NETWORK DESIGN CONSIDERATIONS FOR EXASCALE SUPERCOMPUTERS

Rui Feng¹, Peng Zhang², Yuefan Deng²
¹School of Computer Science and Engineering
Beihang University
Beijing, P.R.China
fengru@act.buaa.edu.cn

²Department of Applied Mathematics and Statistics
Stony Brook University
Stony Brook, NY, United States
{peng.zhang, yuefan.deng}@stonybrook.edu

ABSTRACT
We consider the network design optimization for the Exascale class supercomputers by altering the widely analyzed and implemented torus networks. Our alteration scheme involves interlacing the torus networks with bypass links of lengths 6 hops, 9 hops, 12 hops, and mixed 6 and 12 hops. These bypass links are optimal resulting from exhaustive search of massive possibilities. Our case study is constructed by strategically coupling 288 racks of 6 × 6 × 36 nodes to a full system with 72 × 72 × 72 nodes. The peak performance of such a system is 0.56 Exa-flops when CPU-GPU complexes are adopted as a node module capable of 1.5 Tflops. Our design optimizes, simultaneously, the system performance, performance-cost ratio, and power efficiency. The network diameter and the average node-to-node network distance, regarded as the performance metrics, got reduced, from the original 3D torus network, by 83.3% and 80.4%, respectively. Similarly, the performance-cost ratio and power efficiency are also increased 1.43 and 4.44 times, respectively.

KEY WORDS
Interconnection Network, Parallel Architecture, Network Diameter, Supercomputer

1. Introduction
Designing, constructing, and applying the Exascale supercomputers represents the next frontiers of computer and computational sciences [1]. Engineers must pack 1,000,000 nodes with 64,000,000 GB of memory, and 10×64,000,000 GB of disk storage in a space no bigger than 10,000 square meters and operate it at no more than 20 MW of electricity [2]. With all these constraints, applications developers wish to deliver 10¹⁸ flops of performance.

Many efforts are currently under way in multiple research-and-development centers, are examining multiple technological sectors [3]. Computer developers are focusing on system design optimization in mechanical, thermo, electrical, and electronics subsystems. Others focus on software and application algorithms.

Interconnection networks are the heart of the supercomputers and two categories of interconnection networks [4-6] are potential candidates for Exascale supercomputers. Federated networks, adopted by Tianhe-1A and Nebulae supercomputers [3], connect nodes by off-the-shelf solutions such as Infiniband. Cellular networks, developed for systems as the IBM Blue Gene solutions [7-9] and the Cray XT series supercomputers [10], utilize proprietary interconnection techniques. The latest Blue Gene series, Blue Gene/Q [11], uses a 5D torus network to connect its 17-core nodes. The new network with a higher node degree improves the performance at the expenses of more engineering complications and higher monetary costs [12].

Recent development in adding bypass links to conventional networks has gained attention [13, 14]. The iBT network adds performance and reduces the implementation difficulties in the process of evolving from the 3D torus network, compared with the brand new design of the 5D torus.

This article focuses on analysis of various configurations of the interconnection networks with objective of finding the optimal performance-cost (p/c) ratios. Section 2 introduces the overall system-level design. Section 3 presents the design considerations including expansion schemes for a supercomputer with 373,248 nodes. In Section 4, we compare our design with the original torus systems in terms of network performance, p/c ratios, and power-efficiencies. Conclusions are summarized in Section 5.

2. iBT Interconnection Network
In [15], a new interconnection network called iBT is proposed by interlacing bypass rings to torus networks. A general d-dimensional iBT network constructed from a torus network of dimensions $N_i \times \cdots \times N_d$ can be expressed as

$$iBT(N_1 \times \cdots \times N_d; L = m; b = \langle b_1, \ldots, b_k \rangle),$$

where the bypass link sequence $b = \langle b_1, \ldots, b_k \rangle$ is a vector whose component values increase monotonically, i.e., $b_1 < b_2 < \cdots < b_k$, and it indicates that $b_i$-hop bypass rings ($i = 1, \ldots, k$) are recursively interlaced into any $m$ of the $n$ dimensions ($m \leq n$). The node degree of the resulting iBT network is $2d + 2$ where $2d$ is the node degree for the original torus network and the added 2 is for the bypass connections. To determine the two bypass
connections for a node \( p = (x_1, x_2, \ldots, x_d) \) where \( x_i \in [0, N_i - 1] \) with \( i = 1, \ldots, d \), we introduce three terms: a nodal bypass dimension \( bd(p) \in \{1, 2, \ldots, m\} \), a nodal bypass length

\[
bd(p) = \left( \sum_{i=1}^{m} x_i \right) \mod m + 1 \quad \text{and} \quad bl(p) = b_h,
\]

and thus a nodal bypass species \( bs(p) = (bd(p), bl(p)) \) indicating that two \( bl(p) \)-hop bypass links have been added to the given node \( p \) in each direction along the dimension \( bd(p) \). For example,

\( iBT(32 \times 32 \times 16; b = (4,16)) \)

indicates interlacing 4-hop and 16-hop bypass rings in the \( xy \)-plane of the 3D torus \( T(32 \times 32 \times 16) \). In this network, a node \( p = (1,1,4) \) has \( bd(p) = 1, bl(p) = b_2 = 16 \) and thus \( bs(p) = (1,16) \) indicating that \( p \) has two 16-hop bypass links in each direction along the \( x \) dimension.

Such an interconnection network is a desirable network for an Exascale supercomputer because of its achievable absolute network performance with a given low engineering complexity, high performance-cost ratio, and high power efficiency. Take a 3-D iBT network, a specific network notation:

\( iBT(N_x \times N_y \times N_z; b_x, b_y, b_z) \)

indicates that bypass rings were interlaced to all three dimensions of a 3-D Torus \( T(N_x \times N_y \times N_z) \). The bypass configuration in each dimension is illustrated in \( b_x, b_y \) or \( b_z \). For another example,

\( iBT(6 \times 24 \times 72; b_x = (0), b_y = (6), b_z = (6,12)) \)

indicates that:

1. Base network: 3-D Torus \( T(6 \times 24 \times 72) \);
2. No bypass in \( x \)-dimension indicated by \( b_x = (0) \);
3. Uniform 6-hop bypass in \( y \)-dimension indicated by \( b_y = (6) \);
4. A mix of 6-hop and 12-hop bypasses in \( z \)-dimension indicated by \( b_z = (6,12) \).

3. Design Analysis

Our design follows the common principle in optimizing system performance, enhancing modularity, reducing engineering complexity, and lowering monetary costs [1]. The lowest-level module of our design is a node that performs two functions: computation and communication with its directly connected neighboring nodes. A multiple processing unit or MPU is the 2nd level module with the architecture of \( iBT(6 \times 6 \times 6; b_x, b_y, b_z) \), containing \( 6 \times 6 \times 6 = 216 \) nodes with a given network architecture. The 3rd level module is a rack that is constituted of six MPUs architected as \( iBT(6 \times 6 \times 36; b_x, b_y, b_z) \). The highest level, i.e., the fourth level, is the complete system that can contain an appropriated number of such racks according to budget and performance requirements as well as engineering feasibility. For example, a bi-rack system architected as \( iBT(6 \times 6 \times 72; b_x, b_y, b_z) \) allows full connections of all \( z \)-dimensional inter-rack torus and bypass links. One can further expand to achieve a system of desirable size by connecting such bi-rack entities in \( x \)-or \( y \)-dimension or both. As we will discuss more, a system architected as \( iBT(72^3; b_x, b_y, b_z) \) can achieve an optimal performance at the Exascale. It will have 288 racks of nodes capable of 1.5 Tflops performance per node in 2011.

After extensive studies, we perform analysis on four bypass configurations for the \( iBT(72^3; b_x, b_y, b_z) \) with bypass vectors \( \{(6),(9),(12),(6,12)\} \). With such analysis, we determine the bypass configuration in each dimension.

3.1 Design Considerations

When evaluating a design, we consider two factors: the network characteristics and the p/c ratio. The network characteristics include the network diameter and average node-to-node distance. The network performance is defined as the reciprocal of the average distance (A), i.e., network performance = \( 1/A \).

We use the extra external inter-rack wires in our iBT design over the Blue Gene’s design as our cost metric which is defined as \( C_0 \cdot l \) where \( l \) is the total external wire length and \( C_0 \) is the wires’ unit cost. Thus, the material cost of all of the external wires \( l_i \) \( \forall i \) is

\[ \text{material cost} = \sum_{i} C_0 l_i \]

The p/c ratio, defined as the performance divided by the material cost, is written as

\[ f_{pc} = \frac{\text{network performance}}{\text{material cost}} = \frac{1}{(C_0 \cdot \sum_{i} l_i) \cdot A} \]

By analyzing the relationship of these factors, we identify a class of optimized design plans with specific requirements and trade-offs.

3.2 MPU: Multiple Processing Unit

Fig. 1 shows the multiple processing unit (MPU) architecture which has the following properties:

1. 216 nodes interconnected as a 3-D mesh \( (6\times6\times6) \);
2. 432 external bypass links, two links per node;
3. 216 external torus links, two links per node at the boundaries;
4. 648 external links, in total, for connecting to other MPUs. These 648 links are categorized into 18 groups with 36 links per group.

To classify the external links, we assign each node relative coordinates \( p = (x, y, z) \) where \( x, y, z \in [0,5] \) and let \( s(p) = x + y + z \). Let the \( x \)-dimensional bypass configuration be \( b_x = (b_{xy}, b_{xz}) \). The same notations hold for \( y \)- and \( z \)-dimension. With such, we have

\[ bd(p) = s(p) \mod 3 + 1 \in \{1,2,3\}, \quad bl(p) = b_h, \]

where \( h = \left\lfloor s(p) \mod 6 \right\rfloor + 1 \in \{1,2\} \).
A node is defined as a boundary node if and only if one of its three coordinates is equal to 0 or 5. Accordingly, we classify these external links into 18 groups of which 6 belong to the torus groups and 12 belong to the bypass groups marked in Fig. 2. For example, \( t_x \) contains one link of each node with coordinates \((0, y, z)\). \( b_y \) contains one link of each node with \( b_s(p) = (2, b_z)\), same as other groups. Thirty-six external torus or bypass links are bundled in one group and this group provides the inter-MPU connections.

Fig. 1. External torus and bypass links in three typical planes of a MPU

Fig. 2. External link groups of a MPU

3.3 Rack

A rack consists of six internally connected MPUs with external links for inter-rack connections. The six MPUs can be arranged in two configurations: \( iBT(6 \times 12 \times 18) \) as shown by Fig. 3(a) and \( iBT(6 \times 6 \times 36) \) as shown by Fig. 3(b). No bypass link is added in x-dimension, *i.e.*, \( b_x = (0) \) because of too few nodes in this dimension. Determination of the \( y \)- and \( z \)-bypasses requires numerical experiments and they show results in Table 1.

It appears a bypass scheme \( b_z = (6,12) \) shows an optimal performance among the six possibilities and, thus, \( iBT(6 \times 6 \times 36; b_x, b_y, b_z = (6,12)) \) is selected as our rack configuration.

For identifying a node, we introduce a MPU number and a rack number in addition to the nodal relative coordinates. A MPU number is assigned for identifying the MPU’s position in a rack while a rack number is assigned for identifying the rack’s position in a system.

A rack has 936 external torus links and 1,944 bypass links classified in 18 link sets.

3.4 Z-Expansion

To achieve the final system design goal of having 72 nodes in each dimension, we start assembling the system first from the z-dimension to form a rack pair that has the

![Image](image-url)

**Fig. 3 (a, b)** Internal connections in a rack with the architectures: (a) \( iBT(6 \times 12 \times 18) \) and (b) \( iBT(6 \times 6 \times 36) \)

**Table 1**

<table>
<thead>
<tr>
<th>Base Networks</th>
<th>Configurations</th>
<th>( b_y )</th>
<th>( b_z )</th>
<th>Diameter</th>
<th>Average Distance</th>
<th>Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>( T(6 \times 12 \times 18) )</td>
<td>(6)</td>
<td>(6)</td>
<td>10</td>
<td>5.56</td>
<td>1.64</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(9)</td>
<td>11</td>
<td>6.16</td>
<td>1.92</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>( T(6 \times 6 \times 36) )</td>
<td>(0)</td>
<td>(6)</td>
<td>11</td>
<td>5.94</td>
<td>1.82</td>
</tr>
<tr>
<td></td>
<td>(9)</td>
<td>12</td>
<td>6.23</td>
<td>2.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(12)</td>
<td>13</td>
<td>6.65</td>
<td>2.24</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>( (6,12) )</td>
<td>10</td>
<td>5.69</td>
<td>1.60</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
architecture iBT \((6 \times 6 \times 72; \mathbf{b}_x, \mathbf{b}_y, \mathbf{b}_z = (6,12))\). Inter-rack connections are shown in Fig. 4.

This modular design packs all z-dimensional external links, i.e. 72 torus and 216 bypass links, in one rack pair, allowing us to duplicate the rack pair by connecting its x- and y-dimensional external links, i.e. 864 torus and 1728 bypass links. We classified such links into 12 link sets.

3.5 Y-Expansion

We now expand the system along y-dimension by arranging the 12 rack pairs in a row, resulting in \(\text{iBT}(6 \times 72 \times 72)\). Similarly, we have four bypass configurations in y-dimension: \(\mathbf{b}_y \in \{\langle 6 \rangle, \langle 9 \rangle, \langle 12 \rangle, \langle 6,12 \rangle\}\). Fig. 5 shows the inter-rack connections with the same torus connection but different bypass patterns.

Fig. 5. Y-dimensional external links in the architectures: (a) \(\text{iBT}(\mathbf{b}_y = \langle 6 \rangle)\), (b) \(\text{iBT}(\mathbf{b}_y = \langle 12 \rangle)\) and (a) \(\text{iBT}(\mathbf{b}_y \in \{\langle 9 \rangle, \langle 6,12 \rangle\})\)

From the Fig. 5 we can see that:

1. \(\text{iBT}(\ldots, \mathbf{b}_y = \langle 6 \rangle)\) has the same inter-rack connection pattern as a 3D Torus network but two times more wires per linking bundle.
2. \(\text{iBT}(\ldots, \mathbf{b}_y = \langle 12 \rangle)\) has the longest wirings and the same wires per linking bundle as \(\text{iBT}(\mathbf{b}_y = \langle 6,12 \rangle)\).
3. \(\text{iBT}(\ldots, \mathbf{b}_y = \langle 9 \rangle)\) and \(\text{iBT}(\mathbf{b}_y = \langle 6,12 \rangle)\) use the same inter-rack connection pattern as a mixture of the above two patterns.

Fig. 6 shows the connection of 12 rack-pairs in a row in the implementation of \(\text{iBT}\ (6 \times 72 \times 72; \mathbf{b}_x = (0), \mathbf{b}_y = (6,12))\).

Fig. 7 compares the network performance and p/c ratios of the four \(\mathbf{b}_y\) configurations as the number of racks increases. We found two configurations that warrant further consideration:

1. Configuration with \(\mathbf{b}_y = \langle 6 \rangle\) always achieves the lowest p/c ratio as well as the best network performance when systems size is less than 8 racks;
2. Configuration with \(\mathbf{b}_y = \langle 6,12 \rangle\) reduces the diameter and the average distance by 11.8% and 9.4% than that of \(\mathbf{b}_y = \langle 6 \rangle\) does but it also reduces 13.9% p/c ratio.
Thus, if minimizing the p/c costs is the objective, $b_y = (6)$ is the best choice. If maximizing network performance is the objective, $b_y = (6,12)$ is the best choice. In the following sections we limit our discussions on the bypass configurations with $b_y \in \{ (6) , (6,12) \}$.

### 3.6 X-Expansion

After completing one-row y-expansion, we assemble the system along x-dimension by connecting multiple rows of 12 rack pairs per row to achieve the architecture $\text{iBT}(N_x \times 72 \times 72; b_x, b_y, b_z = (6,12))$, where $b_y, b_x \in \{ (6), (6,12) \}$.

Fig. 8 shows the three architectures of entire system with 288 racks:

- $\text{iBT}(12 \times 72 \times 72; b_x = (6), b_y = b_z = (6,12))$
- $\text{iBT}(24 \times 72 \times 72; b_x = b_y = (6), b_z = (6,12))$
- $\text{iBT}(36 \times 72 \times 72; b_x = b_y = b_z = (6,12))$

We assume the distances between two adjacent rows and between two adjacent racks in a row to be $a = 1.22$ and $b = 1.88$ (meters) [16].

Fig. 9 compares the network performance and the p/c ratios when systems expand with multiple rows. These results show that:
1. Configuration with $b_x = b_y = (6)$ has the highest network p/c ratio but also the lowest performance;
2. Configuration with $b_x = b_y = (6,12)$ has the highest performance but also the lowest p/c ratio, for 2 or more rows;
3. Configuration with $b_x = (6), b_y = (6,12)$ has the moderate performance and moderate p/c ratio and is no better than the above two.

Orchestrating consideration of the y- and x-expansions, we conclude $iBT(b_x = b_y = (6), b_z = (6,12))$ is the optimal architecture if maximizing network p/c ratio is the design objective while $iBT(b_x = b_y = (6), b_z = (6,12))$ is optimal if maximizing network performance is the design objective.

### 3.7 XY-Expansion

In addition to independent y- and x-expansions, we consider simultaneous expansion in both X and Y dimensions. For architecture $iBT(N_x \times N_y \times 72; b_x, b_y, b_z = (6,12))$, we choose the bypass schemes for x- and y-dimension as:

$b_x = \begin{cases} (6), N_x \leq 24 \\ (6,12), \text{otherwise} \end{cases}$
$b_y = \begin{cases} (6), N_y \leq 24 \\ (6,12), \text{otherwise} \end{cases}$

Fig. 10 shows the variations of the performance and p/c ratios as we expand the system from 16 to 288 racks. For example, for 96 racks $iBT(48 \times 36 \times 72)$ outperforms $iBT(72 \times 24 \times 72)$, $iBT(36 \times 48 \times 72)$, and $iBT(24 \times 72 \times 72)$.

### 4. Comparisons with Torus

For further comparisons between the selected iBT configurations and the original 3D torus, we consider the operational cost of electricity to power it up and cool it off. Moving $N_b$ bits on a copper wire consumes [2]

$\text{energy} = rN_b l^2 / a$

where $r$ is the bit rate, $l$ is the wire length, and $a$ is the cross section area of wires. Thus, the energy cost of moving $N_b / N_w$ bits on each of $N_w$ external wires is

$\text{energy} = rN_b \sum_l l^2 / N_w a = \frac{C_1 \sum_l l^2}{N_w}$

where $C_1 = rN_b / a$.

The power efficiency is defined as

$f_{pe} = \frac{\text{network performance}}{\text{energy}} = \frac{N_w}{C_1 \sum_l l^2 \cdot A}$

The $iBT(b_x, b_y, b_z)$ networks, with very minimal design variation over the popular 3D torus, outperforms the later greatly in both categories. Figs. 11 to 13 compare the configurations with the best network performance and the best p/c ratio over the original 3D torus for the y- and x-expansions respectively. Fig. 11 shows the network performance ratios of iBT networks, defined as the diameter of torus divided by the diameter of iBT, a 5-time performance gain. Figs. 12 and 13 show the network performance-cost ratios and power efficiency of iBT networks over torus of the same network size, respectively.
Summarizing the above, we see:
1. Both iBT architectures outperform the 3D Torus of the same node count. The performance-optimized iBT reduces the network diameter and average distance by 83.3% and 80.4% over the 3D Torus, and the cost-optimized iBT reduces the diameter and average distance by 79.6% and 77.0%, respectively;
2. The network performance-cost ratio of the comparable iBT network is 1.43 times that of 3D torus with the same node dimensions of $72 \times 72 \times 72$;
3. The network power efficiency of the comparable iBT network is 4.44 times that of 3D torus with the same node dimensions of $72 \times 72 \times 72$.

A supercomputer with our proposed iBT network coupled with $72 \times 72 \times 72$ CPU-GPU nodes capable peak performance of 1.5 TFlops per node can achieve 0.56 Exa-flops.

5. Conclusion

Through extensive analyses of a technique of alteration of the widely adopted 3D torus networks for supercomputers, we proposed a much more monetary and energy efficient architecture. The new architecture results from adding 6-hop, 9-hop, 12-hop, or 6- and 12-hop bypass links to x-, y-, and z-directions of the corresponding 3D torus network. Our methodology can be applied to general architectures and our case study of a system of $72 \times 72 \times 72 = 373,248$ nodes demonstrates the procedure and the values of our analysis. We analyzed these four bypass configurations, while expanding the system, in terms of network diameter, average distance, distance deviation, length of total external links, number of external links, and relative network costs (average distance $\times$ total link length) for optimal performance and costs. Our comparisons of the performance, price-performance ratios, and power efficiency between iBT architecture and original 3D Torus show that: 1) All four configurations demonstrate significant gain in performance and price-performance ratio; 2) Configurations with parameters $b_x = b_y = (6, 12)$ and $b_x = b_y = (6)$ are optimal for a performance-optimized system and a cost-optimized system, respectively.

Acknowledgements

This work is supported by the National High-Tech Research and Development Plan of China (Project 863) under Grand No. 2009AA012201, and the Shanghai Science and Technology Development Fund under Grand No. 08dz1501600.

References