PAPER

Modified Hierarchical 3D-Torus Network

M.M. Hafizur RAHMAN[†a)], *Student Member*, Yasushi INOGUCHI[††], *and* Susumu HORIGUCHI[†††], *Members*

SUMMARY Three-dimensional (3D) wafer stacked implementation (WSI) has been proposed as a promising technology for massively parallel computers. A hierarchical 3D-torus (H3DT) network, which is a 3D-torus network of multiple basic modules in which the basic modules are 3D-mesh networks, has been proposed for efficient 3D-WSI. However, the restricted use of physical links between basic modules in the higher level networks reduces the dynamic communication performance of this network. A torus network has better dynamic communication performance than a mesh network. Therefore, we have modified the H3DT network by replacing the 3D-mesh modules by 3D-tori, calling it a Modified H3DT (MH3DT) network. This paper addresses the architectural details of the MH3DT network and explores aspects such as degree, diameter, cost, average distance, arc connectivity, bisection width, and wiring complexity. We also present a deadlock-free routing algorithm for the MH3DT network using two virtual channels and evaluate the network's dynamic communication performance under the uniform traffic pattern, using the proposed routing algorithm. It is shown that the MH3DT network possesses several attractive features including small diameter, small cost, small average distance, better bisection width, and better dynamic communication performance.

key words: *MH3DT network, static network performance, wormhole routing, deadlock-free routing, dynamic communication performance*

1. Introduction

Interconnection networks are the key elements for building massively parallel computers [1]. In massively parallel computers with millions of nodes, the large diameter of conventional topologies is intolerable. Hierarchical interconnection networks [2] are a cost-effective way to interconnect a large number of nodes. A variety of hyper-cube based hierarchical interconnection networks have been proposed [3]–[8], but for large-scale multicomputer systems, the number of physical links becomes prohibitively large. To alleviate this problem, several k-ary n-cube-based hierarchical interconnection networks: TESH [9], [10], H3D-torus [11]–[13], and Cube Connected Cycles (CCC) [14], have been presented. However, the dynamic communication performance of these networks is still very low, especially in terms of network throughput.

An H3DT network [12], [13] has been put forward as a new interconnection network for large-scale 3D multicomputers. The H3DT network consists of multiple basic modules (BM) which are 3D-mesh of size $(m \times m \times m)$. The BMs are hierarchically interconnected by a 3D-torus of size $(n \times n \times n)$ to build higher level networks. The restricted use of physical links between basic modules in the higher level networks reduces the dynamic communication performance of this network. Even when the inter-BM links are increased, the network throughput of the H3DT network is still lower than that of the conventional mesh network. It has already been shown that a torus network has better dynamic communication performance than a mesh network [1]. This is the key motivation that led us to replace the 3D-mesh network by a 3D-torus network.

The modified hierarchical 3D-torus (MH3DT) network consists of BMs which are themselves 3D-tori $(m \times m \times m)$, hierarchically interconnected in a 3D-torus $(n \times n \times n)$ networks. In the MH3DT network, both the BMs and the interconnection of higher levels have toroidal interconnections.

In massively parallel computers, an ensemble of nodes works in concert to solve large application problems. The nodes communicate data and coordinate their efforts by sending and receiving messages in the multicomputer through a router, using a routing algorithm. The routing algorithm specifies how a message selects its network path to move from source to destination. Efficient routing is critical to the performance of interconnection networks. In a practical router design, the routing decision process must be as fast as possible to reduce network latency.

Wormhole routing [15], [16] has become the dominant switching technique used in contemporary multicomputers. This is because it has low buffering requirements, and more importantly, it makes latency independent of the message distance. Since wormhole routing relies on a blocking mechanism for flow control, deadlock can occur because of cyclic dependencies over network resources during message routing. Virtual channels [17], [18] were originally introduced to solve the problem of deadlock in wormhole-routed networks.

Deterministic, dimension-order routing is popular in multicomputers because it has minimal hardware requirements and allows the design of simple and fast routers. Although there are numerous paths between any source and destination, dimension-order routing defines a single path from source to destination to avoid deadlock.

Neither the static network performance nor the dynamic communication performance of the MH3DT network have been evaluated yet. The first objective of this paper

Manuscript received May 26, 2004.
Manuscript revised September 30, 2004.

[†] The author is with the Graduate School of Information Science, JAIST, Ishikawa-ken, 923–1292 Japan.

[††] The author is with the Center for Information Science, JAIST, Ishikawa-ken, 923–1292 Japan.

[†††] The author is with the Graduate School of Information Science, Tohoku University, Sendai-shi, 980–8579 Japan.

a) E-mail: rahman@jaist.ac.jp

is to address the architectural details of the MH3DT network, and explore aspects such as the node degree, network diameter, cost, average distance, arc connectivity, bisection width and wiring complexity of the MH3DT network as well as its applicability to several conventional and hierarchical networks. The second objective of this paper is to propose a deadlock-free routing algorithm for the MH3DT network with a minimum number of virtual channels and evaluate the dynamic communication performance of the network.

The remainder of the paper is organized as follows. The basic structure of the MH3DT network and the routing algorithm are explained in Sect. 2 and Sect. 3, respectively. The proof that the proposed routing algorithm is free from deadlock is presented in Sect. 4. Both the static network performance and the dynamic communication performance of the MH3DT network are discussed in Sect. 5. Section 6 presents our conclusions.

2. Interconnection of the MH3DT Network

The MH3DT network consists of basic modules (BMs) that are hierarchically interconnected to form higher level networks. The BM of the MH3DT network is a 3D-torus network of size $(m \times m \times m)$, where m is a positive integer. The BM of $(4 \times 4 \times 4)$ torus, which is shown in Fig. 1, has some free ports at the corners of the xy-plane. These free ports are used for higher level interconnection. All ports of the interior Processing Elements (PEs) are used for intra-BM connections. All free ports of the exterior PEs are used for inter-BM connections to form higher level networks. In this paper, unless specified otherwise, BM refers to a Level-1 network.

Processing Elements (PEs) in the BM are addressed by three base-m digits, the first representing the x-direction, the second representing the y-direction, and the last representing the z-direction. The address of a PE in the BM is expressed by

$$A^{BM} = (a_z)(a_y)(a_x), \quad (0 \le a_z, a_y, a_x \le m-1) \tag{1}$$

All ports of the interior PEs are used for intra-BM connections. Three PEs $(0 \le a_z \le 2)$ have two free ports, as shown in Fig. 1, which are used for inter-BM connections to form higher level networks. Let $a_z = 0$ be the z-direction link, $a_z = 1$ be the y-direction link, and $a_z = 2$ be the x-direction link. We define a gate node as a PE that has free links to interconnect with PEs at the higher level. Thus each *gate* node has two links and is hierarchically interconnected with the PEs at the higher level by a 3D-torus network.

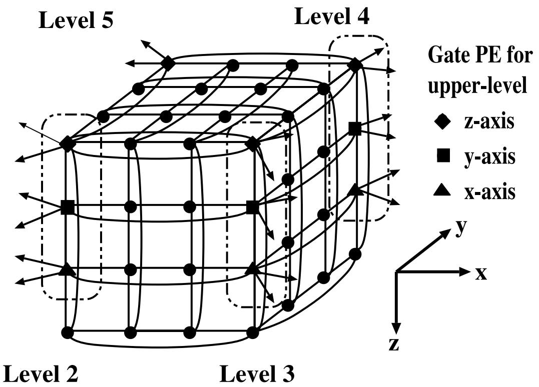


Fig. 1 Basic module.

Successively higher level networks are built by recursively interconnecting lower level subnetworks in a 3D-torus of size $(n \times n \times n)$, where n is also a positive integer. As illustrated in Fig. 2, a Level-2 MH3DT network, for example, can be formed by interconnecting 64 BMs as a $(4 \times 4 \times 4)$ 3D-torus network. Each BM is connected to its logically adjacent BMs. 2^q gate nodes are used higher level interconnection, where q is the inter-level connectivity. As each xy-plane of the BM has 4 gate nodes, $q \in \{0, 1, 2\}$. $q = 0$ leads to minimal inter-level connectivity, while $q = 2$ leads to maximum inter-level connectivity. By using the parameters m, n, L, and q, we can define the MH3DT network as MH3DT(m, n, L, q).

Let N be the total number of nodes in a MH3DT network. Then an MH3DT network with level L has $N = \left[m^3 \times n^{3(L-1)}\right]$. With $q = 0$, for example, Level-5 is the highest possible level to which a $(4 \times 4 \times 4)$ BM can be interconnected. The total number of nodes in a network having $(4\times4\times4)$ BMs and a $(4\times4\times4)$ Level-5 network is $N = 2^{30}$, i.e, more than one billion. A PE in the Level-L is addressed by three base-n numbers as follows:

$$A^L = \left(a_z^L\right)\left(a_y^L\right)\left(a_x^L\right), \quad \left(0 \le a_z^L, a_y^L, a_x^L \le n-1\right) \tag{2}$$

The address of a PE at Level-L is represented by:

$$A^L = \left(a_z^L\right)\left(a_y^L\right)\left(a_x^L\right) \qquad L \text{ is the level number.} \tag{3}$$

More generally, in a Level-L MH3DT network, the node address is represented by:

$$\begin{aligned} A &= A^L A^{L-1} A^{L-2} \;\ldots\ldots\; A^2 A^1 \\ &= a_\alpha\; a_{\alpha-1}\; a_{\alpha-2}\; a_{\alpha-3} \;\ldots\ldots\; a_3\; a_2\; a_1\; a_0 \\ &= a_{3L-1}\; a_{3L-2}\; a_{3L-3}\; a_{3L-4} \;\ldots\ldots\; a_3\; a_2\; a_1\; a_0 \end{aligned}$$

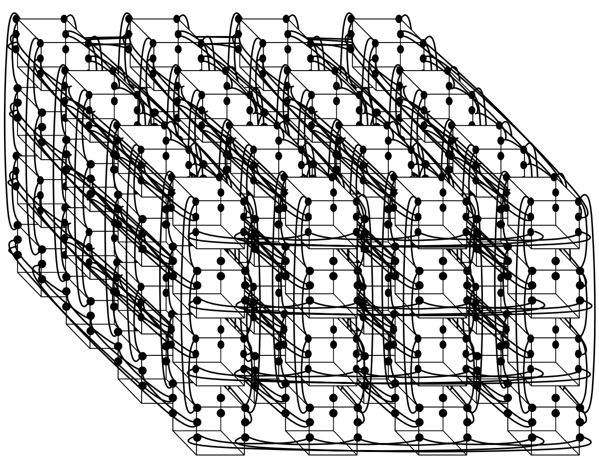


Fig. 2 Interconnection of a Level-2 MH3DT network.

$$= (a_{3L-1}\ a_{3L-2}\ a_{3L-3})\ \ldots\ \ldots (a_2\ a_1\ a_0) \qquad (4)$$

Here, the total number of digits is $\alpha = 3L$, where L is the level number. Groups of digits run from group number 1 for Level-1 (i.e. the BM), to group number L for the L-th level. In particular, i-th group $(a_{3i-1}\ a_{3i-2}\ a_{3i-3})$ indicates the location of a Level-$(i-1)$ subnetwork within the i-th group to which the node belongs; $2 \leq i \leq L$. In a two-level network, for example, the address becomes $A = (a_5\ a_4\ a_3)\ (a_2\ a_1\ a_0)$. The last group of digits $(a_5\ a_4\ a_3)$ identifies the BM to which the node belongs, and the first group of digits $(a_2\ a_1\ a_0)$ identifies the node within that basic module.

3. Routing Algorithm

Routing of messages in the MH3DT network is performed from top to bottom. That is, it is first done at the highest level network; then, after the packet reaches its highest level sub-destination, routing continues within the subnetwork to the next lower level sub-destination. This process is repeated until the packet arrives at its final destination. When a packet is generated at a source node, the node checks its destination. If the packet's destination is the current BM, the routing is performed within the BM only. If the packet is addressed to another BM, the source node sends the packet to the outlet node which connects the BM to the level at which the routing is performed. We have considered a simple deterministic, dimension-order routing algorithm. Routing of messages in the network is performed initially in the z-direction, next in the y-direction, and finally in the x-direction.

Routing in the MH3DT network is strictly defined by the source node address and the destination node address. Let a source node address be $s_\alpha, s_{\alpha-1}, s_{\alpha-2}, \ldots, s_1, s_0$, a destination node address be $d_\alpha, d_{\alpha-1}, d_{\alpha-2}, \ldots, d_1, d_0$, and a routing tag be $t_\alpha, t_{\alpha-1}, t_{\alpha-2}, \ldots, t_1, t_0$, where $t_i = d_i - s_i$. The source node address of the MH3DT network is expressed as $s = (s_{3L-1}, s_{3L-2}, s_{3L-3}), \ldots,\ \ldots, (s_2, s_1, s_0)$. Similarly, the destination node address is expressed as $d = (d_{3L-1}, d_{3L-2}, d_{3L-3}), \ldots,\ \ldots, (d_2, d_1, d_0)$. Figure 3 shows the routing algorithm for the MH3DT network.

As an example, a routing between $PE_{(123)(211)}$ and $PE_{(333)(111)}$ in the MH3DT network is given. First the packet is moved to the gate node in the z-axis at Level-2, $PE_{(123)(000)}$. Then the packet at $PE_{(123)(000)}$ is moved to the node with the same address in the z-axis, $PE_{(323)(000)}$. Next, routing in the y-axis is applied in the same manner. The packet at $PE_{(323)(000)}$ is moved to the gate node $PE_{(323)(100)}$ in the y-axis. The packet at $PE_{(323)(100)}$ is moved to $PE_{(333)(100)}$ with the same address in the y-axis. Finally, the routing is applied within the BM and terminated.

```
Routing MH3DT(s,d);
source node address:$s_\alpha, s_{\alpha-1}, s_{\alpha-2}, ..., s_1, s_0$
destination node address: $d_\alpha, d_{\alpha-1}, d_{\alpha-2}, ..., d_1, d_0$
tag: $t_\alpha, t_{\alpha-1}, t_{\alpha-2}, ..., t_1, t_0$
  for $i = \alpha : 3$
    if ($t_i > 0$ and $t_i \leq \frac{n}{2}$) or ($t_i < 0$ and $t_i = -(n-1)$), movedir = positive; endif;
    if ($t_i > 0$ and $t_i = (n-1)$) or ($t_i < 0$ and $t_i \geq -\frac{n}{2}$), movedir = negetive; endif;
    if (movedir = positive and $t_i > 0$), distance = $t_i$; endif;
    if (movedir = positive and $t_i < 0$), distance = $n + t_i$; endif;
    if (movedir = negative and $t_i < 0$), distance = $t_i$; endif;
    if (movedir = negative and $t_i > 0$), distance = $-n + t_i$; endif;
      $j = i$ mod 3
      while($t_i \neq 0$ or distance $\neq 0$) do
        if ($j = 2$), gate-node = $z$-axis gate-node of Level-$\lceil \frac{i}{3} \rceil$; endif
        if ($j = 1$), gate-node = $y$-axis gate-node of Level-$\lceil \frac{i}{3} \rceil$; endif
        if ($j = 0$), gate-node = $x$-axis gate-node of Level-$\frac{i}{3} + 1$; endif
        if (routedir = positive), move packet to next BM; endif;
        if (routedir = negative), move packet to previous BM; endif;
        if ($t_i > 0$), $t_i = t_i - 1$; endif;
        if ($t_i < 0$), $t_i = t_i + 1$; endif;
      endwhile;
  endfor;
BM_Routing ($t_2, t_1, t_0$);
BM_tag $t_2, t_1, t_0$ = receiving node address ($r_2, r_1, r_0$) – destination ($d_2, d_1, d_0$)
  for $i = 2 : 0$
    if ($t_i > 0$ and $t_i \leq \frac{m}{2}$) or ($t_i < 0$ and $t_i = -(m-1)$), movedir = positive; endif;
    if ($t_i > 0$ and $t_i = (m-1)$) or ($t_i < 0$ and $t_i \geq -\frac{m}{2}$), movedir = negetive; endif;
    if (movedir = positive and $t_i > 0$), distance = $t_i$; endif;
    if (movedir = positive and $t_i < 0$), distance = $m + t_i$; endif;
    if (movedir = negative and $t_i < 0$), distance = $t_i$; endif;
    if (movedir = negative and $t_i > 0$), distance = $-m + t_i$; endif;
 endfor
  while($t_2 \neq 0$ or distance$_2 \neq 0$) do
    if (movedir = positive), move packet to $+z$ node; distance$_2$ = distance$_2 - 1$; endif;
    if (movedir = negetive), move packet to $-z$ node; distance$_2$ = distance$_2 + 1$; endif;
  endwhile;
  while($t_1 \neq 0$ or distance$_1 \neq 0$) do
    if (movedir = positive), move packet to $+y$ node; distance$_1$ = distance$_1 - 1$; endif;
    if (movedir = negetive), move packet to $-y$ node; distance$_1$ = distance$_1 + 1$; endif;
  endwhile;
  while($t_0 \neq 0$ or distance$_0 \neq 0$) do
    if (movedir = positive), move packet to $+x$ node; distance$_0$ = distance$_0 - 1$; endif;
    if (movedir = negetive), move packet to $-x$ node; distance$_0$ = distance$_0 + 1$; endif;
  endwhile;
end
```

Fig. 3 Routing algorithm of the MH3DT network.

4. Deadlock-Free Routing

The most expensive part of an interconnection network is the wire that forms the physical channels; for a particular topology, the physical channel cost is constant. The second most expensive elements are the buffers and switches. Since the networks we consider are wormhole-routed, the main factor in buffer expense is the number of virtual channels. Virtual channels [18] reduce the effect of blocking; they are used widely in parallel computer systems, to improve dynamic communication performance by relieving contention in the multicomputer network and to design deadlock-free routing algorithms. Since the hardware cost increases as the number of virtual channels increases, the unconstrained use of virtual channels is not cost-effective in parallel computers. Therefore, a deadlock-free routing algorithm for an arbitrary interconnection network with a minimum number of virtual channels is preferred. In this section, we discuss the minimum number of virtual channels for deadlock-free routing of the MH3DT network. We also present a proof that the MH3DT network is deadlock free.

To prove the proposed routing algorithm for the MH3DT network is deadlock free, we divide the routing path into three phases, as follows:

- *Phase 1:* Intra-BM transfer path from source PE to the face of the BM.
- *Phase 2:* Higher level transfer path.

sub-phase $2.i.1$: Intra-BM transfer to the outlet PE of

Level $(L-i)$ through the z-link.

sub-phase 2.*i*.2 : Inter-BM transfer of Level $(L-i)$ through the z-link.

sub-phase 2.*i*.3 : Intra-BM transfer to the outlet PE of Level $(L-i)$ through the y-link.

sub-phase 2.*i*.4 : Inter-BM transfer of Level $(L-i)$ through the y-link.

sub-phase 2.*i*.5 : Intra-BM transfer to the outlet PE of Level $(L-i)$ through the x-link.

sub-phase 2.*i*.6 : Inter-BM transfer of Level $(L-i)$ through the x-link.

- *Phase 3:* Intra-BM transfer path from the outlet of the inter-BM transfer path to the destination PE.

The proposed routing algorithm enforces some routing restrictions to avoid deadlocks [17]. Since dimension-order routing is used in the MH3DT network, messages in the network are routed first in the z-direction then in the y-direction, and finally in the x-direction. The interconnection of the BM and the higher levels of the MH3DT network is a toroidal connection. The number of virtual channels required to make the routing algorithm deadlock-free for the MH3DT network is determined using the following lemma.

Lemma 1: If a message is routed in the order $z \rightarrow y \rightarrow x$ in a 3D-torus network, then the network is deadlock free with 2 virtual channels.

Proof: In torus interconnection networks, cyclic dependencies can occur in two ways, first as a result of the inter-dimensional turns made by the messages, or second, as a result of a wrap-around connection in the same direction. To avoid these cyclic dependencies, we need two virtual channels, one for inter-dimensional turns and another for wrap-around connections. Initially, messages are routed over virtual channel 0. Then, if the packet is going to use a wrap-around channel, messages are routed over virtual channel 1. For a 3D-torus network, the channels are allocated as shown in Eq. 5. Enforcing this routing restriction and using virtual channels means cyclic dependencies are avoided. Thus, deadlock freeness is proved.

$$C = \begin{cases} (l, vc, n_2), & z+ \text{ channel,} \\ (l, vc, 4-n_2), & z- \text{ channel,} \\ (l, vc, n_1), & y+ \text{ channel,} \\ (l, vc, 4-n_1), & y- \text{ channel,} \\ (l, vc, n_0), & x+ \text{ channel,} \\ (l, vc, 4-n_0), & x- \text{ channel} \end{cases} \tag{5}$$

Here, $l = \{0,1,2,3,4,5\}$ are the links used in the BM, $l = \{0,1\}$, $l = \{2,3\}$, and $l = \{4,5\}$ are the links used in the z–direction, y–direction, and x–direction, respectively. $vc = \{0,1\}$ are virtual channels, and n_0, n_1, and n_2 are the PE addresses in the BM.

Theorem 1: An MH3DT network with 2 virtual channels is deadlock free.

Proof: Both the BM and the higher levels of the MH3DT network have a toroidal interconnection. In phase-1 and phase-3 routing, packets are routed in the source-BM and destination-BM, respectively. The BM of the MH3DT network is a 3D-torus network. According to Lemma 1, the number of necessary virtual channels for phase-1 and phase-3 is 2. Intra-BM links between inter-BM links are used in sub-phases 2.*i*.1, 2.*i*.3, and 2.*i*.5. Thus, sub-phases 2.*i*.1, 2.*i*.3, and 2.*i*.5 utilize channels over intra-BM links, sharing the channels of either phase-1 or phase-3. The gate nodes, as mentioned earlier, are used for higher level interconnection. The free links in these gate nodes are used in sub-phases 2.*i*.2, 2.*i*.4, and 2.*i*.6, and these links form a 3D-torus network for the higher level network. According to Lemma 1, the number of necessary virtual channels for this 3D-torus network is also 2. The main idea is that messages are routed over one virtual channel. Then, messages are switched over the other virtual channel if the packet is going to use a wrap-around connection.

Therefore, the total number of necessary virtual channels for the whole network is 2.

5. Performance Evaluation

Comparing the performance of different hierarchical interconnection networks such as MH3DT, H3DT [13], TESH [9], and CCC [14] networks is not an easy task, because each network has a different interconnection architecture, which makes it difficult to match the total number of nodes. The total number of nodes in the MH3DT network is $N = \left[m^3 \times n^{3(L-1)}\right]$. If $m = 4$, $n = 4$, and $L = 2$, then the total number of nodes of the MH3DT network is 4096. Level-2 H3DT network with $m = 4$ and $n = 4$, Level-3 TESH, 64×64 mesh, and 12-D hypercube networks also have 4096 nodes. The d-dimensional CCC network has $d \times 2^d$ nodes. If $d = 9$, then the total number of nodes of the CCC network is $9 \times 2^9 = 4608$. According to the structure of the CCC network [14], it is not possible to construct a 4096-node CCC network. We have compared the static network performance and dynamic communication performance of various 4096-node networks. We have also evaluated the static network performance of the CCC network with 4608 nodes.

5.1 Static Network Performance

The topology of an interconnection network determines many architectural features of that parallel computer and affects several performance metrics. Although the actual performance of a network depends on many technological and implementation issues, several topological properties and performance metrics can be used to evaluate and compare different network topologies in a technology-independent manner. Most of these properties are derived from the graph model of the network topology. In this section, we discuss some of the properties and performance metrics that characterize the cost and performance of an interconnection network.

The static network performance of various networks with 4096 nodes, along with that of a CCC network with 4608 nodes, is tabulated in Table 1. The static network performance of the 4608-noded CCC network can not be compared with the other 4096-noded networks. However, its performance is included in Table 1 to show its topological properties.

5.1.1 Node Degree

The *node degree* is defined as the number of physical channels emanating from a node. This attribute is a measure of the node's I/O complexity. For the MH3DT network, the node degree is independent of network size. Since each node has eight channels, its degree is 8. Constant degree networks are easy to expand and the cost of the network interface of a node remains unchanged with increasing size of the network.

5.1.2 Diameter

The *diameter* of a network is the maximum inter-node distance, i.e., the maximum number of links that must be traversed to send a message to any node along a shortest path. As a definition, the distance between adjacent nodes is unity. The diameter is the maximum distance between two nodes. The diameter is commonly used to describe and compare the static network performance of the network's topology. Networks with small diameters are preferable. The smaller the diameter of a network the shorter the time to send a message from one node to the node farthest away from it. In fact, the diameter sometimes (but not always) sets the lower bound for the running time of an algorithm performed on the network. Table 1 shows a comparison of the MH3DT network diameter with several other networks. Clearly, the MH3DT network has a much smaller diameter than the conventional mesh, TESH [9], [10], and H3DT [12], [13] networks but larger than that of the hypercube network.

5.1.3 Cost

Inter-node distance, message traffic density, and fault-tolerance are dependent on the diameter and the node degree. The product (*diameter* × *node degree*) is a good criterion for measuring the relationship between cost and performance of a multiprocessor system [5], [8]. An interconnection network with a large diameter has a very low message passing bandwidth and a network with a high node degree is very expensive. In addition, a network should be easily expandable; there should be no changes in the basic node configuration as we increase the number of nodes. Table 1 shows that the cost of the MH3DT network is smaller than that of the mesh and H3DT [12], [13] networks, equal to that of the hypercube network, and a bit higher than the TESH [9], [10] network.

5.1.4 Average Distance

The *average distance* is the mean distance between all distinct pairs of nodes in a network. A small average distance allows small communication latency, especially for distance-sensitive routing, such as store and forward. But it is also crucial for distance-insensitive routing, such as wormhole routing, since short distances imply the use of fewer links and buffers, and therefore less communication contention. We have evaluated the average distance for different conventional topologies by the corresponding formula and of different hierarchical networks by simulation. As we see in Table 1, the MH3DT network has a smaller average distance than the conventional mesh network and hierarchical TESH [9], [10] and H3DT [12], [13] networks. However, the average distance of the MH3DT network is higher than that of the hypercube network.

5.1.5 Arc Connectivity

Connectivity measures the robustness of a network. It is a measure of the multiplicity of paths between processors. Arc connectivity is the minimum number of links that must removed in order to break the network into two disjoint parts; it is a measure of connectivity. High connectivity (and thus arc connectivity) improves performance during normal operation by avoiding congested links, and also improves fault tolerance. A network is maximally fault-tolerant if its connectivity is equal to the degree of the network. The arc connectivity of several networks is shown in Table 1. Clearly, the arc connectivity of the MH3DT network is higher than a conventional mesh, TESH [9], [10], and H3DT [12], [13] networks and it is closer to the node degree. The arc connectivity of an interconnection networks is independent of its total number of nodes. The arc connectivity of the CCC network is exactly equal to the node degree. Thus, CCC is more fault tolerant than the MH3DT network.

5.1.6 Bisection Width

Bisection width (BW) is another measure of robustness. The BW of a network is defined as the minimum number of communication links that must removed to partition the network into two equal halves. The bisection width of a ring network is two, since any partition cuts across only two communication links. The bisection width of the MH3DT network is calculated by:

$$BW_{(MH3DT)} = 2^{q+1} \times (m \times n) \tag{6}$$

BW is calculated by counting the number of links that need to be removed to partition the highest level (Level-L) torus. This equation is valid for higher level networks. We don't consider the interconnection of basic modules here. The basic module is simply a 3D-torus network so its bisection width is $2m^2$.

Many problems can be solved in parallel using *binary divide-and-conquer:* split the input data set into two halves and solve them recursively on both halves of the interconnection network in parallel, then merge the results from both halves into the final result. Small bisection width implies low bandwidth between the two halves and it can slow down the final merging phase. On the other hand, a large bisection width is undesirable for the VLSI design of the interconnection network, since it implies a lot of *extra chip wires*, such as in hypercube [9]. Table 1 shows that the bisection width of the MH3DT network is higher than that of the conventional mesh and TESH networks [9], [10] and equal to that of the H3DT network [12], [13].

5.1.7 Wiring Complexity

The *wiring complexity* of an interconnection network refers to the number of links needed to be connected to a node as the network is scaled up. The wiring complexity depends on the node degree, which is the number of channels incident to a node. The wiring complexity has a direct correlation to hardware cost and complexity. An n-dimensional torus network has $n \times N$ links, where N is the total number of nodes. The wiring complexity of a Level-L MH3DT network is represented as shown in Eq. 7.

$$\left[3m^3 \times n^{3(L-1)} + \sum_{i=2}^{L} 3(2^q) \left\{ n^{3(L-1)} \right\} \right] \tag{7}$$

Table 1 compares the wiring complexity for a MH3DT network with several other networks. The total number of physical links in the MH3DT network is higher than that in the conventional mesh, TESH [9], [10], and H3DT [12], [13] networks; therefore, the cost of physical links is higher for the MH3DT network. However, it is almost half the wiring complexity of the hypercube network.

5.2 Dynamic Communication Performance

The overall performance of a multicomputer system is affected by the performance of the interconnection network as well as by the performance of the node. Continuing advances in VLSI/WSI technologies promise to deliver more power to individual nodes. On the other hand, low performance of the communication network will severely limit the speed of the entire multicomputer system. Therefore, the success of massively parallel computers is highly dependent on the efficiency of their underlying interconnection networks.

5.2.1 Performance Metrics

The dynamic communication performance of a multicomputer is characterized by message latency and network throughput. Message latency refers to the time elapsed from the instant when the first flit is injected by the source into the network to the instant when the last flit of the message is received at the destination. Network throughput refers to the maximum amount of information delivered per unit of time through the network. For the network to have good performance, low latency and high throughput must be achieved.

Latency is measured in time units. However, when comparing several design choices, the absolute value is not important; because the comparison is performed by computer simulation, latency is measured in simulator clock cycles. Throughput depends on message length and network size. Therefore, throughput is usually normalized, dividing it by message length and network size. When throughput is compared by computer simulation and wormhole routing is used for switching, throughput can be measured in flits per node and clock cycle.

5.2.2 Simulation Environment

To evaluate dynamic communication performance, we have developed a wormhole routing simulator. In our simulation, we use a dimension-order routing algorithm. The dimension-order routing algorithm, which is exceedingly simple, provides the only route for the source-destination pair. Extensive simulations for several networks have been carried out under the uniform traffic pattern, in which each node sends messages to every other node with equal probability. Two virtual channels per physical channel are simulated, and the virtual channels are arbitrated by a round robin algorithm. For all of the simulation results, the packet size is 16 flits. Two flits are used as the header flit. Flits are transmitted at 20,000 cycles; in each clock cycle, one flit is transferred from the input buffer to the output buffer, or from output to input if the corresponding buffer in the next node is empty. Therefore, transferring data between two nodes takes 2 clock cycles.

Table 1 Comparison of 4096 node networks.

	Hyper-Cube	2D-Mesh Network	CCC† 9×2^9	TESH Network	H3DT (4,4,2,0)	MH3DT (4,4,2,0)	H3DT (4,4,2,2)	MH3DT (4,4,2,2)
Node Degree	12	4	3	4	6	8	6	8
Diameter	12	126	22	32	25	20	21	18
Cost	144	504	66	128	150	160	126	144
Average Distance	6	42.67	12.75	17.80	12.74	10.36	10.77	9.37
Arc Connectivity	12	2	3	2	3	6	3	6
Bisection Width	2048	64	256	8	32	32	128	128
Wiring Complexity	24576	8064	6912	6688	9408	12480	9792	12864

†CCC network with 4608 nodes

5.2.3 Dynamic Communication Performance Evaluation

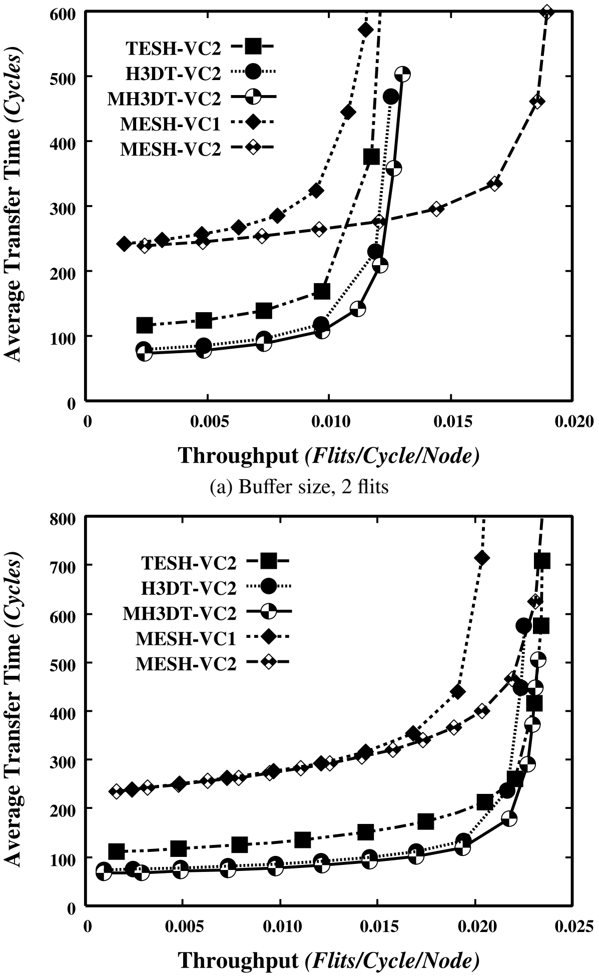
We have evaluated the dynamic communication performance of several 4096 node networks under uniform traffic patterns. As mentioned earlier that it is not possible to construct a CCC network with 4096 nodes. Thus, we can not make a fair comparison between the dynamic communication performance of the 4608-noded CCC network and the other, 4096-noded networks.

Under the assumption of constant bisection width, low-dimensional networks with wide channels provide lower latency, less contention, and higher throughput than higher-dimensional networks with narrow channels [1]. However, if the bisection width is high the network throughput will be high, and vice-versa. The bisection width of hypercube networks is very high. Thus, their throughput is high, with the accompanying cost of large latency [1]. Moreover, it has already been shown that the number of vertical links between silicon wafers in the 3D-WSI implementation for the hypercube network is very large; thus it is not suitable for 3D-WSI implementation [9]–[11]. This is why we have considered conventional mesh networks rather than hypercubes for comparing the dynamic communication performance. We have compared the dynamic communication performance of various interconnection networks, including hierarchical interconnection networks, with low bisection width.

To evaluate the dynamic communication performance of the H3DT and MH3DT networks, we have used maximum inter-level connectivity, i.e., $q = 2$. For a fair comparison, we allocated 2 virtual channels to the router for performance evaluation. Figure 4 (a) depicts the results of simulation under uniform traffic patterns for the various network models. This figure presents the average transfer time as a function of network throughput. Each curve stands for a particular network. As shown in Fig. 4 (a), the average transfer time of the MH3DT network is lower than that of the H3DT, TESH, and mesh networks; its maximum throughput is higher than that of the H3DT network, TESH network, and mesh network with 1 virtual channel, but lower than that of the mesh network with 2 virtual channels.

5.2.4 Effect of Buffer Size

In a large buffer, more flits can be stored in the one channel buffer; therefore, frequency of packet blockings and deadlocks are decreased and the dynamic communication performance of the interconnection network is increased. We have evaluated the dynamic communication performance of various network models using a channel buffer size of 20 flits and the result is portrayed in Fig. 4 (b). As shown in Fig. 4 (b), the average transfer time of the MH3DT network is lower than that of the H3DT, TESH, and mesh networks; the maximum throughput of the MH3DT network is higher than that of the H3DT network and mesh network with 1 & 2 virtual channels and equal to that of the TESH network.



(a) Buffer size, 2 flits

(b) Buffer size, 20 flits

Fig. 4 Dynamic communication performance of dimension-order routing with uniform traffic pattern on various networks: 4096 nodes, 2 VCs, 16 flits.

5.2.5 Effect of Message Length

In this section, we analyzed the effect of message length on dynamic communication performance. We have evaluated the dynamic communication performance of the MH3DT network for short, medium, and long messages to show the effect of message length. Figure 5 shows the average message latency divided by message length for the uniform traffic pattern. It is shown that in the MH3DT network, average transfer time decreases and maximum throughput increases with an increase in message length.

Average message latency is smaller for long messages because wormhole switching is used. Thus, the messages are pipelined in nature. Path setup time is amortized among more flits when messages are long. Moreover, data flits can advance faster than message headers because headers need a routing decision. Hence, headers have to wait for the routing control unit to compute the output channel, and possibly, wait for the output channel to become free. Therefore, when the header reaches the destination node, the data flits advance faster, thus favoring long messages.

5.2.6 Effect of Virtual Channels

We investigated the effect of adding extra virtual channels on the MH3DT network for dimension-order routing. Figure 6 depicts the average transfer time as a function of network throughput under the uniform traffic pattern for different virtual channels. The minimum number of virtual channels for deadlock-free routing is 2. Adding 1 extra virtual channel, for a total of 3, substantially improves the dynamic communication performance of the MH3DT network. We have also evaluated the dynamic communication performance of the MH3DT network using 4 virtual channels. Figure 6 shows that the maximum throughput of the MH3DT network using 3 virtual channels is higher than that using 2 and almost same as that of 4 virtual channels. This striking difference of throughput shows that we can significantly improve the dynamic communication performance of the MH3DT network by adding 1 extra virtual channel over the minimum number.

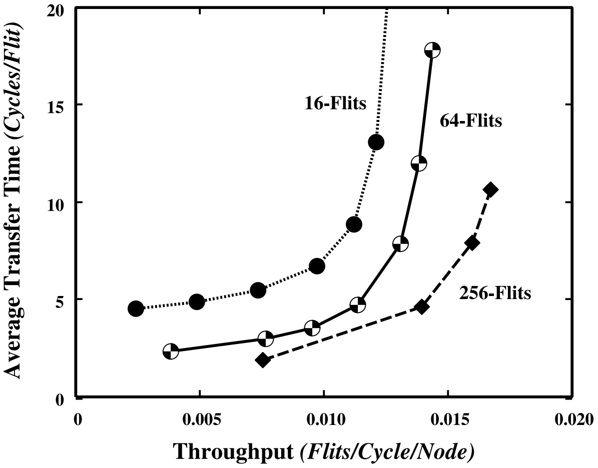


Fig. 5 Average transfer time divided by message length versus network throughput of MH3DT network: 4096 nodes, 2 VCs, 16 flits, buffer size 2 flits.

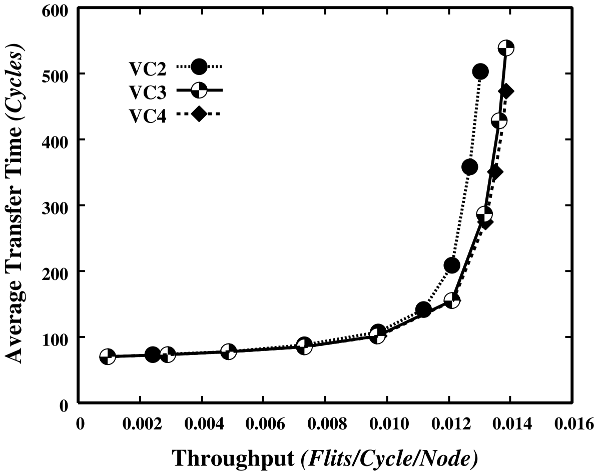


Fig. 6 Dynamic communication performance of dimension order routing with uniform traffic pattern on the MH3DT network: 4096 nodes, various virtual channels, 16 flits, buffer size 2 flits.

The question may arise whether we need massively parallel computers with millions of nodes. The answer is 'yes'. Solving the most challenging problems in many areas of science and engineering such as defense, aerospace, automotive applications, and weather forecasting, requires teraflops performance for more than a thousand hours at a time. This is why, in the near future, we will need computer systems capable of computing at the petaflops level. To achieve this level of performance, we need massively parallel computers with tens of thousands or millions of nodes.

Reconfigurable parallel architectures use the reconfigurability of their interconnection network to establish a communication topology optimized for the problem being solved. However, most reconfigurable LSIs have a large number of logic cells which are tightly connected to each other. To avoid complexity in the wiring area, fixed topology is still exploited.

6. Conclusion

We have modified the H3DT network and presented a new hierarchical interconnection network, called the MH3DT network for massively parallel computers. The architecture of the MH3DT network, routing of messages, and static network performance were discussed in detail. A deadlock-free routing algorithm with a minimum number of virtual channels has been proposed for the MH3DT network. It has been proven that 2 virtual channels per physical channel are sufficient for the deadlock-free routing algorithm of the MH3DT network – 2 is also the minimum required number of virtual channels.

From the static network performance, it has been shown that the MH3DT network possesses several attractive features including constant node degree, small diameter, small cost, high connectivity, small average distance, and better bisection width. By using the routing algorithm described in this paper and the uniform traffic pattern, we have evaluated the dynamic communication performance of the MH3DT network as well as that of several other interconnection networks. The average transfer time of the MH3DT network is lower than that of the H3DT, TESH, and mesh networks. Maximum throughput of the MH3DT network is also higher than that of those networks. A comparison of dynamic communication performance reveals that the MH3DT network achieves better results than the H3DT, TESH, and mesh networks. The MH3DT network yields low latency and high throughput, which are indispensable for high-performance massively parallel computers.

This paper focused on the architectural structure, deadlock-free routing, static network performance, and dynamic communication performance of the MH3DT network using dimension order routing. Other issues should also be kept in mind: (1) assessment of the performance improvement of the MH3DT network with an adaptive routing algorithm and (2) evaluating the system yield by providing hardware redundancy.

Acknowledgment

This research was supported in part by the Grant-in-Aid of Scientific Research (B) 14380138, JSPS, Japan. The authors are grateful to the anonymous reviewers for their constructive comments which helped to greatly improve the clarity of this paper. This paper has benefited greatly from editing by Judith Steeh, Head Editor of the JAIST Technical Communication Program. We would like to thank the associate editor of this paper, Mr. Wei Chen, for arranging a quick review of this manuscript.

References

[1] W.J. Dally, “Performance analysis of *k*-ary *n*-cube interconnection networks,” IEEE Trans. Comput., vol.39, no.6, pp.775–785, June 1990.

[2] Y.R. Potlapalli, “Trends in interconnection network topologies: Hierarchical networks,” Int'l. Conf. on Parallel Processing Workshop, pp.24–29, 1995.

[3] A. El-Amawy and S. Latifi, “Properties and performance of folded hypercube,” IEEE Trans. Parallel Distrib. Syst., vol.2, no.1, pp.31–42, 1991.

[4] A. Esfahanian, L.M. Ni, and B.E. Sagan, “The twisted *n*-cube with application to multiprocessing,” IEEE Trans. Comput., vol.40, no.1, pp.88–93, 1991.

[5] J.M. Kumar and L.M. Patnaik, “Extended hypercube: A hierarchical interconnection network of hypercube,” IEEE Trans. Parallel Distrib. Syst., vol.3, no.1, pp.45–57, 1992.

[6] N.F. Tzeng and S. Wei, “Enhanced hypercube,” IEEE Trans. Comput., vol.40, no.3, pp.284–294, 1991.

[7] S.G. Ziavras, “A versatile family of reduced hypercube interconnection network,” IEEE Trans. Parallel Distrib. Syst., vol.5, no.11, pp.1210–1220, 1994.

[8] L.N. Bhuyan and D.P. Agrawal, “Generalized hypercube and hyperbus structures for a computer network,” IEEE Trans. Comput., vol.33, no.4, pp.323–333, 1984.

[9] V.K. Jain, T. Ghirmai, and S. Horiguchi, “TESH: A new hierarchical interconnection network for massively parallel computing,” IEICE Trans. Inf. & Syst., vol.E80-D, no.9, pp.837–846, Sept. 1997.

[10] V.K. Jain and S. Horiguchi, “VLSI considerations for TESH: A new hierarchical interconnection network for 3-D integration,” IEEE Trans. Very Large Scale Integr. (VLSI) Syst., vol.6, no.3, pp.346–353, 1998.

[11] S. Horiguchi, “New interconnection for massively parallel and distributed systems,” Research Report, Grant-in-Aid Scientific Research, pro. no.09044150, pp.1–72, JAIST, 1999.

[12] S. Horiguchi and T. Ooki, “Hierarchical 3D-torus interconnection network,” Proc. ISPAN'00, pp.50–56, Texas, USA, 2000.

[13] S. Horiguchi and T. Ooki, “Hierarchical 3D-torus interconnection network for massively parallel computers,” JAIST Research Report, IS-RR-2000-022, pp.1–15, ISSN 0918-7553, 2000.

[14] F.P. Preparata and J. Vuillemin, “The cube-connected cycles: A versatile network for parallel computation,” J. ACM, vol.24, no.5, pp.300–309, May 1981.

[15] L.M. Ni and P.K. McKinley, “A survey of wormhole routing techniques in direct networks,” Computer, vol.26, no.2, pp.62–76, 1993.

[16] W.J. Dally and C.L. Seitz, “The torus routing chip,” J. Distrib. Computing, vol.1, no.3, pp.187–196, 1986.

[17] W.J. Dally and C.L. Seitz, “Deadlock free message routing in multiprocessor interconnection networks,” IEEE Trans. Comput., vol.C-36, no.5, pp.547–553, 1987.

[18] W.J. Dally, “Virtual-channel flow control,” IEEE Trans. Parallel Distrib. Syst., vol.3, no.2, pp.194–205, 1992.

[19] J. Duato, “A new theory of deadlock free adaptive routing in wormhole networks,” IEEE Trans. Parallel Distrib. Syst., vol.4, no.12, pp.1320–1331, 1993.

[20] X. Zhang, “System effects of interprocessor communication latency in multicomputers,” IEEE Micro, vol.11, no.2, pp.12–55, 1991.

[21] H.S. Azad, L.M. Mackenzie, and M.O. Khaoua, “The effect of the number of virtual channels on the performance of wormhole-routed mesh interconnection networks,” Proc. 16th Annual UKPEW, Glasgow University, 2000.

[22] W. Feng and K.G. Shin, “The effect of virtual channels on the performance of wormhole algorithms in multicomputer networks,” University of Michigan directed Study Report, May 1994.

[23] A.A. Chien and J.H. Kim, “Planer-adaptive routing: Low-cost adaptive networks for multiprocessors,” J. ACM, vol.42, no.1, pp.91–123, 1995.

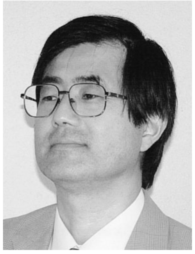
[24] J.H. Kim and A.A. Chien, “An evaluation of planar-adaptive routing (PAR),” Proc. 4th IEEE Symp. Parallel Distrib. Processing, pp.470–478, New York, 1992.

[25] J. Duato, S. Yalamanchili, and L. Ni, Interconnection Network: An Engineering Approach, IEEE Computer Society Press, Los Alamitos, California, USA, 1997.

[26] V. Kumar, V. Kumar, A. Grama, A. Gupta, and G. Karypis, Introduction to Parallel Computing: Design and Analyses of Algorithms, Benjamin/Cummings Pub. Co., Redwood City, California, USA, 1994.

M.M. Hafizur Rahman received his B.Sc. degree in Electrical and Electronic Engineering from Khulna University of Engineering and Technology (KUET), Khulna (erstwhile BIT, Khulna), Bangladesh, in 1996. He obtained his M.Sc. degree in information science from the Japan Advanced Institute of Science and technology (JAIST) in 2003. He is currently pursuing doctoral study in Information Science at JAIST. His current research include interconnection networks, especially hierarchical interconnection networks and optical switching networks. Prior to joining at JAIST, he was a lecturer in Electrical and Electronic Engineering at KUET, Bangladesh from 1996 to 1999. Mr. Rahman is an associate member of IEB of Bangladesh.

Yasushi Inoguchi received his B.E. degree from Department of Mechanical Engineering, Tohoku University in 1991, and received MS degree and Ph.D from JAIST (Japan Advanced Institute of Science and Technology) in 1994 and 1997, respectively. He is currently a Associate Professor of Center for Information Science at JAIST. He was a research fellow of the Japan Society for the Promotion of Science from 1994 to 1997. He is also a researcher of PRESTO program of Japan Science and Technology Agency since 2002. His research interest has been mainly concerned with parallel computer architecture, interconnection networks, and high performance computing on parallel machines. Dr. Inoguchi is a members of IEEE and IPS of Japan.



Susumu Horiguchi received his M.E and D.E degrees from Tohoku University in 1978 and 1981, respectively. He is currently a full professor in the Graduate School of Information Science, Tohoku University. He was a visiting scientist at the IBM Thomas J. Watson Research Center from 1986 to 1987 and a visiting professor at The Center for Advanced Studies, the University of Southwestern Louisiana and at the Department of Computer Science, Texas A&M University in the summers of 1994 and 1997. He was also a professor in the Graduate School of Information Science, JAIST (Japan Advanced Institute of Science and Technology). He has been involved in organizing many international workshops, symposia and conferences sponsored by the IEEE, IEICE and IPS. His research interests have been mainly concerned with interconnection networks, parallel computing algorithms, massively parallel processing, parallel computer architectures, VLSI/WSI architectures, and Multi-Media Integral Systems. Prof. Horiguchi is a senior member of the IEEE Computer Society, and a member of the IPS and IASTED.