LETTER

On Nonuniform Traffic Pattern of Modified Hierarchical 3D-Torus Network

M.M. Hafizur RAHMAN[†a)], **Yukinori SATO**[†], ***and*** **Yasushi INOGUCHI**[†], ***Members***

SUMMARY A **M**odified **H**ierarchical **3D-T**orus (**MH3DT**) network is a 3D-torus network consisting of multiple basic modules, in which each basic module itself is a 3D-torus network. Inter-node communication performance has been evaluated using dimension-order routing and 2 virtual channels (VCs) under uniform traffic patterns but not under non-uniform traffic patterns. In this paper, we evaluate the inter-node communication performance of MH3DT under five non-uniform traffic patterns and compare it with other networks. We found that under non-uniform traffic patterns, the MH3DT yields high throughput and low latency, providing better inter-node communication performance compared to H3DT, TESH, mesh, and torus networks. Also, we found that non-uniform traffic patterns have higher throughput than uniform traffic in the MH3DT network.

key words: *MH3DT network, deadlock-free routing, non-uniform traffic, Inter-node communication performance*

1. Introduction

The **MH3DT** network [1] consists of basic modules (BMs) which are themselves 3D-tori ($m \times m \times m$), hierarchically interconnected in a 3D-torus ($n \times n \times n$) network. In the MH3DT network, both the BMs and the interconnection of higher levels have toroidal interconnections. The H3DT network [2] is modified by replacing the 3D-mesh of its BM by a 3D-torus network.

The inter-node communication performance of the MH3DT with dimension-order routing under uniform traffic was evaluated in a previous study, and it proved to be better than other networks when using a large buffer. With a small buffer, it is not good [1]. However, the inter-node communication performance of the MH3DT under various non-uniform traffic patterns had not yet been evaluated. The main objective of this paper is to investigate the impact of non-uniform traffic patterns on the MH3DT network.

The remainder of the paper is organized as follows. In Sect. 2, we briefly describe the basic structure of the MH3DT network. We review the routing algorithm in Sect. 3, and its freedom from deadlock [1]. The inter-node communication performance of the MH3DT under various non-uniform traffic patterns is discussed in Sect. 4. Finally, Sect. 5 presents the conclusion.

2. Interconnection of the MH3DT Network

The BM is a 3D-torus 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. As shown in Fig. 1, 3 nodes ($0 \le a_z \le 2$) have 2 free ports, 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.

Successively higher level networks are built by recursively interconnecting immediate lower level subnetworks in a 3D-torus of size ($n \times n \times n$), where n is also a positive integer. Figure 2 illustrated a Level-2 MH3DT network consisting of 64 BMs as a ($4 \times 4 \times 4$) 3D-torus. A node that has free links which are used for the interconnection of higher level is known as gate node. 2^q gate nodes are used for higher level interconnection, where q is the inter-level connectivity. As each xy-plane of the BM has 4 gate nodes, $0 \le q \le 2$. By using the parameters m, n, L, and q, we can define the MH3DT network as MH3DT(m, n, L, q). Figure 2 portrays MH3DT$(4, 4, 2, 2)$ network. The address of a node

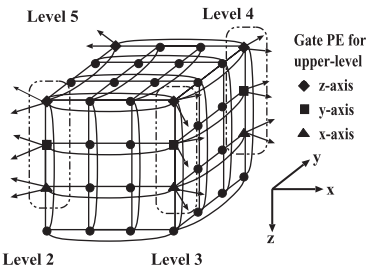

Fig. 1 Basic module.

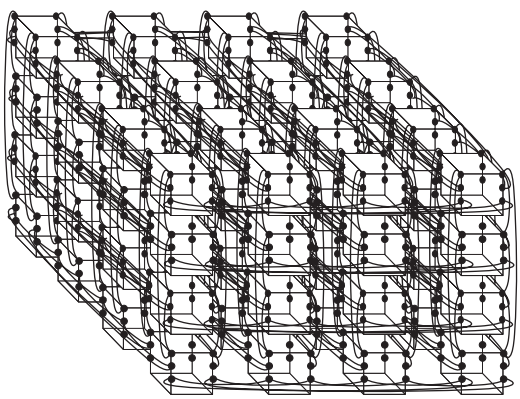

Fig. 2 Interconnection of a Level-2 MH3DT network.

Manuscript received December 7, 2010.
Manuscript revised January 31, 2011.

[†]The authors are with the Center for Information Science, JAIST, Nomi-shi, 923–1292 Japan.
a) E-mail: rahman@jaist.ac.jp

DOI: 10.1587/transinf.E94.D.1109

at Level-L is represented by:

$$A^L = (a_z^L)(a_y^L)(a_x^L) \qquad L \text{ is level number.} \tag{1}$$

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

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

3. Routing Algorithm for MH3DT Network

3.1 Dimension-Order Routing Algorithm

In this section, we review the routing algorithm from our previous study [1] for the convenience of readers. Since dimension-order routing is used in the MH3DT network, messages are routed first in the z-direction then in 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 $s_\alpha, s_{\alpha-1}, s_{\alpha-2}, \ldots \ \ldots, s_1, s_0$, a destination node address be $d_\alpha, d_{\alpha-1}, d_{\alpha-2}, \ldots \ \ldots, d_1, d_0$, and a routing tag be $t_\alpha, t_{\alpha-1}, t_{\alpha-2}, \ldots \ \ldots, t_1, t_0$, 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.

```
Routing MH3DT(s,d);
source node address:s_α, s_{α-1}, s_{α-2}, ..., s_1, s_0
destination node address: d_α, d_{α-1}, d_{α-2}, ..., d_1, d_0
tag: t_α, t_{α-1}, t_{α-2}, ..., t_1, t_0
  for i = α : 3
    if (t_i > 0 and t_i ≤ 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 ≥ −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 ≠ 0 or distance ≠ 0) do
        if (j = 2), gate-node = z-axis gate-node of Level-⌊i/3⌋; endif
        if (j = 1), gate-node = y-axis gate-node of Level-⌈i/3⌉; endif
        if (j = 0), gate-node = x-axis gate-node of Level-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 ≤ 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 ≥ −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 ≠ 0 or distance_2 ≠ 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 ≠ 0 or distance_1 ≠ 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 ≠ 0 or distance_0 ≠ 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.

3.2 Deadlock-Free Routing

Since the hardware cost increases with the increase of VCs, the unconstrained use of VC is not cost-effective in parallel computers. Therefore, deadlock-free routing for an arbitrary network with a minimum number of VCs is preferred. To prove that the routing algorithm for the MH3DT is deadlock-free using minimum number of VCs, we divide it into 3 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 path to the destination node.

The number of VCs required to make the routing algorithm deadlock-free for the MH3DT is determined using Lemma 1. A theorem is also stated below without proof, where the proof was presented in [1].

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 VCs [1].

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

4. Inter-Node Communication Performance

Low communication performance of the underlying interconnection network will severely limit the speed and efficiency of the entire parallel computers.

4.1 Performance Metrics

The inter-node communication performance of a parallel

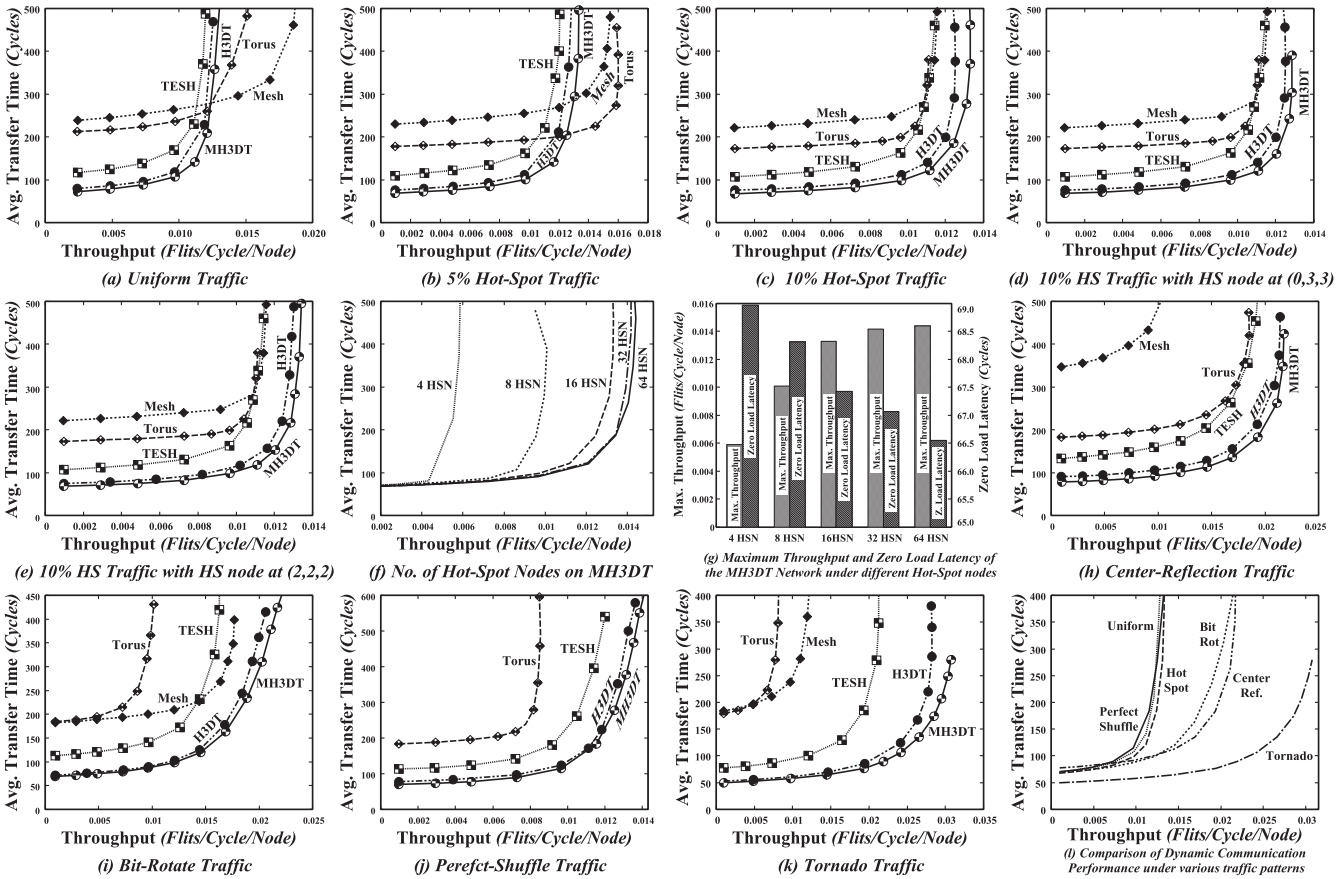

Fig. 4 Inter-node communication performance of dimension-order routing with different traffic patterns on various networks: 4096 nodes, 2 VCs, 16 flits, and $q = 2$.

computers is characterized by *latency* and *throughput*. Message latency refers to the time elapsed from the instant when the first flit (header) is injected to the network from the source, 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.

4.2 Simulation Environment

To evaluate inter-node communication performance, we have developed a wormhole routing simulator. 2 VCs per physical link are simulated, and the VCs are arbitrated by a round robin algorithm. Packet size is 16 flits and 2 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 vice-versa. Extensive simulations for several 4096 node networks have been carried out under hot-spot, tornado, center-reflection, bit-rotate, and perfect shuffle using dimension-order routing. To show the superiority of the MH3DT with non-uniform traffic over the uniform traffic, we have recalled uniform traffic [1].

4.3 Inter-Node Communication Performance Evaluation

When a hot spot occurs due to bursty nature of program communication and data requirements, the entire network may become congested in a remarkably short period of time. Other Bit Permutation and Computation nonuniform traffic patterns are very common in scientific applications and parallel numerical algorithms.

In uniform traffic, every node sends messages to every other node with equal probability. As shown in Fig. 4 (a), recalled from [1], that the zero load latency of the MH3DT is lower than that of the H3DT, TESH, mesh, and torus networks. The maximum throughput of the MH3DT is a significantly higher than that of H3DT and TESH networks and it is far lower than that of mesh and torus networks with a smaller buffer.

In hot-spot traffic, each node first generates a random number and if it is less than a predefined threshold, the message will be sent to the hot-spot node (HSN). Otherwise, it will be sent to other nodes with uniform traffic [4]. We have considered 16 HSNs closer to the center for all networks. The hot-spot flit generation probability are assumed to be $P_h = 0.05$ and 0.10,i.e., 5% and 10% hot-spot traffic. Figure 4 (b) and (c) depict the latency versus throughput curves

for the 5% and 10% hot-spot traffic, respectively. It is shown that the zero load latency of the MH3DT is lower than that of the H3DT, TESH, mesh, and torus networks. The maximum throughput of the MH3DT is higher than that of H3DT and TESH networks, however, it is lower than that of mesh and torus under 5% hot-spot traffic. With the increase of hot-spot traffic, the relative difference in maximum throughput between MH3DT and other networks is decreases as shown in Fig. 4 (c). It is shown that the maximum throughput of the MH3DT network is higher than that of H3DT, TESH, mesh, and torus networks. The question may arise that the throughput may be changed depending on the HSN selection. We have evaluated the inter-node communication performance of the MH3DT network selecting the HSN with co-ordinate at (0,3,3) and (2,2,2) and plotted in Fig. 4 (d) to (e). The same scenario is observed with the change of hot-spot position. To illustrate the effect of number of HSN on performance, we have portrayed it with various number of HSN under 10% hot-spot traffic in Fig. 4 (f) and (g). It is seen that with the increase of number of HSN, the zero load latency is decreasing and maximum throughput is increasing.

In center-reflection traffic [5], a source at (x, y, z) sends a message to a destination at $(k-x-1, k-y-1, k-z-1)$, where k is the number of nodes in one direction. Figure 4 (h) depicts the simulations results under center reflection traffic. It is seen that the zero load latency and the maximum throughput of MH3DT is lower and higher, respectively than that of the H3DT, TESH, mesh, and torus networks. In this traffic, center of the network is congested because all the packets cross the bisection. MH3DT yields better inter-node communication performance than that of other networks even with the adversity of congestion as shown in Fig. 4 (h).

In bit-rotate traffic, the node with binary coordinates $b_{\beta-1}, b_{\beta-2} \ \ldots \ b_1, b_0$ communicates with the Node$(b_0, b_{\beta-1}, \ \ldots \ b_2, b_1)$, i.e., rotate right 1 bit [6]. From the simulation result portrayed in Fig. 4 (i), it is seen that the zero load latency of the MH3DT is lower than that of the H3DT, TESH, mesh, and torus networks. The maximum throughput of the MH3DT is far higher than that of those networks. Therefore, MH3DT yields better inter-node communication performance than of those networks under the bit rotate traffic. The similar scenario is observed in perfect shuffle traffic as shown in Fig. 4 (j). The node with binary coordinates $b_{\beta-1}, b_{\beta-2} \ \ldots \ b_1, b_0$ communicates with the node $\left(b_{\beta-2}, b_{\beta-3}, \ldots b_1, b_0, a_{\beta-1}\right)$, i.e., rotate left 1 bit [7].

In tornado traffic [6], the node (x, y, z) only sends packets to node $\{(x + [k/2] - 1) \bmod k, y, z\}$. This pattern is designed as an adversary of torus network. From the simulation result depicted in Fig. 4 (k), it is seen that the zero load latency and the maximum throughput of MH3DT is lower and higher than that of the H3DT, TESH, mesh, and torus networks, respectively.

It is seen in Fig. 4 (l) that the maximum throughput of the MH3DT under all non-uniform traffic patterns, such as tornado, center reflection, bit rotate, hot-spot, and perfect-shuffle, is higher than uniform traffic. Tornado traffic results the highest throughput and lowest zero load latency. Even with the most congested traffic pattern (center reflection) and most imbalanced traffic pattern (hot-spot), MH3DT yields higher throughput than it does with uniform traffic. Therefore, MH3DT results better inter-node communication performance under non-uniform traffic patterns.

5. Conclusion

Simulation experiments for inter-node communication performance reveal that the MH3DT outperforms the H3DT, TESH, mesh, and torus networks, achieving low latency and high throughput which are indispensable for high-performance parallel computers. The inter-node communication performance of the MH3DT network under non-uniform traffic is higher than its performance when a uniform traffic pattern is used. Therefore, MH3DT is a suitable network for non-uniform traffic. The future work focuses on the replacement of long electronic links by optical links, i.e., to study of opto-electronic (hybrid)-MH3DT network [8].

Acknowledgment

The preliminary version of this paper has been presented in the 2^{nd} Int'l. workshop UPDAS, 2010. This work is supported in part by JSPS fellowship program and Grand-in-Aid for Scientific Research, 21-09058, 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 Mary Ann Mooradian, Lecturer, JAIST, Technical Comm. Program.

References

[1] M.M. Hafizur Rahman, Y. Inoguchi, and S. Horiguchi, "Modified hierarchical 3D-torus network," IEICE Trans. Inf. & Syst., vol.E88-D, no.2, pp.177–186, Feb. 2005.

[2] S. Horiguchi and T. Ooki, "Hierarchical 3D-torus interconnection network," ISPAN'00, pp.50–56, USA, 2000.

[3] 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.

[4] G.F. Pfister and V.A. Norton, "Hot spot contention and combining in multistage interconnection networks," IEEE Trans. Comput., vol.34, no.10, pp.943–948, 1985.

[5] L. Schwiebert and R. Bell, "Performance tuning of adaptive wormhole routing through selection function choice," JPDC, vol.62, no.7, pp.1121–1141, 2002.

[6] W.J. Dally and B. Towles, Principles and Practices of Interconnection Networks, MK Publishers, 2004.

[7] H.H. Najaf-abadi and H. Sarbazi Azad, "The effects of adaptivity on the performance of the OTIS-hypercube under different traffic patterns," Proc. IFIP Int'l. Conf. NPC2004, LNCS, pp.390–398, Springer, 2004.

[8] L. Xiao and K. Wang, "Reliable opto-electronic hybrid interconnection network," Proc. 9th I-SPAN, pp.239–244, 2008.