Dynamic Communication Performance of the TESH Network under Nonuniform Traffic Patterns

M.M. Hafizur Rahman*, Yasushi Inoguchi†, Yukinori Sato†, Yasuyuki Miura††, Susumu Horiguchi‡

*Dept. of Computer Science and Engineering, KUET, Khulna-9203, Bangladesh
E-mail: hafiz90305@gmail.com
†Center for Information Science, JAIST, Nomi-Shi, Ishikawa 923-1292, Japan
E-mail: inoguchi@jaist.ac.jp, yukinori@jaist.ac.jp
††Shonan Institute of Technology, Fujisawa, Kanagawa 251-8511, Japan
E-mail: miu@info.shonan-it.ac.jp
‡GSIS, Tohoku University, Aoba 6-3-09, Aramaki, Sendai 980-8579, Japan
E-mail: susumu@ecei.tohoku.ac.jp

***Abstract*— Interconnection networks play a crucial role in the performance of massively parallel computer systems. Hierarchical interconnection networks provide high performance at low cost by exploring the locality that exists in the communication patterns of massively parallel computer systems. The *Tori-connected mESH (TESH) Network* is a 2D-torus network of multiple basic modules, in which the basic modules are 2D-mesh networks that are hierarchically interconnected for higher level networks. In this paper, we evaluate the dynamic communication performance of the TESH network using dimension order routing algorithm under various nonuniform traffic patterns. We present a proof that the routing algorithm for the TESH network is deadlock free using 2 virtual channels – 2 being the minimum number for dimension-order routing. We evaluate the dynamic communication performance of TESH, mesh, and torus networks by computer simulation. It is shown that the dynamic communication performance of the TESH network is better than that of the mesh and torus networks.**

***Index Terms*— TESH network, deadlock-free routing, nonuniform traffic patterns, dynamic communication performance.**

I. INTRODUCTION

Interconnection networks are the key elements for building massively parallel computers [1]. In such computers, with millions of nodes, the large diameter of conventional topologies is completely infeasible. Hierarchical interconnection networks [2] are a cost-effective way to interconnect a large number of nodes. A variety of hypercube-based hierarchical interconnection networks have been proposed [3]- [7], but for massively parallel computer systems, the number of physical links becomes prohibitively large. To alleviate this problem, a k-ary n-cube based hierarchical interconnection network, called TESH network have been proposed [8].

A Tori connected mESH (TESH) network [8] is an interconnection network aiming for large-scale 3D massively parallel computers, consisting of multiple basic modules (BMs) which are 2D-mesh networks. The BMs are hierarchically interconnected by a 2D-torus to build higher level networks. While the details of TESH networks are available in the literature [8]–[11], its major features can be summarized as follows: it is hierarchical, thus allowing exploitation of computation locality as well as easy scalable up to thousands or tens of thousands of processors, it permits efficient VLSI realization, it is designed for fault tolerance by making use of redundancy for defect circumvention, it is well suited for 3-D stacked implementation, and it provides good dynamic communication performance under uniform traffic pattern with 4 virtual channels [11]. In this paper, we are specifically interested to investigate the effect of nonuniform traffic patterns in the TESH network with minimum number of virtual channels.

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 massively parallel computer through a router, using a routing algorithm. The routing algorithm specifies how a message selects its 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.

Deterministic, dimension-order routing has been popular in massively parallel computers because it has minimal hardware requirements and allows the design of simple and fast routers [12]. Although there are numerous paths between any source and destination, dimension-order routing defines a single path from source to destination. If that selected path is congested, the traffic between that source and the destination is delayed, despite the presence of uncongested alternative paths. However, it is still very popular because of its low cost and simple router design. This is why most existing parallel computers, such as J-machine, Touchstone, Ametek 2010, and Cosmic cube, use deterministic routing.

doi:10.4304/jnw.4.10.941-951

Wormhole routing [13], [14] has become the dominant switching technique used in contemporary massively parallel computer systems. 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 [12], [15] were originally introduced to solve the problem of deadlock in wormhole-routed networks. 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. In this study, we use a deadlock-free routing algorithm for the TESH network using 2 virtual channels – 2 being the minimum number for dimension-order routing.

The dynamic communication performance of the TESH network with a dimension-order routing algorithm under uniform traffic pattern using 2 and 4 virtual channels and hot-spot traffic pattern using 4 virtual channels was evaluated [11], [16], and it is proved to be better than that of conventional mesh network, but not that of torus network. However, the dynamic communication performance of the TESH network under the various nonuniform traffic patterns has not yet been evaluated. The main objective of this paper is to investigate the impact of various nonuniform traffic patterns on the TESH network using dimension order routing algorithm with minimum number of virtual channels.

The remainder of the paper is organized as follows. In Section II, we briefly describe the basic structure of the TESH network. The dynamic routing algorithm is proposed in Section III and its freedom from deadlock is also proved using minimum number of virtual channels. The dynamic communication performance of the TESH network under the various nonuniform traffic patterns is discussed in Section IV. Finally, Section V presents the conclusion.

II. Interconnection of the TESH network

A. Architecture of TESH Network

The *Tori-connected mESH (TESH) Network* is a hierarchical interconnection network consisting of Basic Modules (BM) that are hierarchically interconnected to form a higher level network. The BM of the TESH network is a 2D-mesh network of size ($2^m \times 2^m$). In this paper, unless specified otherwise, BM refers to a Level-1 network. Successively higher level networks are built by recursively interconnecting immediately lower level subnetworks in a 2D-torus network. A higher-level network is built using a 2D-toroidal connection among (2^{2m}) immediate lower level subnetworks.

If $m = 2$, the size of the BM is (4×4), and in this paper, we focus attention on $m = 2$ i.e., (4×4) BMs. A BM of (4×4) is shown in Figure 1. As seen in the figure, the BM has some free ports in the periphery 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, either one or two, are used for inter-BM connections to form higher level networks.

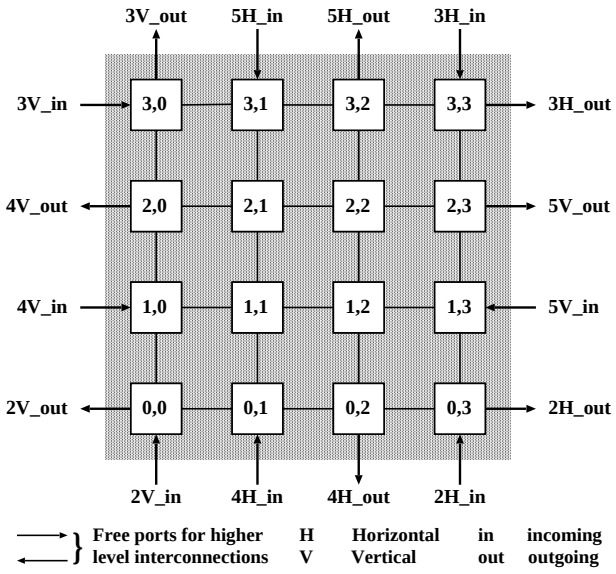


Figure 1. Basic module of the TESH network

In principle, m could be any positive integer value. However, if $m = 1$, then the network degenerates to a hypercube network. Hypercube is not a suitable network, because its node degree increases along with the increase of network size. If $m = 2$, then it is considered the most interesting case, because it has better granularity than the large BMs. In the rest of this paper we consider $m = 2$, therefore, we focus on a class of $\mathrm{TESH}(2, L, q)$ networks.

A $2^m \times 2^m$ BM has 2^{m+2} free ports at the contours for higher level interconnection. For each higher level interconnection, a BM uses $4 \times (2^q) = 2^{q+2}$ of its free links, $2(2^q)$ free links for vertical interconnections and $2(2^q)$ free links for horizontal interconnections. Here, $q \in \{0, 1, ..., m\}$, is the inter-level connectivity. $q = 0$ leads to minimal inter-level connectivity, while $q = m$ leads to maximum inter-level connectivity.

With $q = 0$, for example, $L_{max} = \left(2^{2-0} + 1\right) = 5$. Level-5 is the highest possible level for (4×4) BM interconnection. The limitation of having a maximum level of hierarchy is not a serious constraint. For the case of (4×4) BM with $q = 0$, a network built with the highest level, Level-5, consists of 1 million PEs. Successive higher level networks are built recursively by interconnecting the immediately lower level sub-networks. A higher-level network having $(2^m \times 2^m)$ BM is built using a (2^{2m}) subnetworks using a 2D-toroidal connection . For example, considering $(m = 2)$ a Level-2 subnetwork, can be formed by interconnecting $2^{2\times 2} = 16$ BMs. Similarly, a Level-3 network can be formed by interconnecting 16 Level-2 subnetworks, and so on. This phenomena is illustrated in Figure 2, a Level-2 TESH network can be formed by interconnecting 16 BMs as a (4×4) 2D-torus network. Each BM is connected to its logically adjacent BMs. To avoid clutter, the wraparound links of the BMs are not shown.

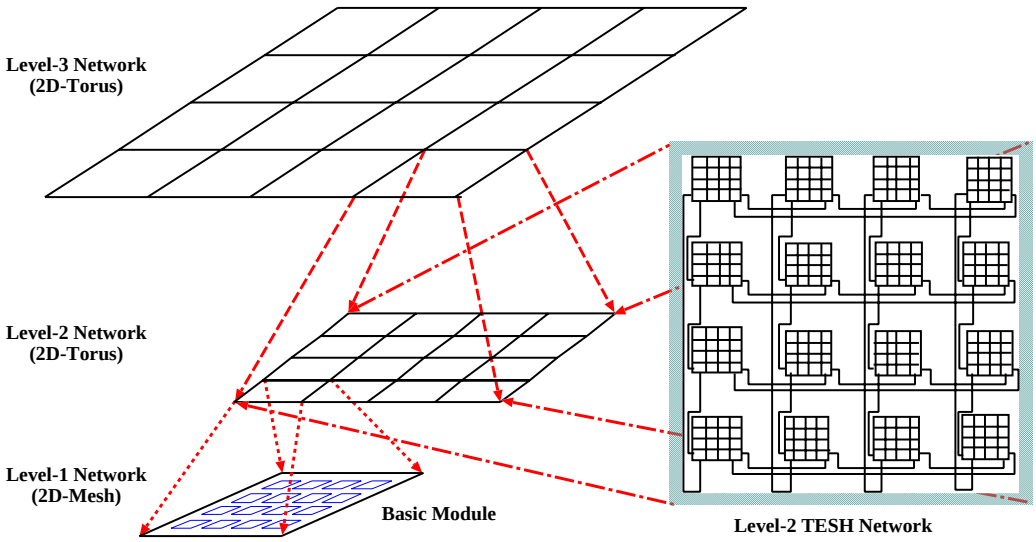


Figure 2. Interconnection of a TESH network

B. Addressing of Nodes

Base-4 numbers are used for convenience of address representation. As seen in Figure 1, nodes in the BM are addressed by two digits, the first representing the row index and the next representing the column index. More generally, in a Level-L TESH network, the node address is represented by:

$$\begin{aligned} A &= A^L A^{L-1} A^{L-2} \ \ldots\ \ldots\ \ldots\ A^2 A^1 \\ &= a_{n-1}a_{n-2}a_{n-3} \ \ldots\ \ldots\ \ldots\ a_2 a_1 a_0 \\ &= a_{2L-1}a_{2L-2}\ a_{2L-3}a_{2L-4} \ \ldots\ \ldots\ \ldots\ a_3 a_2\ a_1 a_0 \\ &= (a_{2L-1}\ a_{2L-2})\ (a_{2L-3}\ a_{2L-4})\ \ldots\ \ldots \\ &\quad \ldots\ \ldots\ \ldots (a_3\ a_2)\,(a_1\ a_0) \end{aligned} \tag{1}$$

Here, the total number of digits is $n = 2L$, where L is the level number. A^L is the address of level L in row-major scheme, and $(a_{2L-1}a_{2L-2})$ is the co-ordinate position of Level-$(L-1)$ for Level-L network. Pairs of digits run from group number 1 for Level-1, i.e., the BM, to group number L for the L-th level. Specifically, i-th group $(a_{2i-1}\ a_{2i-2})$ indicates the location of a Level-$(i-1)$ subnetwork within the i-th group to which the node belongs; $1 \le i \le L$. In a two-level network the address becomes $A = (a_4\ a_3)\ (a_1\ a_0)$. The first pair of digits $(a_4\ a_3)$ identifies the BM to which the node belongs, and the last pair of digits $(a_1\ a_0)$ identifies the node within that basic module.

The assignment of inter-level ports for the higher level networks has been done quite carefully so as to minimize the higher level traffic through the BM. The address of a node n^1 encompasses in BM_1 is represented as $n^1 = \left(a^1_{2L-1}a^1_{2L-2} \ \ldots\ \ldots\ \ldots\ a^1_3 a^1_2\ a^1_1 a^1_0\right)$. The address of a node n^2 encompasses in BM_2 is represented as $n^2 = \left(a^2_{2L-1}a^2_{2L-2} \ \ldots\ \ldots\ \ldots\ a^2_3 a^2_2\ a^2_1 a^2_0\right)$. The node n^1 in BM_1 and n^2 in BM_2 are connected by a link if the following condition is satisfied.

$$\begin{aligned} &\exists i\{a^1_i = (a^2_i \pm 1) \bmod 2^m \\ &\wedge \forall j(j \neq i \ \rightarrow\ a^1_j = a^2_j)\} \end{aligned} \tag{2}$$

where $i, j \ge 2$

III. Routing Algorithm for TESH Network

A. Dynamic Routing Algorithm

A routing algorithm determines the path a packet takes as it travels through the network from its source to its destination. Routing of messages in the TESH 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. For intra-BM transfer, there are two directions in both x-direction and y-direction. One of those is $+$ direction and the other is $-$ direction and they called x+, x−, y+, y−.

We have considered the dimension order routing algorithm for the TESH network. We use the following strategy: at each level, vertical routing is performed first. Once the packet reaches the correct row, then horizontal routing is performed. Routing in the TESH network is strictly defined by the source node address and the destination node address. Let a source node address be $s = (s_{2L-1}, s_{2L-2}), (s_{2L-3}, s_{2L-4}),\ \ldots\ , (s_3, s_2), (s_1, s_0)$, a destination node address be $d = (d_{2L-1}, d_{2L-2}), (d_{2L-3}, d_{2L-4}),\ \ldots\ , (d_3, d_2), (d_1, d_0)$, and a routing tag be $t = (t_{2L-1}, t_{2L-2}), (t_{2L-3}, t_{2L-4}),\ \ldots\ , (t_1, t_0)$, where $t_i = d_i - s_i$. The function

get_group_number is the function to get group number. Arguments of this function are source PE address s, destination PE address d, and direction. The function outlet_x and outlet_y are the function to get x coordinate a_1 and y coordinate a_0 of the PE n that link $(g, l, d\delta)$ exists. Each links are labeled as $(g, l, d\delta)$ by get_group_number, level $l(2 \leq l \leq L)$, dimension $d(d \in \{\text{V,H}\})$ and direction $\delta(\delta \in \{+, -\})$. Here the $+$ and $-$ direction of vertical and horizontal links are represented by V+, V−, H+, and H−, respectively. Figure 3 shows the routing algorithm for the TESH network and Figure 4 portrayed an example of routing message from source to destination.

```
Routing Algorithm for a Level-L TESH:

Routing(s,d);
source;s={s2L-1,s2L-2,...,s0 };  destination;d={d2L-1,d2L-2,...,d0 };
tag;t2L-1,t2L-2,...,t0 ;     group;g ;

  for i = 2L-1:2;
    if (di-si+2^m) mod 2^m <= 2^m/2 then
            routedir = plus; ti= (di-si+2^m) mod 2^m;
    else  routedir = minus; ti= 2^m- (di-si+2^m) mod 2^m; endif;

    g = get_group_number(s,d,routedir);

    while(ti != 0) do
      if i is even number then
              outlet_nodex= outlet_x(g,i/2+1,H,routedir);
              outlet_nodey= outlet_y(g,i/2+1,H,routedir);endif;
      if i is odd number then                                        } (a)
              outlet_nodex= outlet_x(g,i/2+1,V,routedir);
              outlet_nodey= outlet_y(g,i/2+1,V,routedir);endif;
      BM_routing(outlet_nodex, outlet_nodey);

      if routedir = plus then send packet to next BM;
      else             send packet to previous BM; endif;        } (b)

      ti= ti- 1;
    endwhile;
  endfor;

    BM_routing(d1,d0);  ------------------------------------------> (c)
end.


BM_routing(dx, dy);
source;sx,sy; destination;dx,dy;
tag;tx,ty;

  tx= dx- sx;
  ty= dy- sy;
  while(ty != 0) do
    if ty> 0 then move packet to upper node; ty= ty- 1; endif;
    if ty< 0 then move packet to lower node; ty= ty+ 1; endif;
  endwhile;
  while(tx != 0) do
    if tx> 0 move packet to right node; tx= tx- 1; endif;
    if tx< 0 move packet to left node; tx= tx+ 1; endif;
  endwhile;
end.
```

Figure 3. Routing algorithm of the TESH network

B. Deadlock-Free Routing

Deadlock is catastrophic to an interconnection network. After a few resources (buffers or channels) are occupied by deadlocked packets, other packets block on these resources, paralyzing network operation. To prevent this situation, networks must either use deadlock avoidance or deadlock recovery. Almost all modern network use deadlock avoidance, usually by imposing an order on the resources and insisting that packets acquire these resources in order.

A deadlock free routing algorithm can be constructed for an arbitrary interconnection network by introducing

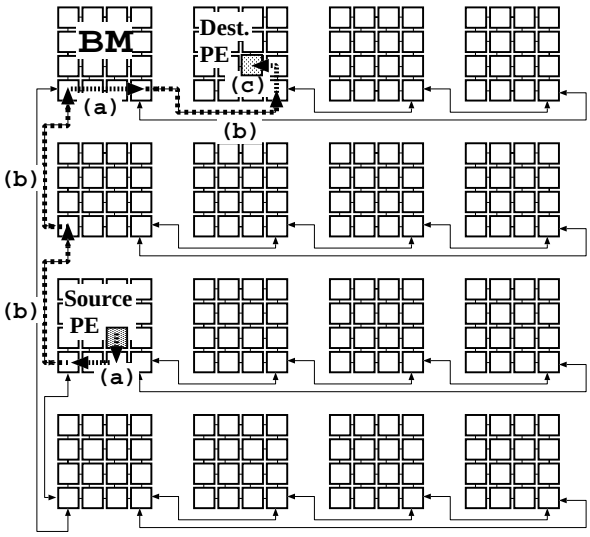


Figure 4. An Example of Routing

virtual channels. Since the hardware cost increases as the number of virtual channels increases, the unconstrained use of virtual channels is cost-prohibitive in parallel computers. Therefore, a deadlock free routing algorithm with a minimum number of virtual channels is needed. In this section, we discuss the minimum number of virtual channels for deadlock free routing of the TESH network. We also present a proof that the TESH network is deadlock free. To prove the proposed routing algorithm for the TESH network is deadlock free, we divide the routing path into three phases, as follows:

- *Phase 1:* Intra-BM transfer path from source node to the outlet node 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 y-link.
  - sub-phase $2.i.2$: Inter-BM transfer of Level $(L-i)$ through the y-link.
  - sub-phase $2.i.3$: Intra-BM transfer to the outlet PE of Level $(L - i)$ through the x-link.
  - sub-phase $2.i.4$: 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 phase 1, phase 2, and phase 3 routing of message are shown in the loop (a), loop (b), and loop (c), respectively in Figure 3.

The proposed routing algorithm enforces some routing restrictions to avoid deadlocks [12]. Since dimension-order routing is used in the TESH network, routing of message first performed in the vertical direction and then in the horizontal direction. The interconnection of the BM is a mesh network and the higher level network is a toroidal connection. Deterministic, dimension order routing is deadlock-free in a network if and only if the channel dependency graph is acyclic. As an example, we have considered a simple 4-node ring network on

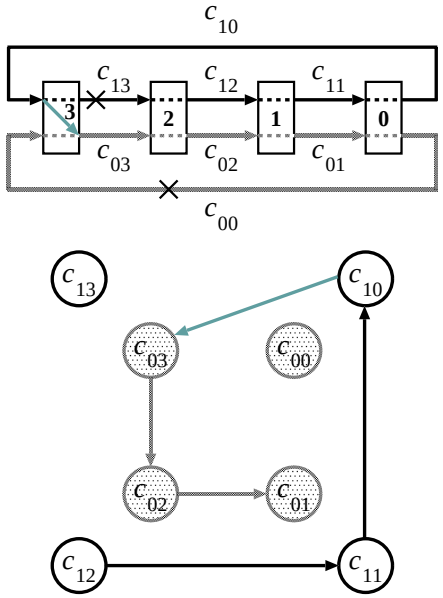


Figure 5. Deadlock-free routing in a 4-node ring network using 2 virtual channels and the corresponding acyclic channel dependency graph

Figure 5. Using 2 virtual channels is fairly easy here. Whenever a packet is destined for a node which is in the descending segment, the lower virtual channels are used. If the destination is in the upper segment with respect to the source, then the packet starts moving left using upper channels until it reaches an end-to-end (wrap-around channel connected) node, when it switches to lower channels and keep using them until it reaches the destination.

Two lemmas are stated below without proof. The proof is straight forward and it is allocation of channels and links during routing messages [1]. By using the following two lemmas, we will prove that the proposed routing algorithm for the TESH network is deadlock-free using 2 virtual channels.

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

Proof: If the channels are allocated according to Eq. 3 for a 2D-mesh network and the messages are routed according to the above mentioned phenomena, then cyclic dependency will not occur. Therefore, freedom from deadlock is proved.

$$C = \begin{cases} (l, a_1), & \text{y+ channel,} \\ (l, 2^m - a_1), & \text{y− channel,} \\ (l, a_0), & \text{x+ channel,} \\ (l, 2^m - a_0), & \text{x− channel.} \end{cases} \quad (3)$$

Here, $l = \{l_0, l_1, l_2, l_3\}$ are the links used in the BM, $l = \{l_0\}$, $l = \{l_1\}$, $l = \{l_2\}$, and $l = \{l_3\}$ are the links used in the y+ direction, y− direction, x+ direction, x− direction interconnection, respectively. 2^m is the size of the BM, and a_1 and a_0 are the node addresses in the BM.

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

Proof: If the channels are allocated as shown in Eq. 4 for a 2D-torus network, then deadlock freeness is proved.

$$C = \begin{cases} (l, vc, a_{2L-1}), & y+ \text{ channel,} \\ (l, vc, n - a_{2L-1}), & y- \text{ channel,} \\ (l, vc, a_{2L-2}), & x+ \text{ channel,} \\ (l, vc, n - a_{2L-2}), & x- \text{ channel} \end{cases} \quad (4)$$

Here, $l = \{l_4, l_5\}$ are the links used for higher-level interconnection, l_4 is used in the interconnection of the vertical directions and l_5 is used in the interconnection of the horizontal directions. $vc = \{VC_0, VC_1\}$ are the virtual channels, 2^m is the size of the higher level networks, and a_{2L-1} and a_{2L-2} are the node addresses in the higher level, where L is the level number.

Theorem 1: A TESH network with 2 virtual channels is deadlock free.

Proof: In phase-1 and phase-3 routing, packets are routed in the source-BM and destination-BM, respectively. The BM of the TESH network is a 2D-mesh network. According to Lemma 1, the number of necessary virtual channels for phase-1 and phase-3 is 1. The routing of the message in source-BM and destination-BM is carried out separately. The virtual channels required in *phase*-1 and *phase*-3 can share each other. Intra-BM links between inter-BM links are used in sub-phases $2.i.1$ and $2.i.3$. Thus, sub-phases $2.i.1$ and $2.i.3$ utilize channels over intra-BM links, sharing the channels of either phase-1 or phase-3. The free links of the BM are used in inter-BM routing, i.e., sub-phases $2.i.2$ and $2.i.4$, and these links form a 2D-torus network for the higher level network. According to Lemma 2, the number of necessary virtual channels for this 2D-torus network is also 2. According to the structure of the TESH network, the inter-BM links other than the wrap-around links in the end-to-end node of the higher level networks used in sub-phases $2.i.2$ and $2.i.4$ routes messages like mesh network. Thus, the routing of messages in this part can share the channels of either phase-1 or phase-3. And the wrap-around links uses the another links of the 2D-torus in the higher level networks.

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

IV. Dynamic Communication Performance

The overall performance of a massively parallel computer system is affected by the performance of the interconnection network, as well as by the performance of the nodes. Continuing advances in VLSI 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 system. Therefore, the success of massively parallel computers is highly dependent on the efficiency of their underlying interconnection networks. The evaluation of dynamic communication performance of the TESH network, along with several other networks, is described in this section.

A. Performance Metrics

The dynamic communication performance of a massively parallel computer is characterized by *message latency* and *network throughput*. Message latency is the time required for a packet to traverse the network from source to destination. It refers to the time elapsed from the instant when the first flit is injected to the network from the source, to the instant when the last flit of the message is received at the destination. In wormhole routing, it is the average value of the time elapsed between injection of the header flit into the network from the source, and reception of the last unit of the data flit at the destination. 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.

Network throughput is the rate at which packets are delivered by the network for a particular traffic pattern. It refers to the maximum amount of information delivered per unit of time through the network. It also can be defined as the maximum traffic accepted by the network. Throughput depends on message length and network size. Therefore, throughput is usually normalized, dividing it by message length and network size. When throughput of various networks are compared by computer simulation and wormhole routing is used for switching, throughput can be measured in flits per node and per clock cycle.

For the network to have good performance, low latency and high throughput must be achieved. Zero-load latency is a lower bound on the average latency of a packet through the network. In zero-load, it is assumed that a packet never contends for network resources with other packets. Under this assumption, the average latency of a packet is its serialization latency plus its hop latency. Throughput is a property of the entire network, and depends on routing and flow control as much as on the topology. Maximum throughput is the upper bound throughput through a network. A resource is in saturation when the demands being placed on it are beyond its capacity for servicing those demands. A channel becomes saturated when the amount of data to be routed over the channel exceeds the bandwidth of the channel. The saturation throughput of a network is the smallest rate of traffic for which some channel in the network becomes saturated. If no channels are saturated, the network can carry more traffic. We also call this saturation throughput "maximum throughput".

B. Traffic Patterns

Traffic patterns are pairs of nodes that communicate. In an interconnection network, sources and destinations for messages form the traffic pattern. Traffic characteristics such as message length, message arrival time, and destination distribution have significant performance implications. Message destination distributions vary a great deal depending on the network topology and the application's mapping onto different nodes. Depending on the characteristics of an application, some nodes may communicate with each other more frequently than with others. Consequently, non-uniform traffic patterns are frequent, and cause uneven usage of traffic resources, significantly degrading the dynamic communication performance of the network.

When a hot spot occurs, a particular communication link experiences a much greater number of requests than the rest of the links – more than it can service. In a remarkably short period of time, the entire network may become congested. Hot spots are particularly insidious because they may result from the cumulative effects of very small traffic imbalances. Hot spots often occur because of the burst nature of program communication and data requirements and, therefore can provide a benchmark for interconnection networks. Bit permutation and computation [20] is a class of non-uniform traffic patterns which are very common in scientific applications, where the source node sends messages to a predefined destination. Both dimension order routing and bit permutation & communication create significant congestion under dimension order routing in the network, and when congestion occurs, the network throughput decreases precipitously. BPC distribution of traffic achieves the maximum degree of temporal locality and are also considered as benchmarks for interconnection networks. To observe the effect of nonuniformity, in this paper, we are specifically interested in how the nonuniform traffic patterns affect the dynamic communication performance of the TESH network.

1) Hot Spot traffic Pattern: A hot spot is a node that is accessed more frequently than other nodes in the uniform traffic distribution. In hot-spot traffic pattern, each node generates a random number. If that number is less than a threshold, the message is sent to the hot spot node. Otherwise it is sent to other nodes with a uniform distribution. Here, in uniform distribution, message destinations are chosen randomly with equal probability among the nodes in the network.

2) Complement Traffic: The binary representation of the node address is $b_{\beta-1}, b_{\beta-2}$ ... ... b_1, b_0. In complement traffic, the $\mathrm{node}\,(b_{\beta-1}, b_{\beta-2}$ ... ... $b_1, b_0)$ communicates with the $\mathrm{node}\,(\overline{b_{\beta-1}}, \overline{b_{\beta-2}},$ ... ... $\overline{b_2}, \overline{b_1}, \overline{b_0})$.

3) Perfect Shuffle Pattern: The node with binary coordinates $b_{\beta-1}, b_{\beta-2}$ ..... b_1, b_0 communicates with the $\mathrm{node}\,(b_{\beta-2}, b_{\beta-3},$ ..... $b_0, a_{\beta-1})$. That is, rotate left 1 bit.

C. Simulation Environment

The total number of nodes in a TESH is $N = 2^{2mL}$. If $m = 2$ and $L = 3$ then the total number of nodes of the TESH is 4096. That is, Level-3 TESH network has 4096 nodes. 64×64 mesh network and 64×64 torus network have 4096 nodes. We have evaluated the dynamic communication performance of TESH, mesh, and torus networks with 4096-nodes.

To evaluate dynamic communication performance, we have developed a wormhole routing simulator. In our simulation, we use a dimension-order routing algorithm

and various nonuniform traffic patterns. The dimension-order routing algorithm, which is exceedingly simple, provides the only route for the source-destination pair. 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.

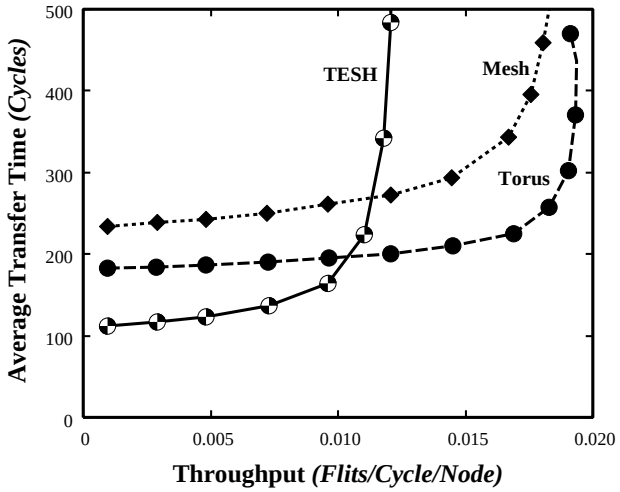
In the evaluation of dynamic communication performance, flocks of messages are sent in the network to compete for the output channels. For each simulation run, we have considered that the message generation rate is constant and the same for all nodes. Flits are transmitted at $20,000$ cycles; in each 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.

D. Dynamic Communication Performance Evaluation

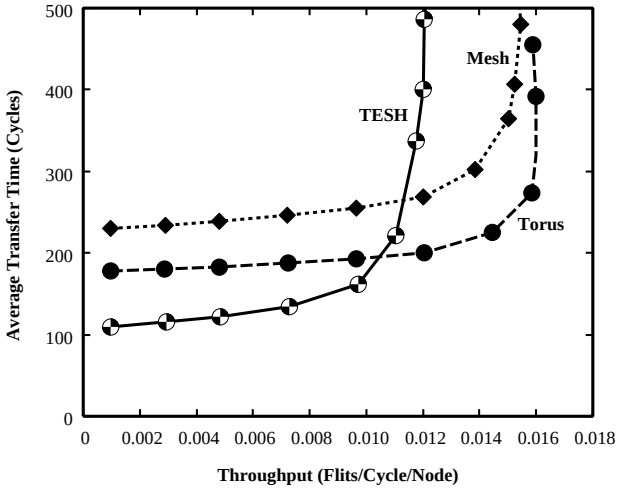
We have evaluated the dynamic communication performance of several 4096-node networks under various traffic patterns. We have evaluated the dynamic communication performance using dimension-order routing algorithm under three different traffic patterns: hot-spot, complement, and perfect shuffle.

1) Hot-Spot Traffic Pattern: For generating hot spot traffic we used a model proposed by Pfister and Norton [19]. According to this model, each node first generates a random number. If that number is less than a predefined threshold, the message will be sent to the hot-spot node. Otherwise, the message will be sent to other nodes, with a uniform distribution. Here, in uniform distribution, message destinations are chosen randomly with equal probability among the nodes in the network. However, in real application, it may happen that there are some packets (hot-spot packets) which remain in the network, and request rates are very high. Here, the simulations were carried out under the condition that, for TTN and TESH network, $\mathrm{Node}(n_5, n_4)(n_3, n_2)(n_1, n_0)$ are source nodes and $\mathrm{Node}(n_5, n_4)(0, 0)(0, 0)$ are hot-spot nodes. For mesh and torus networks, we divide the networks into 4×4 matrix. From each part, the node which is closest to the center is assumed to be the hot-spot node. The hot-spot flit generation probability are assumed to be $P_h = 0.02, 0.05, 0.10, 0.20$, and 0.30, i.e., the hot-spot percentages are assumed to be 2%, 5%, 10%, 20%, and 30% for all networks. For mesh and torus networks, we divide the networks into 4×4 matrix. From each part, the node which is closest to the center is assumed to be the hot-spot node, with same hot-spot percentages as listed above.

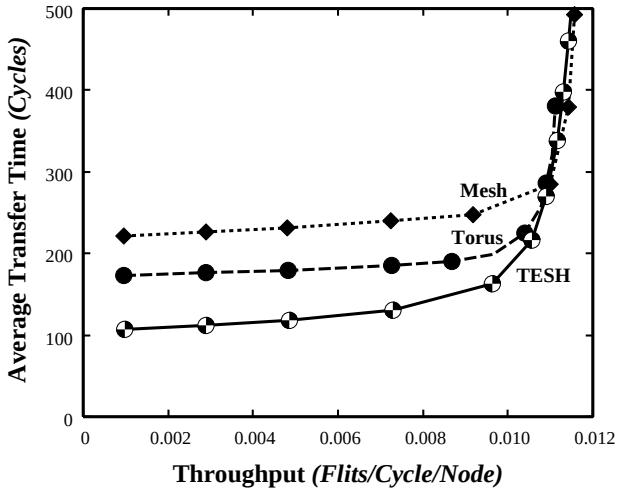
Figure 6 depicts the message latency versus network throughput curves for the hot-spot traffic pattern with different hot-spot percentage. Figure 6(a) represents the result of simulations using 2% hot-spot traffic. It is shown that the average transfer time of the TESH network is far lower than that of the mesh and torus networks. The maximum throughput of the TESH network is lower than



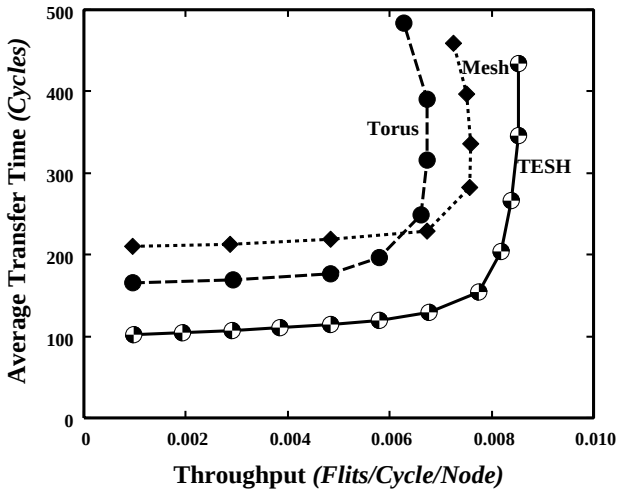
(a) 2% Hot-spot traffic



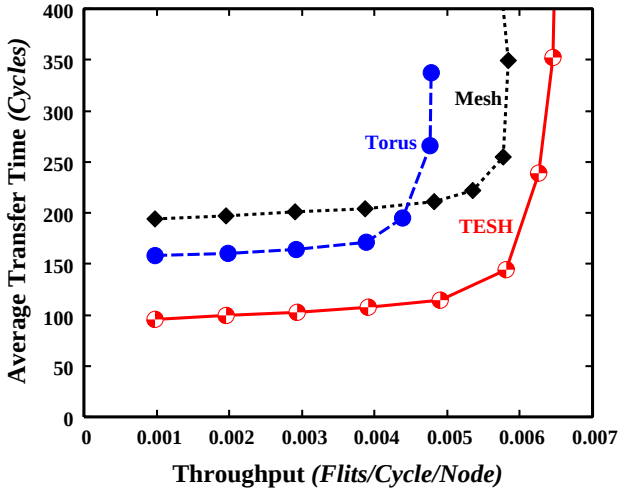
(b) 5% Hot-spot traffic



(c) 10% Hot-spot traffic



(d) 20% Hot-spot traffic



(e) 30% Hot-spot traffic

Figure 6. Dynamic communication performance of dimension-order routing with hot spot traffic pattern on various networks: 4096 nodes, 2 VCs, 16 flits, and $q = 0$

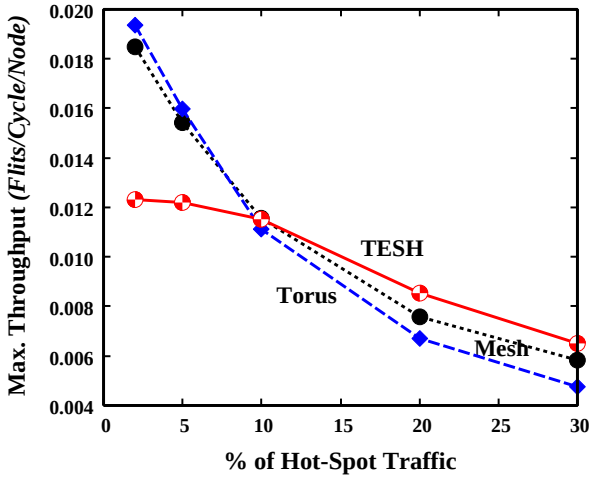


Figure 7. Effect of hot-spot traffic on maximum throughput of various networks: 4096 nodes, 2 VCs, 16 flits, and $q = 0$

that of those networks. Figure 6(b) represents the result of simulations using 5% hot-spot traffic. It is shown that the average transfer time of the TESH network is far lower than that of the mesh and torus networks. The maximum throughput of the TESH network is still lower than that of the mesh and torus networks. However, the relative difference in maximum throughput between TESH and mesh and torus networks is decreases with the increase of hot-spot traffic from 2% to 5%. Figure 6(c) represents the result of simulations using 10% hot-spot traffic. It is also shown that the average transfer time of the TESH network is far lower than that of the mesh and torus networks. The maximum throughput of the TESH network is higher than that of torus network and almost equal to that of mesh network. Figure 6(d) represents the result of simulations using 20% hot-spot traffic. It is shown that the average transfer time of the TESH network is far lower than that of the mesh and torus networks. The maximum throughput of the TESH network is far higher than that of the mesh and torus networks. Figure 6(e) represents the result of simulations using 30% hot-spot traffic. As usual, it is shown that the average transfer time of the TESH network is far lower than that of the mesh and torus networks. The maximum throughput of the TESH network is far higher than that of the mesh and torus networks. Therefore, with the increase of hot-spot traffic percentage in hot-spot traffic pattern, the hierarchical TESH network yields better dynamic communication performance than that of the conventional mesh and torus networks.

One interesting point to be noted here is that the relative difference in maximum throughput between torus and the hierarchical TESH network decreases with the increase of hot-spot traffic, and it is shown in Figure 6(d) and 6(e) that with 20% and 30% hot spot traffic the maximum throughput of the TESH networks are higher than that of the mesh and torus network.

To illustrate that interesting nature of TESH network, we have portrayed the maximum throughput with respect to percentage of hot-spot traffic in Figure 7. Each curve stands for a particular network. It is worthless to mention that with the increase of hot-spot traffic, the throughput of a particular network will decrease. It is shown in Figure 7 that with 10% hot-spot traffic, the maximum throughput of different network are almost equal. After 10% hot-spot traffic the maximum throughput of TESH network is higher than that of mesh and torus network. And the relative difference is increasing with the increase of hot-spot traffic.

To illustrate the effect of number of hot-spot node on dynamic communication performance, we have portrayed the dynamic communication performance of the TESH network with various number of hot-spot node in Figure 8. Each curve stands for the dynamic communication performance of the TESH network with a particular number of hot-spot node. It has already shown in figure 7 that after 10% hot-spot traffic, the dynamic communication

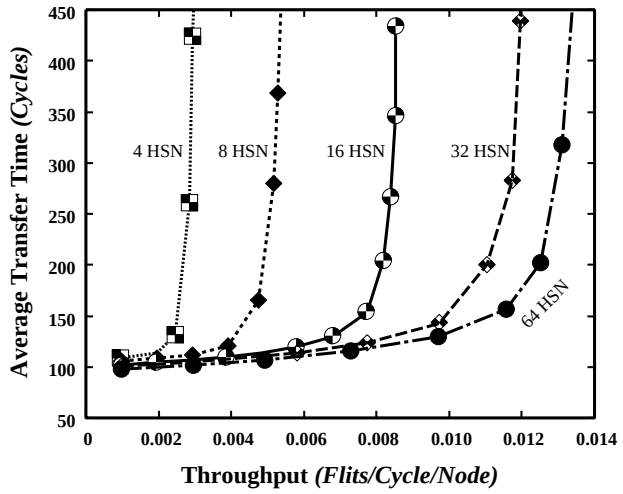


Figure 8. Effect of number of hot-spot node on dynamic communication performance of TESH network: 4096 nodes, 2 VCs, 16 flits, and $q = 0$

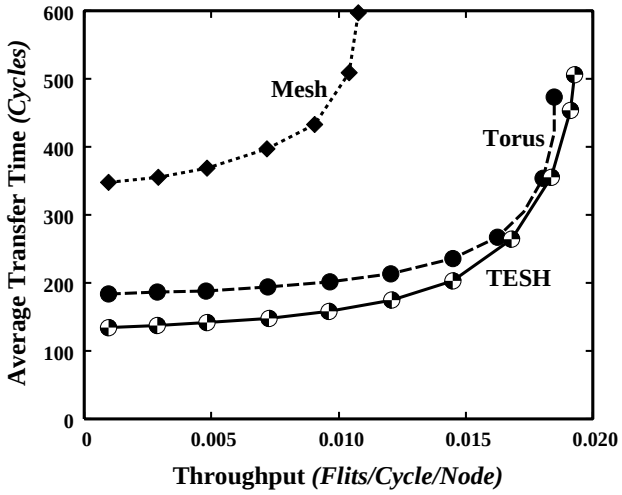


Figure 10. Dynamic communication performance of dimension-order routing with complement traffic pattern on various networks: 4096 nodes, 2 VCs, 16 flits, and $q = 0$

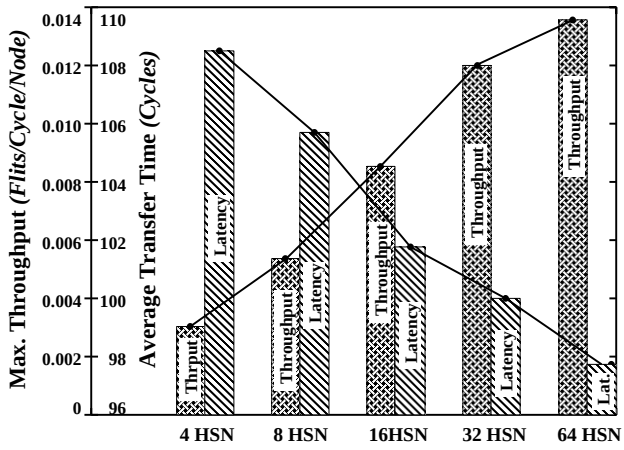


Figure 9. Maximum throughput and average transfer time at zero load of the TESH network with various number of hot-spot node: 4096 nodes, 2 VCs, 16 flits, and $q = 0$

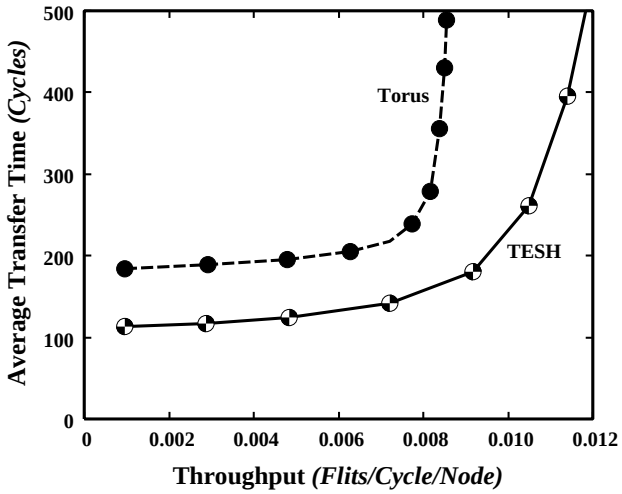


Figure 11. Dynamic communication performance of dimension-order routing with perfect shuffle traffic pattern on various networks: 4096 nodes, 2 VCs, 16 flits, and $q = 0$

performance of the TESH network is better than that of other networks. This is why, we have considered, 20% hot-spot traffic to show the effect of number of hot-spot node on dynamic communication performance. It is shown in Figure 8 that maximum throughput of the TESH network with 4 hot-spot node is lowest and with 64 hot-spot node is the highest than that of others. Also the average transfer time at zero load of the TESH network with 4 hot-spot node is highest and with 64 hot-spot node is the lowest than that of others. The maximum throughput and average transfer time at zero load of the TESH network with various number of hot-spot node is presented in Figure 9. It is seen that with the increase of number of hot-spot node, the average transfer time at zero load is decreasing and maximum throughput is increasing.

2) Bit Permutation and Communication (BPC) Traffic Pattern: Bit Permutation and Computation (BPC) [20] is a class of non-uniform traffic pattern, which are very common in scientific applications. BPC communication patterns take into account the permutations that are usually performed in parallel numerical algorithms [21], [23]. These distributions achieve the maximum degree of temporal locality and are also considered as benchmarks for interconnection networks. Among various BPC traffic patterns, in this paper, we have considered complement and perfect shuffle traffic pattern.

The complement is a particularly difficult traffic pattern, since it requires all packets to cross the network bisection. Therefore, the middle of the network is congested. Figure 10 portrays the results of simulations for the various network models under complement traffic pattern. From Figure 10, it is seen that the average transfer time of the TESH network is far lower than that of the mesh network and significantly lower than that of torus network. The maximum throughput of the TESH

network is far higher than that of mesh network and noticeably higher than that of torus network. Therefore, TESH network achieves better dynamic communication performance than the conventional mesh and torus network under the complement traffic pattern.

Figure 11 shows the comparison of dynamic communication performance between TESH and torus network under perfect shuffle traffic pattern. It is shown that the average transfer time of the TESH network is far lower than that of the torus network. And, the maximum throughput of the TESH network is far higher than that of the torus network. Therefore, TESH network yields better dynamic communication performance than that of torus network.

Torus is a suitable network for parallel computers, due to its symmetry and regularity. However, the length of the longest wire is a limiting factor for a network with thousands or millions of nodes. The operating speed of a network is limited by the physical length of links. With the cost of some additional short length links in the TESH network, the dynamic communication performance is better than that of torus network under various non-uniform traffic patterns.

In all the traffic patterns presented in this paper, the average transfer time at zero load of the TESH network is remarkably lower than that mesh and torus networks. The maximum throughput of the TESH network is higher than that of those networks. In other BPC traffic patterns, the average transfer time at zero load of the TESH network is remarkably lower than that mesh and torus networks. However, the maximum throughput of the TESH network is a little bit worse than that of those networks. The main reason is how the communication takes place for a particular traffic pattern in an interconnection network. This low latency and high throughput of the TESH network will make it a good choice for future generation massively parallel computer systems.

V. Conclusion

A deadlock free routing algorithm using dimension-order routing with a minimum number of virtual channels has been proposed for the TESH network. It has been proved that 2 virtual channels per physical channel are sufficient for the deadlock free routing algorithm of the TESH network; 2 is also the minimum number of virtual channels for dimension-order routing. By using the routing algorithm described in this paper and various nonuniform traffic pattern, we have evaluated the dynamic communication performance of the TESH network as well as that of several other commonly used networks. The average transfer time at zero load of TESH network is lower than that of the mesh and torus networks in all non-uniform traffic patterns. In hot-spot traffic pattern, the maximum throughput of the TESH network is higher than that of those networks with the increase of hot-spot traffic. The maximum throughput is also higher than that of those networks under various bit permutation and communication traffic patterns. A comparison of dynamic communication performance reveals that the TESH network outperforms the mesh and torus networks because it yields low latency and high throughput, which are indispensable for high performance massively parallel computers.

The performance degradation with the presence of large number of long length links is removed by replacing the long length electronic links by optical links. The dynamic communication performance evaluation of this opto-electronic (hybrid) TESH network is kept in mind as future works.

Acknowledgment

The authors are grateful to the anonymous expert reviewers for their extremely valuable comments, which helped to greatly improve the clarity of this paper.

References

[1] W.J. Dally, "Performance Analysis of k-ary n-cube Interconnection Networks", *IEEE Trans. on Computers*, vol. 39, no. 6, (1990) 775–785.

[2] Y.R. Potlapalli, "Trends in Interconnection Network Topologies: Hierarchical Networks", *Int'l. Conf. on Parallel Processing Workshop*, (1995) 24–29.

[3] A. El-Amawy and S. Latifi, "Properties and Performance of Folded Hypercube", *IEEE Trans. on Parallel and Distributed System*, vol. 2, no. 1, (1991) 31–42.

[4] A. Esfahanian, L.M. Ni, and B.E. Sagan, "The Twisted n-Cube with Application to Multiprocessing", *IEEE Trans. on Computers*, vol. 40, no. 1, (1991) 88–93.

[5] J.M. Kumar and L.M. Patnaik, "Extended Hypercube: A Hierarchical Interconnection Network of Hypercube", *IEEE Trans. on Parallel and Distributed System*, vol. 3, no. 1, (1992) 45–57.

[6] N.F. Tzeng and S. Wei, "Enhanced Hypercube", *IEEE Trans. on Computers*, vol. 40, no. 3, (1991) 284–294.

[7] S.G. Ziavras, "A Versatile Family of Reduced Hypercube Interconnection Network", *IEEE Trans. on Parallel and Distributed System*, vol. 5, no. 11, (1994) 1210–1220.

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

[9] V.K. Jain and S. Horiguchi, "VLSI Considerations for TESH: A New Hierarchical Interconnection Network for 3-D Integration", *IEEE Trans on VLSI Systems*, vol.6, no. 3, pp. 346-353, 1998.

[10] M. Maziarz, and V.K. Jain, "Automatic reconfiguration and yield of the TESH multicomputer network", *IEEE Trans. on Computers*, pp. 963-972, 2002.

[11] Y. Miura, "Wormhole Routing for Hierarchical Interconnection Networks", *Ph.D. Dissertation*, School of Information Science, JAIST, 2002.

[12] W.J. Dally and C.L. Seitz, "Deadlock Free Message Routing in Multiprocessor Interconnection Networks", *IEEE Trans. on Computers*, vol.36, no.5, pp. 547–553, 1987.

[13] L.M. Ni and P.K. McKinley, "A Survey of Wormhole Routing Techniques in Direct Networks", *IEEE Computer*, vol.26, no.2, pp.62–76, 1993.

[14] W.J. Dally and C.L. Seitz, The Torus Routing Chip, *Journal of Distributed Computing*, vol. 1, no. 3, pp. 187–196, 1986.

[15] W.J. Dally, "Virtual-Channel Flow Control", *IEEE Trans. on Parallel and Distributed System*, vol.3, no.2, pp.194-205, 1992.

[16] Y. Miura and S. Horiguchi, "A Deadlock-Free Routing for Hierarchical Interconnection Network: TESH", *Proc. of the 4th Int'l. Conf. on HPC Asia*, pp.128-133, 2000.
[17] M.M. Hafizur Rahman and Susumu Horiguchi, "A deadlock-free routing algorithm using minimum number of virtual channels and application mappings for Hierarchical Torus Network", *International Journal of High Performance Computing and Networking*, vol.4, no.3/4, pp.174-187, 2006.
[18] M.M. Hafizur Rahman and Susumu Horiguchi, "High Performance Hierarchical Torus Network under Matrix Transpose Traffic Patterns", *Proc. of the 7th Int'l. Symposium on Parallel Architectures, Algorithms and Networks (ISPAN04)*, pp.111-116, 2004.
[19] G.F. Pfister and V.A. Norton, "Hot Spot Contention and Combining in Multistage Interconnection Networks", *IEEE Trans. on Computers*, vol. 34, no. 10, pp. 943-948, 1985.
[20] M. Grammatikakis, D.F. Hsu, M. Kratzel and J.F. Sibeyn, "Packet routing in fixed connection networks: a survey", *Journal of Parallel and Distributed Computing*, vol. 54, no. 2, pp. 77–132, 1998.
[21] Andrew A. Chien and Jae H. Kim, "Planer-Adaptive Routing: Low-cost Adaptive Networks for Multiprocessors", *Journal of the ACM*, vol.42, no.1, pp.91–123, 1995.
[22] K. Bolding, M. Fulgham, and L. Synder, "The Case of Chaotic Adaptive Routing", *IEEE Trans. on Computers*, vol. 46, no. 12, pp. 1281–1292, 1997.
[23] P.R. Miller, "Efficient Communications for Fine-Grain Distributed Computers", *Ph.D. Dissertation, Southampton University*, U.K., 1991.

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 received his M.Sc. and Ph.D. degree in Information Science from the Japan Advanced Institute of Science and Technology (JAIST) in 2003 and 2006, respectively. He is currently serving as an assistant professor in the Dept. of CSE at KUET. He was also a visiting researcher in the School of Information Science at JAIST in 2008. His current research include interconnection networks, especially hierarchical interconnection networks and optical switching networks. Dr. Rahman is member of IEICE of Japan and 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 Japan Advanced Institute of Science and Technology (JAIST) in 1994 and 1997, respectively. He is currently an 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 from 2002 to 2006. His research interest has been mainly concerned with parallel computer architecture, interconnection networks, GRID architecture, and high performance computing on parallel machines. Dr. Inoguchi is a member of IEEE and IPS of Japan.

Yukinori Sato received the BS, MS, and Ph.D. degree in Information Science from Tohoku University in 2001, 2003, 2006 respectively. From 2006, he engaged in embedded processor system design in Sendai Software Development center of FineArch Inc. and also became a joint research member at Tohoku University. From 2007, he has been working at JAIST as an assistant professor. His research interests include high-speed and low-power computer architectures and reconfigurable computing. Dr. Sato is a member of the IEEE, ACM, IEICE and IPS of Japan.

Yasuyuki Miura received the bachelor's degree in Tohoku University in 1997, and the MS and Ph.D degree in Japan Advanced Institute of Science and Technology (JAIST) in 1999 and 2002. Then he had worked in the NICT, Japan until December 2004. From January 2005 to March 2006, he was researcher of the Japan Science and Technology Agency (JST). Since April 2006, he is lecturer in the Shonan Institute of Technology. His research interests include parallel processing, interconnection network, and MPEG video streaming. He is a member of the IEEE and the IEEE Computer Society.

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 CS, and a member of the IPS and IASTED.