facilitates the routing of messages through the network. In addition, the regular and symmetric nature of the network provides fault tolerance. Despite its small diameter, high connectivity, simple routing scheme, and fault tolerance, the hypercube topology is rarely adopted in the most recent projects for massively parallel computers, such as the Intel Paragon, the Cray Research MPP Model, the Caltech Mosaic C, the MasPar MP-1, the Tera Computer Tera Multiprocessor, and the Stanford Dash Multiprocessor, which are based on the torus/mesh topology.[1,8] One major reason is its lack of size scalability. As the dimension of the hypercube is increased by one, one additional link needs to be added to every node in the network. In addition to the changes in the node configuration, at least a doubling of the number of existing nodes is required for the regular hypercube network to expand and to remain as a hypercube.

Torus networks (henceforth, the mesh is referred to as a torus if the mesh has wraparound connections in the rows and columns) are easily implemented because of the simple regular connection and the small number of links (four) per node. Because of its constant node degree, the torus network is highly size scalable. With a network size of N nodes the minimal incremental size is approximately $N^{1/2}$ for a perfectly balanced network. However, the torus network also suffers from a major limitation, which is its large diameter ($N^{1/2}$ for an N-node network), along with its limited connectivity. Despite the fact that the mesh/torus topology has limited connectivity and a large diameter, many recent projects for massively parallel computers targeting Tflops use this topology for the interconnection network.

Motivated by these limitations, we explored a novel topology for optical interconnection networks, called the optical multimesh hypercube (OMMH), which combines the advantages of both the hypercube (small diameter, high connectivity, symmetry, simple control and routing, and fault tolerance) and the mesh (constant node degree and size scalability) topologies yet circumvents their disadvantages (the lack of size scalability of the hypercube and the large diameter of the mesh/torus). The topology of the OMMH network is size scalable. Time scalability is provided by the optics-based interconnection architecture. We developed a three-dimensional optical design methodology that exploits the advantage of both space-invariant free-space and multiwavelength fiber-based optical interconnect technologies. The proposed implementation is also analyzed for an example OMMH network with a ten-cube module as a basic building block. The analysis includes examination of the power flow, the efficiency, and the system volume. It is shown that a ten-cube (1024-node) module is containable within a 25.4 mm × 25.4 mm × 203.2 mm volume with a power efficiency of 16% and that the torus subnetwork has a power efficiency of 23.6%.

The distinctive advantages of the proposed design methodology include the following: (1) an efficient and scalable interconnection network, (2) better utilization of the space–bandwidth product of optical imaging systems, (3) full exploitation of the parallelism of free-space optics and the high bandwidth of fiber optics, and (4) compatibility with the emerging two-dimensional optical logic and switching and the optoelectronic integrated-circuit technologies.

2. Topology of the Optical Multimesh Hypercube Network

In this section we define the structure of the OMMH. We then compare and contrast structural properties of the OMMH with the hypercube network.

A. Definition of the Optical Multimesh Hypercube Network

An OMMH is characterized by a triplet (l, m, n), where l represents the row dimension of a torus, m represents the column dimension of the torus, and n represents the dimension of a binary hypercube.

An (l, m, n)-OMMH network is constructed as follows. For two nodes (i_1, j_1, k_1) and (i_2, j_2, k_2), where $0 \le i_1 < l, 0 \le i_2 < l, 0 \le j_1 < m, 0 \le j_2 < m, 0 \le k_1 < 2^n$, and $0 \le k_2 < 2^n$, the following holds:

(1) There is a link (called a torus link) between two nodes if (i) $k_1 = k_2$ and (ii) two components, i and j, differ by 1 in one component while the other component is identical.

(2) There also exists a torus link for the wraparound connection in the row if (i) $k_1 = k_2$ and (ii) $i_1 = i_2$, $j_1 = 0$, $j_2 = m - 1$, or for the wraparound connection in the column if (i) $k_1 = k_2$ and (ii) $j_1 = j_2$, $i_1 = 0$, and $i_2 = l - 1$.

(3) There is also a link (called a hypercube link) between two nodes if and only if (i) $i_1 = i_2$, (ii) $j_1 = j_2$, and (iii) k_1 and k_2 differ by one bit position in their binary representation (Hamming distance of 1).

Figure 1 shows a (4, 4, 3)-OMMH interconnection, in which solid lines represent hypercube links and dashed lines represent torus links. A (4, 4, 3)-OMMH consists of $(4)(4)(2^3) = 128$ nodes. Filled circles represent nodes of the OMMH network which are, in this paper, abstractions of PE's, which consist of electronic processing modules for computation and optical sources and detectors for communication. Both ends of torus links, shown by dashed lines, are connected for wraparound connections of the torus if they have the same labels. The size of the OMMH can grow without a change in the number of links per node by expansion of the size of the torus, for

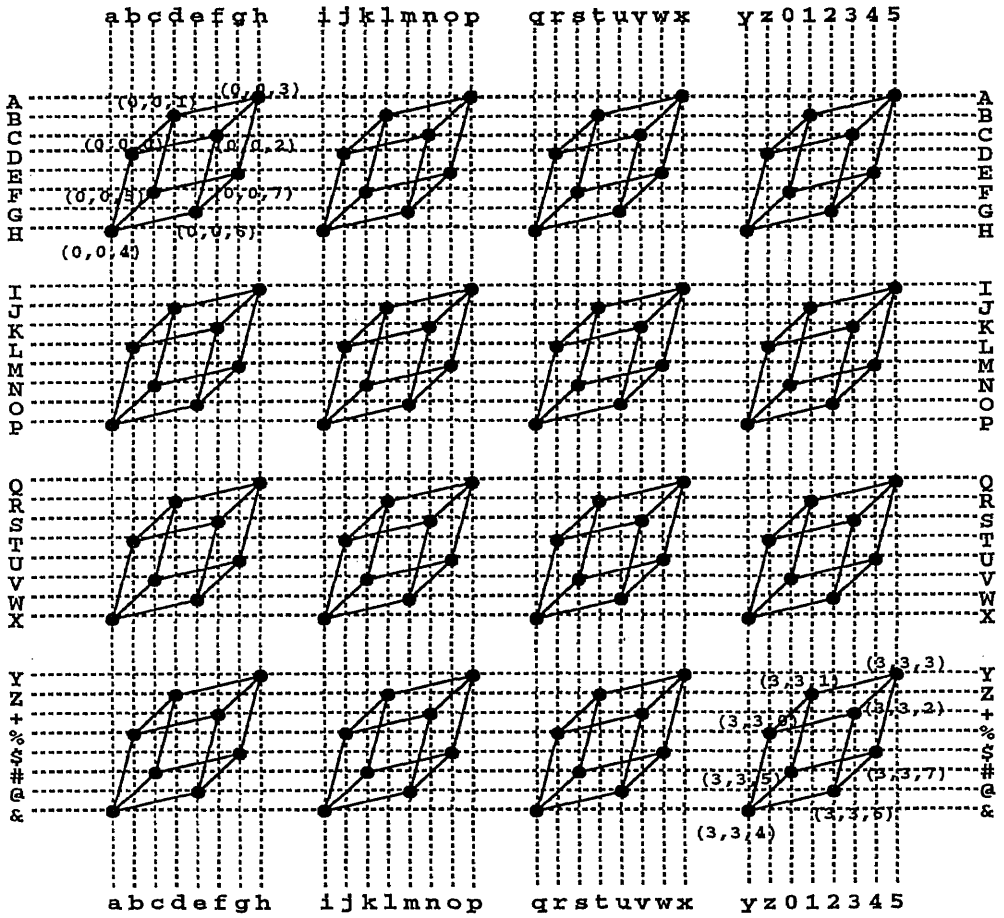

Fig. 1. Example of the optical multimesh hypercube network: a (4, 4, 3)-OMMH (128-node) interconnection is shown. Two links with the same labels are connected for the wraparound connections of the torus. Only a few addresses are shown in parentheses for clarity. Solid lines represent hypercube connections, and dashed lines represent torus connections.

example, by insertion of three-cube modules in the row or the column of the torus in Fig. 1. This feature permits the OMMH to be size scalable. More discussion on the scalability issue follows in Subsection 3.A.

An interesting isomorphic network is shown in Fig. 2. The same network is redrawn as a 4 × 4 torus-clustered three-cube network. It can be viewed as eight concurrent toruses, where eight nodes having identical torus addresses form one three-cube module. It can also be viewed as 16 concurrent three-cube modules in which 16 nodes having identical hypercube addresses form a 4 × 4 torus. The (4, 4, 3)-OMMH in Fig. 1 appears similar to a three-cube-clustered 4 × 4 torus. Depending on the problem at hand, the OMMH can be configured as torus-clustered hypercubes or as hypercube-clustered toruses.

B. Message Routing in the Optical Multimesh Hypercube

Because of the regularity of the structure, a distributed routing scheme can be implemented without global information. Since the OMMH is a point-to-point network, packet communication is assumed in the message-routing scheme. For an (l, m, n)-OMMH network, let the addresses of two arbitrary nodes S and T be (i_s, j_s, k_s) and (i_t, j_t, k_t), respectively, where $0 \le i_s < l, 0 \le i_t < l, 0 \le j_s < m, 0 \le j_t < m$, $0 \le k_s < 2^n$, and $0 \le k_t < 2^n$. The message-routing scheme from S to T is that of an n-cube network, an $l \times m$ torus network, or a combination of the two, depending on the relative locations of the nodes:

(1) *Routing within a Hypercube.* If $i_s = i_t$ and $j_s = j_t$, then S and T are within the same hypercube. The routing scheme for this case is exactly the same as that of the n-cube network.

(2) *Routing within a Torus.* If $k_s = k_t$, then S and T are within the same torus. The routing scheme for this case is exactly the same as that of the $l \times m$ torus network.[9]

(3) *Routing through Toruses and Hypercubes.* If none of the above two cases are true, S and T share neither a hypercube nor a torus. There are several options available for this case. One option uses the hypercube routing scheme until the message arrives at the same torus at which T resides, and then it uses the torus routing scheme for the message to arrive at T. In another option the torus routing scheme can first be applied to forward the message to the same hypercube at which T resides, and then the message can reach T with the hypercube routing scheme.

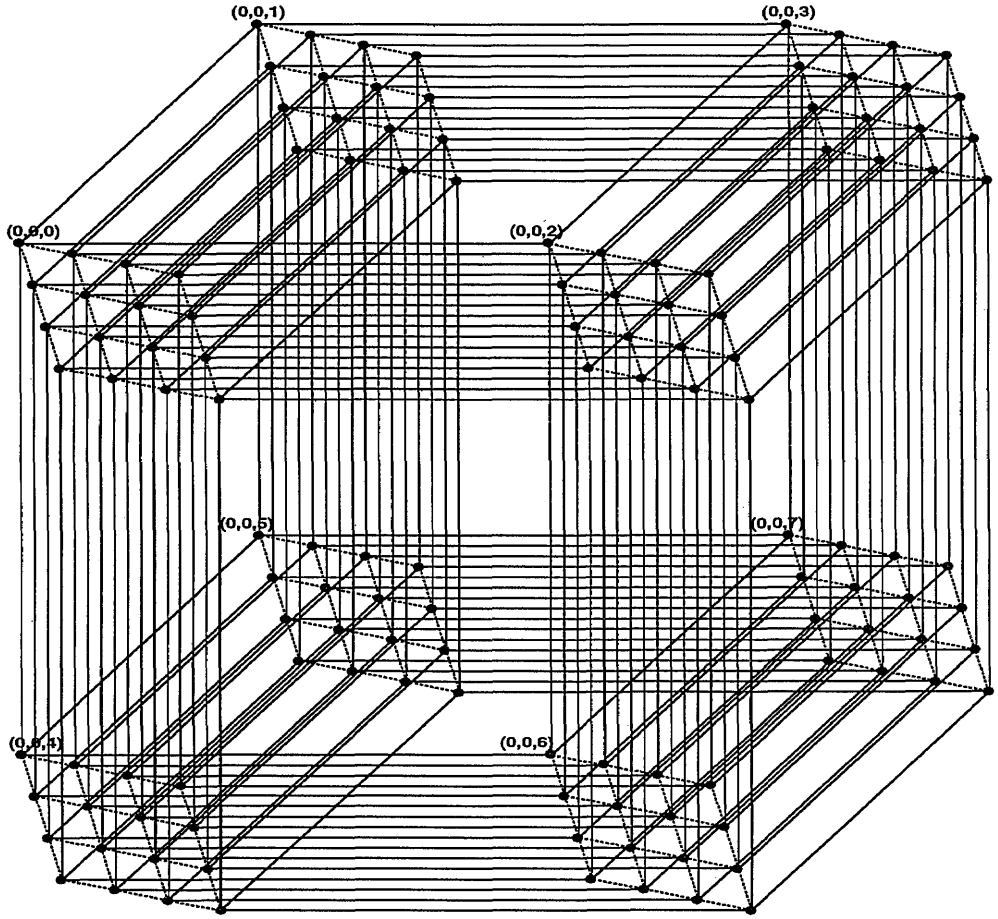

Fig. 2. (4, 4, 3)-OMMH interconnection network represented in an isomorphic view. Wraparound connections of the torus are omitted, and only a few addresses are shown in parentheses for clarity. Solid lines represent hypercube connections, and dashed lines represent torus connections.

We can also mix the hypercube and the torus routing until the message is forwarded to the same hypercube or to the same torus at which T resides, and then we can forward the message to T using the hypercube or the torus routing scheme, respectively.

C. Diameter and Link Complexity

The distance between two nodes in a network is defined as the number of links connecting these two nodes. The diameter of a network is defined as the maximum of all the shortest distances between any two nodes. The diameter of the network is of great importance since it determines the maximum number of hops that a message may have to take. For two extreme cases the diameter of a linear array with N nodes is $(N - 1)$, while that of a completely connected network is unity. An $l \times m$ torus has diameter $(\lfloor l/2 \rfloor + \lfloor m/2 \rfloor)$. The diameter of a hypercube with N nodes is $\log_2(N)$. Thus the diameter of an (l, m, n)-OMMH is $(\lfloor l/2 \rfloor + \lfloor m/2 \rfloor + n)$.

Link complexity or node degree is defined as the number of links per node. The higher the link complexity, the greater is the hardware complexity and, consequently, the cost of the network. The node degree of a hypercube with N nodes is $\log_2(N)$ and that of an (l, m, n)-OMMH is $(n + 4)$. N is equal to $(l)(m)(2^n)$ if the hypercube and the OMMH have the same network size. A comparison of diameters should be accompanied by a comparison link complexity because a higher connectivity resulting from a higher link complexity is expected to lead to smaller diameters. Figure 3(a) compares the diameters of the hypercube and the OMMH, in which (16, 16, n)-OMMH means the size of the torus in the OMMH is fixed and the size of the hypercube in the OMMH is changed to have the same network size for comparison purposes. Similarly, $(l, m, 4)$-OMMH implies the size of the hypercube in the OMMH is fixed and that of the torus is changed. Figure 3(b) compares link complexities, or node degrees, of the hypercube and those of the OMMH. Figure 3(c) depicts the growth of the total number of links in the network as the network size increases. For a network size of 10^6 nodes the hypercube network contains $\sim 10.5 \times 10^6$ links, the $(l, m, 4)$-OMMH has $\sim 4.2 \times 10^6$ links, and the (16, 16, n)-OMMH has approximately $\sim 8.4 \times 10^6$ links. Since one link implies one physical path, electrical or optical, between two nodes, the OMMH network is cost efficient compared with the hypercube network in terms of hardware requirements.

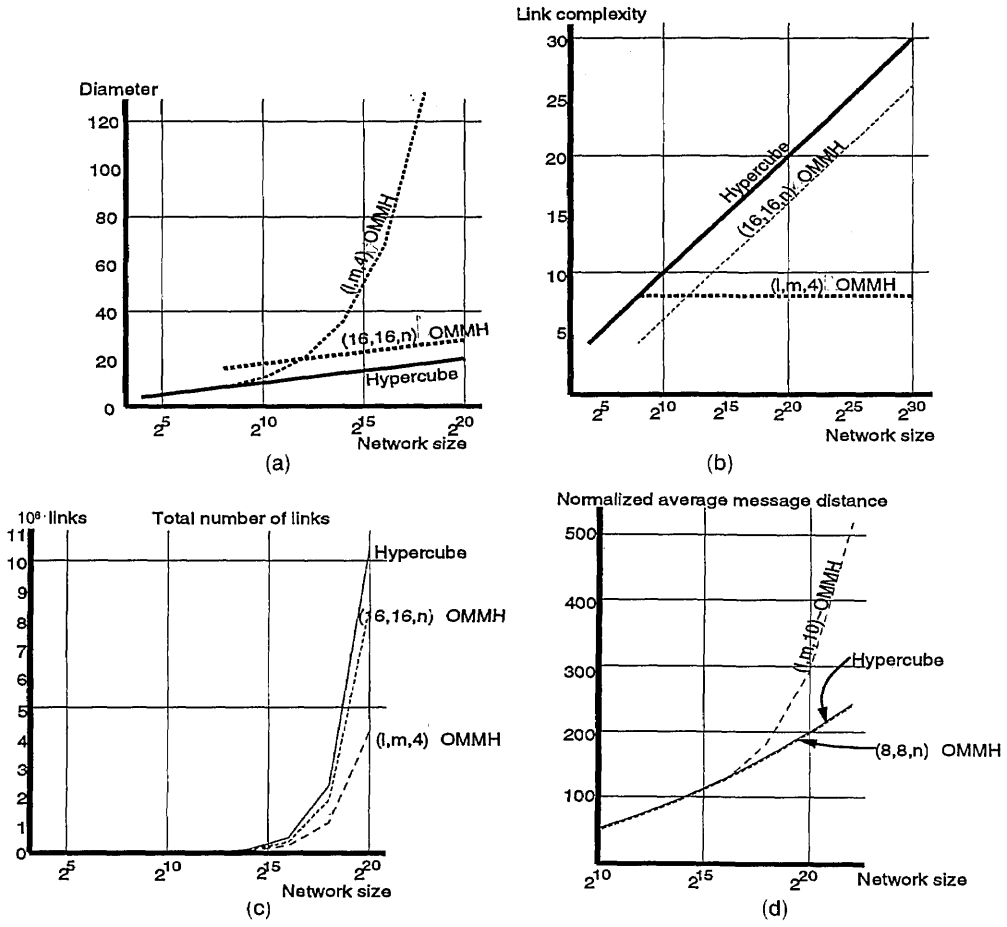

Fig. 3. Comparison of (a) diameter, (b) link complexity, (c) total number of links, and (d) normalized average message distance of the hypercube and the OMMH when the two networks have the same number of nodes.

D. Fault Tolerance of the Optical Multimesh Hypercube Network

As the number of components in a system grows, the probability of the existence of faulty components increases. For a large-scale system we cannot always expect that all components in such a system are free from any failures. However, we need to expect such a system to continue to operate correctly in the presence of a reasonable number of failures.

Because of the concurrent presence of toruses and hypercubes in the OMMH network, rerouting of messages in the presence of a single faulty link or a single faulty node can easily be done with little modification of existing fault-free routing algorithms. In the OMMH network any single faulty link or any single faulty node can be bypassed by only two additional hops as long as this particular node is not involved in the communication; namely, the node is neither the source nor the destination for any message. Briefly, this can be proven as follows. When the torus routing scheme is being applied in the presence of a faulty link or node, one additional hop is needed to forward the message to a neighboring torus subnetwork through a hypercube link [n such neighboring toruses exist in an (l, m, n)-OMMH], and another hop is needed to return the message to the original torus subnetwork. Similarly, when the hypercube routing scheme is being applied, a message can detour a faulty link or node with two additional hops, one to forward the message to a neighboring hypercube subnetwork and another to return the message to the original hypercube subnetwork.

3. Modular and Scalable Optical Interconnection Architecture of the Optical Multimesh Hypercube Network and Its Optical Implementation

In this section we discuss modularity and scalability issues of the OMMH interconnection architecture, and we present an optical implementation of the OMMH network. Then we discuss the rationale and the performance of the proposed OMMH implementation.

A. Scalable Interconnection Architecture of the Optical Multimesh Hypercube Network

1. *Size Scalability of the Optical Multimesh Hypercube Network*

Size-scalable networks have the property that the size of the system (e.g., the number of communicating nodes) can be increased with nominal changes in the existing configuration. Also, the increase in system

size is expected to result in an increase in performance to the extent of the increase in size. As the dimension of the hypercube is increased by one, one more link needs to be added to every node in the network. In addition to the changes in the node configuration, at least a doubling of the size is required for the hypercube network to expand. This implies that the hypercube does not permit an incremental expansion of small sizes. Thus the hypercube network is not scalable according to the above definition. We should note that the hypercube network may be scalable at a greater cost. Moreover, it is not modular.[6] The lack of size scalability and modularity have limited the application of the hypercube topology to large-scale massively parallel systems.

As can be seen in Fig. 3(b), the OMMH with a constant cube as a basic building block [e.g., an $(l, m, 4)$-OMMH] has a constant node degree, which means that the size of the OMMH is ready to be scaled up by expansion of the size of the torus without the link complexity of existing nodes being affected; as is the case in expansion of the size of the hypercube network. However, we cannot just add one node to the OMMH. For an (l, m, n)-OMMH we need to add at least $(l)(2^n)$ nodes (if $l < m$) when the torus subnetwork needs to balanced.

An OMMH network is constructed from simple building blocks (hypercubes) in a modular and incremental fashion. These building blocks, once constructed, are left undisturbed when the network grows in size. The OMMH can be viewed as a two-level interconnection network: high-density, local connections for hypercube links (within a basic module) and high-bit-rate, low-density, long connections for the torus links connecting the basic building blocks. Thus one can increase the size of the OMMH by adding hypercube modules, which provides modularity and size scalability.

2. *Generation Scalability of the Optical Multimesh Hypercube Network*

Generation-scalable architectures are designed with consideration of what the future implementations may be. Such architectures will survive throughout generations. A generation-scalable network can be implemented in a new technology, and the interconnection bandwidth of the network should grow at the same rate as processing speed and memory. Without increasing interconnection bandwidth, we cannot fully exploit the increased speed of evolutionary processing elements. Generation scalability in the OMMH interconnection architecture is provided by the use of high-bandwidth optics, which would match communications bandwidth requirements of future processing elements.

B. Optical Implementation of the Optical Multimesh Hypercube Network

In this subsection we present an optical implementation of the hypercube networks for constructing basic building blocks. Then we show how to design torus links to connect such hypercube modules to construct the OMMH network.

1. *Optical Implementation of Space-Invariant Hypercubes Using Binary Phase Gratings*

We discuss an optical implementation of the three-dimensional space-invariant hypercube network; the design methodology is proposed in Ref. 10. The design methodology is based on an observation that nodes in an interconnection network can be partitioned into two sets of nodes such that any two nodes in a set do not have a direct link (except for completely connected networks). This is a well-known problem of bipartitioning a graph if the interconnection network is represented as a graph. For a binary n-cube network, nodes whose addresses differ by more than a Hamming distance of 1 can be in the same partition, since no link exists between two nodes if their Hamming distance is greater than 1. Besides bipartitioning the graph, we arrange the nodes in each partition onto the plane such that interconnection between two planes becomes space invariant. Two partitions of nodes are called Plane_L and Plane_R, respectively, in Ref. 10. The methodology uses a free-space optical multiple imaging technique to replicate and to spatially shift the image of one partition of nodes. Multiple images are then simultaneously incident upon the other partition. The locations of multiple image spots on the receiving partition are determined by the required connection patterns.[10]

There is a wide variety of optical means to generate multiple images. These include phase gratings,[11,12] beam splitters,[13] multiple split lenses,[14] lenslet arrays,[15] arrays of mirrors,[16] and holographic techniques.[17] A hypercube-based architecture for cellular image processing has been demonstrated with holograms,[18] and a new concept of grid patterns for the layout of an optoelectronic integrated-circuit chip has been proposed.[19]

In the following we discuss the design of binary phase gratings (BPG's) for the five-cube implementation as an example. Figure 4 describes a hardware arrangement of optical components for a space-invariant five-cube network. For clarity, only a two-dimensional view is shown. A BPG is added at the pupil plane between two imaging lenses to provide necessary beam-steering operations. This type of

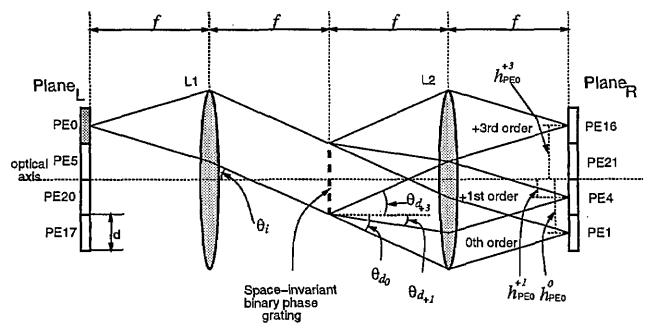

Fig. 4. Space-invariant optical implementation of a five-cube network with a binary phase grating: L1, L2, lenses.

arrangement was first proposed in Ref. 20. We extended it here for the implementation of space-invariant hypercube networks. Since the interconnection patterns are space invariant, any beam-steering operation performed on one of the beams must be performed on all of the beams that pass through the BPG. The beam-steering operation of the BPG is dictated by the grating equation shown in Eq. (1), which describes the relationships among the angle of the incident beam θ_i, the period of the grating p, the wavelength of the light λ, the grating order m, and the angle of the mth order's diffracted beam θ_{d_m}:

$$p(\sin \theta_{d_m} - \sin \theta_i) = m\lambda. \quad (1)$$

Assume that the size of a node in one dimension is d and that the focal length of each lens is f. Let h^m_{PE0} be the distance of an image spot in Plane$_R$ from the optical axis made by the mth-order diffracted beam from PE0. Then,

$$h^m_{\mathrm{PE0}} = f(\tan \theta_{d_m}). \quad (2)$$

Given that $\theta_i = \tan^{-1}(1.5d/f)$, Eq. (2) can be rewritten as

$$h^m_{\mathrm{PE0}} = f \tan\left(\sin^{-1}\left\{\frac{m\lambda}{p} + \sin\left[\tan^{-1}\left(\frac{1.5d}{f}\right)\right]\right\}\right). \quad (3)$$

We assume that the structure of the grating is designed such that the power of the incident beam is equally distributed into the zeroth, the positive and negative first, and the positive and negative third orders of diffracted beams, and others are suppressed. We can have different amounts of optical power from the original beam routed into the different orders by changing the periodic structure of the grating. To have different angular spacings, we should change the period of the grating.[17] Since PE0 is supposed to be connected with PE1, PE4, and PE16 for the five-cube network, the following conditions should be satisfied:

$$\begin{aligned} h^0_{\mathrm{PE0}} &= 1.5d, \\ h^{+1}_{\mathrm{PE0}} &= 0.5d, \\ h^{+3}_{\mathrm{PE0}} &= -1.5d, \\ h^{-1}_{\mathrm{PE0}} &> 2.0d, \\ h^{-3}_{\mathrm{PE0}} &> 2.0d. \end{aligned} \quad (4)$$

Note that the conditions for h^{-1}_{PE0} and h^{-3}_{PE0} make negative-first- and negative-third-order diffracted beams fall outside Plane$_R$ to avoid unwanted conditions.

Similarly, the beam from PE5 generates multiple spots in Plane$_R$, for which the distances from the optical axis are

$$h^m_{\mathrm{PE5}} = f \tan\left(\sin^{-1}\left\{\frac{m\lambda}{p} + \sin\left[\tan^{-1}\left(\frac{0.5d}{f}\right)\right]\right\}\right). \quad (5)$$

To make connections from PE5 to PE1, PE4, and PE21, we need the following set of conditions;

$$\begin{aligned} h^0_{\mathrm{PE5}} &= 0.5d, \\ h^{+1}_{\mathrm{PE5}} &= -0.5d, \\ h^{-1}_{\mathrm{PE5}} &= 1.5d, \\ h^{+3}_{\mathrm{PE5}} &< -2.0d, \\ h^{-3}_{\mathrm{PE5}} &> 2.0d. \end{aligned} \quad (6)$$

Note that conditions for h^{+3}_{PE5} and h^{-3}_{PE5} make positive-third and negative-third diffracted beams fall outside Plane$_R$. Since PE0 and PE5 are placed symmetrically with respect to the optical axis, with PE17 and PE20, we can determine the period of the grating p to provide the required connections for the five-cube network by solving relations (4) and (6) given the size of a node, the focal length of the lens, and the wavelength of the light source. However, we cannot have an exact solution since image spots generated by both PE0 and PE5 cannot be placed on uniform spacings in Plane$_R$. An approximate solution could be determined by a computer program that optimizes conditions in relations (4) and (6). By optimization we mean minimization of errors in each condition. For example, given that the node size in one dimension is 5 mm, the wavelength of the light source 960 nm, and the focal length of the lens 50.8 mm, then the optimum period of the grating is computed to be 19.6 μm, which causes maximum misalignment of 9.0 μm at PE21 from the PE5 connection. The feature size of a node for the construction of massively parallel computers will depend mainly on the area required for PE's and memories for a single node, not on the size of the light source and detector. With wafer-scale integration the size of a node would be small enough (assuming fine-grain or medium-grain parallelism) so as not to make imaging lenses impractically large.[21]

The size of a basic n-cube module that can be implemented is determined primarily by the number of fan-outs that can be managed by the BPG since an n-cube implementation requires $2n - 1$ fan-outs. The BPG must be able to generate $2n - 1$ beams of equal power. We note that a hypercube of relatively small size could be implemented easily with electronics if the bandwidth requirement were not large, but as the network size grew, optics would be more advantageous than electronics both in design complexity and bandwidth.

2. *Design of Torus Links to Connect Hypercube Modules*

An (l, m, n)-OMMH can be constructed as follows:

(1) $l \times m$ n-cube modules, as described in Subsection 3.B.1, are placed in an $l \times m$ matrix form.

(2) $l \times m$ nodes, each of which is from the same location of the n-cube modules, are connected to form a torus of dimensions $l \times m$ (A node consists of an

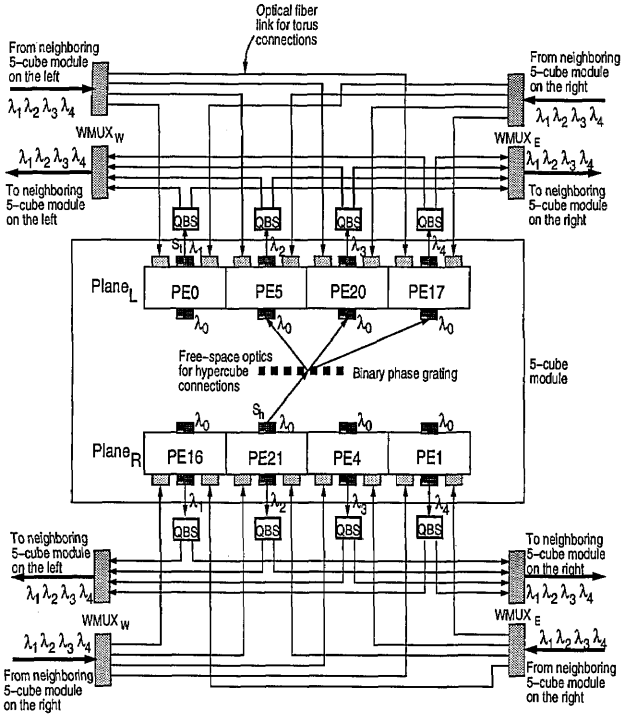

Fig. 5. Two-dimensional side view of a five-cube module to interface with torus links for the construction of the $(l, m, 5)$-OMMH network. (See Subsection 3.B.2 for a description of the components.)

n-cube module with a torus link interface. An example of a five-cube module with a torus interface is shown in Figs. 5 and 6.)

(3) Step 2 is repeated until every node is connected, resulting in 2^n toruses of size $l \times m$.

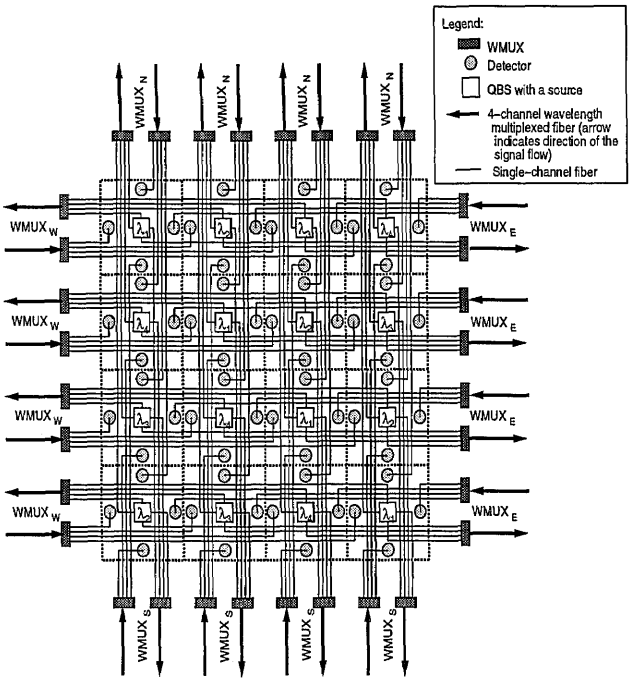

Fig. 6. Two-dimensional top view of the five-cube module shown in Fig. 5.

Since two adjacent n-cube modules are connected by 2^n torus links, the number of optical fibers required grows exponentially as n increases. A possible solution for reducing the number of optical fibers required is the use of a wavelength-multiplexing technique. However, a straightforward use of the wavelength multiplexing also requires a prohibitively large number of different wavelengths. For example, to connect two ten-cube modules, we need $2^{10} = 1024$ different wavelengths. A wavelength-node assignment technique[22] can alleviate this problem as follows.

Referring to Ref. 10, we can see that an n-cube layout (Plane_L or Plane_R) consists of $2^{\lfloor (n-1)/2 \rfloor}$ nonempty rows and $2^{\lceil (n-1)/2 \rceil}$ nonempty columns. For Plane_L and Plane_R we assign the following wavelengths to the nodes in the first row: $\lambda_1, \lambda_2, \ldots, \lambda_{2^{\lceil (n-1)/2 \rceil}}$. Then we assign $\lambda_2, \ldots, \lambda_{2^{\lceil (n-1)/2 \rceil}}, \lambda_1$ as wavelengths to the nodes in the seconds row. In general, wavelength assignment in a row is achieved by rotation of the wavelength assignment of the previous row by one column. This wavelength assignment results in no two nodes in the same row or column having an identical wavelength. Figure 7 shows a wavelength assignment for a five-cube module. We then use a $2^{\lceil (n-1)/2 \rceil}$-channel wavelength-multiplexed fiber to connect two rows in the adjacent two n-cube modules. Similarly, a $2^{\lfloor (n-1)/2 \rfloor}$-channel fiber is used to connect two columns in the adjacent two n-cube modules. Thus an implementation of an (l, m, n)-OMMH with the above wavelength-assignment method requires no more than $2^{\lceil (n-1)/2 \rceil}$ different wavelengths. In addition, no more than $2^{\lceil (n-1)/2 \rceil}$ optical fibers are required for the connections between any two adjacent n-cube modules.

Now we consider an optical implementation of the (l, m, n)-OMMH network. We assume the availability of two optical components: A quadrant beam splitter (QBS), which splits a single beams into four beams, and an i-channel wavelength multiplexer (WMUX), which multiplexes beams with i different wavelengths into a single beam (and also demultiplexes since it is bidirectional). The realization of these two components with current technology is discussed in detail in Subsection 3.B.3. Figures 5 and 6 shows an example of a five-cube basic module

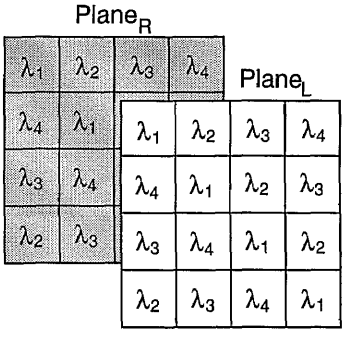

Fig. 7. Wavelength assignment for a five-cube module. Wavelengths are assigned such that no two nodes in the same row or column have an identical wavelength.

construction, including the torus link interface. We assume that each node has two light sources; one source, S_h, illuminates the BPG to generate the required hypercube links, and the second source, S_t, is coupled with an optical fiber for the torus links. The number of detectors per node is dependent on the incoming channel distinction technique.[23] If we use the space-division multiplexing technique, we need four detectors per node for the torus connection and 2^n detectors per node for the n-cube connection (the number of detectors in the n-cube module can be reduced greatly if a channel-encoding technique is used.[23,24] A QBS is attached to every S_t to provide the four fan-outs, S_{t_N}, S_{t_S}, S_{t_E}, and S_{t_W} (north, south, east, and west). A WMUX is located at both ends of each row and each column. Let each WMUX at the right end of a row be WMUX_E, each WMUX at the left end of a row be WMUX_W, each WMUX at the top of a column be WMUX_N, and each WMUX at the bottom of a column be WMUX_S. In a given row a WMUX_E multiplexes light from the S_{t_E} sources of that row into a single fiber, which is then connected to a WMUX_W in the neighboring n-cube module. Similarly, S_{t_N}s, S_{t_S}s, and S_{t_W}s sources are multiplexed by WMUX_N, WMUX_S, and WMUX_W, respectively. In the receiving module, these signals are demultiplexed by the WMUX and routed to the corresponding nodes. Figure 5 illustrates a side view of a five-cube module with a torus link interface. For clarity, only the two-dimensional view is shown, and thus only two fan-outs by a QBS is given. Figure 6 shows the corresponding top view of the module.

3. *Optical Hardware Required for Torus Links*

In this subsection we discuss the functionality and the limits of two optical components used in the implementation of torus links.

Quadrant Beam Splitter. The function of the QBS is to split one beam into four beams. An optical arrangement of the QBS that uses graded-index (GRIN) lenses[25] is illustrated in Fig. 8(a). Four small GRIN lenses are placed on the end facet of the large GRIN lens. The large lens is used to collimate a beam from a single trunk fiber, and the aperture of the collimated beam is divided into four by the smaller lenses. The small lenses then focus the beams onto fibers. Beam combination or merging is performed, but in the opposite direction. Figure 8(b) illustrates the geometry of the QBS with GRIN lenses for the purpose of calculating power loss occurring at the connection between the large GRIN lens and the small GRIN lenses. Since four small GRIN lenses do not cover the entire end-facet area of the large GRIN lens, some portion of the beam aperture from the large GRIN lens cannot be captured by four small GRIN lenses, resulting in power loss. Suppose that the radius of a small lens is r. The smallest possible radius of the large lens that can cover four small lenses is then $r + \sqrt{2}r$. Thus $4\pi r^2/[\pi(1 + \sqrt{2})^2 r^2] = 68.6\%$ of the end-facet area of the large GRIN lens is covered by the four small lenses. Therefore approximately 31.4% of power is lost from the large GRIN lens to the four small lenses during the beam-splitting process.

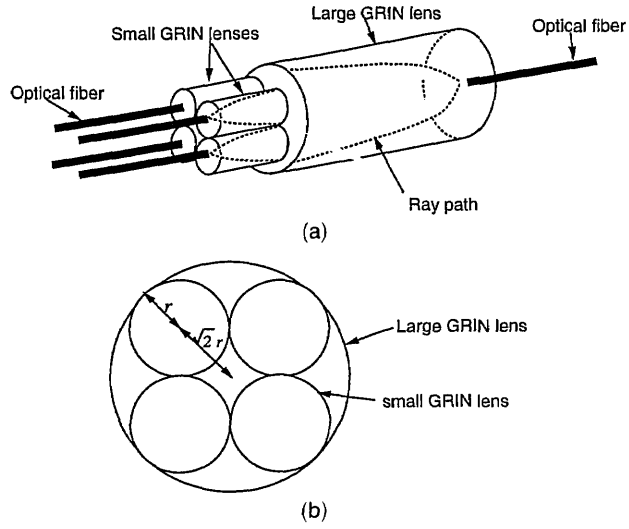

Fig. 8. (a) Quadrant beam splitter with GRIN lenses, (b) geometry of the quadrant beam splitter.

A more power-efficient (less than 20% power loss) QBS has been reported in Ref. 26 with substrate-mode holograms to reduce mechanical alignment and chromatic sensitivity. The QBS design with substrate-mode holograms is better than the design with GRIN lenses in terms of power efficiency, feature size, alignment, and fiber coupling efficiency. However, substrate-mode multiplexed holograms for the QBS's are not commercially available at this time.

Wavelength Multiplexer. A wavelength multiplexer and demultiplexer (WMUX) with a GRIN lens and a blazed grating is discussed in Ref. 27. WMUX's of this type permit more of the total bandwidth of the optical fiber to be used, and more than ten channels are currently available. Typical values of the insertion loss and the cross talk in available WMUX's are generally 1 to 2 dB and less than −30 dB, respectively. Since an (l, m, n) OMMH requires $2^{\lceil (n-1)/2 \rceil}$-channel WMUX's, with eight-channel WMUX's, it is possible to implement any size of OMMH network if $n \leq 7$.

4. *Rationale for the Two-Level Design Approach*

As discussed in Subsection 3.B.3, the optical implementation of the OMMH network consists of two levels: free-space space-invariant optics for the construction of basic building blocks and multiwavelength fibers for the torus links. The rationale for the two-level design approach is as follows: The use of space-invariant free-space optics would result in compact and simple building blocks that could be easily reproduced.[23,28] However, it would not be easy to implement scalable optical interconnects with totally space-invariant optics only, since a single space-invariant optical component such as a hologram is used to image multiple nodes for totally space-invariant interconnects. Thus it would be necessary to redesign the component in order to increase the number of nodes. However, since the minimum

incremental size of the OMMH is one hypercube module (a basic building block), the use of space-invariant optics within the basic building block would not limit the scalability of the OMMH. We use multiwavelength fiber optics to connect the basic building blocks because fiber optics would provide affordable scalable interconnects and the wavelength multiplexing technique would make a better utilization of the transmission capacity of an optical fiber.[29–32] The breakdown of functional requirements for the OMMH network is consistent with the advantages of free-space and optical fiber technologies.

5. *Evaluation of Optical Multimesh Hypercube Implementation*

To demonstrate the feasibility of the OMMH implementation, we performed an evaluation of the power efficiency and system volume with the proposed design based on practical component sizes. The design consists of ten-cube modules with 25.4-mm-diameter lenses, each with a focal length of 50.8 mm. A detector diameter of 50 μm was calculated to provide the connection density to produce a ten-cube module while permitting space-division techniques to be incorporated to avoid signal overlap.

With this system design the system volume of the hypercube module is 25.4 mm × 25.4 mm × 203.2 mm. The efficiency of the hypercube was calculated to be approximately 16%. This efficiency is due mostly to the phase hologram (which is theoretically 33% of maximum) and to the unwanted fan-out beams that result from the space-invariant nature of the design. The unwanted fan-out contributes an efficiency η_h, given by

$$\eta_h = \frac{n}{2n-1}, \tag{7}$$

where n is the hypercube dimension. For the ten-cube module this is 53%.

As for the OMMH torus subnetwork, a power analysis of the fiber-optic system was performed. Figure 9 shows a single unidirectional link of the mesh. For this link a −1-dB loss occurs from the insertion of the laser signal into the fiber. The QBS suffers a −0.97-dB loss, while each WMUX loses −1 dB of power (this calculation is for total system efficiency rather than per channel power flow, so the QBS loss does not include fan-out). Furthermore, the fiber is assumed to be at most 1 m in length, and a mean operating wavelength of 960 nm is assumed for the loss calculation. At this wavelength the fiber has an attenuation of −3.5 dB/km. Thus the fiber loss for the system is −0.0035 dB. Furthermore, the detector loss is −1 dB. Connection losses in the system were calculated based on reflection at the fiber interface. For the QBS each connection suffers a −0.45 dB loss. As for the WMUX's, an index-matching oil was assumed to be used between the fiber and the GRIN lens of the WMUX to ease the index transition. With this setup the connection losses around the WMUX's are −0.1 dB each. The

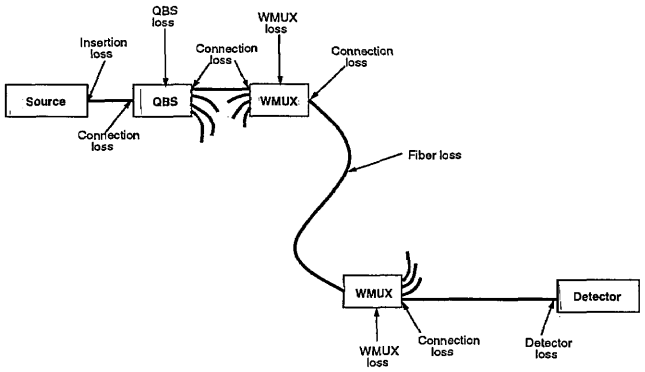

Fig. 9. Model for the optical power budget calculation of the torus subnetwork in the OMMH network.

final result is a total of −6.27 dB of system losses. This is equivalent to a system efficiency of 23.6%.

To gain a sense of the bandwidth of the proposed implementation, note that a leading-edge vertical-cavity surface-emitting laser source is capable of producing 6.5 Gbits/s with an output power of 2.2 mW.[33] When used with this system, a detector must have a sensitivity of 0.13 mW for the same bandwidth. This sensitivity can be reached with a detector having a bit error rate of 10^{-17}.[34]

4. Conclusions

Scalable networks and architectures are becoming more and more desirable for massively parallel computers since they can grow in size without major changes of the existing system configuration (size scalability), and they are also able to employ new evolving technologies (generation scalability). In fact, scalable network topologies are becoming the preferred choice for the computer industry despite their inherently limited topological characteristics such as low connectivity, large diameters, long average distances, and lack of fault tolerance. For example, many recent projects for the development of ultracomputers (Intel Paragon, Cray Research MPP Model, Caltech Mosaic C, MasPar MP-1, Kendall Square Research KSR-1, Stanford Dash Multiprocessor, Tera Computer Tera Multiprocessor, and Thinking Machine Corporation CM-5) are based on the scalable topologies such as the mesh/torus, ring, or tree topologies. Interconnection networks that are not only scalable but also possess good topological characteristics such as small diameter, high connectivity, constant node degree, simple routing scheme, and fault tolerance would greatly enhance the performance of massively parallel computers.

We have presented in this paper a new interconnection network called the optical multimesh hypercube (OMMH) for massively parallel computers. The distinctive features of the OMMH network are its scalability, both in size and generation, and its modularity while retaining positive features of both the hypercube (high connectivity, small diameter, simple message routing, and fault tolerance) and the mesh (constant node degree and scalability) topologies.

We have also proposed an optical interconnection architecture of the OMMH and its three-dimensional implementation. The proposed implementation is divided into two levels: space-invariant free-space optical interconnects for localized high-density hypercube modules and high-bandwidth multiwavelength optical fiber links for global low-density torus connections. This breakdown of functional requirements for the OMMH implementation is intended to exploit fully the advantages of free-space space-invariant optics (parallelism, simple and compact design, high connectivity, and cost efficiency) as well as wavelength-multiplexed fiber-based optics (full utilization of transmission bandwidth and scalability). In addition, the breakdown is intended to provide modularity and scalability both in size and generation. The two-level design methodology enables the construction of the OMMH network in a modular, incremental fashion (size scalability); the use of high-bandwidth wavelength-multiplexed optics in the OMMH can satisfy communication bandwidth requirements of current or near-future processing elements (generation scalability). We also have analyzed the proposed implementation. The implementation demonstrates good feasibility by showing a reasonable optical power efficiency and a volume size capable for inclusion within the case of a massively parallel computer.

This research was supported by National Science Foundation grant MIP 9310082 and a grant from USWest. The authors thank Matthew Derstine, the topical editor, and anonymous reviewers for their comments to improve the quality of this paper. They also express their gratitude to Michael Major, at the University of Arizona, for his help in analyzing the performance of the proposed implementation.

References

1. K. Hwang, *Advanced Computer Architecture: Parallelism, Scalability, Programmability* (McGraw-Hill, New York, 1993).
2. G. Bell, "Ultracomputers: a teraflop before its time," Commun. ACM **35,** 27–47 (1992).
3. H. S. Stone and J. Cocke, "Computer architecture in the 1990's," Computer **24**(9), 30–38 (1991).
4. L. N. Bhuyan and D. P. Agrawal, "Generalized hypercube and hyperbus structures constructing massively parallel computers," IEEE Trans. Comput. **C-33,** 323–333 (1984).
5. N.-F. Tzeng and S. Wei, "Enhanced hypercubes," IEEE Trans. Comput. **40,** 284–294 (1991).
6. J. R. Goodman and C. H. Sequin, "Hypertree: a multiprocessor interconnection topology," IEEE Trans. Comput. **C-30,** 923–933 (1981).
7. C. L. Seitz, "The cosmic cube," Commun. ACM **28,** 22–33 (1985).
8. D. Lenoski, J. Laudon, K. Gharachorloo, W. D. Weber, A. Gupta, J. Hennessy, M. Horowitz, and M. Lam, "The Stanford dash multiprocessor," Computer **25**(3), 63–79 (1992).
9. D. Nassimi and S. Sahni, "An optimal routing algorithm for mesh connected parallel computers," J. Assoc. Comput. Mach. **27**(1), 6–29 (1980).
10. A. Louri and H. Sung, "A design methodology for three-dimensional space-invariant hypercube networks using graph bipartitioning," Opt. Lett. **18,** 2050–2052 (1993).
11. L. P. Boivin, "Multiple imaging using various types of simple phase gratings," Appl. Opt. **11,** 1782–1792 (1972).
12. F. B. McCormick, "Generation of large spot arrays from a single laser beam by multiple imaging with binary phase gratings," Opt. Eng. **28,** 299–304 (1989).
13. K. H. Brenner and A. Huang, "Optical implementation of the perfect shuffle interconnection," Appl. Opt. **27,** 135–137 (1988).
14. A. S. Kumar and R. M. Vasu, "Multiple imaging and multichannel optical processing with split lenses," Appl. Opt. **26,** 5345–5349 (1987).
15. N. Streibl, U. Nolscher, J. Jahns, and S. Walker, "Array generation with lenslet arrays," Appl. Opt. **30,** 2739–2742 (1991).
16. Y. Sheng, "Space invariant multiple imaging for hypercube interconnections, " Appl. Opt. **29,** 1101–1105 (1990).
17. K. S. Urquhart, S. H. Lee, C. C. Guest, M. R. Feldman, and H. Farhoosh, "Computer aided design of computer generated holograms for electron beam fabrication," Appl. Opt. **28,** 3387–3396 (1989).
18. K.-S. Huang, A. A. Sawchuk, B. K. Jenkins, P. Chavel, J.-M. Wang, and A. G. Weber, "Digital optical cellular image processor (DOCIP): experimental implementation," Appl. Opt. **32,** 166–173 (1993).
19. M. W. Haney, "Self-similar grid patterns in free-space shuffle-exchange networks," Opt. Lett. **18,** 2047–2049 (1993).
20. T. J. Cloonan and M. J. Herron, "Optical implementation and performance of one-dimensional and two-dimensional trimmed inverse augmented data manipulator networks for multiprocessor computer systems," Opt. Eng. **28,** 305–314 (1989).
21. C. M. Habiger and R. M. Lea, "Hybrid-WSI: a massively parallel computing technology?" Computer **26**(4), 50–61 (1993).
22. 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).
23. A. Louri and H. Sung, "Efficient implementation methodology for three-dimensional space-invariant hypercube-based free-space optical interconnection networks," Appl. Opt. **32,** 7200–7209 (1993).
24. A. Louri, H. Sung, and Y. Moon, "An efficient channel encoding scheme for space-invariant optical interconnection networks," submitted to *Applied Optics.*
25. J. P. E. Green, *Fiber Optic Networks* (Prentice-Hall, Englewood Cliffs, N. J.; 1993).
26. M. Kato, Y.-T. Huang, and R. K. Kostuk, "Multiplexed substrate-mode holograms," J. Opt. Soc. Am. A **7,** 1441–1447 (1990).
27. B. D. Metcalf and J. F. Providakes, "High-capacity wavelength demultiplexer with a large-diameter graded-index rod lens," Appl. Opt. **21,** 794–796 (1982).
28. G. E. Lohman and K. H. Brenner, "Space-invariance in optical computing systems," Optik **89,** 123–134 (1992).
29. R. J. Vetter and D. H. C. Du, "Distributed computing with high-speed optical networks," Computer **26**(2), 8–18 (1993).
30. M. G. Hluchyj and M. J. Karol, "ShuffleNet: an application of generalized perfect shuffles to multihop lightwave networks," J. Lightwave Technol. **9,** 1386–1397 (1991).
31. T. S. Wailes and D. G. Meyer, "Multiple channel architecture: a new optical interconnection strategy for massively parallel computers," J. Lightwave Technol. **9,** 1702–1716 (1991).
32. G. R. Hill, "Wavelength domain optical network techniques," Proc. IEEE **77,** 121–132 (1989).
33. G. Shtengel, H. Temkin, P. Brunsenbach, T. Uchida, M. Kim, C. Parsons, W. E. Quinn, and S. E. Swirhun, "High-speed vertical-cavity surface emitting laser," IEEE Photon. Technol. Lett. **5,** *1359–1362 (1993).*
34. R. D. Smith and S. D. Personick, "Receiver design for optical fiber communication systems," in *Semiconductor Devices for Optical Communications,* H. Kressel, ed. (Springer-Verlag, Berlin, 1987), Chap. 4.