A New Derivative of Midimew-Connected Mesh Network

Md. Rabiul Awal[1,⋆], M.M. Hafizur Rahman[1], Rizal Bin Mohd. Nor[1], Tengku Mohd. Bin Tengku Sembok[2], and Yasuyuki Miura[3]

[1] Department of Computer Science, KICT, IIUM, Malaysia
rabiulawal1@gmail.com, {hafizur,rizalmohdnor}@iium.edu.my
[2] Cyber Security Center, National Defense University Malaysia, Malaysia
tmtsembok@gmail.com
[3] Graduate School of Technology, Shonan Institute of Technology, Japan
miu@info.shonan-it.ac.jp

Abstract. In this paper, we present a derivative of **M**idimew connected **M**esh **N**etwork (**MMN**) by reassigning the free links for higher level interconnection for the optimum performance of the MMN; called Derived MMN (DMMN). We present the architecture of DMMN, addressing of nodes, routing of message and evaluate the static network performance. It is shown that the proposed DMMN possesses several attractive features, including constant degree, small diameter, low cost, small average distance, moderate bisection width, and same fault tolerant performance than that of other conventional and hierarchical interconnection networks. With the same node degree, arc connectivity, bisection width, and wiring complexity, the average distance of the DMMN is lower than that of other networks.

Keywords: Massively Parallel Computers, Interconnection Network, DMMN, MMN, and Static Network Performance.

1 Introduction

Interconnection network is one of the crucial parameters for modern high performance computing. After the introduction of packet switching [1], it has become the "Performance determining factor" for massively parallel computers (MPC) [2] and dominates the performance of a computing system [3,4]. Current research suggests that MPCs of next decade will contain 10 to 100 millions of nodes [5] with computing capability at the tens of petaflops or exaflops level. With this huge amount of nodes conventional topologies for MPC possess a large diameter, hence completely infeasible for next generation MPCs. The hierarchical interconnection network (HIN) provides a cost-effective way in which several network topology can be integrated together [6]. Therefore, HIN is a plausible alternative way to interconnect the future MPC [6] systems. A variety of hypercube based HINs found in the literature, however, its huge number of physical

⋆ Corresponding author.

S. Boonkrong et al. (eds.), *Recent Advances in Information and Communication Technology,* Advances in Intelligent Systems and Computing 265,
DOI: 10.1007/978-3-319-06538-0_20,

links make it difficult to implement. To alleviate this problem, several k-ary n-cube based HIN have been proposed [7,8]. Nevertheless, the performance of these networks does not yield any obvious choice of a network for MPC. No one is clear winner in all aspect of MPC design. As the performance improvement of an interconnection network is likely related to smaller diameter, the problem of designing interconnection network with low diameter with scalability of network size is still desirable [8,9]. A TESH network [10,11] is a k-ary k-cube HIN aiming for large-scale 3D MPC systems, consisting of multiple basic modules (BMs) which are 2D-mesh networks. The BMs are hierarchically interconnected by a 2D-torus network to build higher level networks. Additionally, MInimal DIstance MEsh with Wrap-around links (midimew) network is an optimal topology [12,13]. With this key motivation, to find a network which is suitable for interconnecting a large number of nodes while keeping small diameter, we have replaced the higher level 2D-torus of a TESH network by a 2D midimew network. To use the free ports in the periphery of the 2D-mesh network for higher level interconnection, we kept the basic module as 2D-mesh network same as TESH network. Hence the TESH network becomes Midimew-connected Mesh Network (MMN). The horizontal free ports of the BM are distributed for symmetric tori connection and vertical free links are used for diagonal wrap-around connection. This new HIN, thus allowing exploitation of computation locality, as well as providing scalability up to a million of nodes. In our previous research [15] we arranged the free links of Basic Module (BM) in a specific manner. To the thirst of more efficient way to interconnect the higher level networks for better performance, we derived the free links of BMs in a particular way. Hence we call it Derived MMN (DMMN).

The remainder of the paper is organized as follows. In Section 2, we present the basic architecture of the DMMN. Addressing of nodes and the routing of messages are discussed in Section 3 and Section 4, respectively. The static network performance of the DMMN is discussed in Section 5. Finally, in Section 6, we conclude this paper.

2 Architecture of the DMMN

Derived Midimew connected Mesh Network (DMMN) is a hierarchical interconnection network consisting of multiple basic modules (BM) that are hierarchically interconnected to form a higher level network. Basically the DMMN has two major parts of its architecture, the basic module (BM) and higher level networks. The BMs act as the basic building blocks of DMMN whereas higher level networks determines the construction of DMMN from BMs.

2.1 Basic Module of DMMN

Basic Module of DMMN is a 2D-mesh network of size $(2^m \times 2^m)$. BM consists of 2^{2m} processing elements (PE) with 2^m rows and 2^m columns, where m is a positive integer. Considering $m = 2$, a BM of size (4×4) is portrayed in Figure 1.

Each BM has $2^{(m+2)}$ free ports at the contours for higher level interconnection. The usability of free ports of DMMN is defined by the number of higher levels and denoted by q. All ports of the interior nodes are used for intra-BM connections. All free ports of the exterior nodes, either one or two, are used for inter-BM connections to form higher level networks. In this paper, BM refers to a Level-1 network.

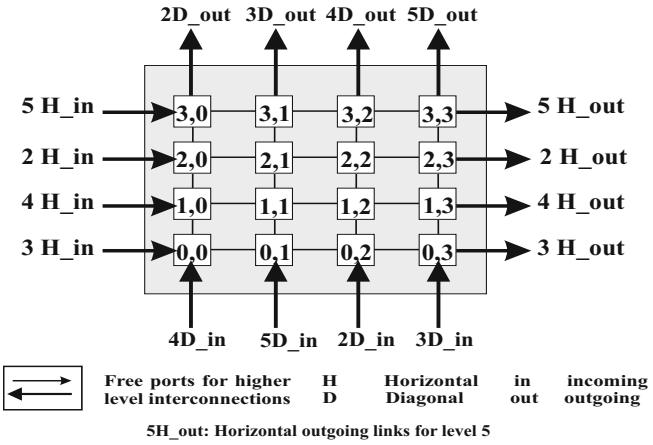

Fig. 1. Basic Module

2.2 Higher Level DMMN

Successive higher level networks are built by recursively interconnecting 2^{2m} immediate lower level subnetworks in a $(2^m \times 2^m)$ midimew network. In a midimew network, one direction (either horizontal or vertical) is symmetric tori connected and other direction is diagonally wrap-around connected. We have assigned the vertical free links of the BM for symmetric tori connection and horizontal free links are used for diagonal wrap-around connection. As portrayed in Figure 2, considering (m = 2) a Level-2 DMMN can be formed by interconnecting $2^{(2\times2)} = 16$ BMs. Similarly, a Level-3 network can be formed by interconnecting 16 Level-2 sub-networks, and so on. Each BM is connected to its logically adjacent BMs. It is useful to note that for each higher level interconnection, a BM uses $4 \times (2^q) = 2^{q+2}$ of its free links, $2^{(2q)}$ free links for diagonal interconnections and $2^{(2q)}$ free links for horizontal interconnections. Here, $q \in \{0, 1, ..., m\}$,, is the inter-level connectivity. $q = 0$ leads to minimal interlevel connectivity, while $q = m$ leads to maximum interlevel connectivity. For example the (4×4) BM has $2^{(2\times2)} = 16$ free ports as shown in Figure 1. If we chose $q = 0$, then $4 \times (2^0) = 4$ of the free ports and their associated links are used for each higher level interconnection, 2 for horizontal and 2 for diagonal interconnection. Among these 2 links, one is used for incoming link and another one for used for outgoing link, i.e., a single links is used for diagonal in, diagonal out, horizontal in, and horizontal out.

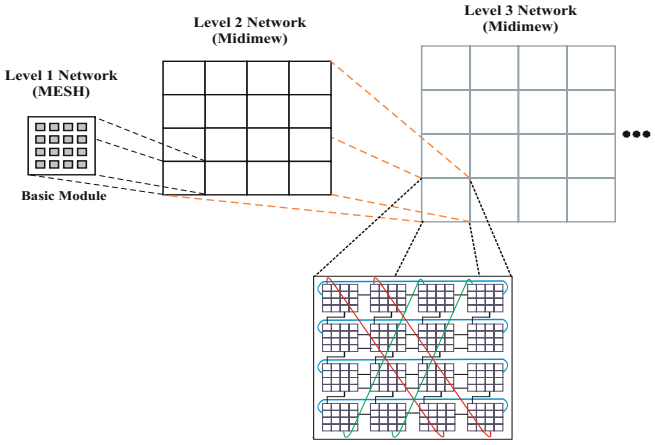

Fig. 2. Higher Level Network

A DMMN(m, L, q) is constructed using $(2^m \times 2^m)$ BMs, has L levels of hierarchy with inter-level connectivity q. 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. If $m \geq 3$, the granularity of the family of networks is coarse. If $m = 3$, then the size of the BM becomes (8×8) with 64 nodes. Correspondingly, the Level-2 network would have 64 BMs. In this case, the total number of nodes in a Level-2 network is $N = 2^{2\times3\times2} = 4096$ nodes, and Level-3 network would have 262144 nodes. Clearly, the granularity of the family of networks is rather coarse.In the rest of this paper we consider m = 2, therefore, we focus on a class of DMMN(2,L,q) networks.

The highest level network which can be built from a $(2^m \times 2^m)$ BM is $L_{max} = 2^{m-q} + 1$ with $q = 0$ and $m = 2$, $L_{max} = 5$, Level-5 is the highest possible level. The total number of nodes in a network having $(2^m \times 2^m))$ BMs is $N = 2^{2mL}$. If the maximum hierarchy is applied then number of total nodes which could be connected by DMMN(m,L,q) is $N = 2^{2m(2^{m-q}+1)}$. For the case of (4×4) BM with $q = 0$, a DMMN network consists of over 1 million nodes.

3 Addressing of Nodes

Nodes in the BM are addressed by an address block, consisting of two digits, the first is representing the horizontal coordinate and the next is representing the vertical coordinate. The address of the nodes are expressed by the base-4 numbers. In case of higher levels, 1 address block is used for each level.

Again the blocks are consists of two digits with base-4 numbers. More generally, in a Level-L MMMN, the node address is represented by:

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

Here, the total number of digits is $n = 2L$, where L is the level number. A^L is the address of level L 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, l-th group $(a_{2l-1}a_{2l-2})$ indicates the location of a Level-$(l-1)$ subnetwork within the l-th group to which the node belongs; $1 \leq l \leq L$. In a two-level network the address becomes $A = (a_4a_3)(a_1a_0)$. The first pair of digits (a_4a_3) identifies the BM to which the node belongs, and the last pair of digits (a_1a_0) identifies the node within that BM.

The address of a node n^1 encompasses in BM_1 is represented as $n^1 = (a^1_{2L-1}\ a^1_{2L-2}\ a^1_{2L-3}\ a^1_{2L-4} \ldots \ldots a^1_3\ a^1_2\ a^1_1\ a^1_0)$. The address of a node n^2 encompasses in BM_2 is represented as $n^2 = (a^2_{2L-1}\ a^2_{2L-2}\ a^2_{2L-3}\ a^2_{2L-4} \ldots \ldots a^2_3\ a^2_2\ a^2_1\ a^2_0)$. In DMMN, the node n^1 in BM_1 and n^2 in BM_2 are connected by a link if the following condition is satisfied.

$$\exists i\{a^1_i(a^2_i \pm 1) mod\ 2^m \wedge \forall j(j \neq i \rightarrow a^1_j = a^2_j)\}$$
$$where\ i\%2 = 0, i, j \geq 2\ ;$$
$$\exists i\{a^1_i = (a^2_i \pm 1) \wedge \forall j(j \neq i \rightarrow a^1_j = a^2_j)\}$$
$$where\ a^1_i = 2^m - 1, i\%2 = 1, i, j \geq 2\ ;$$
$$\exists i\{a^1_i = (a^2_i \pm 1) mod\ 2^m \wedge \forall j(j \neq i \rightarrow a^1_j = a^2_j + 2)\}$$
$$where\ i\%2 = 1, i, j \geq 2$$

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 n1 encompasses in BM1 is represented as . The address of a node n2 encompasses in BM2 is represented as . The node n1 in BM1 and n2 in BM2 are connected by a link if the following condition is satisfied.

4 Routing of DMMN

Routing of messages in the DMMN is simple, top to bottom fashion in order[10,11]. Routing of highest level is done first, the lower level routing at last. BM has outlet/inlet port for higher levels. When a particular transaction of packet is set up from a source to destination, first the shortest path is calculated. Based on the shortest path, outlet port for source and inlet port for destination are fixed. The packet uses the outlet port to reach at highest level

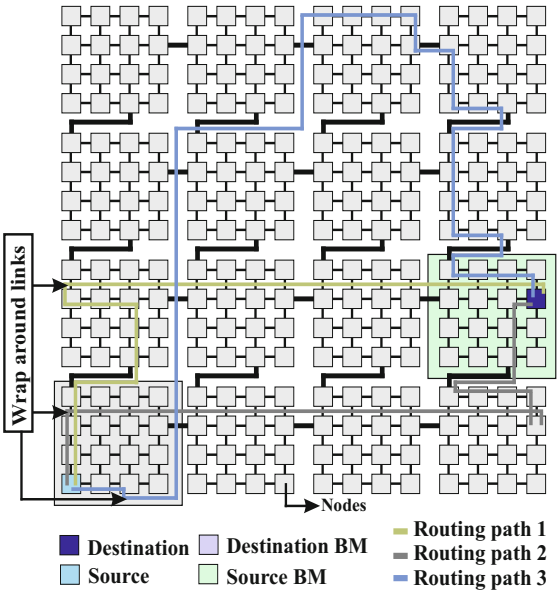

Fig. 3. Routing of DMMN

sub-destination and continue to move through the sub-network to lower level sub-destination until it reaches its final destination. Horizontal routing is performed first, once the packet matches the destination column then diagonal routing starts.

Figure 3 illustrates a routing between level-2 DMMN. Let us consider, a packet is to be routed from source node 0000 to destination node 1323. For this transaction, three routing path is shown. For routing path 1, first the packet moves to outlet node 0030 for level 2. Then it enters to node 1002 and the horizontal BM address is matched. Now the packet will move to the vertical direction using a wraparound link and match vertical BM address by entering destination node 1323. Hence the order of routing is followed by deterministic strategy. The other routing paths are followed by same manner. 9 hops, 11 hops and 22 hops are needed for the packet to reach destination through routing path 1, 2 and 3 respectively. Thus the routing path 1 is shortest and will be followed by packet for routing. Here we assumed that, all the links of DMMN are bidirectional full-duplex links.

5 Static Network Performance

Several topological properties and performance metrics of interconnection network are closely related to many technological and implementation issues. The static network performances do not reflect the actual performance but they have a direct impact on network performance. In this section we discuss about several performance metrics. For the performance evaluation, we have considered

mesh, torus, TESH network, MMN and the proposed DMMN. Some performance metrics like diameter and average distance of MMN, DMMN and TESH were evaluated by simulation, the other metrics like Wiring Complexity, cost were evaluated by their corresponding equations.

5.1 Node Degree

Node degree is the maximum number of neighbor nodes are directly connected with a node. It refers to the number of links at a node. Constant node degree is preferable for networks. Network with constant degree is easy to expand. Also the cost is related to the node degree proportionally. For fair comparison, we have consider degree 4 network. It is shown in Table 1 that the degree of the mesh, torus, TESH, MMN, and DMMN are equal, it is 4 are independent of network size.

5.2 Diameter

Diameter refers to the maximum distance between any pair of source and destination. In other words the number of maximum links to cross for any transaction with a pair of nodes in a given network. Diameter indicates the locality of the network. Latency and message passing time depend on the diameter. Small diameter gives better locality to the network. Hence smaller diameter is convenient. We have evaluated the diameter of the TESH, MMN and DMMN network by simulation and mesh and torus network by their static formula and the results are presented in Table 1. Clearly, the DMMN has a much smaller diameter than that of TESH and mesh networks; equal to MMN and a slightly large diameter than that of torus networks.

5.3 Cost

Cost is one of the important parameter for evaluating an interconnection network. Node degree and diameter effect the performance metrics of the interconnection network including message traffic density, faulttolerance and inter-node distance. Therefore, the product ($diameter \times node\,degree$) is a good criterion to indicate the relationship between cost and performance of a parallel computer systems [16]. The cost of different networks is plotted in Table 1. The DMMN is less costly than mesh and TESH, same to MMN and a slightly higher than torus network.

5.4 Average Distance

The average distance is the average of all distinct paths in a network. Average distance reflects the ease of communication within the network i.e. average network latency. A small average distance results small communication latency.

Table 1. Comparison of Static Network Performance of Various Networks

Network	Node Degree	Diameter	Cost	Average Distance	Ark Connectivity	Bisection Width	Wiring Complexity
2D-Mesh	4	30	120	10.67	2	16	480
2D-Torus	4	16	64	8	4	32	512
TESH(2,2,0)	4	21	84	10.47	2	8	416
MMN(2,2,0)	4	17	68	9.07	2	8	416
DMMN(2,2,0)	4	17	68	8.56	2	8	416

In store and forward communication which is sensitive to the distance, small average distance tend to favor the network [17]. But it is also crucial for distance-insensitive routing, such as wormhole routing, since short distances imply the use of fewer links and buffers, and therefore less communication contention. We have evaluated the average distances for DMMN, MMN, and TESH network by simulation and mesh and torus networks by their corresponding formulas and the results are tabulated in Table 1. It is shown that the average distance of DMMN is lower than that of MMN, mesh and TESH networks, and slightly higher than that of torus networks.

5.5 Bisection Width

The Bisection Width (BW) refers to the minimum number of communication links that must be removed to partition or segment the network into two equal halves. Small BW impose low bandwidth between two parts. Nevertheless large BW requires lots of wires and is difficult for VLSI design. Hence moderate BW is highly desirable. BW is calculated by counting the number of links that must to be eliminated from Level-L DMMN. Table 1 is showing that, BW of the DMMN is exactly equal to that of the MMN and TESH network and lower than that of mesh and torus network.

5.6 Arc Connectivity

The arc connectivity of a network suggest the minimum number of arcs that must be removed from the network to break it into two disconnected networks. It measures the robustness of a network and the multiplicity of paths between nodes over the network. High arc connectivity improves performance during normal operation, and also improves fault tolerance. A network is maximally fault-tolerant if its connectivity is equal to the degree of the network. From Table 1 it is clear that for DMMN, MMN and TESH, the arc connectivity is exactly equal. Nonetheless arc connectivity of torus is equal to its degree, thus more fault tolerant than others.

5.7 Wiring Complexity

The wiring complexity of an interconnection network refers to the total number of links required to form the network. It has a direct correlation to hardware

cost and complexity. A (16×16) 2D-mesh and 2D-torus networks have $\{N_x \times (N_y 1) + (N_x 1)Ny = 16 \times (161) + (161) \times 16 = 480\}$ and $(2 \times N_x \times N_y = 2 \times 16 \times 16) = 512$ links, respectively. Ni represents the number of nodes in the ith dimension. The wiring complexity of a Level-L DMMN, MMN, and TESH networks is $[\#\ of\ links\ in\ a\ BM \times k_{2(L1)} + \sum_{x=2}^{L} 2(2^q) \times k^{2(L1)i}]$. Considering, $m = 2$, a BM of DMMN, MMN, and TESH network have 24 links. Hence the total number of links of a Level-2 DMMN, MMN, and TESH are 416. Table 1 is showing that the total number of links of DMMN is lower than that of mesh and torus network and exactly equal to that of MMN and TESH network.

The static network performance is claiming that torus network is better than DMMN except in the term of wiring complexity. Now, torus network has $N_x + N_y$ long wrap-around links, where $N_x \times N_y$ is the network size. In case of DMMN, from Level-2 to Level-L, each level contains $(2^m/2) + (2^m/2)$ wrap-around links. Also the wrap-around links of DMMN do not increase with network size, instead they increase with higher levels. But in torus they increase with network size. Hence the implementation of DMMN is easier than torus.

6 Conclusion

A new derivative of MMN, called DMMN, is proposed for the future generation MPC systems. The architecture of the DMMN, addressing of nodes, and routing of message have been discussed in detail. We have evaluated the static network performance of the DMMN, as well as that of several other networks. From the static network performance, it has been shown that the DMMN possesses several attractive features, including constant node degree, small diameter, low cost, small average distance, and better bisection width. We have seen that with the same node degree, arc connectivity, bisection width, and wiring complexity, the average distance of the DMMN is lower than that of MMN, TESH, and mesh networks. Also, DMMN has slightly higher diameter and average distance to that of torus network. The DMMN yields better static network performance with low cost, which are indispensable for next generation MPC systems. This paper focused on the architectural structure and static network performance. Issues for future work include the following: (1) evaluation of static network performance considering different value of m; L; and q and (2) evaluation of dynamic communication performance using dimension order routing algorithm.

Acknowledgments. The authors are grateful to the anonymous reviewers for their constructive comments which helped to greatly improve the clarity of this paper. This work is partly supported by FRGS13-065-0306, Ministry of Education, Government of Malaysia.

References

1. Baran, P.: On distributed communications networks. IEEE Transactions on Communications Systems 12, 1–9 (1964)
2. Wu, C.L., Feng, T.Y.: Tutorial: Interconnection networks for parallel and distributed processing. IEEE Computer Society, Los Alamitos (1984)
3. Yang, Y., Funahashi, A., Jouraku, A., Nishi, H., Amano, H., Sueyoshi, T.: Recursive diagonal torus: an interconnection network for massively parallel computers. IEEE Transactions on Parallel and Distributed Systems 12, 701–715 (2001)
4. Rahman, M.H., Jiang, X., Masud, M.A., Horiguchi, S.: Network performance of pruned hierarchical torus network. In: 6th IFIP Int'l Conf. on Network and Parallel Computing, pp. 9–15 (2009)
5. Beckman, P.: Looking toward exascale computing. In: 9th Int'l Conf. on Parallel and Distributed Computing, Applications and Technologies, p. 3 (2008)
6. Abd-El-Barr, M., Al-Somani, T.F.: Topological properties of hierarchical interconnection networks: a review and comparison. J. Elec. and Comp. Engg. 1 (2011)
7. Lai, P.L., Hsu, H.C., Tsai, C.H., Stewart, I.A.: A class of hierarchical graphs as topologies for interconnection networks. J. Theoretical Computer Science, Elsevier 411, 2912–2924 (2010)
8. Liu, Y., Li, C., Han, J.: RTTM: a new hierarchical interconnection network for massively parallel computing. In: Zhang, W., Chen, Z., Douglas, C.C., Tong, W. (eds.) HPCA 2009. LNCS, vol. 5938, pp. 264–271. Springer, Heidelberg (2010)
9. Camarero, C., Martinez, C., Beivide, R.: L-networks: A topological model for regular two-dimensional interconnection networks. IEEE Transactions on Computers 62, 1362–1375 (2012)
10. Jain, V.K., Ghirmai, T., Horiguchi, S.: TESH: A new hierarchical interconnection network for massively parallel computing. IEICE Transactions on Information and Systems 80, 837–846 (1997)
11. Jain, V.K., Horiguchi, S.: VLSI considerations for TESH: A new hierarchical interconnection network for 3-D integration. IEEE Transactions on Very Large Scale Integration (VLSI) Systems 6, 346–353 (1998)
12. Beivide, R., Herrada, E., Balcazar, J.L., Arruabarrena, A.: Optimal distance networks of low degree for parallel computers. IEEE Transactions on Computers 40, 1109–1124 (1991)
13. Puente, V., Izu, C., Gregorio, J.A., Beivide, R., Prellezo, J., Vallejo, F.: Improving parallel system performance by changing the arrangement of the network links. In: Proceedings of the 14th Int'l Conf. on Supercomputing, pp. 44–53 (2000)
14. Lau, F.C., Chen, G.: Optimal layouts of midimew networks. IEEE Transactions on Parallel and Distributed Systems 7, 954–961 (1996)
15. Awal, M.R., Rahman, M.H., Akhand, M.A.H.: A New Hierarchical Interconnection Network for Future Generation Parallel Computer. In: 16th Int'l Conf. on Computers and Information Technology (2013)
16. Kumar, J.M., Patnaik, L.M.: Extended hypercube: A hierarchical interconnection network of hypercubes. IEEE Transactions on Parallel and Distributed Systems 3, 45–57 (1992)
17. Lonka, O., Naralchuk, A.: Comparison of interconnection networks. LUT (2008)