Interlacing Bypass Rings to Torus Networks for More Efficient Networks

Peng Zhang, Reid Powell, and Yuefan Deng, *Member, IEEE*

Abstract — We introduce a new technique for generating more efficient networks by systematically interlacing bypass rings to torus networks (iBT networks). The resulting network can improve the original torus network by reducing the network diameter, node-to-node distances, and by increasing the bisection width without increasing wiring and other engineering complexity. We present and analyze the statement that a 3D iBT network proposed by our technique outperforms 4D torus networks of the same node degree. We found that interlacing rings of sizes 6 and 12 to all three dimensions of a torus network with meshes 30×30×36 generates the best network of all possible networks, including 4D torus and hypercube of approximately 32,000 nodes. This demonstrates that strategically interlacing bypass rings into a 3D torus network enhances the torus network more effectively than adding a fourth dimension, although we may generalize the claim. We also present a node-to-node distance formula for the iBT networks.

Index Terms— Network topology, torus networks, bypass ring, network diameter, node-to-node distance, routing

1 Introduction

ADVANCED networking architectures [1-5] have helped enable supercomputers such as RoadRunner [6] to break the petaflop barrier and such progress has stimulated the parallel computing community's ambitions to invent more scalable interconnection networks to accommodate the ever-increasing demands of performance and functionalities by incorporating millions of powerful processor cores. A scalable interconnection network, of a fixed node degree, must satisfy most of the performance requirements including small diameter, large bisection width, topological simplicity, high-degree symmetry, design modularity, and engineering feasibility, as well as expandability. For example, a 3D torus network such as those in the IBM's Blue Gene and Cray's T3E [1-4] with up to 20 thousand nodes and several small-scale hypercube network supercomputers [5, 7-9] satisfy several of the requirements. However, the network diameters grow as $\Theta(\sqrt{n})$ for a torus and as $\Theta(\log n)$ for a hypercube and at a similarly rapid rate for many of their derivatives [7-12], where n is the network size. This defect of rapidly growing diameters greatly limits the expandability of these networks. Mesh networks of fixed dimension provide an alternative with relatively low node-degree and low engineering complexity but with large network diameter and small overall bandwidth. Other efforts to increase bandwidth without increasing network diameters include that of the hybrid fat-tree [13], a low-cost, low-degree network with irregular node degree; however, it is suceptible to faulty links and to message contension towards roots. Other proposals have also been introduced, such as the incomplete torus and its derivatives [14] that reduce node degree at the expense of losing symmetry and topological simplicity. Honeycomb mesh and torus networks [15] received considerable early attention that faded quickly due to implementation obstacles, among other difficulties. Hexagonal networks introduced in [16] also boast a small diameter but carry a burden of a high node degree. Modifications of the traditional torus including the PEC [17], SRT [18], TESH [19], and RDT [20] networks all build upon the simplicity of mesh and torus networks, achieving improved network properties with unfavorable expandability and network cost. However, these variants demonstrated that interlacing rings of various lengths to a torus network is a profitable practice for improving network performance without adding significant engineering complexity.

Motivated by this, we propose the iBT network. The iBT network is constructed by interlacing bypass rings evenly into a torus network. We preserve the simplicity of grid-like layout and improve the performance of the network with minimal number of bypass links. Our model allows generalization of the bypass construction of the base torus to arbitrary dimensions for much larger and scalable networks, rather than in 2D as in [17, 18, 20]. This new network achieves a low network diameter, high bisection width, short node-to-node distances, and low engineering complexity in terms of network cost. Furthermore, the iBT network has much lower node degree and lower network cost than a hypercube of the similar network size does. To ensure network symmetry and modularity, we interlace rings into the torus network consistently. To analyze the topological properties for achieving an optimal network, we present the node-to-node hop distance distributions.

The paper is organized as follows: We first define the iBT interconnection model and its generation scheme in

- *Peng Zhang is with the Department of Applied Mathematics, Stony Brook University, NY 11794. E-mail: Peng.Zhang@StonyBrook.edu .*
- *Reid Powell is with the Department of Applied Mathemetics, Stony Brook University, NY 11794. E-mail: RPowell@ams.sunysb.edu .*
- *Yuefan Deng is with the Department of Applied Mathematics, Stony Brook University, NY 11794. E-mail: Yuefan.Deng@StonyBrook.edu .*

Manuscript received June 2009.

Section 2.1. In Sections 2.2 and 2.3, we describe the evaluation criteria for networks and the performance comparisons with torus and hypercube networks. A formula for a node-to-node shortest distance is discussed in Section 3. A conclusion is drawn in the last section.

2 iBT Interconnection Model

The iBT network, generated by interlacing bypass rings into a torus network, is an n-dimensional composite network that contains the original n-dimensional torus network and the added bypass rings. In this section, we formally define the iBT network and discuss its topological properties. We also demonstrate the procedure to generate an optimal 3D iBT network from a 3D torus network. For a concrete example, we show the detailed procedure to generate an optimal iBT network with approximately 32,400 nodes and to compare it to a 4D torus network of 32,768 nodes with identical node degree of eight and to a hypercube with $2^{15} = 32{,}768$ nodes.

2.1 Definition of iBT Networks

An iBT$(N_1 \times \cdots \times N_n; L = m; l = \langle l_1, \cdots, l_k \rangle)$ network outgrows from an n-dimensional torus network $T(N_1 \times \cdots \times N_n)$ by interlacing l_i-hop bypass rings $(i = 1 \ldots, k)$ recursively into any m of the n dimensions $(m \le n)$. The m dimensions with bypass rings are referred to as the bypass dimensions and the remaining $n - m$ dimensions without bypass rings are referred to as plain dimensions. The terms $L = m; l = \langle l_1, \cdots, l_k \rangle$ are referred to as a bypass scheme for generating the iBT network. This interconnection model results in a node degree of $2n + 2$ where $2n$ is from the original torus connections and the additional 2 from the bypass connections. To determine the two bypass connections for a node $p = (x_1, x_2, \cdots, x_n)$, where $x_i \in [0, N_i - 1]$, $i = 1 \ldots, k$, we introduce three terms: a node bypass dimension $d(p) \in \{1,2, \cdots m\}$ and a node bypass length $l(p) \in \{l_1, \cdots, l_k\}$ which can be expressed as:

$$d(p) = \left[\left(\sum_{i=1}^{m} x_i \right) (\mathrm{mod}\, m) \right] + 1 \quad \text{and} \quad l(p) = l_h,$$

where

$$h = \left\lfloor \frac{(\sum_{i=1}^{m} x_i)(\mathrm{mod}\, mk)}{m} \right\rfloor + 1 \in \{1, \ldots, k\}.$$

Thus, a node bypass species $s(p) = \langle d(p), l(p) \rangle$, indicating that two $l(p)$-hop bypass links have been added to the given node p in each direction along the dimension $d(p)$. For example, iBT$(32 \times 32 \times 16; L = 2; l = \langle 4,16 \rangle)$ indicates the interlacing of 4-hop and 16-hop bypass rings in the first two dimensions, i.e., xy-plane, of the 3D torus $T(32 \times 32 \times 16)$. A node $p = (1,1,4)$ has a bypass dimension $d(p) = 1$, a bypass length $l(p) = l_2 = 16$, and thus its bypass species is $\langle 1,16 \rangle$, implying that p has two 16-hop bypass links in each direction along dimension x, i.e., the first dimension. In order for the node sets of any two bypass rings to be disjoint, both $l_1, \cdots, l_k$ and $N_1, \cdots, N_m$ must be divisible by mk.

In Fig. 1, we show the hop distance distribution and bypass configurations for a family of 1-dimensional iBT networks written as iBT$(32; L = 1; l = \langle l_1, l_2 \rangle)$ generated from a 1-dimension torus network with 32 nodes. A 1-dimesional torus network $T(32)$ is itself a ring. This 1D scheme is easy to follow and to generalize for illustration at higher dimensions. The node-to-node distance within a 32-node 2D torus network $T(4 \times 8)$ is $D_{T(4\times8)} = (Average\ Distance) \pm (Standard\ Deviation) = 3.00 \pm 1.41$. As shown by Fig. 1, the node-to-node distances of the resulting iBT networks are statistically smaller than $D_{T(4\times8)}$, for example, $D_{iBT(32,;L=1;\, l=\langle 4,8 \rangle)} = 2.41 \pm 0.95$ and $D_{iBT(32;L=1;l=\langle 8 \rangle)} = 2.94 \pm 1.32$. In Fig. 2, we draw all of the links for the 2D iBT network iBT$(8 \times 8; L = 2; l = \langle 4 \rangle)$. In Fig. 3, we illustrate the bypass arrangement for the 3D iBT network iBT$(30 \times 30 \times 36; L = 3; l = \langle 6,12 \rangle)$. In this figure, we only draw the bypass links along the three easily visible edges to identify the nodes that are connected by the appropriate bypass links. Other links are omitted for the purpose of clarity.

2.2 Evaluation Criteria for Networks

To evaluate a network model, we consider its diameter, average node-to-node hop distance, bisection width, and more importantly, the node-to-node hop distance distribution [21-23]. The network diameter, defined as the longest node-to-node hop distance, indicates the worst-case communication latency, while the average distance, defined as the average of the node-to-node hop distances, represents the expected communication latencies over the network. These two measures provide some information about the network while the hop distance distribution provides a richer representation of the network properties including maximum, average, and standard deviation of node-to-node distances. Additionally, we consider the bisection width to measure the aggregate network capacity and the network cost, defined as a product of diameter and node degree, for network comparison [24].

2.3 Topological Optimization

In this section, we compare the topological properties of 3D iBT networks with a number of nodes closest to that of a 4D torus network with exactly 32,768 nodes [5]. We will analyze three iBT networks: (1) the 3D iBT network with bypass rings in two dimensions iBT$(32 \times 32 \times 32; L = 2)$; (2) iBT$(64 \times 64 \times 8; L = 2)$; and (3) iBT$(30 \times 30 \times 36; L = 3)$, a 3D iBT network with bypass rings in all three dimensions. For each iBT network, we evaluate them with bypass rings of the same length or a mix of two different lengths. For comparison, we also analyze two other networks with a similar number of nodes: a 4D torus network $T(16 \times 16 \times 16 \times 8)$ and a 15D hypercube $H(2^{15})$. Due to the fact that the node numbers of various network configurations are usually non-contiguous integers, it is unlikely one can find two configurations with exactly the same number of nodes. We compare two closest: a 32,768-node $T(16 \times 16 \times 16 \times 8)$ and $H(2^{15})$ with a 32,400-node iBT$(30 \times 30 \times 36; L = 3)$.

Our analysis starts from numerical experiments. Figures 4 to 8 illustrate the numerical results for the 3D iBT$(L = 2)$ and iBT$(L = 3)$ networks, compared with 4D torus and hypercube. Table 1 presents such network topological properties as node-to-node hop distance distribu-

tion, network diameter, and bisection width. These experiments show, as expected, the dependence of the network properties on the bypass scheme; network properties behave relatively poorly at bypass extremes: too short or too long. Starting from a torus network $T(N_x \times N_y \times N_z)$, we study the following cases:

1. For $N_x \times N_y \times N_z = 32 \times 32 \times 32 = 32{,}768$, we bypass in two dimensions with uniform bypass length to generate a new network, iBT$(32 \times 32 \times 32; L = 2; l = \langle l_1 \rangle)$ with $l_1 \in \{2, 4, 6, 8, 16\}$. We found the resulting networks have relatively poor network properties for extreme bypassing lengths such as $l_1 \in \{2, 16\}$, but have better properties with middle-sized bypass lengths such as $l_1 = 6$, resulting in that iBT$(32 \times 32 \times 32; L = 2; l = \langle 6 \rangle)$ is the optimal iBT network with uniform bypass length in Fig. 5.
2. For $N_x \times N_y \times N_z = 32 \times 32 \times 32$ (same as above), we bypass in two dimensions with a mixture of two bypass lengths to generate a network, iBT$(32 \times 32 \times 32; L = 2; l = \langle l_1, l_2 \rangle)$, with $l_1, l_2 \in \{4, 8, 16, 32\}$. With all possible combinations, the bypassing parameter $l = \langle 4{,}16 \rangle$ generates the optimal iBT network with two bypass lengths in Fig. 6. We also found that iBT$(32 \times 32 \times 32; L = 2; l = \langle 4{,}16 \rangle)$ excels over iBT$(32 \times 32 \times 32; L = 2; l = \langle 6 \rangle)$ in Table 1.
3. For $N_x \times N_y \times N_z = 64 \times 64 \times 8 = 32{,}768$, we bypass in two of the longest dimensions to generate a new network, iBT$(64 \times 64 \times 8; L = 2; l = \langle l_1, l_2 \rangle)$. Under the same bypass scheme $L = 2; l = \langle l_1, l_2 \rangle$, we found iBT$(64 \times 64 \times 8; L = 2)$ always outperforms iBT$(32 \times 32 \times 32; L = 2)$ in Fig. 4 and iBT$(64 \times 64 \times 8; L = 2; l = \langle 4{,}16 \rangle)$ is the best of all possibilities where $L = 2$ in Table 1.
4. For $N_x \times N_y \times N_z = 30 \times 30 \times 36 = 32{,}400 \approx 32{,}768$, we consider bypassing in all three dimensions to generate a network iBT$(30 \times 30 \times 36; L = 3)$. As shown by Fig. 7 and Fig. 8, iBT$(30 \times 30 \times 36; L = 3; l = \langle l_1 \rangle)$ with $l_1 \in \{6, 9, 12\}$ demonstrated the same behavior as those for iBT$(64 \times 64 \times 8; L = 2; l = \langle l_1, l_2 \rangle)$. We also found iBT$(30 \times 30 \times 36; L = 3; l = \langle 6{,}12 \rangle)$ is the best among all of the iBT networks, better than 4D torus and similar to 15D hypercube in Fig. 8 and Table 1.
5. In Table 1, we found most networks have the same network diameter but different average distances and various standard deviations and a network with a larger network diameter may have a smaller average distance such as iBT$(30 \times 30 \times 36; L = 3; l = \langle l_1 \rangle)$ with $l_1 \in \{9{,}12\}$.

From these experiments, we make the following claims:

1. For the iBT networks with uniform bypass length, extreme bypass length achieves poorer network properties than a middle-sized bypass lengths do;
2. An appropriate mixture of bypass lengths is favored in the interlacing arrangement for iBT networks over uniform bypass length;
3. For iBT networks with plain dimensions, a plain dimension size should be shrunk to scale to bypass dimensions for optimized performance;
4. The most efficient bypass scheme for a 3D iBT network is without plain dimensions. It is shown that, among all the possibilities of a system with approximately 32,000 nodes, iBT$(30 \times 30 \times 36; L = 3; l = \langle 6{,}12 \rangle)$ is the best network. It performs much better than the simple 4D torus $T(16 \times 16 \times 16 \times 8)$ with 32,768 nodes and it performs similarly to the 15D hypercube $H(2^{15})$ with 32,768 nodes with degree 15. Its network cost of value 96 is much smaller than the 4D torus's 224 and the hypercube's 225;
5. The node-to-node hop distance distribution is efficient and precise in its depiction of topological details for the comparison of networks.

2.4 Performance Comparisons

Through exhausive numerical search, we found the optimal iBT network of approximately 32,000 nodes to be iBT$(30 \times 30 \times 36; L = 3; l = \langle 6{,}12 \rangle)$. We further compare this with other networks and graphs in Table 1 and Fig. 9. All networks, in the table, except the 3D torus, hypercube, the CCC network [12] and the scalable Barrel Shifter [25] have a node degree of 8. As shown in Fig. 9, these networks are grouped into three categories by their sizes. Of all networks of size 32,000 nodes and of degree 8, the iBT network has the shortest average distance.

3 Distance Formulas

For optimal routing, we always need to search for a path with the shortest node-to-node distance [1, 2, 16]. There are many possible paths for linking a pair of source and destination nodes for iBT networks. It is much less obvious to recognize such paths for the iBT networks than for the torus network. This appears to be one of the few disadvantages of the iBT network. To overcome this, we have derived a closed form node-to-node distance formula for the iBT networks with the bypass scheme $L \in \{2{,}3\}; l = \langle l_1 \rangle$. Other more complex cases can also be derived.

3.1 Terminology

In iBT$(N_1 \times \cdots \times N_n; L = m; l = \langle l_1, \cdots, l_k \rangle)$ networks, the number of hops in a shortest path between a node-pair can be partitioned into two parts: one from the first m bypass dimensions and the other from the remaining $n - m$ plain dimensions, which are defined as $B(p_1, p_2)$ and $T(p_1, p_2)$, respectively. Since the plain dimensions have no bypass connections, the procedure for calculating $T(p_1, p_2)$ is identical to that of a traditional torus network. Thus, we concentrate on calculating $B(p_1, p_2)$ by assuming that $m = n$. Considering this, we abbreviate the iBT networks with a uniform-length bypass connection in the first two or three dimensions as iBT$(N_x \times N_y; L = 2; l = \langle l_1 \rangle)$ and iBT$(N_x \times N_y \times N_z; L = 3; l = \langle l_1 \rangle)$ respectively.

In the iBT$(N_x \times N_y; L = 2; l = \langle l_1 \rangle)$ network, consider two points $p_1 = (x_1, y_1)^T$ and $p_2 = (x_2, y_2)^T$, where

$x_i \in [0, N_x - 1]$ and $y_i \in [0, N_y - 1]$. The $sign(x)$ function is a standard sign function, while $sgn(x)$ is the signum function, defined as:

$$sign(x) = \begin{cases} -1, & x < 0 \\ 1, & x \geq 0 \end{cases}$$

$$sgn(x) = \begin{cases} -1, & x < 0 \\ 0, & x = 0 \\ 1, & x > 0 \end{cases}$$

If x is a vector, the same operation applies to each of its components. For example, suppose $v = (x, y)^T$, then

$$sgn(v) = \begin{pmatrix} sgn(x) \\ sgn(y) \end{pmatrix} \quad \text{and} \quad |v| = \begin{pmatrix} |x| \\ |y| \end{pmatrix}.$$

Let $\delta(x, \alpha)$ be a two-point function defined as

$$\delta(x, \alpha) = \begin{cases} 1, & x = \alpha \\ 0, & \text{otherwise.} \end{cases}$$

Let the vector $t = (t_x, t_y)^T$ be referred to as the fundamental torus distance. The magnitude of each element t_i represents the number of hops along dimension i in the $sign(t_i)i$ direction on a non-bypass shortest path from p_1 to p_2. For example, $t_x \gtrless 0$ indicates that the message traverses $|t_x|$ basis torus links in the positive or negative x-dimension, making the fundamental torus distance similar to the distance formula of a traditional torus. The definition of t_x is written as

$$t_x = \Delta x + \frac{N_x}{2} sgn(\Delta x)\{sign(N_x - 2\|\Delta x\|) - 1\},$$

in which $\Delta x = x_2 - x_1$. We similarly define and express t_y.

In iBT networks, the set of links on a shortest path from p_1 to p_2 can be partitioned into two subsets: bypass rings and residual torus links. The bypass ring subset is referred to as the bypass distance $b = (b_x, b_y)^T$, a vector in which the magnitude of each component b_i is the number of bypass hops in dimension i in the $sign(b_i)i$ direction on a shortest path from p_1 to p_2. For example, $b_x \gtrless 0$ indicates that a message from p_1 to p_2 traverses $|b_x|$ hops of bypass rings in the positive or negative x-dimension. Thus, b_x is written as

$$b_x = round\left(\frac{t_x}{l}\right) - sgn(t_x) \cdot \delta\left(\left|frac\left(\frac{t_x}{l}\right)\right|, 0.5\right),$$

in which $round(x)$ rounds x to the nearest integer and $frac(x)$ returns the fractional part of x.

As stated previously, in addition to the bypass distance, a shortest path from p_1 to p_2 also has a residual torus link component. The residual torus distance is referred to as a vector $\hat{t} = (\hat{t}_x, \hat{t}_y)^T$ in which the magnitude of each component $\hat{t}_i$ is the number of torus hops in dimension i in the $sign(\hat{t}_i)i$ direction on a shortest path from p_1 to p_2. For example, $\hat{t}_x \gtrless 0$ indicates that the message routes $|\hat{t}_x|$ hops of torus links in the positive or negative dimension x. Thus,

$$\hat{t} = t - l \cdot b \Leftrightarrow \begin{pmatrix} \hat{t}_x \\ \hat{t}_y \end{pmatrix} = \begin{pmatrix} t_x - l \cdot b_x \\ t_y - l \cdot b_y \end{pmatrix}.$$

In addition to the distance vectors, we also consider the bypass species of a node p_1 - $s(p_1)$: a vector defined as

$$s(p_1) = e_{[x_1 + y_1 (mod\ L)] + 1},$$

where "1" in $s(p_1)$ indicates the dimension in which the node adds bypass connections. For example, $s(p_1 = (1,2)^T) = e_2$ means p_1 adds bypass rings to the second dimension, i.e., dimension y.

The relationship among coordinates of p_1, p_2 and $\hat{t}$ is:

$$x_1 + y_1\ (mod\ L) = x_2 + y_2 + \hat{t}_x + \hat{t}_y\ (mod\ L)$$

The stated definitions in iBT$(N_x \times N_y; L = 2; l = \langle l_1 \rangle)$ can all be extended to iBT$(N_x \times N_y \times N_z; L = 3; l = \langle l_1 \rangle)$.

3.2 iBT$(N_x \times N_y; L = 2; l = \langle l_1 \rangle)$

In iBT$(N_x \times N_y; L = 2; l = \langle l_1 \rangle)$ in which $l_1 = 2(k+1), k \in Z^+$, the distance between p_1 and p_2 is given by

$$D(p_1, p_2) = \|b\|_1 + \|\hat{t}\|_1 + \varphi_{xy}(p_1, p_2),$$

$$\text{where } \varphi_{xy}(p_1, p_2) = \begin{cases} 2, & \alpha = 0, \beta = 2; \\ 2, & \alpha = 0, \gamma \in \{5,10\}; \\ 0, & \text{otherwise,} \end{cases}$$

in which $\alpha = \|\hat{t}\|_2^2$, $\beta = \|sgn(b)\|_2^2$ and $\gamma = \||sgn(b)| + s_1 + s_2\|_2^2$. The notation $\alpha = 0, \gamma \in \{5,10\}$ means that if $\alpha = 0, \gamma = 5$ or if $\alpha = 0$, $\gamma = 10$, we have $\varphi_{xy}(p_1, p_2) = 2$.

In this equation, the terms $\|b\|_1$ and $\|\hat{t}\|_1$ represent the bypass and torus hops a message needs to traverse under the assumption that a single node is on bypass rings across each of the bypass dimensions. After this, the penalty term $\varphi_{xy}(p_1, p_2)$ accounts for the interlacing of bypass rings, where a given node has bypass rings in exactly one bypass dimension. For example, $\alpha = 0$ indicates no residual torus links are required, implying that $s(p_1) = s(p_2)$; meanwhile, $\beta = 2$ tells that a message has to traverse bypass rings in two dimensions. In this case, whichever bypass species a message eminates from, an additional torus hop is required to reach a node of a different species to traverse bypass hops in that bypass dimension. Then, a second torus hop is required to return to a node of the original bypass species. Thus, a valid shortest path always requires a positive number of torus hops, meaning $\varphi_{xy}(p_1, p_2) = 2$. The term γ is more subtle, implying the relationship between the bypass dimensions a message must traverse and its source/destination bypass species. Here, $\alpha = 0$ holds the same meaning. For example, consider a case where $s(p_1) = s(p_2) = e_1$. However, $\gamma \in \{5,10\}$ implies that the message has bypass hops in the second dimension, i.e., dimension y. Similarly, the message also needs a positive number of torus hops to complete a valid shortest path in the iBT network.

3.3 iBT$(N_x \times N_y \times N_z; L = 3; l = \langle l_1 \rangle)$

In iBT$(N_x \times N_y \times N_z; L = 3; l = \langle l_1 \rangle)$ in which $l_1 = 3(k+1), k \in Z^+$, the distance between p_1 and p_2 is given by

$$D(p_1, p_2) = \|b\|_1 + \|\hat{t}\|_1 + \varphi_{xyz}(p_1, p_2)$$

$$\text{where, } \varphi_{xyz}(p_1, p_2) = \begin{cases} 4, & \alpha = 0, \gamma \in \{6,11\}; \\ 2, & \alpha \in \{0,1\}, \gamma = (5 - 2\alpha)\beta; \\ 2, & \alpha = 2, \hat{t} \cdot 1 = 0, \gamma = \beta^2 + 2; \\ 0, & \text{otherwise.} \end{cases}$$

4 Conclusions

In this paper, we proposed a scheme to generate efficient networks by interlacing bypass rings to torus. We analyzed the topological properties of many possible networks for systems with approximately 32,000 nodes. Among the networks considered, we found iBT$(30 \times 30 \times 36; L = 3; l = \langle 6,12 \rangle)$ is to be a superior network in nearly

all cases, according to our metrics of network diameter, node-to-node distances and their distribution, bisection width etc. This network is much more efficient than a 4D torus network with a similar level of engineering difficulties and it is slightly better than a hypercube which contains an excessive number of links. We also introduced a node-to-node distance formula for facilitating message routing with our new network. The methodology can be easily generalized to studying systems with other node numbers, even to analyzing networks of higher dimensions.

REFERENCES

[1] Adiga, N.R., et al. An Overview of the BlueGene/L Supercomputer. in Supercomputing, ACM/IEEE 2002 Conference. 2002.

[2] Blumrich, M.A., et al., Design and Analysis of the BlueGene/L Torus Interconnection Network, in IBM Research Report. 2003.

[3] S.L. Scott, G.M.T., The Cray T3E Network: Adaptive Routing in a High Performance 3D Torus. Proc. Symp. High Performance Interconnects (Hot Interconnects 4), 1996: p. 157–160.

[4] Anderson, E., et al., Performance of the CRAY T3E multiprocessor, in Proceedings of the 1997 ACM/IEEE conference on Supercomputing (CDROM). 1997, ACM: San Jose, CA.

[5] TOP500. Top 500 Supercomputer Sites. Available from: http://www.top500.org.

[6] Barker, K.J., et al., Entering the petaflop era: the architecture and performance of Roadrunner, in Proceedings of the 2008 ACM/IEEE conference on Supercomputing. 2008, IEEE Press: Austin, Texas.

[7] Bhuyan, L.N. and D.P. Agrawal, Generalized Hypercube and Hyperbus Structures for a Computer Network. Computers, IEEE Transactions on, 1984. C-33(4): p. 323-333.

[8] Harary, F., J.P. Hayes, and H.-J. Wu, A survey of the theory of hypercube graphs. Computers & Mathematics with Applications, 1988. 15(4): p. 277-289.

[9] Efe, K., A Variation on the Hypercube with Lower Diameter. IEEE Trans. Comput., 1991. 40(11): p. 1312-1316.

[10] Kaushal, R.P. and J.S. Bedi, Comparison of hypercube, hypernet, and symmetric hypernet architectures. SIGARCH Comput. Archit. News, 1992. 20(5): p. 13-25.

[11] Heun, V. and E.W. Mayr, Efficient Embeddings into Hypercube-like Topologies. The Computer Journal, 2003. 46(6): p. 632-644.

[12] Preparata, F.P. and J. Vuillemin, The cube-connected cycles: a versatile network for parallel computation. Commun. ACM, 1981. 24(5): p. 300-309.

[13] Harwood, A., Hong Shen, A Low Cost Hybrid Fat-tree Interconnection Network. Proceedings of International Conference on Parallel and Distributed Processing and Applications. 1998.

[14] Parhami, B. and D.-M. Kwai, Incomplete k-ary n-cube and its derivatives. Journal of Parallel and Distributed Computing, 2004. 64(2): p. 183-190.

[15] Stojmenovic, I., Honeycomb Networks: Topological Properties and Communication Algorithms. IEEE Trans. Parallel Distrib. Syst., 1997. 8(10): p. 1036-1042.

[16] Decayeux, C. and D. Seme, 3D hexagonal network: modeling, topological properties, addressing scheme, and optimal routing algorithm. Parallel and Distributed Systems, IEEE Transactions on, 2005. 16(9): p. 875-884.

[17] Kirkman, W.W. and D. Quammen. Packed exponential connections-a hierarchy of 2-D meshes. in Parallel Processing Symposium, 1991. Proceedings., Fifth International. 1991.

[18] Inoguchi, Y. and S. Horiguchi, Shifted Recursive Torus Interconnection for High Performance Computing, in Proceedings of the High-Performance Computing on the Information Superhighway, HPC-Asia '97. 1997, IEEE Computer Society.

[19] Jain, V.K. and S. Horiguchi, VLSI considerations for TESH: a new hierarchical interconnection network for 3-D integration. Very Large Scale Integration (VLSI) Systems, IEEE Transactions on, 1998. 6(3): p. 346-353.

[20] Yang, Y., et al., Recursive Diagonal Torus: An Interconnection Network for Massively Parallel Computers. IEEE Trans. Parallel Distrib. Syst., 2001. 12(7): p. 701-715.

[21] Dally, W.J., Performance Analysis of k-ary n-cube Interconnection Networks. IEEE Trans. Comput., 1990. 39(6): p. 775-785.

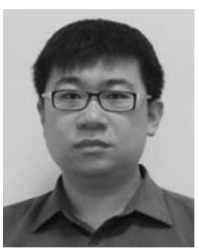
[22] Duato, J., S. Yalamanchili, and N. Lionel, Interconnection Networks: An Engineering Approach. 2002: Morgan Kaufmann Publishers Inc. 650.

[23] Dally, W. and B. Towles, Principles and Practices of Interconnection Networks. 2003: Morgan Kaufmann Publishers Inc.

[24] Parhami, B., Swapped interconnection networks: Topological, performance, and robustness attributes. Journal of Parallel and Distributed Computing, 2005. 65(11): p. 1443-1452.

[25] Chaki, N., et al., A New Logical Topology Based on Barrel Shifter Network over an All Optical Network, in Proceedings of the 28th Annual IEEE International Conference on Local Computer Networks. 2003, IEEE Computer Society.

[26] Samatham, M.R. and D.K. Pradhan, The de Bruijn multiprocessor network: a versatile parallel processing and sorting network for VLSI. Computers, IEEE Transactions on, 1989. 38(4): p. 567-581.



Peng Zhang is a Ph.D. student in the Applied Mathematics Department at Stony Brook University. He received his B.S. degree in mathematics from Nankai University in 2003 and M.S. degree in parallel computing from Nankai Institute of Scientific Computing in 2006.

Reid Powell is a Ph.D. student in the Applied Mathematics Department at Stony Brook University, where he received his B.S. in mathematics in 2001.

Yuefan Deng is a professor of applied mathematics at Stony Brook University with 20 years of experience of HPC research. He received his B.S. in physics from Nankai University in 1983 and M.A., M. Phil., and Ph.D. degrees from Columbia University in 1985, 1986 and 1989, respectively.

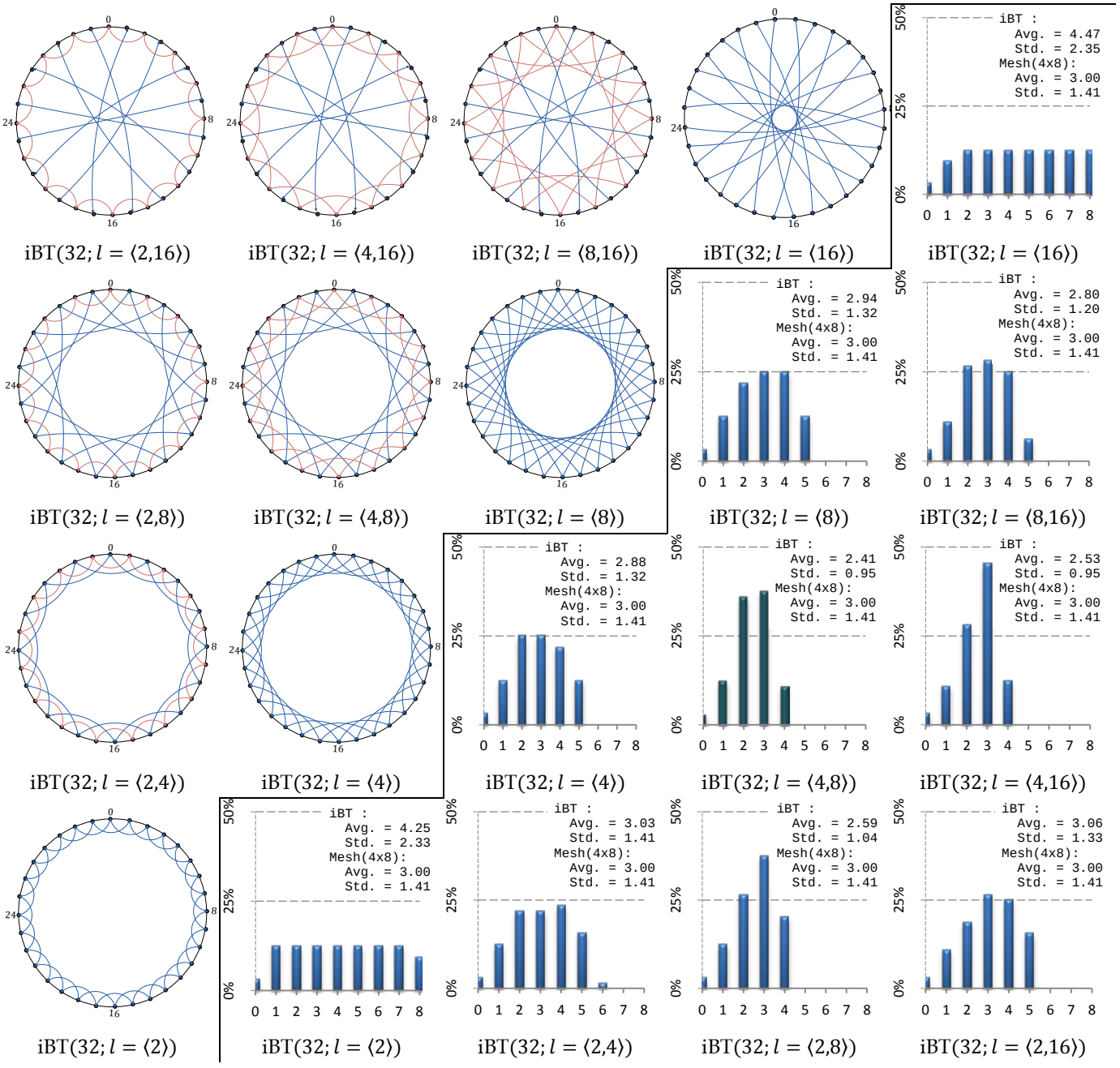


Fig. 1. The bypass scheme and the node-to-node hop distance distribution for iBT$(32; L = 1; l = \langle l_1, l_2 \rangle)$

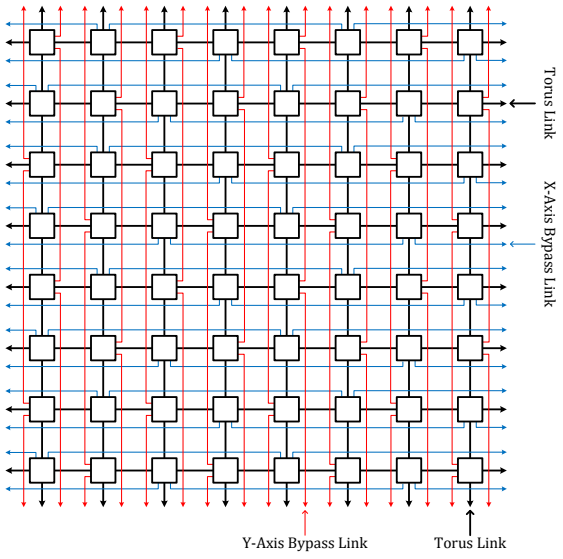


Fig. 2. All of the links in 2D iBT($8 \times 8; L = 2; l = \langle 4 \rangle$)

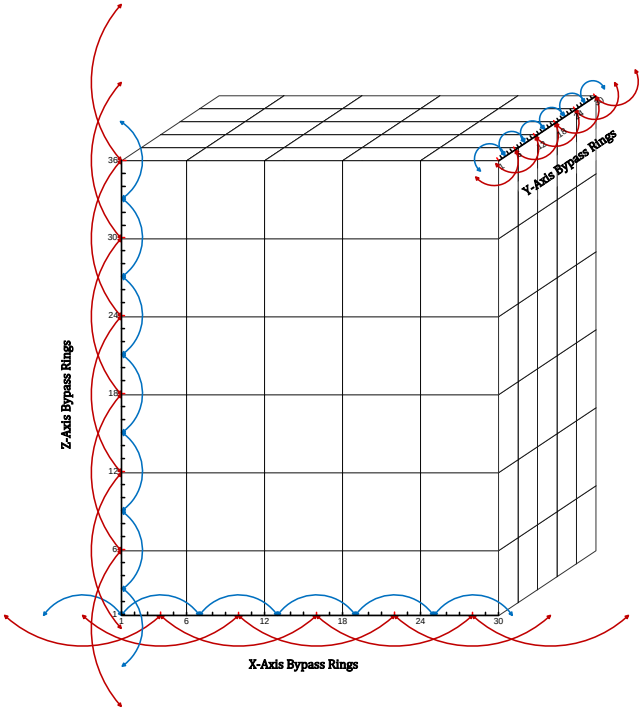


Fig. 3. The bypass scheme in 3D iBT($30 \times 30 \times 36; L = 3; l = \langle 6,12 \rangle$)

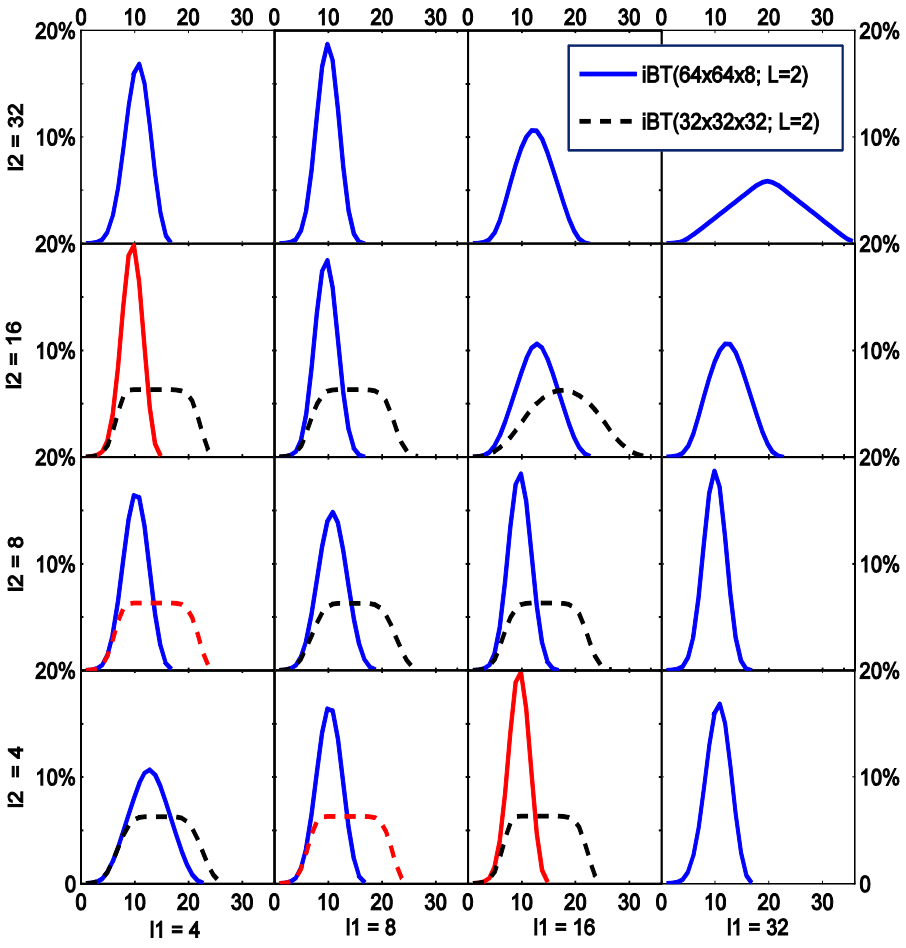


Fig. 4. Hop distance distribution for 3D iBT(L=2)

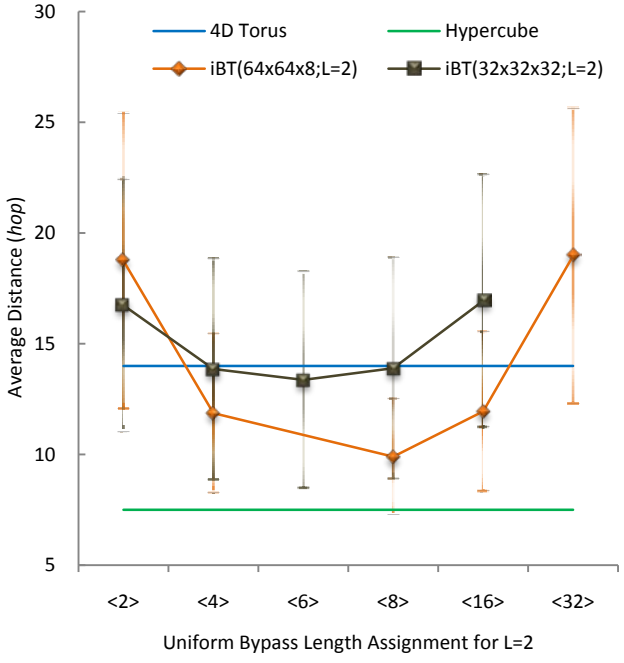


Fig. 5. Average distances and standard deviations for 3D iBT(L=2) with uniform bypass length, $T(16 \times 16 \times 16 \times 8)$ and $H(2^{15})$ networks

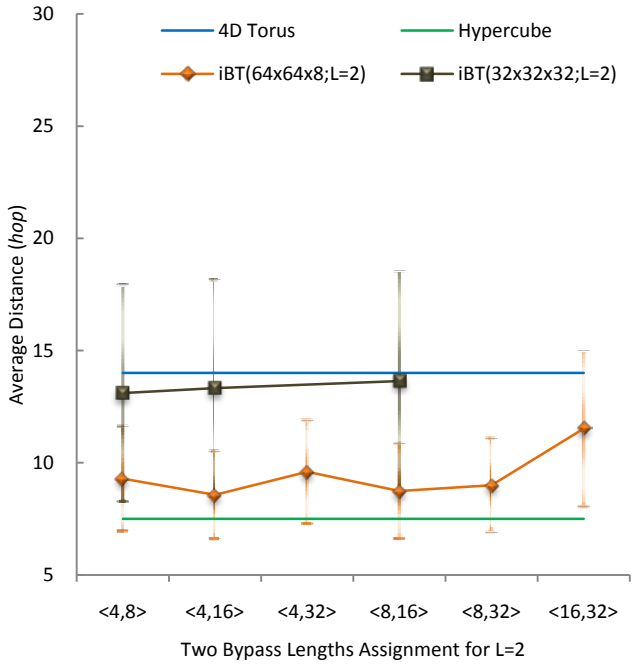


Fig. 6. Average distances and standard deviations for 3D iBT(L=2) with two bypass lengths, $T(16 \times 16 \times 16 \times 8)$ and $H(2^{15})$ networks

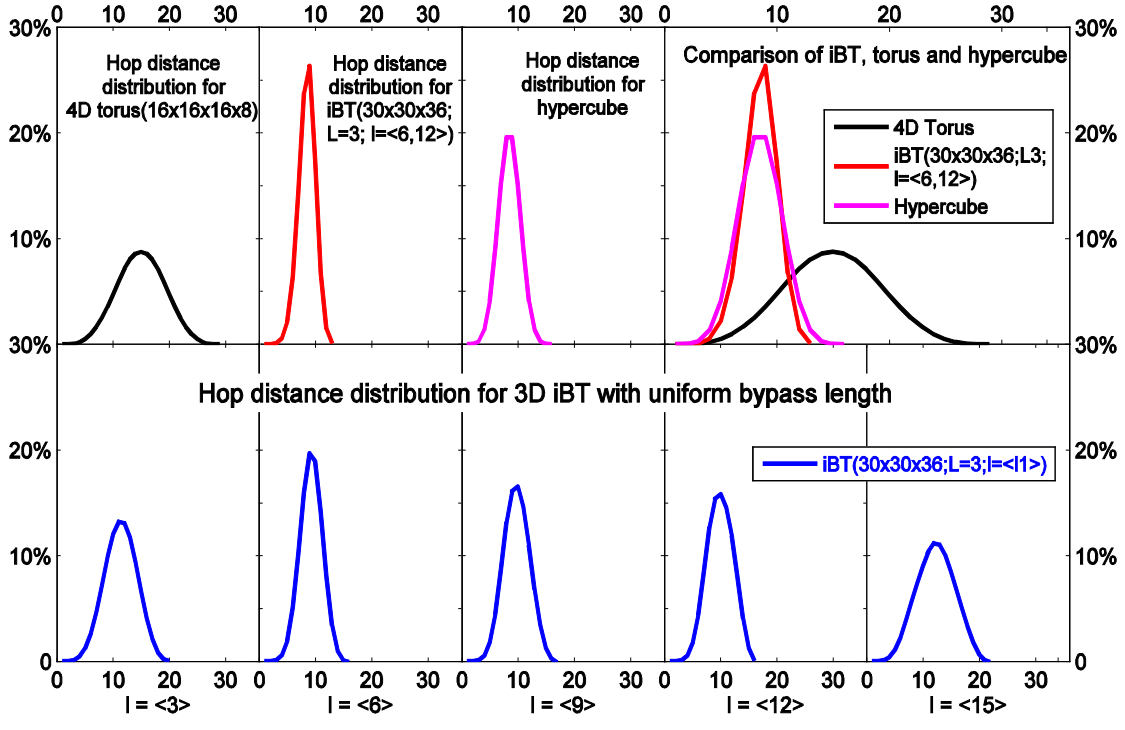


Fig. 7. Hop distance distribution for 3D iBT(L=3), $T(16 \times 16 \times 16 \times 8)$ and $H(2^{15})$ networks

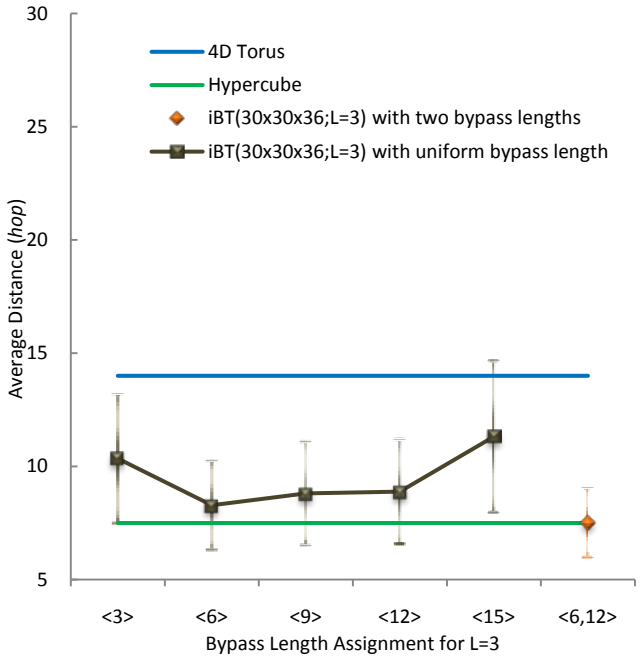


Fig. 8. Average distances and standard deviations for 3D iBT(L=3), $T(16 \times 16 \times 16 \times 8)$ and $H(2^{15})$ networks

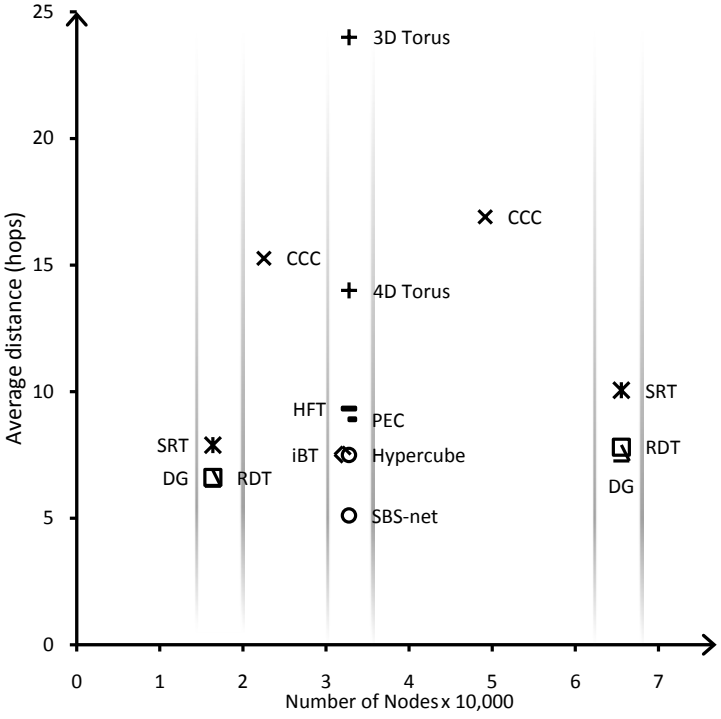


Fig. 9. Performance comparisons in average distance among different networks

TABLE 1
TOPOLOGICAL PROPERTIES OF IBT, TORUS AND HYPERCUBE INTERCONNECTION MODELS

iBT interconnection models			Node Degree	Hop Distance (*hop*)		Network Diameter (*hop*)	Bisection Width (*link*)	Network Cost
$N_x \times N_y \times N_z$	L	$l = \langle l_1, l_2 \rangle$		Average	Standard Deviation			
$32 \times 32 \times 32$	2	$\langle 2 \rangle$		16.7344	5.6961	33	2048	264
		$\langle 4 \rangle$		13.8593	4.9855	26		208
		$\langle 6 \rangle$		13.3730	4.8800	26		208
		$\langle 8 \rangle$		13.8984	4.9872	26		208
		$\langle 16 \rangle$		16.9414	5.6963	32		256
	2	$\langle 4,8 \rangle$		13.1035	4.8235	24		192
		$\langle 4,16 \rangle$		13.3257	4.8199	24		192
		$\langle 8,16 \rangle$		13.6416	4.9104	26		208
$64 \times 64 \times 8$	2	$\langle 2 \rangle$	8	18.7422	6.6664	37	2048	296
		$\langle 4 \rangle$		11.8672	3.5835	22	3072	176
		$\langle 8 \rangle$		9.9023	2.6172	18	6120	144
		$\langle 16 \rangle$		11.9434	3.5945	22	8192	176
		$\langle 32 \rangle$		18.9697	6.6687	36	8192	288
	2	$\langle 4,8 \rangle$		9.2908	2.3277	16	4096	128
		$\langle 4,16 \rangle$		8.5679	1.9477	14	6144	112
		$\langle 4,32 \rangle$		9.5942	2.2946	16	6144	128
		$\langle 8,16 \rangle$		8.7402	2.1133	16	7168	128
		$\langle 8,32 \rangle$		8.9987	2.0954	16	7168	128
		$\langle 16,32 \rangle$		11.5198	3.4786	22	8192	176
$30 \times 30 \times 36$	3	$\langle 3 \rangle$		10.3464	2.8542	19	3600	152
		$\langle 6 \rangle$		8.2800	1.9675	15	5400	120
		$\langle 9 \rangle$		8.8034	2.2895	16	7200	128
		$\langle 12 \rangle$		8.8827	2.3044	15	9000	120
		$\langle 15 \rangle$		11.3114	3.3441	21	7560	168
	3	$\langle \mathbf{6, 12} \rangle$		**7.5152**	**1.5288**	**12**	**7200**	**96**
3D Torus$(32 \times 32 \times 32)$ [1-4]			6	24.0000	8.0312	48	2048	288
4D Torus$(16 \times 16 \times 16 \times 8)$			8	14.0000	4.2426	28	4096	224
Hypercube(2^{15})			15	7.5000	1.9365	15	16384	225
PEC(256×128) (32,768 nodes) [17]			8	8.9068	1.8342	15	1920	120
2D SRT(128×128) (16,384 nodes) [18]			8	7.8904	1.6543	13	1664	104
2D SRT(256×256) (65,536 nodes) [18]			8	10.0529	1.9974	16	3840	128
RDT$(2,4,1)/\alpha(128 \times 128)$(16,384 nodes) [20]			8	6.6113	1.2340	10	5632	80
RDT$(2,4,1)/\alpha(256 \times 256)$(65,536 nodes) [20]			8	7.8076	1.4521	12	23552	96
CCC 11-11 (22,528 nodes) [12]			3	15.2685	2.8432	25	1024	75
CCC 12-12 (49,152 nodes) [12]			3	16.9020	2.9487	28	2048	84
Scalable Barrel Shifter (32,768 nodes) [25]			29	5.1111	1.1000	8	49150	232
de Bruijn Graph $DG(4,7)$ (16,384 nodes) [26]			8	6.0287	0.9025	7	32768	56
de Bruijn Graph $DG(4,8)$ (65,536 nodes) [26]			8	7.0145	0.9134	8	131072	64
Hybrid Fat Tree (32,768 nodes) [13]			4 [1]	9.3334	1.6922	15	3	60

[1]: 32,768-node hybrid fat tree [13] has the average node degree of 4 while its minmum and maximum degrees are 2 and 29.