A Hypercube-based Optical Interconnection Network: A Solution to the Scalability Requirements for Massively Parallel Computers[1]

Ahmed Louri Hongki Sung

Department of Electrical and Computer Engineering
The University of Arizona, Tucson, AZ 85721, U.S.A

Abstract

An important issue in the design of interconnection networks for massively parallel computers is scalability. Size-scalability refers to the property that the number of nodes in the network can be increased with negligible effect on the existing configuration and generation-scalability implies that the communication capabilities of a network should be large enough to support the evolution of processing elements through generations. The lack of size-scalability has limited the use of certain types of interconnection networks (e.g., hypercube) in the area of massively parallel computing. This paper presents a new optical interconnection network, called an Optical Multi-Mesh Hypercube (OMMH), which is both size- and generation-scalable while combining positive features of both the hypercube (small diameter, high connectivity, symmetry, simple routing, and fault tolerance) and the mesh (constant node degree and scalability) networks. Also presented is a three-dimensional optical implementation methodology of the OMMH network.

1 Introduction

The quest for Teraflops (10^{12} floating point operations per second) supercomputers fueled by the launching of the High Performance Computing and Communication Program is putting major emphasis on exploiting massive parallelism with greater than one thousand processing elements (PEs) networked to form massively parallel computers(MPCs)[1, 2]. Several companies and universities have announced MPC projects. An important issue in the design of MPCs is scalability. The term *scalability* has several meanings. In the light of algorithms, a scalable computer should be able to perform well as the problem size increases, which is called problem scalability. In the light of architecture, scalability has two aspects; size and generation. A size-scalable computer is designed to have a scaling range from a small to large number of resource components. Linearly increased performance is expected as system size grows. As important as size-scalability is generation-scalability which is the adaptability of the architecture to the rapid evolution of technologies. Since microprocessors become obsolete every three years and the time to find efficient algorithms for a new system is long, a significant portion of investment on a new architecture should be preserved throughout generations.

The key to size scalability of MPCs is the interconnection network (IN) which is also a deciding factor in terms of performance and cost of the entire system[3]. A size-scalable IN has the property that the number of communicating nodes can be increased with minor or no change in the existing configuration. A generation-scalable IN can be implemented in a new technology, and interconnection bandwidth of the IN 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.

Numerous topologies have been explored for parallel computers[1, 4, 5, 6, 7]. However, the lack of size-scalability of some of these networks have limited their use in MPCs despite their many other advantages. For example, one of the most popular network for parallel computers is the binary n-cube topology, 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. For a binary n-cube network, the diameter is identical to the degree of a node $n = \log_2 N$. Each node is numbered in such a way that there is only one binary digit difference between any node and its $\log_2 N$ neighbors that are directly connected to it. This property greatly fa-

[1]This research was supported by an NSF grant No. MIP 9310082 and a grant from USWest.

0-8186-5830-4/94 $3.00

cilitates 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 is not used in the most recent MPC projects. 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 small number of links (four) per node. Due to 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 have limited connectivity and a large diameter, many recent MPC projects such as Intel Paragon[8], Cray Research MPP Model[9], Caltech Mosaic C[10], MasPar MP-1[11], Stanford Dash Multiprocessor[12], and Tera Computer Tera Multiprocessor[1], use this topology for the IN.

Motivated by these limitations, we have explored a new network topology, called *Optical Multi-Mesh Hypercube* (OMMH), which combines the advantages of both the hypercube (small diameter, high connectivity, symmetry, simple control and routing, fault tolerance, etc.) and the mesh (constant node degree and scalability) topologies, while circumventing their disadvantages (lack of scalability of the hypercube, and large diameter of the mesh). We have also developed a three-dimensional (3-D) optical implementation methodology which exploits the advantage of both space-invariant free-space and multiwavelength fiber-based optical interconnects technologies.

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

2 Structure of optical multi-mesh hypercube network

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

2.1 Topological definition of OMMH network

An (l, m, n)-OMMH, where l, m, and, n are integers, network consists of $l \times m \times 2^n$ nodes and an address of a node has three components; (i, j, k), where $0 \leq i < l$, $0 \leq j < m$, $0 \leq k < 2^n$, and i, j, k are integers. The topology of an (l, m, n)-OMMH network is defined by five interconnection functions for a node (i, j, k) as follows:

- $f_{m_1}(i, j, k) = ((i + 1) \bmod l,\ j,\ k)$
- $f_{m_2}(i, j, k) = ((l + i - 1) \bmod l,\ j,\ k)$
- $f_{m_3}(i, j, k) = (i,\ (j + 1) \bmod m,\ k)$
- $f_{m_4}(i, j, k) = (i,\ (m + j - 1) \bmod m,\ k)$
- $f_{c_d}(i, j, k_{n-1} \cdots k_{d+1} k_d k_{d-1} \cdots k_0) = (i, j, k_{n-1} \cdots k_{d+1} \bar{k}_d k_{d-1} \cdots k_0)$, for $d = 0, 1, \cdots, n-1$, where $k_{n-1} \cdots k_{d+1} k_d k_{d-1} \cdots k_0$ is a binary representation of integer k.

The first four interconnection functions, f_{m_1}, f_{m_2}, f_{m_3}, and f_{m_4}, define torus connections of the OMMH network. We refer to links generated by these four interconnection functions as *torus links*. The last interconnection function, f_{c_d}, for $d = 0, 1, \cdots, n-1$, determines the binary n-cube interconnection. We refer to links generated by this interconnection function as *hypercube links*.

2.2 OMMH structure characteristics

2.2.1 OMMH interconnection structure

Fig.1 shows a $(4, 4, 3)$-OMMH interconnection where solid lines represent hypercube links and dashed lines represent torus links. A $(4, 4, 3)$-OMMH consists of $4 \times 4 \times 2^3 = 128$ nodes. Small black circles represent nodes of the OMMH network which are, in this paper, abstractions of PEs which consist of electronic processing modules for computation and optical sources/detectors for communication. Both ends of torus links, dashed lines, are connected for wraparound connections of the torus if they have the

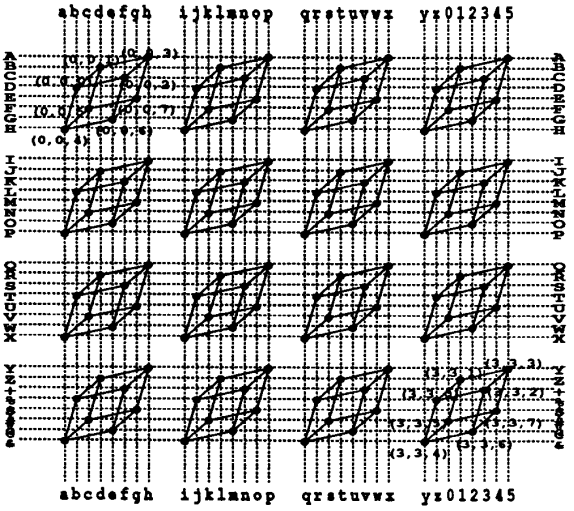

Figure 1: An example of the optical multi-mesh hypercube network: a $(4,4,3)$-OMMH (128 nodes) interconnection is shown. Two links with the same labels are connected for the wraparound connections of the torus. Only a few addresses are shown in the parenthesis for clarity. Solid lines represent hypercube connections and dashed lines torus connections.

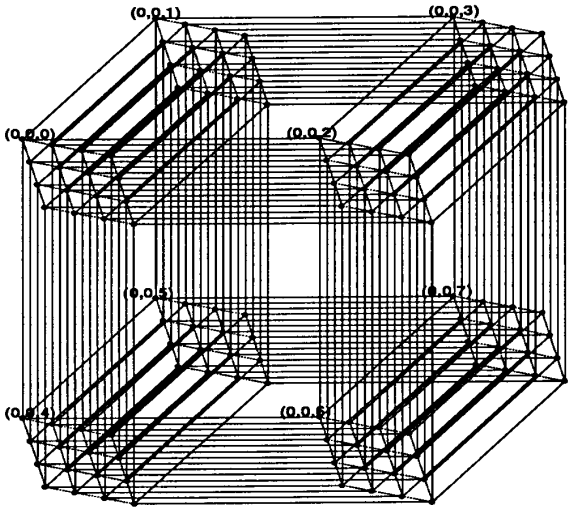

Figure 2: A $(4,4,3)$-OMMH interconnection network, another isomorphic view. Wraparound connections of the torus are omitted and only a few addresses are shown in the parenthesis for clarity. Solid lines represent hypercube connections and dashed lines torus connections.

same labels. The size of the OMMH can grow without altering the number of links per node by expanding the size of the torus; for example, by inserting 3-cubes in the row or column of the torus in Fig. 1. This feature allows the OMMH to be size-scalable. More discussion on the scalability issue will follow in subsection 2.3.2. An interesting isomorphic network is shown in Fig. 2. The same network is redrawn as a 4×4 torus-clustered 3-cube. It can be viewed as 8 concurrent toruses where 8 nodes having identical torus addresses form one 3-cube. It can also be viewed as 16 concurrent 3-cubes in which 16 nodes having identical hypercube addresses form a 4×4 torus.

2.2.2 Message routing in OMMH

Due to 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 or that of an $l \times m$ torus network or a combination of the two depending upon 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 regular 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 regular $l \times m$ torus network [13].

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

forward the message to T using the hypercube or the torus routing scheme, respectively.

The OMMH is less sensitive to performance degradation due to faults in links or nodes because the routing scheme in the OMMH has no preferred path, meaning all alternative paths have the same number of hops between any two nodes. This is an important advantage over other networks which have preferred paths such as Hypernet[7], Enhanced hypercube[5], or Extended hypercube[14].

2.2.3 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. 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 (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 (l, m, n)-OMMH is $(n + 4)$. N is equal to $(l \times m \times 2^n)$ if the hypercube and the OMMH have the same network size. A comparison of diameters should be accompanied by a comparison of link complexity, because a higher connectivity resulting from a higher link complexity is expected to lead to smaller diameters. Fig.3(a) compares the diameters of the hypercube and the OMMH, where $(16, 16, n)$-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. Fig.3(b) compares link complexities or node degrees of the hypercube and that of the OMMH. It should be noted that $(l, m, 4)$-OMMH has constant link complexity over the network size. This feature enables OMMH network to be scalable; that is, the growth of the network size does not affect the link complexity. Fig.3(c) depicts the growth of the total number of links in the network as the network size increases. For a network size of one million nodes, the hypercube network contains about 10.5 million links while the $(l, m, 4)$-OMMH has about 4.2 million links and $(16, 16, n)$-OMMH has approximately 8.4 million links.

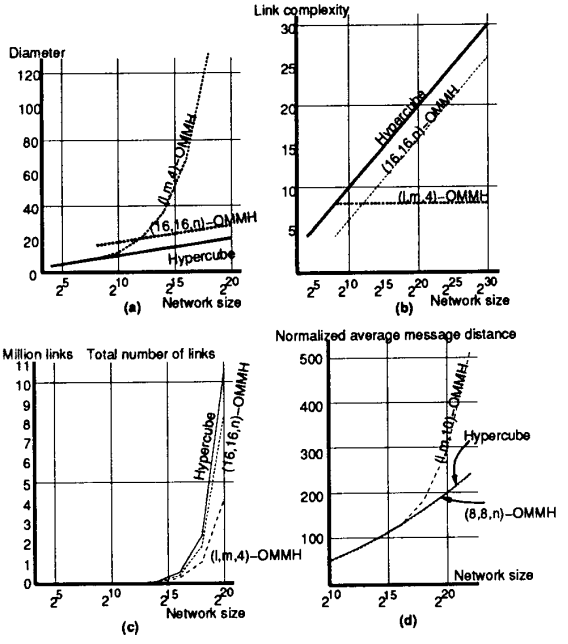

Figure 3: Comparison of (a) diameter, (b) link complexity, (c) total number of links, and (d) normalized average message distance of the hypercube and the OMMH when the two networks have the same number of nodes.

Since one link implies one physical path, electrical or optical, between two nodes, the OMMH network is cost-efficient compared to the regular hypercube network in terms of hardware requirement.

2.3 OMMH network properties

2.3.1 Communication efficiency

It seems reasonable to assume that an efficient and realistic multicomputer system will show much heavier traffic over short distances than over long communication paths since tasks which can be partitioned into smaller subtasks would usually be assigned to neighboring processors. To characterize the locality of messages in multicomputer systems, the *Geometric Distribution Model* have been suggested and used to show performance of computer networks[7, 6, 15]. The geometric distribution model is defined as follows. For every source S, the nodes of the network are divided into regions $R_1, R_2, \cdots$ of increasing distance from S. A fraction β of all messages is destined for region R_1 of S, β of the remaining messages go to region R_2, and so on. Within each region, the distribution is uniform.

Fig.4 shows the normalized average message distance using the geometric distribution model where each region is 4-hop wide. Normalized average distance is defined to be the average message distance multiplied by the number of links at the node[6]. We

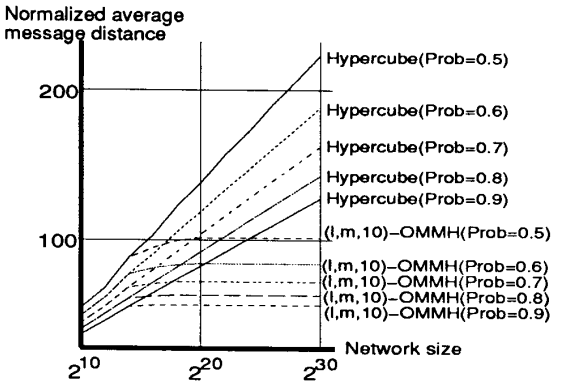

Figure 4: Normalized average message distance using geometric distribution model with 4-link wide region. Probability within each region is 0.5, 0.6, 0.7, 0.8, or 0.9.

compare normalized average message distances of the hypercube and the (l, m, n)-OMMH when the two networks have the same number of nodes. With N nodes as the network size, the dimension of the hypercube is $\log_2 N$ and $l \times m \times 2^n$ nodes in the OMMH must be equal to N. The size of the torus in the OMMH is chosen as square as possible. Fig.4 reveals that the increase of the normalized average message distance of the OMMH with constant cube with respect to the growth of the network size is negligible (constant in the graph) while that of the hypercube grows logarithmically with respect to the network size. This implies that the OMMH can be scaled up with little increase of the normalized average message distance when the message destination distribution can be predicted by the geometric distribution model.

2.3.2 Size-scalability

As can be seen in Fig.3(b), the OMMH with a constant cube as a basic building block has a constant node degree, which means that the size of the OMMH is ready to be scaled up by expanding the size of the torus without affecting the link complexity (number of links per node) of existing nodes as is the case in expanding 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 $l \times 2^n$ nodes (if $l < m$). In addition, in Fig.4, the normalized average distance of the OMMH under geometric message distribution remains constant as the network size grows. This implies that the OMMH can be scaled up without increasing the normalized average distance. On the contrary, the regular hypercube can only be scaled up with logarithmic increase in the normalized average distance.

3 Scalable Optical Design of OMMH Network

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 and long connections for the torus links connecting the basic building blocks. The optical implementation also 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 can be easily reproduced[16]. 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 will 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 division multiplexing technique would make a better utilization of the transmission capacity of an optical fiber[17, 18, 19]. The breakdown of functional requirements for the OMMH network is consistent with the advantages of free-space and optical fiber technologies. The size of the OMMH can be increased by adding hypercube modules, which provides modularity and size-scalability. Generation-scalability is provided by the use of high-bandwidth wavelength multiplexed optics which would match communication bandwidth re-

quirements of future processing elements.

In the following, we first summarize a design methodology to implement optical space-invariant hypercube networks (for more details see Ref. [20]) and then, propose an optical implementation of hypercubes as basic building blocks using binary phase gratings. Finally we show how to connect these building blocks with multiwavelength optical fibers for the construction of OMMH networks.

3.1 Design methodology for optical space-invariant hypercubes

A model of a 3-D free-space optical interconnection network architecture is depicted in Fig. 5. It is assumed in the model that processing elements (PEs) are partitioned into two sets of equal size and there is no inter-PE communication link between any two PEs on the same set. Instead, free-space optics provides all inter-PE communication links. This eliminates all electrical links between PEs on the same plane, resulting in much denser realization of arrays of PEs on the plane. Each plane contains PEs and optical tranceivers (sources and detectors) that could be located throughout the entire plane and allow circuit designers more of a three-dimensional (3-D) layout flexibility rather than being limited to the periphery of the plane as with electrical interconnections.

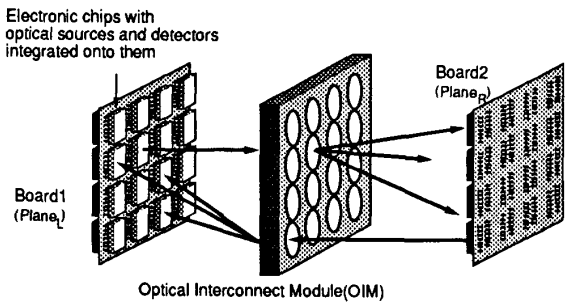

Figure 5: A model for 3-D free-space optical interconnect architectures

The construction of an arbitrary n-cube network is based on the space-invariant $(n-1)$-cube network. Fig. 6 presents the 3-D optical implementation of n-cube networks, for $n = 2, \cdots, 5$. These implementations represent basic modules to be used for the implementation of larger network sizes. To facilitate the description of the generalized embedding algorithm, we define a *connection rule* to be the amount of row-wise (upward or downward) or column-wise (left or right) spatial shifts required to achieve hypercube connections. Thus, each connection rule for an n-cube contains two entities, $Row(n)$ and $Col(n)$. For example, a shift rule such as $Row(3) = 0, \pm 1$, $Col(3) = \pm 1$ states that in order to implement a 3-cube network, each plane is to be replicated into 5 images. For a $Row(3) = 0$, the corresponding plane is imaged into the opposite plane straight without any row-wise or column-wise shift. The remaining 4 replicas are then shifted as follows. For a $Row(3) = +1$, the corresponding image is shifted upwards by one row. Similarly, a $Row(3) = -1$ indicates a shift by one row downwards. A $Col(3) = +1$ means a shift to the right by one column, and $Col(3) = -1$ is a shift to the left by one column. The shifted images need to be simultaneously superposed on the opposite plane to achieve the required connections.

We summarize symbols and their meanings to be used in the generalized algorithm.

- $Plane_L$ (or $Plane_R$): A plane on which one of the two partitions of nodes is placed.
- $\mathcal{E}_r(n)$ (or $\mathcal{E}_c(n)$): the number of empty rows (or columns) that are inserted between two layouts of $(n-1)$-cubes on the same plane to construct an n-cube.
- $\mathcal{D}_r(n)$ (or $\mathcal{D}_c(n)$): the number of rows (or columns) of the resulting 3-D n-cube on each plane.
- $\mathcal{R}_r(n)$ (or $\mathcal{R}_c(n)$): the amount of upward rotation of an $(n-1)$-cube layout on each plane to construct an n-cube.
- $Row(n)$ (or $Col(n)$): the amount of row-wise (or column-wise) shifts for implementing an n-cube.

An Algorithm for Constructing a 3-D Space-invariant n-cube from an $(n-1)$-cube

The following three-step algorithm constructs a 3-D space-invariant n-cube ($n > 5$) from a 3-D space-invariant $(n-1)$-cube network.

Step one: Given $Plane_L$ and $Plane_R$ of a space-invariant $(n-1)$-cube, we rotate each plane to the left by $2^{\frac{n-4}{2}}$ columns if n is even or $2^{\frac{n-5}{2}}$ rows upwards if n is odd.

Step two: The rotated plane is then placed at the right side of the original $(n-1)$-cube of the opposite plane if n is even, or underneath if n is odd. During the rotation, no empty columns or rows that already exist in the $(n-1)$-cube plane are counted as the shift amount. If n is even, we insert $\mathcal{E}_c(n) = 2^{\frac{n-6}{2}}$

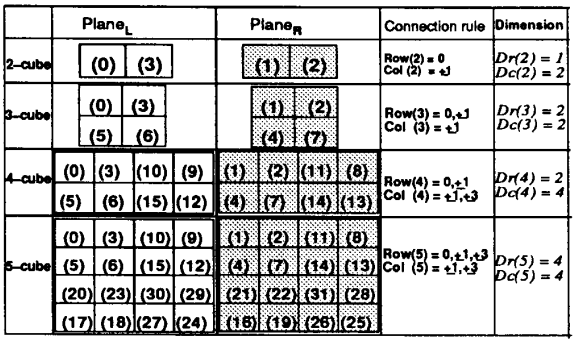

Figure 6: 3-D space-invariant hypercube networks of dimension n, where $2 \leq n \leq 5$.

$+ \sum_{i=1}^{\frac{n-6}{2}} \mathcal{E}_c(2i+4)$ empty columns between the two planes, the original and the rotated one. Or $\mathcal{E}_r(n) = 2^{\frac{n-7}{2}} + \sum_{i=1}^{\frac{n-7}{2}} \mathcal{E}_r(2i+5)$ empty rows if n is odd. Note that this insertion is done for $Plane_L$ of $(n-1)$-cube and the rotated version of $Plane_R$ of $(n-1)$-cube, and for $Plane_R$ of $(n-1)$-cube and the rotated version of $Plane_L$ of $(n-1)$-cube.

Step three: We prefix 0 as the most significant bit in all addresses of nodes on the resulting $Plane_L$, and 1 as the most significant bit in all addresses of nodes on the resulting $Plane_R$.

When n is even, $Plane_L$ and $Plane_L$ for the space-invariant n-cube have the same row dimensions as those of the $(n-1)$-cube and column dimensions are $2\times$ (column dimension of the $(n-1)$-cube) + (the number of empty columns inserted in step two). By row dimension and column dimension we mean the number of rows including empty rows and the number of columns including empty columns, respectively. Thus, $\mathcal{D}_r(n) = \mathcal{D}_r(n-1)$ and $\mathcal{D}_c(n) = 2 \times \mathcal{D}_c(n-1) + \mathcal{E}_c(n)$.

When n is odd, $Plane_L$ and $Plane_L$ for the space-invariant n-cube have row dimensions that are equal to $2\times$ (row dimension of the $(n-1)$-cube) + (the number of empty rows inserted in step 2), and the same column dimensions as those of the $(n-1)$-cube. Thus, $\mathcal{D}_r(n) = 2 \times \mathcal{D}_r(n-1) + \mathcal{E}_r(n)$ and $\mathcal{D}_c(n) = \mathcal{D}_c(n-1)$.

If n is even, the connection rule of the resulting n-cube is: $Row(n) = Row(n-1)$ and $Col(n) = Col(n-1)$, $\pm[\mathcal{D}_c(n) - \mathcal{D}_c(n-3)]$. If n is odd, the connection rule of the resulting n-cube is: $Row(n) = Row(n-1)$, $\pm[\mathcal{D}_r(n) - \mathcal{D}_r(n-3)]$, and $Col(n) = Col(n-1)$. Since we assume bidirectional communication between two planes, the connection rule applies to both planes.

3.2 Optical implementation of space-invariant hypercubes using binary phase gratings

In this section, we discuss an optical implementation of 3-D space-invariant hypercube networks using binary phase gratings (BPGs). Fig. 7 describes a hardware arrangement of optical components which could implement a space-invariant 5-cube network For

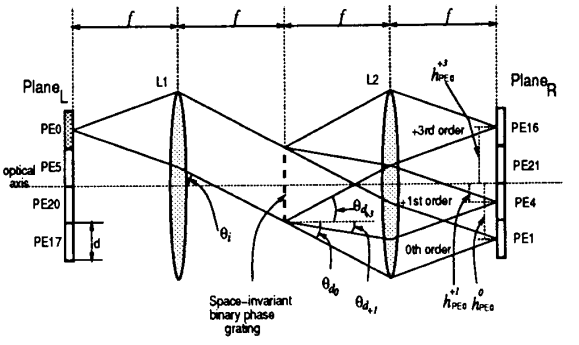

Figure 7: Space-invariant optical implementation of a 5-cube network using a binary phase grating.

clarity, only a 2-D view is shown. A BPG is added at the pupil plane between two imaging lenses to provide necessary beam steering operations. This type of arrangement was first proposed in Ref. [21]. 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 relationship between 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 m-th order's diffracted beam(θ_{d_m}).

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

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

$$h^m_{PE0} = f \times \tan\theta_{d_m} \qquad (2)$$

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

$$h^m_{PE0} = f \times \tan\{\sin^{-1}[\frac{m\lambda}{p} + \sin(\tan^{-1}(\frac{1.5d}{f}))]\} \qquad (3)$$

We assume that the structure of the grating is designed such that the power of the incident beam is equally distributed into 0th-order, ±1st order, and ±3rd order of diffracted beams and others are suppressed. We can have different amounts of optical power from the original beam to be 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[22]. Since PE0 is supposed to be connected with PE1, PE4, and PE16 for the 5-cube network, the following conditions should be satisfied.

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

Note that conditions for h^{-1}_{PE0} and h^{-3}_{PE0} make −1st and −3rd order diffracted beams fall outside $Plane_R$ to avoid unwanted connections.

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

$$h^{m}_{PE5} = f \times \tan\{\sin^{-1}[\frac{m\lambda}{p} + \sin(\tan^{-1}(\frac{0.5d}{f}))]\} \quad (5)$$

To make connections from PE5 to PE1, PE4, PE21, the following set of conditions should hold:

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

Note that conditions for h^{+3}_{PE5} and h^{-3}_{PE5} make +3rd and −3rd diffracted beams fall outside $Plane_R$. Since PE0 and PE5 are symmetrically placed, with respect to the optical axis, with PE17 and PE20, we can determine the period of grating(p) to provide the required connections for the 5-cube network by solving Eqs. 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 which optimizes conditions in Eqs. 4 and 6. By optimization we mean minimization of errors in each condition. For example, given that the node size in one dimension is $5mm$, the wavelength of the light source $785nm$, and the focal length of the lens $175mm$, the optimum period of the grating is computed to be $27.5\mu m$ which causes maximum misalignment of $4.0\mu m$ at PE21 from the PE5 connection.

The size of a basic n-cube module that can be implemented is primarily determined by the number of fanouts that can be managed by the BPG since an n-cube implementation requires $2n-1$ fanouts. The BPG must be able to generate $2n-1$ beams of equal power.

Power Loss: The proposed hardware setup suffers approximately 44% power loss (4 out of 9 diffracted beams from each source fall outside the receiving plane as discussed in Sec. 3.1) which is not because of the BPG but because of the totally space-invariant implementation methodology. Power efficiency is traded for low design complexity and for better use of SBWP of optical components required. We are currently investigating other optical means such as lenslet arrays for implementing the basic modules which are power efficient and more robust.

3.3 OMMH construction: Design of torus links to connect hypercube modules

3.3.1 Design methodology for torus links with fiber optics

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

(1) $l \times m$ n-cube modules as described in Sec. 3.2 are placed in a $l \times m$ 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 dimension $l \times m$.

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

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 division multiplexing (WDM) technique. However, a straightforward use of the WDM also requires a prohibitively large number of different wavelengths. For example, to connect two ten-cube modules, we need $2^{10} = 1024$ different wavelengths. In this paper, we use a wavelength-node assignment technique which alleviates this problem.

Referring to Sec. 3.1, an n-cube layout ($Plane_L$ or $Plane_R$) consists of $2^{\lfloor (n-1)/2 \rfloor}$ non-empty rows

and $2^{\lceil (n-1)/2 \rceil}$ non-empty columns. For $Plane_L$ and $Plane_R$, we assign the following wavelengths to the nodes in the first row; λ_1, λ_2, $\cdots$, $\lambda_{2^{\lceil (n-1)/2 \rceil}}$. Then, we assign λ_2, $\cdots$, $\lambda_{2^{\lceil (n-1)/2 \rceil}}$, λ_1, as wavelengths to the nodes in the second row. In general, wavelength-assignment in a row is achieved by rotating the wavelength-assignment of previous row by one column. This wavelength assignment results in no two nodes in the same row or column having an identical wavelength. 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 (l, m, n)-OMMH using 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 will be 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 beam into four beams (the QBS also combines four beams into one since it is bi-directional) and an i-channel wavelength multiplexor (WMUX) which multiplexes beams with i different wavelengths into a single beam (also demultiplexes since it is bi-directional). The realization of these two components with current technology will be discussed in detail in the following subsection. We also 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. A QBS is attached to every S_t to provide the four fanouts, 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 lights 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 are multiplexed by WMUX_N, WMUX_S, and WMUX_W, respectively. Figure 8 illustrates a five-cube module with torus link interface. For clarity, only the 2-D view is shown and, thus, only two fanouts by a QBS is given. Figure 9 shows a full size representation of the $(5, 4, 5)$-OMMH implementation emphasizing the torus links. For clarity, only links among $Plane_L$s are depicted.

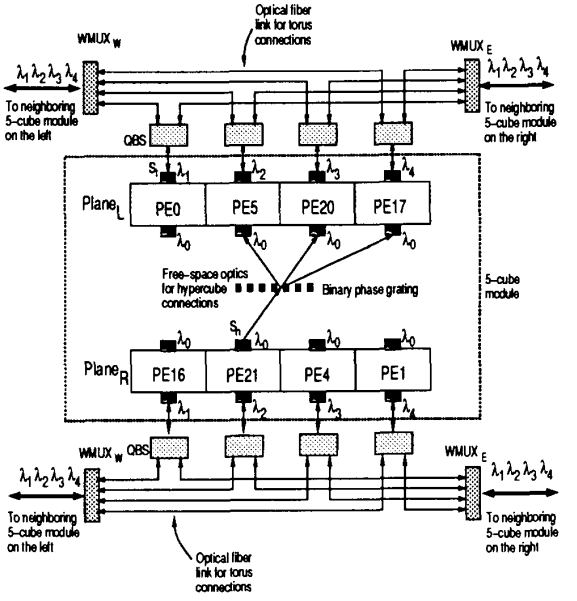

Figure 8: A 2-D view of a 5-cube module to interface with torus links for the construction of the $(l, m, 5)$-OMMH network.

3.3.2 Optical hardware required for torus links

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

Quadrant Beam Splitter(QBS):

The function of the QBS is to either split one beam into four beams or combine four beams into a single beam. An optical arrangement of the QBS using graded index (GRIN) lenses[23] is illustrated in Fig. 10.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 10.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 small GRIN lenses. Since four small GRIN lenses do not cover the entire end-facet area of the large GRIN lens, some portion of beam aperture from the large GRIN

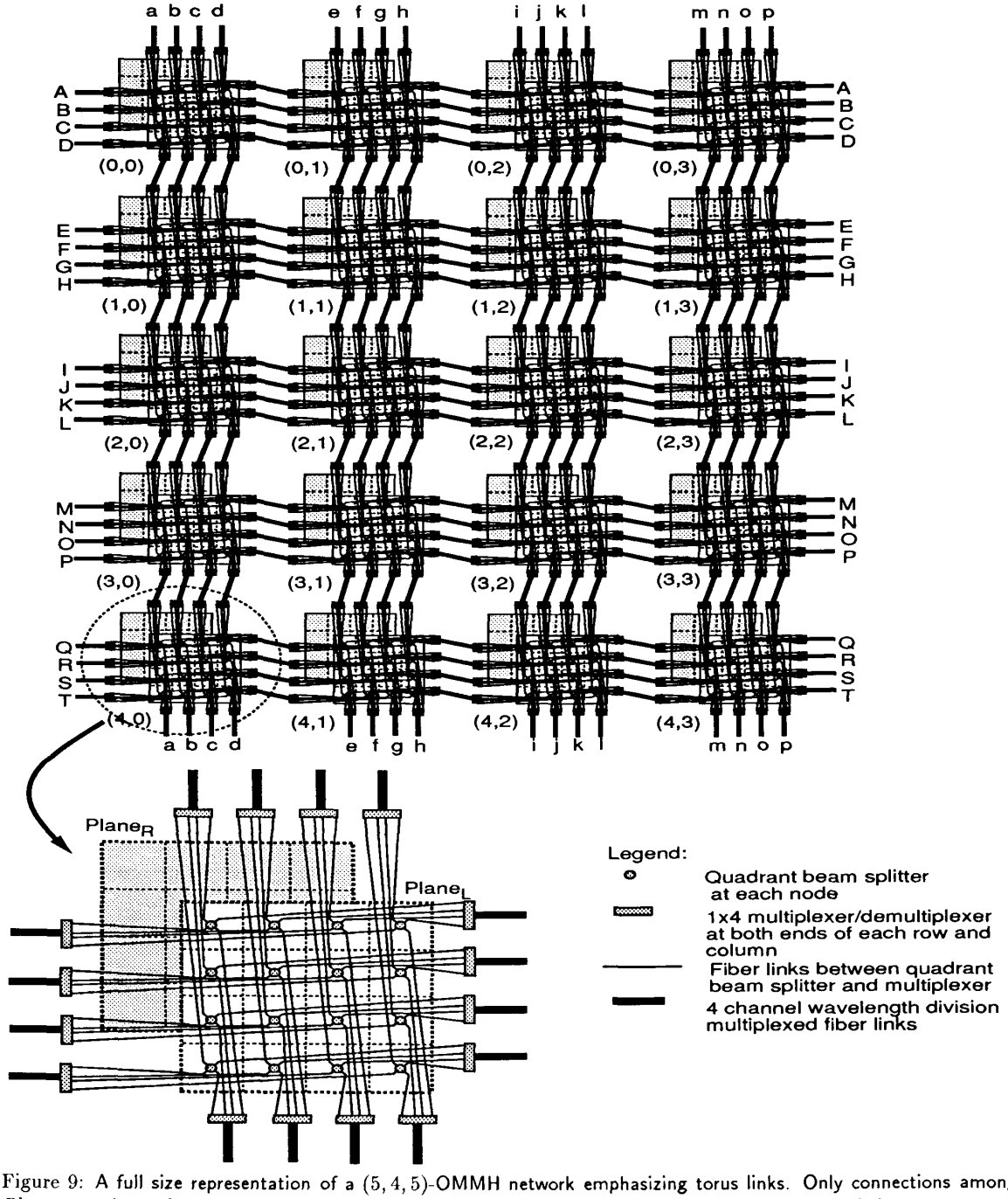

Figure 9: A full size representation of a $(5, 4, 5)$-OMMH network emphasizing torus links. Only connections among $Plane_L$ are shown for clarity. Similar connections among $Plane_R$ exist. Two links with the same labels are connected for wraparound connections. In the parenthesis, first two address components of $(5, 4, 5)$-OMMH (i.e., torus address components) are shown.

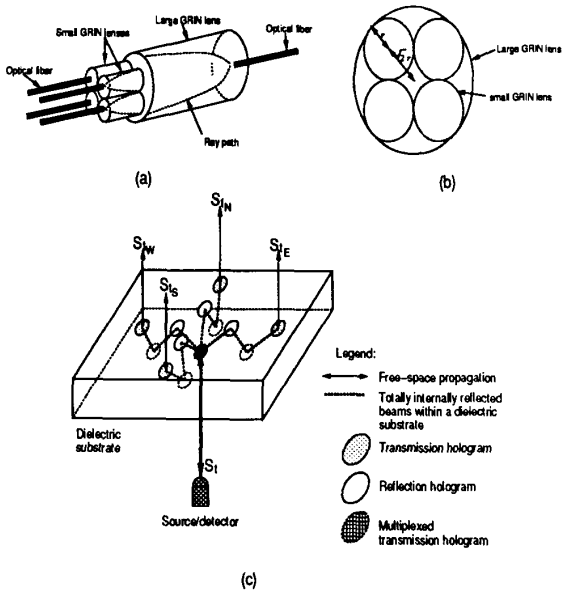

Figure 10: (a) A quadrant beam splitter using GRIN lenses. (b) Geometry of the quadrant beam splitter. (c) A quadrant beam splitter using substrate mode holograms [24]

lens cannot be captured by four small GRIN lenses, resulting in power loss. Suppose that a 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 beam splitting process. However, negligible power is lost during beam combination process since the end-facets of the four small lenses are completely covered by the large GRIN lens.

A more power efficient (less than 20% power loss) QBS has been reported in Ref. [24]. It uses substrate mode holograms to reduce mechanical alignment and chromatic sensitivity. Figure 10.c illustrates the functionality of a QBS using multiplexed substrate-mode volume holograms. Incident beam S_t is beam split by the multiplexed transmission hologram consisting of four superimposed gratings. The gratings diffract the beam into four orthogonal directions. The divided beams propagate through the waveguide by total internal reflection provided by the reflection holograms. Finally, four beams are coupled out of the substrate guide through transmission holograms shown as $S_{t_N}, S_{t_S}, S_{t_E}$, and S_{t_W}.

The QBS design with substrate mode holograms is better than the design with GRIN lenses in terms of power efficiency, alignment, and fiber coupling efficiency. However, substrate mode multiplexed holograms for the QBSs are not commercially available at this time.

Wavelength Multiplexor(WMUX):

An i-channel wavelength multiplexor(or demultiplexor) which separates (or combines) i channel wavelength division multiplexed beams using a GRIN lens and a blazed grating was demonstrated in Ref. [25]. Demultiplexing is performed as follows; the i channel wavelength multiplexed beam enters the GRIN lens. The beam is then reflected and divided into i different angles, depending on the wavelength, from the reflection grating at the other end of the GRIN lens. The i separated beams are then coupled into the appropriate fibers. Multiplexing is performed in a similar fashion but in the reverse way. WMUXs of this type allow 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 crosstalk in available WMUXs are generally $1 \sim 2$ dB and less than -30 dB, respectively. Since (l, m, n)-OMMH requires $2^{\lceil (n-1)/2 \rceil}$-channel WMUXs, with 16-channel WMUXs, it is possible to implement any size of OMMH networks if $n \leq 9$.

4 Conclusions

Size-scalable network topologies such as mesh/torus, ring, and tree, are becoming the preferred choice for the computer industry in the design of massively parallel computers 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, etc) are based on the such topologies. Interconnection networks which 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 Multi-Mesh Hyper-

cube (OMMH), for massively parallel computers. The distinctive features of the OMMH network are its scalability, both in size and generation, and 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 a three-dimensional optical implementation method of the OMMH. 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 fully exploit 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 and incremental fashion (size-scalability) and the use of high bandwidth wavelength multiplexed optics in the OMMH can satisfy communication bandwidth requirements of the current or near future processing elements (generation-scalability). We have also discussed functionality and limitations of possible optical hardware which implements the OMMH network.

References

[1] K. Hwang, *Advanced Computer Architecture: Parallelism, Scalability, Programmability.* New York, NY: McGraw-Hill, 1993.

[2] G. Bell, "Ultracomputers: A Teraflop Before Its Time," *Communications of the ACM*, vol. 35, pp. 27–47, Aug 1992.

[3] H. S. Stone and J. Cocke, "Computer Architecture in the 1990s," *Computer*, pp. 30–38, Sep. 1991.

[4] L. N. Bhuyan and D. P. Agrawal, "Generalized Hypercube and Hyperbus Structures Constructing Massively Parallel Computers," *IEEE Trans. Comput.*, vol. C-33, pp. 323–333, 1984.

[5] N.-F. Tzeng and S. Wei, "Enhanced Hypercubes," *IEEE Trans. Comput.*, vol. 40, pp. 284–294, Mar 1991.

[6] J. R. Goodman and C. H. Sequin, "Hypertree: a Multiprocessor Interconnection Topology," *IEEE Trans. Comput.*, vol. C-30, pp. 923–933, 1981.

[7] K. Hwang and J. Ghosh, "Hypernet: a Communication-Efficient Architecture for Constructing Massively Parallel Computers," *IEEE Trans. Comput.*, vol. C-36, pp. 1450–1466, 1987.

[8] Supercomputer Systems Division, Intel Corporation, Beaverton, OR, *Paragon XP/S Product Overview*, 1991.

[9] Cray Research Inc., Eagan, MN, *Cray/MPP Announcement*, 1992.

[10] C. L. Seitz, "Mosaic C: An Experimental Fine-Grain Multicomputer," tech. rep., California Institute of Technology, Pasadena, CA, 1992.

[11] "The MasPar Family Data-Parallel Computer." Technical summary, 1991.

[12] D. Lenoski, J. Laudon, K. Gharachorloo, W. D. Weber, A. Gupta, J. Hennessy, M. Horowitz, and M. Lam, "The Stanford Dash Multiprocessor," *IEEE Computer*, pp. 63–79, Mar 1992.

[13] D. Nassimi and S. Sahni, "An Optimal Routing Algorithm for Mesh Connected Parallel Computers," *Journal of the ACM*, vol. 27, pp. 6–29, January 1980.

[14] J. M. Kumar and L. M. Patnaik, "Extended Hypercube: A Hierarchical Interconnection Network of Hypercubes," *IEEE Trans. Parallel and Distributed Systems*, vol. 3, pp. 45–57, Jan. 1992.

[15] D. A. Reed and H. D. Schwetman, "Cost Performance Bounds for Multi-microcomputer Networks," *IEEE Trans. Comput.*, vol. C-32, pp. 83–95, 1983.

[16] G. E. Lohman and K. H. Brenner, "Space-invariance in Optical Computing Systems," *Optik*, vol. 89, pp. 123–134, 1992.

[17] M. G. Hluchyj and M. J. Karol, "ShuffleNet: An Application of Generalized Perfect Shuffles to Multihop Lightwave Networks," *IEEE Journal of Lightwave Technology*, vol. 9, pp. 1386–1397, Oct 1991.

[18] T. S. Wailes and D. G. Meyer, "Multiple Channel Architecture: A New Optical Interconnection Strategy for Massively Parallel Computers," *IEEE Journal of Lightwave Technology*, vol. 9, pp. 1702–1716, Dec 1991.

[19] G. R. Hill, "Wavelength Domain Optical Network Techniques," *Proceedings of the IEEE*, vol. 77, pp. 121–132, Jan 1989.

[20] A. Louri and H. Sung, "A Design Methodology for three-dimensional space-invariant hypercube networks using graph bipartitioning," *Optics Letters*, vol. 18, pp. 2050–2052, Dec 1993.

[21] 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," *Optical Engineering*, vol. 28, pp. 305–314, 1989.

[22] 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," *Applied Optics*, vol. 28, pp. 3387–3396, 1989.

[23] J. P. E. Green, *Fiber Optic Networks*. Englewood Cliffs, NJ: Prentice Hall, 1993.

[24] M. Kato, Y.-T. Huang, and R. K. Kostuk, "Multiplexed Substrate-mode Holograms," *J. Opt. Soc. Am. A*, vol. 7, pp. 1441–1447, 1990.

[25] B. D. Metcalf and J. F. Providakes, "High-capacity Wavelength Demultiplexer with a Large-diameter GRIN Rod Lens," *Applied Optics*, vol. 21, pp. 794–796, 1982.