HAL
archives-ouvertes.fr

On Dynamic Communication Performance of a Hierarchical 3D-Mesh Network

M. Rahman, Asadullah Shah, Yasushi Inoguchi

▶ **To cite this version:**

M. Rahman, Asadullah Shah, Yasushi Inoguchi. On Dynamic Communication Performance of a Hierarchical 3D-Mesh Network. James J. Park; Albert Zomaya; Sang-Soo Yeo; Sartaj Sahni. 9th International Conference on Network and Parallel Computing (NPC), Sep 2012, Gwangju, South Korea. Springer, Lecture Notes in Computer Science, LNCS-7513, pp.180-187, 2012, Network and Parallel Computing. <10.1007/978-3-642-35606-3_21>. <hal-01551354>

HAL Id: hal-01551354
https://hal.inria.fr/hal-01551354

Submitted on 30 Jun 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution| 4.0 International License

On Dynamic Communication Performance of a Hierarchical 3D-Mesh Network

M. M. Hafizur Rahman*, Asadullah Shah* and Yasushi Inoguchi‡

*Dept. of Computer Science, KICT, IIUM, Gombak-53100, Malaysia
‡Center for Information Science, JAIST, Ishikawa 923-1292, Japan
{hafizur,asadullah}@iium.edu.my,inoguchi@jaist.ac.jp

Abstract. A Hierarchical 3D-Mesh (H3DM) Network is a 2D-mesh network of multiple basic modules (BMs), in which the basic modules are 3D-torus networks that are hierarchically interconnected for higher-level networks. In this paper, we evaluate the dynamic communication performance of a Hierarchical 3D-Mesh (H3DM) network using a deadlock-free routing algorithm with minimum number of virtual channels under the uniform and non-uniform traffic patterns; and compare it with other networks to show the superiority of the H3DM network over other networks. We have also evaluated the dynamic communication performance of the mesh and torus networks. It is shown that H3DM network yields low average transfer time than that of mesh and torus networks. The trade-off between throughput and latency of these networks shown that H3DM network provide better dynamic communication performance than that of mesh and torus networks before saturation.

Keywords: Interconnection network, H3DM network, Deadlock-free routing algorithm, Traffic patterns, Dynamic communication performance.

1 Introduction

High-performance computing is necessary in solving the grand challenge problems in many areas such as development of new materials and sources of energy, development of new medicines and improved health care, strategies for disaster prevention and mitigation, weather forecasting, and for scientific research including the origins of matter and the universe. This makes the current supercomputer changes into massively parallel computer (MPC) systems with thousands of node (Kei, Cray XT5-HE), that satisfy the insatiable demand of computing power. In near future, we will need computer systems capable of computing at the petaflops or exaflops level. To achieve this level of performance, we need MPC with tens of thousands or millions of nodes. Interconnection networks play a crucial role in the performance of MPC systems [1]. Many recent experimental and commercial parallel computers use direct networks for low latency and high bandwidth of interprocessor communication. For future MPC with millions of nodes, the large diameter of conventional topologies is intolerable. Hence, the hierarchical interconnection network (HIN) provides an alternative efficient way in which several network topologies can be integrated [2] together to construct the future MPC [2] . A variety of hypercube based HINs found in the litera-

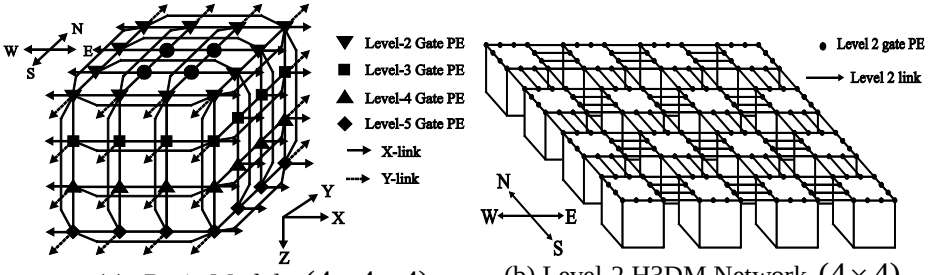
ture, however, its huge number of physical links make it difficult to implement. To alleviate this problem, *k*-ary *n*-cube based HIN [3, 4] is a plausible alternative way.

A Hierarchical 3D-Mesh (H3DM) Network [5] is a 2D-mesh network $(n \times n)$ of multiple basic modules (BMs), in which the BMs are 3D-torus networks $(m \times m \times m)$ that are hierarchically interconnected for higher-level networks. Wormhole routing [6] has become the dominant switching technique used in contemporary multicomputers. This is because it has low buffering requirements and it makes latency independent of the message distance. Deterministic, dimension-order routing is popular in MPC because it has minimal hardware requirements and allows the design of simple and fast routers. 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 (VCs) [7] are used to solve the problem of deadlock in wormhole-routed networks. Since the hardware cost increases as the number of VCs increases, the unconstrained use of VCs is not cost-effective in MPC systems. The static network performance of the H3DM network is evaluated and presented in [5]. And in our another study, we have presented a deadlock-free routing algorithm for the H3DM network using 2 VCs [8]. The main objective of this paper is to study the dynamic communication performance of the H3DM network.

The remainder of this paper is organized as follows. In Section 2, we briefly describe the basic structure of the H3DM network. In Section 3, we recall the deadlock-free dimension order routing for the H3DM network. Section 4 discusses the evaluation of dynamic communication performance. Finally, in Section 5, we conclude the results presented in this paper.

2 Interconnection of the H3DM network

The *H3DM* network [5] is a HIN consisting of multiple BM that are hierarchically interconnected for higher level networks. The BM of the H3DM network is a 3D-torus network of size $(m \times m \times m)$, where *m* is a positive integer. *m* can be any value, however the preferable one is $m = 2^p$, where *p* is a positive integer. The BM of a $(4 \times 4 \times 4)$ torus, as depicted in Figure 1(a), has some free ports at the contours of the *xy*-plane. A $(m \times m \times m)$ BM has $(4 \times m^2)$ free ports for higher level interconnection. All free ports, typically one or two, of the exterior Processing Elements (PEs) are used for inter-BM connections to form higher level networks. Successively higher level networks are built by recursively interconnecting lower level subnetworks in a 2D-mesh of size $(n \times n)$, where *n* is also a positive integer. As portrayed in Figure 1(b), a Level-2 H3DM network can be formed by interconnecting 16 BMs as a (4×4) 2D-mesh network. Similarly, a Level-*3* network can be formed by interconnecting n^2 Level-*2* subnetworks, and so on. Each BM is connected to its logically adjacent BMs.



(a) Basic Module $(4\times4\times4)$ (b) Level-2 H3DM Network (4×4)

Fig. 1. Interconnection of a H3DM Network

For each higher level interconnection of H3DM network, a BM must use $4m(2^q)$ of its free links: $2m(2^q)$ free links for y-direction and $2m(2^q)$ free links for x-direction interconnections. Here, $q \in \{0,1,..., p\}$, is the inter-level connectivity, where $p = \lfloor \log_2^m \rfloor$. *q = 0* leads to minimal inter-level connectivity, while *q = p* leads to maximum inter-level connectivity. It is depicted in Figure 1(a) that the $(4\times4\times4)$ BM has $(4\times4^2 = 64)$ free ports. With *q = 0,* $(4\times4\times2^0 = 16)$ free links are used for each level interconnection, 8 for y-direction and 8 for x-direction interconnections as portrayed in Figure 1(b). The highest level network which can be built from a $(m\times m\times m)$ BM is $L_{\max} = 2^{p-q}+1$. With q = 0, Level-5 is the highest possible level to which a $(4\times4\times4)$ BM can be interconnected. The total number of nodes in a network having $(m\times m\times m)$ BMs and $(n\times n)$ higher level is $N = [m^3 \times n^{2(L_{\max}-1)}]$. Thus, the maximum number of nodes which can be interconnected by the H3DM network is $N = [m^3 \times n^{2(2^{p-q})}]$. If *m = 4*, *n = 4*, and *q = 0*, then $N = 4^3 \times 4^8 = 4194304$, i.e., about 4.2 million.

The address of a PE at Level-*L* H3DM network is represented by Eq. 1.

$$A^L = \begin{cases} (a_z)(a_y)(a_x) & \text{if } L = 1 \\ (a_y{}^L)(a_x{}^L) & \text{if } L \geq 2 \end{cases} \tag{1}$$

More generally, in a Level-L H3DM, the node address is represented by:

$$\begin{aligned} A &= A^L A^{L-1} A^{L-2} .....A^3 A^2 A^1 = a_\alpha a_{\alpha-1} a_{\alpha-2} a_{\alpha-3} .....a_2 a_1 a_0 \\ &= (a_{2L} a_{2L-1})(a_{2L-2} a_{2L-3}).....(a_4 a_3)(a_2 a_1 a_0) \end{aligned} \tag{2}$$

Here, the total number of digits is $\alpha = 2L+1$, where L is the level number. In particular, ith group $(a_{2i}\ a_{2i-1})$ indicates the location of a Level-*(i-1)* subnetwork within the ith group to which the node belongs; $2 \leq i \leq L$.

3 Routing Algorithm for H3DM Network

3.1 Routing Algorithm

Routing of messages in the H3DM network 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 [10]. For messages routing using dimension-order routing in H3DM network, first find the nonzero offset in the most significant position by subtracting the current address from the destination. Then make a step towards nullifying the offset by sending the packet in descending order. When the offset along a dimension is zero, then the routing message is switched over to the next dimension. Routing dimension is strictly followed in the dimension order routing. Routing at the higher level H3DM is performed first in the *y*-direction and then in the *x*-direction. In a BM, the routing order is *z*-direction, *y*-direction, and *x*-direction, respectively.

```
Routing H3DM (s,d);
source node address: s_α, s_{α-1}, ...s_2, s_1, s_0
destination node address: d_α, d_{α-1}, ... d_1, d_0
tag: t_α, t_{α-1}, t_{α-2}, ...t_2, t_1, t_0
 for i = n : 3
  if (i /2 = 0 and t_i > 0), routedir = North, endif;
  if (i/2 = 0 and t_i < 0), routedir = South, endif;
  if (i%2 = 1 and t_i > 0), routedir = East, endif;
  if (i%2 = 1 and t_i < 0), routedir = West, endif;
   while (t_i ≠ 0) do
    Nz = outlet_z(s, d, L, routedir)
    Ny = outlet_y(s, d, L, routedir)
    Nx = outlet_x(s, d, L, routedir)
    BM_Routing (Nz, Ny, Nx)
    if routedir = North or East
                move packet to next BM;
    if routedir = South or West
                move packet to previous BM;
    t_i = t_i – 1;
   endwhile
  endfor
   BM_Routing (t_z, t_y, t_x)
end
```

```
BM_Routing ();
BM_tag t_2, t_1, t_0 = (r_2, r_1, r_0) – (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 = 1 - m)
     movedir = positive; endif;
 if(t_i > 0 and t_i = m - 1) or (t_i < 0 and t_i ≥ - m/2)
     movedir = negative; endif;
 if(movedir = positive and t_i > 0), dist = t_i; endif;
 if(movedir = positive and t_i < 0), dist = m+t_i; endif;
 if(movedir = negative and t_i < 0), dist = t_i; endif;
 if(movedir = negative and t_i > 0), dist = t_i-m; endif;
endfor
while (t_2 ≠ 0 or dist_2 ≠ 0) do
 if movedir = positive, move packet to +z node
            dist_2 = dist_2 - 1; endif;
 if movedir = negative, move packet to -z node
            dist_2 = dist_2 + 1; endif;  endwhile
while (t_1 ≠ 0 or dist_1 ≠ 0) do
 if movedir = positive, move packet to +y node
            dist_1 = dist_1 - 1; endif;
 if movedir = negative, move packet to -y node
            dist_1 = dist_1 + 1; endif; endwhile
while (t_0 ≠ 0 or dist_0 ≠ 0) do
 if movedir = positive, move packet to +x node
            dist_0 = dist_0 - 1; endif;
 if movedir = negative, move packet to -x node
            dist_0 = dist_0 + 1; endif; endwhile
end
```

Fig. 2. Dimension-Order Routing Algorithm of the H3DM Network

Routing in the H3DM network is strictly defined by the source node address and the destination node address. Let a source node address be $s_\alpha, s_{\alpha-1}, s_{\alpha-2}, \ldots, \ldots s_2, s_1, s_0$, a destination node address be $d_\alpha, d_{\alpha-1}, d_{\alpha-2} \ldots\ldots d_2, d_1, d_0$, and a

routing tag be $t_\alpha, t_{\alpha-1}, t_{\alpha-2}.....t_2, t_1, t_0$, where $t_i = d_i - s_i$. The source node address of H3DM is expressed as $s = (s_{2L}, s_{2L-1}), (s_{2L-2}, s_{2L-3}).....(s_2, s_1, s_0)$. Similarly, the destination address is expressed as $d = (d_{2L}, d_{2L-1}), (d_{2L-2}, d_{2L-3})$ $.....(d_2, d_1, d_0)$. Figure 2 shows the routing algorithm for the H3DM network.

3.2 Deadlock-Free Routing

A deadlock-free routing algorithm can be constructed for a wormhole routed interconnection network by introducing VCs [7]. Since the hardware cost increases as the number of VCs increases, the unconstrained use of VCs is prohibited for cost-effective parallel computers. A deadlock-free routing algorithm with a minimum number of VCs is preferred. In our previous study [8], we proved that the dimension-order routing algorithm on H3DM network is deadlock-free using 2 VCs and 2 is the minimum number of VCs for the H3DM network.

Theorem 1: *A H3DM network is deadlock-free with 2 virtual channels [8].*

4 Dynamic Communication Performance

The overall performance of a MPC system is affected by the performance of the interconnection network as well as by the performance of the node. Low performance of the underlying interconnection network will severely limit the speed of the entire MPC system. Therefore, the success of a MPC is highly dependent on the efficiency of their interconnection networks.

4.1 Performance Metrics

The dynamic communication performance of a MPC system is characterized by message latency and network throughput. Message latency refers to the time elapsed from the instant when the first flit (header) is injected into the network from the source to the instant when the last data 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

We have developed a wormhole routing simulator using C language to evaluate the dynamic communication performance. We use a dimension-order routing and uniform and bit-flip traffic patterns. In the evaluation of performance, flocks of messages are sent through the network to compete for the output channels. Packets are transmitted by the request-probability *r* during *T* clock cycles and the number of flits which reached at destination node and its transfer time is recorded. Then the average transfer time and throughput are calculated and plotted as average transfer time in the horizon-

tal axis and throughput in the vertical axis. The process of performance evaluation is carried out with changing the request-probability *r*. We have considered that the message generation rate is constant and the same for all nodes. Flits are transmitted at 20,000 cycles i.e., T = 20000. In each clock cycle, one flit is transferred from the input buffer to the output buffer, or vice versa if the corresponding buffer in the next node is empty. Thus, transferring data between two nodes takes 2 clock cycles. The message length is considered as short (16 flits), medium (64 flits), and long (256 flits); and the buffer length of each channel is 2 flits. For fair comparison of dynamic communication performance, two VCs per physical link are simulated, and the VCs are arbitrated by a round robin algorithm.

4.3 Dynamic Communication Performance Evaluation

We have evaluated the dynamic communication performance of several networks using deadlock-free dimension order routing with minimum number of virtual channels under the uniform and bit-flip traffic patterns. For fair comparison we should have equal number of nodes for all the considered network. If $m = 4$, $n = 4$, and $L = 2$, then the total number of nodes in the H3DM network is 1024. 32 × 32 mesh and 32 × 32 torus networks also have 1024 nodes.

Uniform Traffic Pattern

The most frequently used, simplest, and most elegant traffic pattern is the uniform traffic pattern where the source and the destination are randomly selected, i.e., every node sends messages to every other node with equal probability [9]. Figure 3 (a), (b), and (c) show the average transfer time as a function of network throughput under uniform traffic pattern for different networks. The average transfer time at no load is called zero load latency. As shown in Figure 3, for all message length, the zero load latency of the H3DM network is lower than that of the mesh and torus networks.

The throughput and latency of a network is increased with the increase of load. Because the links and VCs become congested and the message competes to each other for the network resources, links and channels. With the injection of more and more messages and in course of time, the network becomes saturated. After saturation, the message latency is increasing dramatically while the network throughput will not increase anymore. Up to saturation the trade-off between throughput and latency of the H3DM network is better than that of mesh and torus networks as illustrated in Fig. 3. The limited connectivity of higher level links of 2D-mesh network becomes congested with the increase of packet in the network. On top of this 2D-mesh network saturates earlier due to lack of symmetry. However, the maximum throughput of the mesh and torus network is higher than that of H3DM network as shown in Fig. 3. The number of channels required deadlock-free routing for mesh network is one; however, we used two VCs for fair comparison. With this additional channel the congestion of the mesh network is relief and the throughput is increased. It is portrayed in Figure 3(b) and (c) that the relative difference of maximum throughput between H3DM and mesh network is diminishing with the increase of message length. In torus network all the end-to-end nodes are connected by long length wrap-around links. These links provides a by-pass path for messages which in turns increase the throughput.

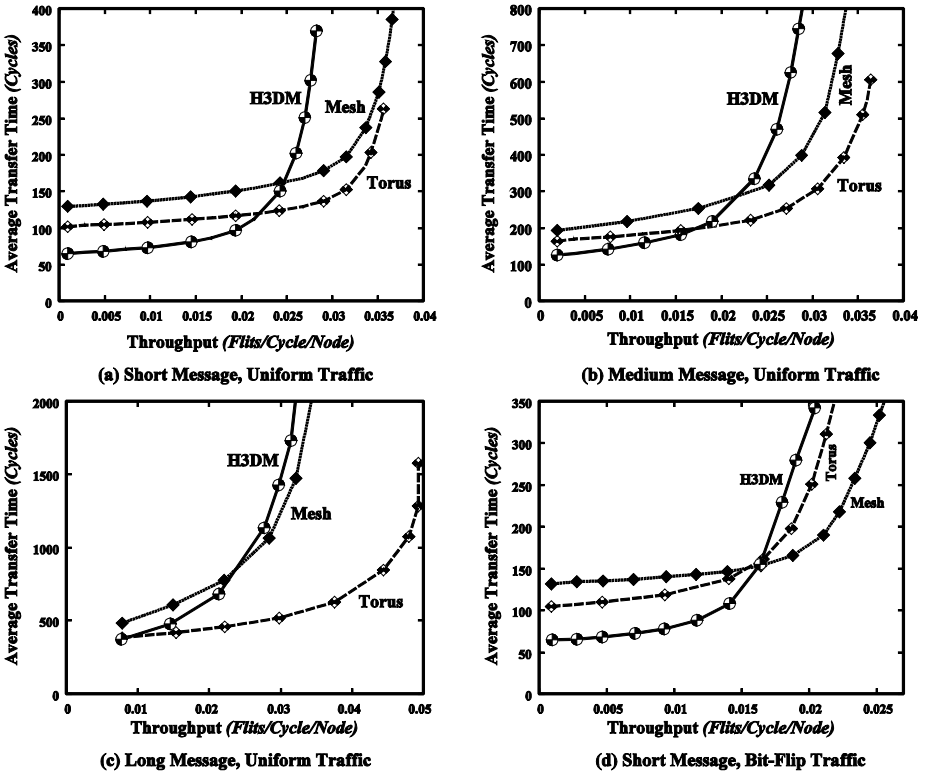


Fig. 3. Dynamic communication performance of various networks using dimension-order routing: 1024 nodes, 2 VCs, and 2 flits buffers

Bit-Flip Traffic Pattern.

In a bit flip traffic, a node with binary address $\mathrm{Node}(b_{\beta-1}, b_{\beta-2} \ldots\ldots b_1, b_0)$ sends messages to $\mathrm{Node}(\overline{b_0}, \overline{b_1}, \ldots\ldots \overline{b_{\beta-2}}, \overline{b_{\beta-1}})$. Figure 3(d) portrays the result of simulations under bit-flip traffic pattern for the various networks for short message. It is seen that the average transfer time at zero load of the H3DM network far lower than that of the mesh and torus networks. Up to saturation the trade-off between throughput and latency under bit-flip traffic of the H3DM network is better than that of mesh and torus networks. However, the maximum throughput of the mesh and torus network is higher than that of H3DM network as depicted in Figure 3(d).

5 Conclusion

A deadlock-free routing algorithm using dimension order routing with a minimum number of VCs was proposed for the H3DM network. It is proven that 2 VCs per physical link are sufficient for the deadlock-free routing algorithm of the H3DM net-

work; 2 is also the minimum number of VCs for dimension order routing. By using the deadlock-free dimension-order routing and the uniform and bit-flip traffic patterns, we have evaluated the dynamic communication performance of the H3DM, mesh, and torus networks. The average transfer time of H3DM network is lower than that of the mesh and torus networks. Maximum throughput of the H3DM network is also higher than that of those networks. A comparison of dynamic communication performance reveals that the H3DM outperforms mesh and torus networks because it yields low latency and high throughput, which are indispensable for next generation high performance massively parallel computer systems. The important issue of assessing the dynamic communication performance improvement of the H3DM network by the adaptive routing algorithm remains a subject for further exploration.

Acknowledgment

This work is supported in part by IIUM Endowment-B research fund EDW B11-169-0647, RMC, IIUM, Malaysia and Postdoctoral Fellowship by Japan Society for the Promotion of Science, No. P09058. The authors are grateful to the anonymous reviewers for their constructive 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, pp. 775-785, 1990.
2. M. Abd-El-Barr and T.F. Al-Somani, Topological Properties of Hierarchical Interconnection Networks: A Review and Comparison, Journal of Electrical and Computer Engineering, Hindawi Publishing Corporation, Vol. 2011, 12 pages.
3. P.L. Lai, H.C. Hsu, C.H. Tsai, I.A. Stewart, A class of hierarchical graphs as topologies for interconnection networks, Theor. Comp. Science, Elsevier, Vo. 411, pp. 2912--2924, 2010.
4. Youyao Liu, Cuijin Li, and Jungang Han, RTTM: A New Hierarchical Interconnection Network for Massively Parallel Computing, Proc. of the HPCA, LNCS 5938, pp. 264--271, 2010.
5. S. Horiguchi, New Interconnection for massively Parallel and Distributed System, Research Report, 09044150, JAIST, pp. 47-57, 1999.
6. 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.
7. W.J.Dally, Virtual-Channel Flow Control, IEEE Trans. on Parallel and Distributed Systems, vol.3, no.2, pp. 194-205, 1992.
8. M. M. Hafizur Rahman, Asadullah Shah, and Yasushi Inoguchi, A Deadlock-Free Dimension Order Routing for Hierarchical 3D-Mesh Network, Proc. of the ICCIS'12, 2012.
9. H.H. Najaf-abadi and H. Sarbazi Azad, The Effects of Adaptivity on the Performance of the OTIS-Hypercube Under Different Traffic Patterns, Proc. of IFIP Int'l. Conf. NPC2004, LNCS, Springer, pp. 390--398, 2004.
10. R. Holsmark, S. Kumar, M. Palesi, and A. Mekia, HiRA: A Methodology for Deadlock Free Routing in Hierarchical Networks on Chip, Proc. of the 3rd ACM/IEEE NOCS, pp. 2-11, 2009