Available online at www.sciencedirect.com
SCIENCE DIRECT
ELSEVIER
Mathematical and Computer Modelling 40 (2004) 701–720
MATHEMATICAL AND COMPUTER MODELLING
www.elsevier.com/locate/mcm

Comparing Four Classes of Torus-Based Parallel Architectures: Network Parameters and Communication Performance

B. PARHAMI† AND DING-MING KWAI
Department of Electrical and Computer Engineering
University of California
Santa Barbara, CA 93106-9560, USA
parhami@ece.ucsb.edu

(Received April 2003; revised and accepted August 2004)

Abstract—The relative communication performance of low- versus high-dimensional torus networks (k-ary n-cubes) has been extensively studied under various assumptions about communication patterns and technological constraints. In this paper, we extend the comparison to torus networks with incomplete, but regular, connectivities. Taking an nD torus as the basis, we show that a simple pruning scheme can be used to reduce the node degree from $2n$ to 4, while preserving many of the desirable properties of the intact network. Orienting the torus links (removing half of the channels) provides a second form of pruning that leads to (multidimensional) Manhattan street networks. Finally, combined pruning and orientation yields the fourth class of toroidal networks studied here. We compare the static performance parameters of these networks and evaluate their dynamic communication performance under the assumptions of virtual cut-through switching and constant pin count. The 3D case, leading to networks that are efficiently realizable with current technology, is used to demonstrate and quantify the performance benefits. Our results reinforce, extend, and complement previous studies that have demonstrated the performance advantages of low-dimensional k-ary n-cubes over higher-dimensional ones. For example pruned 3D tori provide additional design points that fall between 2D and 3D tori in terms of implementation complexity but can outperform both of these standard architectures. Thus, from a practical standpoint, pruning introduces additional flexibility in implementation options and trade-offs available to designers.

Keywords—Analytic performance evaluation, Incomplete torus, k-ary n-cube, Manhattan street network, Pruned torus, Symmetric network, Virtual cut-through.

1. INTRODUCTION

Advances in electronic technology have led to unprecedented processing power and storage capacity being packed in a single microchip. Even though the long awaited GHz processors emerged right on schedule [1], future technology may prove incapable of keeping pace with a tradition of progress which has led to five orders of magnitude in performance increase over the past three decades. As an alternative to reliance on faster hardware, large-scale parallelism has become the universal hope of high-performance computing [2].

†Author to whom all correspondance should be addressed.

0895-7177/04/$ - see front matter Typeset by AMS-TEX
doi:10.1016/

After experimentation with many different architectures, particularly the hypercube and its various derivatives, practical highly parallel architectures have converged on designs based on 2D and 3D arrays. Several parallel machines have adopted such networks for interprocessor communication, including two that surpassed teraFLOPS performance with 3D configurations as early as 1996 [3,4]. There are several reasons for this convergence. One key factor is the difficulty of realizing higher-dimensional architecture with what essentially amounts to a 2.5D implementation technology that doubly penalizes such structures due to their need for denser connectivity and longer wires [5]. Even when optical interconnections become economically viable, going beyond 3D connectivity poses challenging problems in layout and packaging. On the other hand, networks that are structured in two or three dimensions can be mapped naturally to the physical space, thus simplifying their hardware realizations.

A second key factor is that when implementation cost is normalized, low-dimensional arrays have been shown to have an edge in performance. This is due to a combination of short, regular connections that allow higher clock rate, simpler logic for routing decisions, and wider data paths that are possible with the same pin count. Together, these factors more than offset the negative effects of less favorable topological parameters such as diameter and bisection bandwidth.

Our results add another dimension to such trade-offs. Take a 3D torus, for example. The connectivity of this architecture can be reduced in two ways:

(1) replacing it by a 2D structure, and
(2) removing some of the links from each node.

Both approaches simplify the physical realization and can potentially lead to improved performance due to the considerations outlined in the preceding paragraph. However, there is a real possibility that pruning a torus leads to complications in routing and algorithm implementation due to irregular or less regular structure. We will show that in fact pruning can be done in such a way that the resulting network remains node- and edge-transitive. Such symmetry properties are critical to the simplicity of routing algorithms and their amenability to analytical evaluation.

Our approach leads to a unified view of nD toroidal networks whose in- and out-degrees are regularly decreased from $2n$ to a small constant number; for example, from six to four, three, or two in the case of 3D torus. We formulate these networks as algebraic graphs, prove that they remain both node- and edge-transitive, study their topological properties, and evaluate them both in terms of static measures (e.g., diameter or average internode distance) as well as dynamic communication performance under various traffic loads. Though not studied in this paper, our unified view covers a variety of other networks that have been found useful in the past, including torus variants such as honeycomb and diamond networks [6].

Our presentation is organized as follows. In Section 2, we describe the four classes of toroidal networks under consideration, with their symmetry properties based on the Cayley digraph model derived in Section 3. Section 4 compares the static performance parameters such as diameter, average internode distance, and bisection (band)width. Section 5 contains a delay throughput relationship for performance comparisons in a dynamic context. Comparative performance results are presented in Section 6. Section 7 contains our conclusions.

2. FOUR TYPES OF TOROIDAL NETWORKS

To increase the efficiency in utilizing the available communication bandwidth, which may be considered a constant under the assumption of limited I/O resources, three variations of the complete torus with bidirectional links have been considered. We call these variants "toroidal" because they are derived from tori and each is a subgraph of a torus of the same size. Although all of these networks were known previously, proof of their symmetry properties, derivation of some topological parameters, unified formulation, and performance comparisons are new here.

The first variant is the directed version derived by *orienting* links of a torus. Orientation is standard graph-theory terminology for converting an undirected edge to a directed one. Choosing

a uniform orientation along each dimension permits simple dimension-order routing and is often discussed in the literature on performance analysis by virtue of its simplicity [7–9]. A more efficient orientation, in the sense of causing minimal increase in the longest and average internode distances [10], is to alternately assign opposite directions to the links of each dimension. The 2D special case of the latter scheme, known as Manhattan street network [11], has been studied extensively [12–15].

The second variant is the pruned version derived by removing some of the links [16–18]. The 3D toroidal network of the Tera MTA (multithreaded architecture) [19] is obtained by selecting one dimension, keeping its links intact, and alternately removing links of the remaining two dimensions along the chosen dimension. When drawn as graphs, such pruned networks bear a superficial resemblance to the bus-based structures proposed by Wittie [20], even though the two classes of networks are quite different, both topologically and from algorithmic and performance standpoints. Incorporating both the orientation and pruning strategies, so as to yield a pruned directed torus, has also been proposed [21].

We can, therefore, classify the various toroidal networks into four categories based on the two dichotomies of undirected versus directed and unpruned versus pruned (Figure 1). For each of the torus variants, the trade-offs between cost and performance have been previously justified by the fact that the longest and average internode distances are only slightly larger than those of the complete torus. Thus, at least for light traffic loads, performance comparable to that of the complete torus can be obtained with the pruned versions at lower cost.

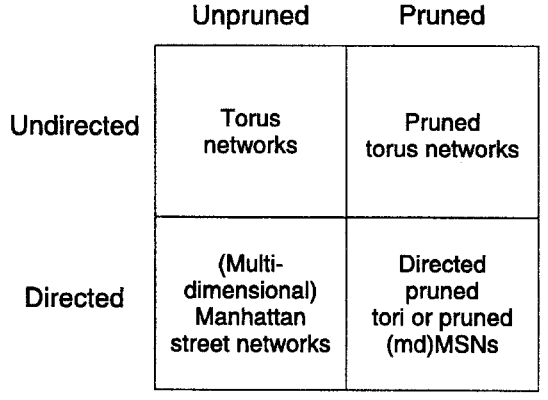


Figure 1. The four classed of toroidal networks studied in this paper.

We express node indices as column vectors. In an unpruned undirected nD torus, each node $(a_0, a_1, \ldots, a_i, \ldots, a_{n-1})$ is adjacent to $2n$ neighbors: $(a_0, a_1, \ldots, a_i \pm 1, \ldots, a_{n-1})$, $0 \leq i \leq n-1$. Here and throughout, it is understood that all node-index expressions are evaluated modulo k in view of the wraparound connections.

By selecting dimension 0 as the basis and alternately removing dimension-i links, $1 \leq i \leq n-1$, along dimension 0, each node $(a_0, a_1, \ldots, a_i, \ldots, a_{n-1})$ in the pruned undirected nD torus will be adjacent to four neighbors (Figure 2)

$$(a_0 \pm 1, a_1, \ldots, a_i, \ldots, a_{n-1}),$$
$$(a_0, a_1, \ldots, a_i \pm 1, \ldots, a_{n-1}), \qquad \text{if} \quad a_0 \equiv (i-1) \bmod (n-1).$$

In the directed nD torus, each node $(a_0, a_1, \ldots, a_i, \ldots, a_{n-1})$ is adjacent to n out-neighbors

$$(a_0, a_1, \ldots, a_i + 1, \ldots, a_{n-1}), \quad \text{if} \sum_{j=0(j\neq i)}^{n-1} a_j = \text{even},$$

$$(a_0, a_1, \ldots, a_i - 1, \ldots, a_{n-1}), \quad \text{if} \sum_{j=0(j\neq i)}^{n-1} a_j = \text{odd}.$$

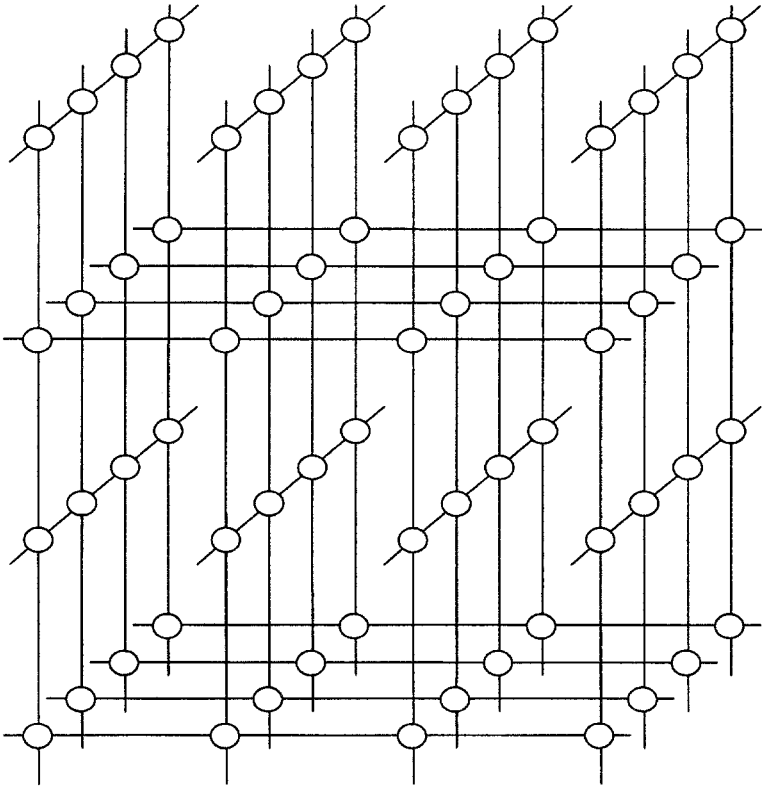


Figure 2. Three-dimensional pruned 4-torus (pruned, undirected). To avoid clutter, wraparound links are partially drawn.

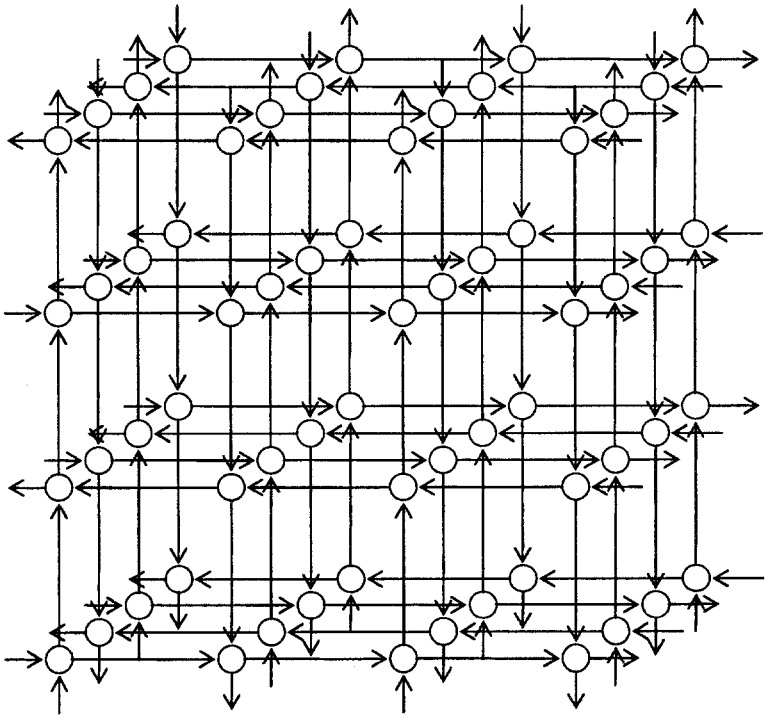


Figure 3. Three-dimensional Manhattan street network (unpruned, directed). To avoid clutter, wraparound links are partially drawn.

The resulting unpruned, directed toroidal network, also known as (multidimensional) Manhat tan street network, is depicted in Figure 3 for $n = 3$ and $k = 4$.

In a similar manner, by alternately removing the network links along dimension 0, each node $(a_0, a_1, \ldots, a_i, \ldots, a_{n-1})$ in the pruned, directed nD torus is adjacent to two out-neighbors

$$(a_0 + 1, a_1, \ldots, a_i, \ldots, a_{n-1}), \quad \text{if } \sum_{j=1}^{n-1} a_j = \text{even},$$

$$(a_0 - 1, a_1, \ldots, a_i, \ldots, a_{n-1}), \quad \text{if } \sum_{j=1}^{n-1} a_j = \text{odd},$$

$$(a_0, a_1, \ldots, a_i + 1, \ldots, a_{n-1}), \quad \text{if } \sum_{j=0\ (j\neq i)}^{n-1} a_j = \text{even and } \quad a_0 \equiv (i-1) \bmod (n-1),$$

$$(a_0, a_1, \ldots, a_i - 1, \ldots, a_{n-1}), \quad \text{if } \sum_{j=0\ (j\neq i)}^{n-1} a_j = \text{odd and } \quad a_0 \equiv (i-1) \bmod (n-1).$$

Such a pruned directed toroidal network can be visualized by applying the orientations of the links in Figure 3 to the links of Figure 2.

3. CAYLEY-GRAPH FORMULATIONS

Next, using Cayley digraphs of abelian groups, we show that the four networks under study (namely, unpruned/pruned, undirected/directed tori) are node- and edge-transitive. These results will allow us to devise efficient distributed routing algorithms and to obtain simple closed-form expressions for the average internode distance for each architecture. They also allow us to make assumptions about uniform traffic in each node and link, thereby facilitating the analysis of network performance.

Given an identity element ι from some finite group Γ, define a subset Ω, such that $\iota \notin \Omega$; i.e., there is no self-loop in the resulting digraph. The Cayley digraph is formed with the node set Γ and an edge from $a \in \Gamma$ to $b \in \Gamma$ whenever $b = ag$ for some $g \in \Omega$. The cardinality $|\Omega|$ of the generator set determines the out-degree, which is exactly the same as the in-degree.

In our case, we have $\Gamma = \{(a_0, a_1, \ldots, a_{n-1}) | 0 \leq a_i \leq k-1, \text{ for all } 0 \leq i \leq n-1\}$ and the identity element $\iota = (0, 0, \ldots, 0)$. If $a = (a_0, a_1, \ldots, a_{n-1})$ is adjacent to $b = (b_0, b_1, \ldots, b_{n-1})$, their index vectors are related by a semidirect product. In the following subsections, we specify the product form and the associated generator set Ω for each of the four architectures of interest.

3.A. Unpruned, Undirected Torus

The adjacency relationship of the unpruned, undirected torus corresponds to the expression $b = a + g$ and the generator set

$$\Omega = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} k-1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ k-1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \ldots, \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ k-1 \end{bmatrix} \right\}.$$

Note that because the addition is modulo k, adding $k-1$ is the same as subtracting 1.

3.B. Pruned, Undirected Torus

The adjacency relationship of the pruned, undirected torus is specified as $b = a + \Psi^{a_0} g$, where Ψ is an $n \times n$ matrix defined as

$$\Psi = \begin{bmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 1 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix}.$$

We call Ψ the "pruning matrix", because it specifies the connectivity reduction scheme employed. In particular, pruning along dimension 0 is represented by Ψ^{a_0}. Note that the i^{th} power Ψ^i of Ψ is obtained by cyclically left-shifting the rightmost $n-1$ elements of the bottom $n-1$ rows of the identity matrix by i positions. Hence, Ψ possesses the periodic property $\Psi^{i+(n-1)j} = \Psi^i$ with a period of $n-1$. The generator set becomes

$$\Omega = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} k-1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ k-1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \right\}.$$

EXAMPLE 1. The pruned undirected 3D torus has the 3×3 permutation matrix

$$\Psi^{a_0} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}^{a_0} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \langle a_0+1 \rangle_2 & \langle a_0 \rangle_2 \\ 0 & \langle a_0 \rangle_2 & \langle a_0+1 \rangle_2 \end{bmatrix},$$

where $\langle x \rangle_2$ denotes $x \bmod 2$. ∎

3.C. Unpruned, Directed Torus

The adjacency relationship of the unpruned, directed torus corresponds to the expression $b = a + \prod_{i=0}^{n-1} \Phi_i^{a_i} g$, where Φ_i is an $n \times n$ diagonal matrix with all entries being -1, except for the i^{th} entry which is 1. We call Φ_i the orientation matrix associated with dimension i; it leads to the assignment of opposite directions to all other links for any two nodes that are adjacent along dimension i. The generator set, having $|\Omega| = n$, is obtained by removing the inverses from the generator set associated with the unpruned undirected torus

$$\Omega = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \dots, \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \right\}.$$

EXAMPLE 2. The unpruned, directed 3D torus has the 3×3 permutation matrix

$$\prod_{i=0}^{2} \Phi_i^{a_i} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix}^{a_0} \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}^{a_1} \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}^{a_2}$$

$$= \begin{bmatrix} (-1)^{a_1+a_2} & 0 & 0 \\ 0 & (-1)^{a_0+a_2} & 0 \\ 0 & 0 & (-1)^{a_0+a_1} \end{bmatrix}.$$ ∎

3.D. Pruned, Directed Torus

In a similar manner, we can apply the preceding formulations to the pruned, directed torus, leading to the adjacency relationship described by the expression $b = a + \Psi^{a_0} \prod_{i=0}^{n-1} \Phi_i^{a_i} g$. The generator set, which is the intersection of the sets corresponding to the pruned, undirected, and unpruned, directed tori, is

$$\Omega = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \right\}.$$

EXAMPLE 3. The pruned, directed 3D torus has the 3×3 permutation matrix:

$$\Psi^{a_0} \prod_{i=0}^{2} \Phi_i^{a_i} = \begin{bmatrix} (-1)^{a_1+a_2} & 0 & 0 \\ 0 & (-1)^{a_0+a_1} \langle a_0+1 \rangle_2 & (-1)^{a_0+a_2} \langle a_0 \rangle_2 \\ 0 & (-1)^{a_0+a_1} \langle a_0 \rangle_2 & (-1)^{a_0+a_2} \langle a_0+1 \rangle_2 \end{bmatrix}. \quad \blacksquare$$

For the aforementioned toroidal networks to be node-transitive, the pruned versions must satisfy $k \bmod (n-1) \equiv 0$ and the directed versions must satisfy $k \bmod 2 \equiv 0$ (obviously, both conditions must hold for the pruned, directed version). These restrictions are not serious for the 3D case, where it is only required that k be even.

3.E. Symmetry and Other Properties

The preceding observations establish the node-transitivity of the toroidal networks. The edge-transitivity of these networks follows from the fact that the generators are exchangeable.

Note that the pruning scheme outlined in Section 3.B above is not the only viable one. Replacing the permutation matrix Ψ^{a_0} by the more general $\Psi^{f(a_0,a_1,a_2)}$ yields a variety of pruning schemes. As an example, $f(a_0, a_1, a_2) = a_0 + a_1 + a_2$ yields the diamond network, which can be viewed as the 3D version of honeycomb network [6], [17], [22]. However, it has been shown, somewhat surprisingly, that the simpler pruning scheme used here is also the best in terms of both regularity and performance [23].

4. STATIC NETWORK PROPERTIES

At a very coarse level, networks can be characterized by certain static properties or topological parameters [5]. The most important such parameters are the diameter D, average internode distance Δ, fault diameter D_F, and bisection bandwidth B. We review these in the following sections. We also briefly review the issues of scalability and packageability. Theorems are given without proofs.

4.A. Network Diameter

The diameter of a network, defined as the length of the longest among the shortest paths between all pair of nodes, is clearly important with packet routing because it dictates the worst-case communication latency. Whereas it has become fashionable to downplay the importance of diameter by stating that in wormhole routing (the dominant routing scheme in modern parallel computers), latency is insensitive to the diameter, there are counter arguments that show that diameter is in fact still important, even with wormhole routing, when performance penalties of blocking and deadlock are factored in. This is especially true with very long messages.

THEOREM 1. *The diameter of an nD pruned, undirected k-torus is [18]*

$$\begin{aligned} &(n-1)\lfloor k/2 \rfloor + \max(2n-4, \lfloor k/2 \rfloor), && \text{if } k \geq 2(n-1), \\ &(n-1)\lfloor k/2 \rfloor + \max(n-3+\lfloor k/2 \rfloor, k), && \text{if } k = n-1. \end{aligned}$$

Recall that k is assumed to be a multiple of $n-1$. $\blacksquare$

Diameters of various toroidal networks with $n = 3$ are shown in Table 1, assuming k is even and $k \geq 4$. The diameter of the pruned, directed k-torus is easily obtained for $n = 3$, though we do not have a counterpart to Theorem 1 for this case. From Table 1, it is quite evident that for systems of practical sizes, the diameter is in fact unimportant in distinguishing these networks.

Table 1. Static performance parameters of various 3D toroidal networks.

nD k-Torus Variant	In/Out Degree d	Diameter D	Average Internode Distance Δ	Fault Diameter D_F	Bisection Width B
Unpruned, Undirected	6	$1.5k$	$0.75k$	$D+1$	$2k^2$
Pruned, undirected	4	$1.5k$	$0.75k + 2/k - 2/k^2$	$\leq D + \lceil k/2 \rceil - n + 2$	k^2
Unpruned, directed	3	$1.5k + 1$	$0.75k + 1 - 4/k^3$	n/a	k^2
Pruned, directed	2	$1.5k + 3$	$\cong 0.75k + 3.5 - 4/k$	n/a	$0.5k^2$

4.B. Average Internode Distance

Whereas the network diameter is an indicator of worst-case network latency under light load, the average internode distance has a similar significance for the average communication latency with randomly destined messages. In fact, in symmetric networks, network diameter and average internode distance are closely related, so that either parameter can be used in comparative static evaluation of networks.

For our four node-symmetric networks, the average internode distance can be derived by computing the sum of distances from a given node, say node (0, 0, ... , 0) to all other nodes and dividing the result by $k^3 - 1$ or k^3. While the first option more accurately reflects the intuitive notion of internode distance, we take the second option (which also counts the distance of a node to itself), because it leads to simpler expressions. Results based on the first option can be easily obtained by multiplying our results by $k^3/(k^3 - 1)$.

The average internode distance of torus and Manhattan street networks have previously been derived. For the 3D pruned, undirected torus, we have the following result, which is easily derived as explained in the preceding paragraph.

THEOREM 2. *The average internode distance of a 3D pruned, undirected k-torus is* $0.75k + 2/k - 2/k^2$. ∎

The average internode distances of various 3D toroidal networks are shown in Table 1. In the case of 3D pruned, directed torus, no closed form formula has been found for the average internode distance; however, the expression $0.75k + 3.5 - 4/k$ is a good fit to numerically derived results for $k \geq 8$. Figure 4 plots the average internode distance as a function of the radix k, with k assuming all even values from 4 to 32. Note that, as was the case for network diameter, pruning has a lesser effect on the average internode distance than orientation. Figure 5 compares unpruned and pruned tori of four and five dimensions with respect to average internode distance. As we observed in the case of network diameter, Figures 4 and 5 indicate that for networks of practical sizes, the average internode distance variation among these networks is small enough not to be of major concern.

4.C. Fault Diameter

The fault diameter of a network, defined as the diameter of the surviving part after the occurrence of a small number of faults (fewer than the network's connectivity) is an indicator of network resilience [24]. Provided that routing in the incomplete or "injured" network is not significantly more difficult than in the intact network, a small fault diameter might allow the network to function close to full performance despite the occurrence of faults. This is the case, for example, when adaptive routing is used on the intact network for performance reasons.

The fault diameter of an nD torus (with $n - 1$ or fewer faults) is known to be no more than one hop greater than its fault-free diameter [25]. In a companion paper [18], we have proven the

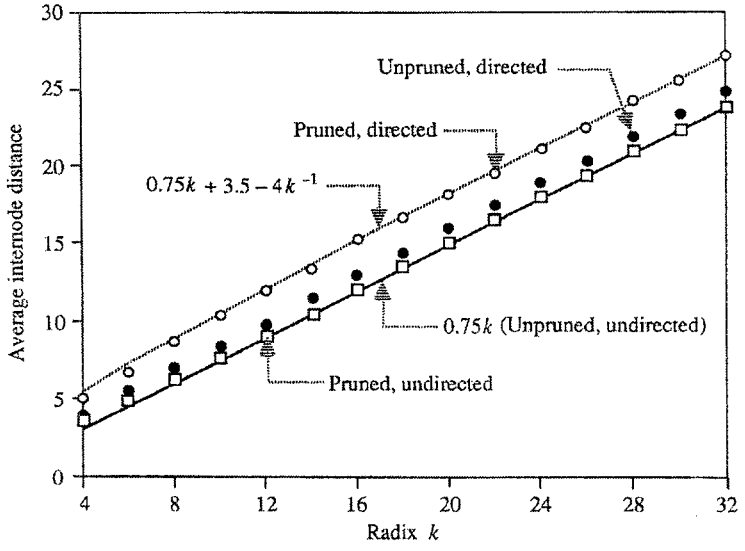


Figure 4. The average internode distance for the four types of 3D toroidal networks.

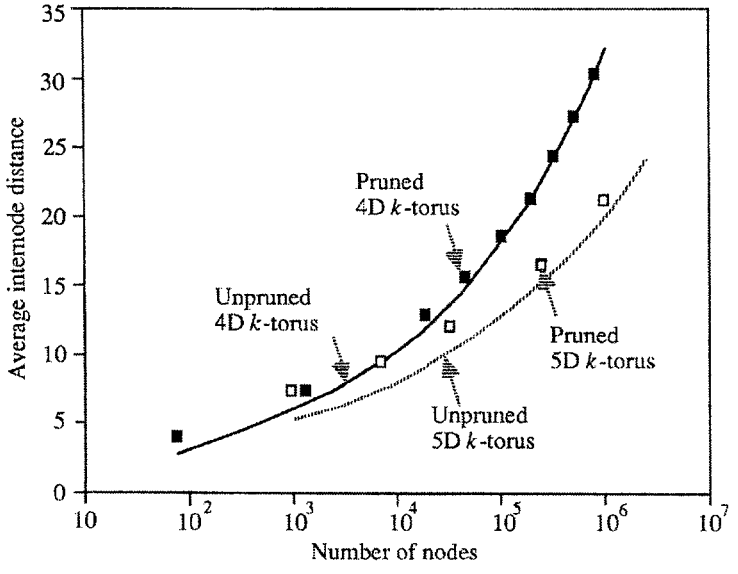


Figure 5. Average internode distance of some higher-dimensional toroidal networks.

following result about the fault diameter of pruned torus networks. This is done constructively by deriving four node disjoint paths between node 0 and an arbitrary node and noting the length of the longest of the four paths.

THEOREM 3. *For $k \geq 4(n-1)$, the fault diameter of an nD pruned, undirected k-torus is no greater than $(n-1)\lfloor k/2 \rfloor + k - n + 2$.* ∎

No corresponding result on the fault diameter of Manhattan street networks and, consequently, for pruned directed tori, is known to us. Note that due to node symmetry, which allows the largest possible number of node-disjoint paths between any pair of nodes [26], we expect the fault diameters of these networks to also be close to those of the undirected variants.

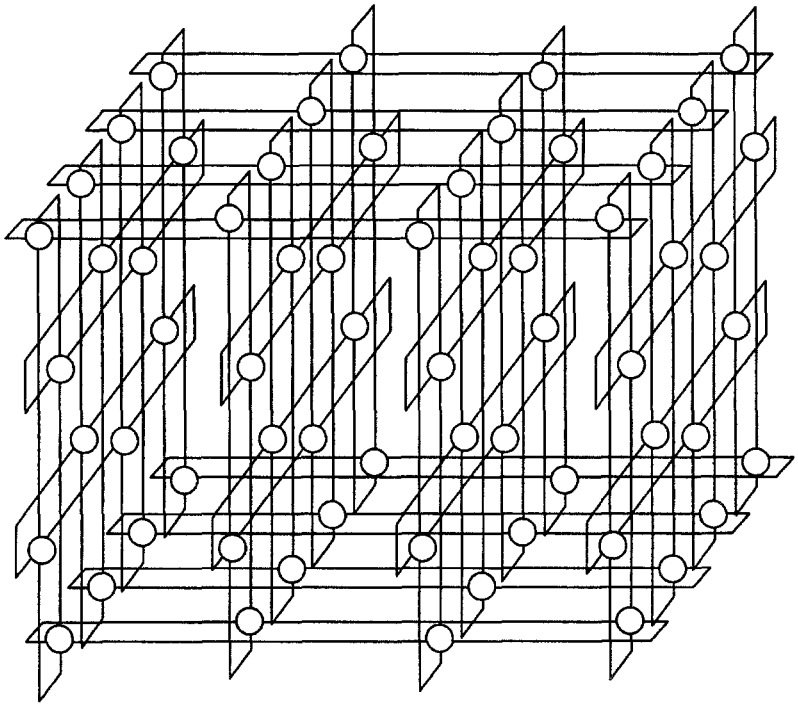
4.D. Bisection Width

The bisection width of a network, defined as the minimum number of links whose removal divides the network into two equal halves, is a good indicator of the network's ability to efficiently

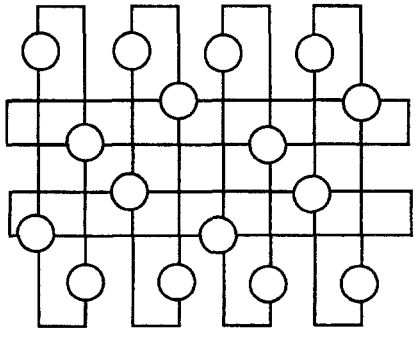
run communication intensive applications that may involve frequent invocation of many-to-many communication primitives. Bisection bandwidth is similarly defined, except that the sum of link capacities, rather than the number of links, is considered.

THEOREM 4. *The bisection width of an nD pruned, undirected k-torus is $2k^{n-1}/(n-1)$, where a bidirectional link is assumed to consist of two unidirectional links.* ∎

The bisection width of an nD pruned, directed k-torus is half of that given in Theorem 4. Bisection widths of the 3D toroidal networks are provided in Table 1. If all links in the various networks have unit capacity, then bisection bandwidths of these networks are similarly related. However, if we assume that the sparser architectures allow the use of wider links, with the same total network cost, then B must be appropriately weighted for each scheme for a fair comparison of the bisection bandwidths. A reasonable weight might be the inverse of node degree d, also shown in Table 1, because with the same pin count per node, wider links can be accommodated when the node degree is smaller. This is further discussed in Section 4.E. below.



(a) 3D view.



(b) Side view.

Figure 6. Folded layout of a pruned 3D 4-torus network.

4.E. Scalability and Packageability

We limit our discussion in this section to 3D structures. A 3D k-torus, folded in 3D space to allow implementation with only short wires occupies a $2 \times 2 \times 2$ subgrid per node, for a total space of $8k^3$. It is easily seen that pruning reduces the required volume (or layout area, in the case of 2D layout of the 3D structure). In fact, the improvement factor is greater than that suggested by the halving of the bisection width; another factor of 4/3 improvement results from the fact that folding needs to occur along only one of the two pruned dimensions (Figure 6). Thus, even after the increase in link width by a factor of 3/2 is factored in (due to node degree being reduced from 6 to 4), the layout space or area is still smaller for the pruned variant.

The preceding informal argument, combined with the observation that the pruned links along some of the dimensions make the network easier and less costly to partition (e.g., by slicing Figure 2 vertically rather than horizontally), should be enough to convince the reader that when cost is also taken into account, pruned and/or directed tori exhibit additional benefits. The scalability of pruned nD pruned torus network is a direct result of its combining the node complexity of 2D torus with the diameter and average internode distance of nD torus. We will report on cost-effectiveness and scalability issues separately in the near future.

5. DYNAMIC PERFORMANCE MODEL

With the same clock speed and I/O pin count for the four types of toroidal networks, reducing the connectivity allows an increase in the link width and, thus, improved link bandwidth. To be concrete, we exemplify our comparisons assuming the link width to be 16, 24, 32, and 48 bits, inversely proportional to the in- and out-degrees of 6, 4, 3, and 2 (a total of 96 I/O pins per node). All estimates are based on conservative packaging assumptions [27] and with reference to Cray 3TE [4]. The effect of various link widths can be specified by a factor $F = L/W$ denoting that each L-bit message must be broken into F flits for transmission over a W-bit link.

As the toroidal networks under study were shown to be both node- and edge-transitive, it is reasonable to assume that in a dynamic routing context, each node generates requests on a uniform basis and each link encounters the same traffic load. With no contention, the network performance can simply be described by the static parameters listed in Table 1.

For switches that use store-and-forward routing, the average message latency is $s\Delta F$, where s is the switch delay (or fall-through time) in clock cycles. For switches that use virtual cut-through routing [28], the average message latency becomes $F + s(\Delta - 1)$. Note that trading connectivity for wider links is always beneficial for store-and-forward routing. We will not pursue this further in this paper. For virtual cut-through routing, the trade-off might be worthwhile if the message is long enough to compensate for the increased average internode distance. The exact crossover point depends on the switch delay s.

As was noted in Section 4, the pruned and/or directed versions of the torus network have relatively minor differences among themselves, and with standard tori, in terms of maximum and average internode distances, especially as the network size grows. One can thus draw an early conclusion that pruning and/or directing a torus is worthwhile because of the smaller degree-diameter and degree-average distance products. We will show that this is indeed the case, even after conflicts and other routing complexities are factored in.

Let the message generation rate at a node be m packets/cycle. In any cycle, the probability that a packet injected from the local resource travels along a particular link is m/d, where d is the node degree. A packet on average takes Δ hops to arrive at its destination. When the network reaches steady state, the arrival rate or utilization ρ of an arbitrary link is given by

$$\rho = (m/d)F\Delta,$$

where $F = L/W$ is the message length as defined earlier.

In the absence of contention, the average latency experienced by a message is $F + \Delta - 1$ steps or, more generally, $F + s(\Delta - 1)$ cycles when each routing step within a switch is pipelined to take s cycles. For instance, the router design of Tera MTA uses $s = 3$, with two cycles spent in node logic and one on the wire leading to the next node. In this paper, we take $s = 3$ as a default value, but let s grow to as large as seven for architectures that require more complex routing decisions (i.e., up to three times as complex in terms of latency).

To model contention, the length of the queue associated with each outgoing link is treated as a discrete random variable $b \in [0, \infty)$; i.e., we assume no message loss due to buffer overflow. Note that a finite buffer can support near-optimal performance, given that the probability of b significantly exceeding its mean value β is negligibly small. Hence, the assumption of unlimited buffer space is commonly used to make the analysis of different types of networks and switching schemes tractable [7, 29–34].

The arrival rate ρ can be decomposed into two components, depending on how a packet proceeds in the network

$$\begin{aligned} m/d &= \rho/(F\Delta), && \text{the probability the packet enters/exits the network,} \\ \rho - m/d &= \rho(1 - 1/(F\Delta)), && \text{the probability that the packet stays in the network.} \end{aligned}$$

Consider an intermediate node along the route. The packet is never sent back to the node where it came from and the remaining $d - 1$ neighboring nodes are equiprobable to be used as the next hop. The probability $p(i)$ of i packets being sent from $d - 1$ input links to a particular output link has a binomial distribution of the general form $p(i) = \binom{k}{i}(1-\lambda)^{k-i}\lambda^i$, with $k = d-1$ in our case and $\lambda = \rho(1 - 1/(F\Delta))/(d-1)$. Thus

$$p(i) = \binom{d-1}{i}\left[1 - \frac{\rho}{d-1}\left(1 - \frac{1}{F\Delta}\right)\right]^{d-i-1}\left[\frac{\rho}{d-1}\left(1 - \frac{1}{F\Delta}\right)\right]^{i}.$$

The preceding probability is based only on the packets that stay in the network. If we include the packets that enter the network at the intermediate node, we have

$$q(i) = \left(1 - \frac{\rho}{F\Delta}\right)p(i) + \frac{\rho}{F\Delta}p(i-1),$$

where $q(i)$ is the probability of i packets requiring to use the link at the same time.

The probability $r(i)$ of i packets simultaneously contending for the same outgoing link follows a Markovian process. The state, indicating the current number of contenders, changes based on the number of incoming packets, and possibly one outgoing packet. Enumerating all possible combinations, we get the state distribution $r(i)$

$$r(i) = \underset{\substack{0 \text{ waiting,} \\ i \text{ arrive}}}{r(0)\,q(i)} + \sum_{j=1}^{i+1} \underset{\substack{j \text{ waiting, } 1 \text{ forwarded,} \\ i+1-j \text{ arrive}}}{r(j)\,q(i+1-j)}.$$

The preceding equation can be written as a recurrence for ease of evaluation

$$r(i+1) = \frac{1}{q(0)}\left[r(i) - r(0)\,q(i) - \sum_{j=1}^{i} r(j)\,q(i+1-j)\right],$$

with $r(0) = 1 - \rho$. Figure 7 shows typical probability mass functions of $r(i)$ for low, medium, and heavy utilizations ($\rho = 0.1, 0.5, 0.9$).

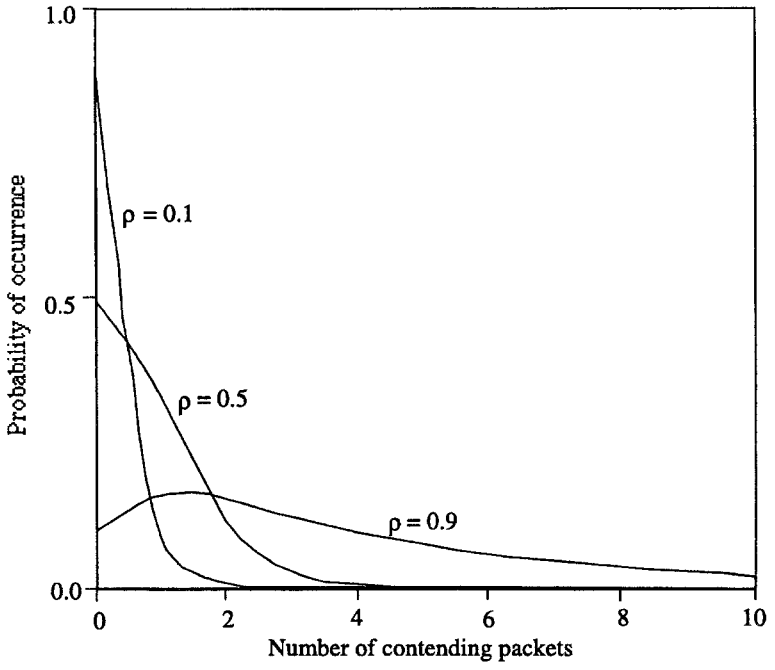


Figure 7. Probability $r(i)$ of i packets simultaneously contending for the same output link.

When i packets contend for the same outgoing link, all but one must be queued, leading to the requirement for a buffer of size $(i-1)F$. The average queue length can be shown to be

$$\beta = \sum_{i=2}^{\infty} (i-1) F r(i) = \frac{\rho^2}{2(1-\rho)} \frac{(d-2) F\Delta + 2 - d/(F\Delta)}{(d-1)\Delta}.$$

Using Little's identity, the average queue length β at steady state is equal to the product of the mean time β/ρ waiting for a link and its utilization ρ. As a packet on average goes through Δ links, the delay T_c attributable to contention is $\beta\Delta/\rho$.

$$T_c = \frac{\rho}{2(1-\rho)} \frac{(d-2) F\Delta + 2 - d/(F\Delta)}{d-1}.$$

The latency-throughput relationship of the directed network can be similarly developed. The only difference is that all d, rather than $d-1$, output links are equiprobable for forwarding an incoming packet at the intermediate node. The delay due to contention in this case is:

$$T_c = \frac{\rho}{2(1-\rho)} \frac{(d-1) F\Delta + 2 - (d+1)/(F\Delta)}{d}.$$

We are now in a position to find the average message latency in the presence of contention. Let p_w denote the probability of waiting at some buffer [28]. Then, cut-through occurs with probability

$$1 - p_w = r(0) + \sum_{i=1}^{\infty} \frac{r(i)}{i}.$$

Thus, including the effect of contention, the average message latency is:

$$T = [F + s(\Delta - 1)](1 - p_w) + T_c.$$

This completes the construction of our analytic performance model.

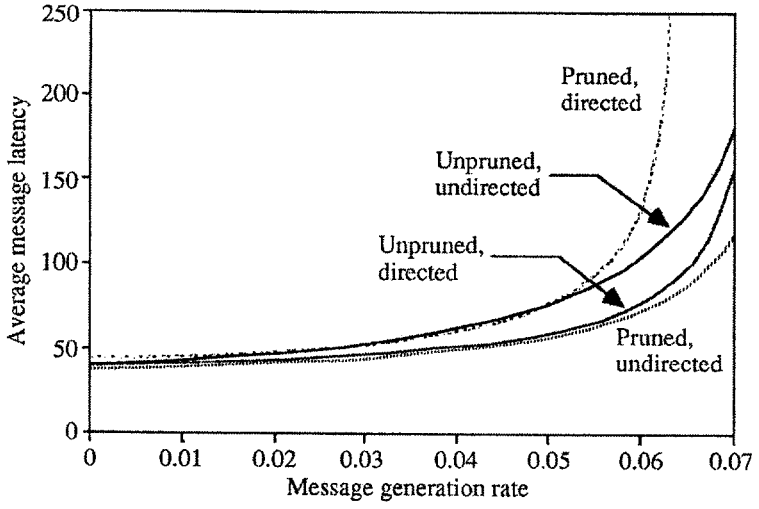


Figure 8. Average message latency in toroidal networks with side length (radix) $k = 16$, message length $L = 96$, and switching delay $s = 3$.

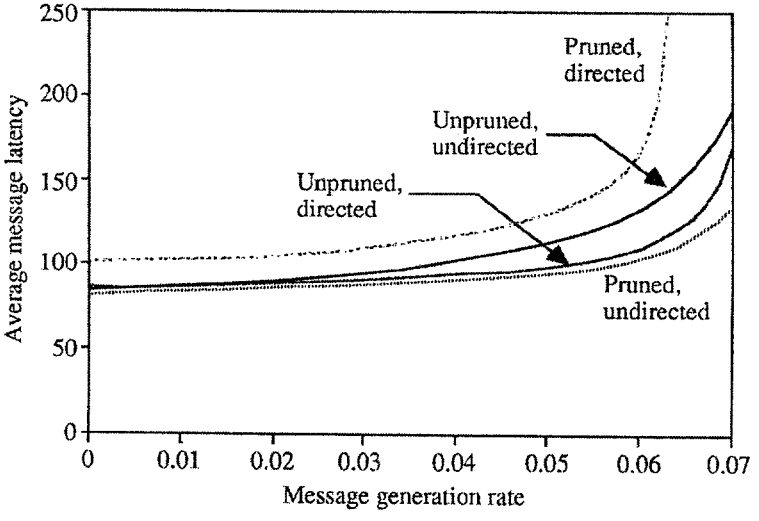


Figure 9. Average message latency in toridal networks with side length (radix) $k = 16$, message length $L = 96$, and switching delay $s = 7$.

6. PERFORMANCE COMPARISONS

We begin by assuming relatively short messages of length $L = 96$ bits and compare the four types of 3D toroidal networks with radix $k = 16$ (4096 nodes). The switch delay of $s = 3$ cycles represents state of the art router designs. Figure 8 shows the results. The immediate conclusion here is that as long as the network does not operate close to saturation, either pruning or orientation can improve the average message latency. The combination of pruning and orientation, however, is not an attractive option. At heavy loads, pruning is more effective than orientation. Note that the routing algorithm for pruned torus is simpler than that of MSN; thus, the foregoing comparison, which assumes $s = 3$ in either case, is somewhat unfair. More on this later. Increasing the switch delay to $s = 7$ (Figure 9) leads to similar results, but tends to magnify small differences in the average internode distance. Overall, pruning appears to be the best option under the assumption of short messages.

Figures 10 and 11 present the corresponding results for somewhat longer messages of $L = 384$ bits. Again, we see that pruning and orientation provide comparable performance benefits; what

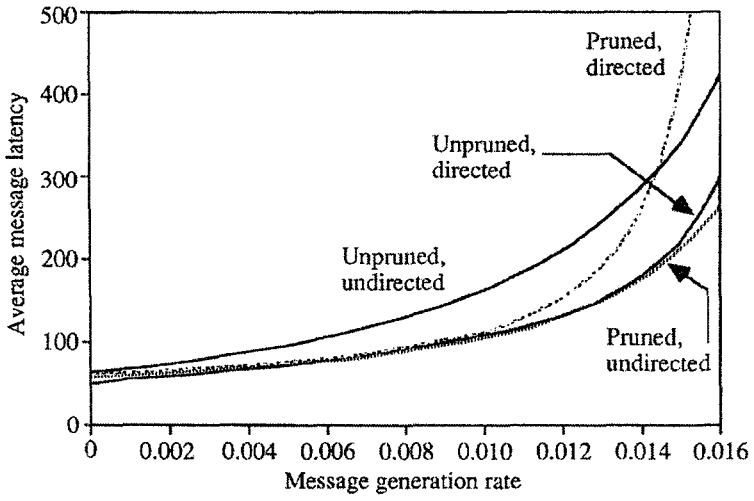


Figure 10. Average message latency in toroidal networks with side length (radix) $k = 16$, message length $L = 384$, and switching delay $s = 3$.

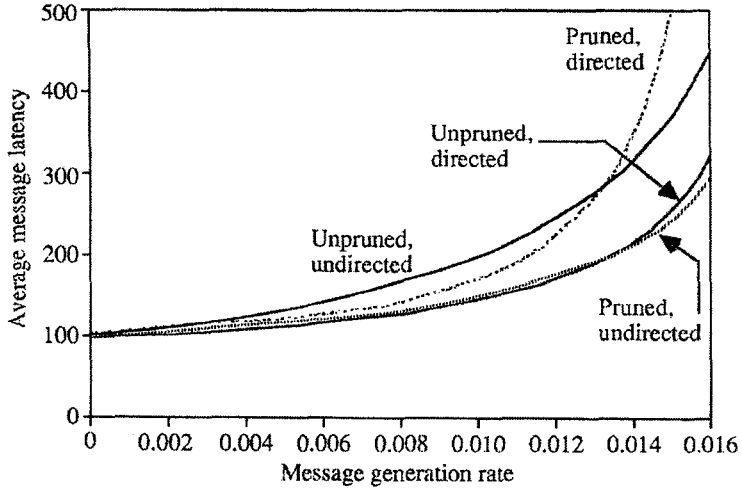


Figure 11. Average message latency in toroidal networks with side length (radix) $k = 16$, message length $L = 384$, and switching delay $s = 3$.

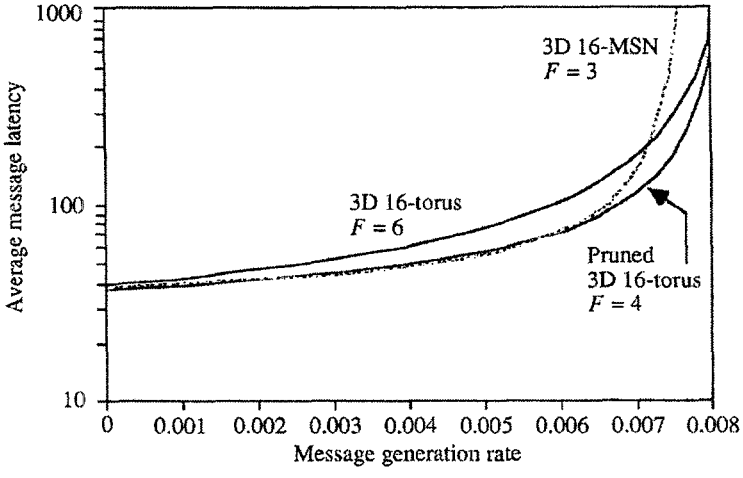


Figure 12. Virtual cut-through switching performance of 3D toroidal networks with 4096 nodes and constant pin count.

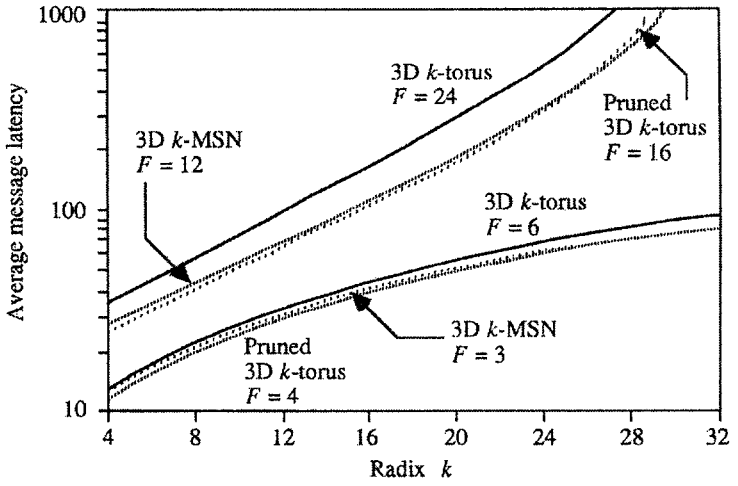


Figure 13. Virtual cut-through switching performance of 3D toroidal networks with message generation rate $m = 0.01$.

has changed most dramatically relative to Figures 8 and 9 is that the combination of pruning and orientation is no longer as bad as in the case of shorter messages, particularly when s is small.

Having established that pruning and orientation have comparable performance benefits, let us examine the differences between the two schemes in more detail. The average latency T is shown as a function of the message generation rate m for $k = 16$ ($N = 4096$) in Figure 12. With equal message length L and number of pins, the message length F will be proportional to the node degree d (three for MSN, four for pruned torus, and six for ordinary torus). Because current VLSI packages are limited to several hundred pins [1], our assumption of $F = 3$ for 3D k-MSN implies that each packet contains no more than a single word of data. We note that pruning improves the latency uniformly, except when operating close to saturation. The saturation point corresponds to full utilization or $\rho = 1$; thus the message generation rate m must be strictly less than $d/(F\Delta)$. Given that, as argued above, d/F is a constant for our toroidal networks, it is easy to see that the average internode distance Δ, which grows with network size, limits the scalability. The preceding problem is exacerbated for long messages.

Figure 13 shows the message latency as a function of the radix k for $m = 0.01$. The two sets of curves correspond to different message lengths, again assuming a constant pin count. The aforementioned concern with scalability not withstanding, we note that the advantages of pruning over orientation are even more pronounced for longer messages. Our tentative conclusion is that the pruned torus outperforms MSN for large networks and heavy loads, whereas MSN does better for smaller networks and lighter loads. Given that the difference in latencies is relatively smaller in the latter case, pruned tori can be considered superior overall.

To study the effects of varying both the network size $N = k^3$ and the message length F on the communication performance of pruned 3D k-torus, we plot the average latency relative to its unpruned counterpart in Figure 14. The message generation rate is again fixed at $m = 0.01$ which is below the saturation level for both networks. At some point, the average internode distance Δ and the message length F cause the average latency to grow quadratically rather than linearly. Figure 14 clearly shows that the improvement due to pruning is more significant for larger network sizes and longer messages.

To explain the foregoing, let us simplify the equation for the average latency T by setting $p_w = \rho$ and $F\Delta \gg 1$.

$$T = [F + s(\Delta - 1)](1 - \rho) + \frac{\rho}{1 - \rho}\frac{d - 2}{2d - 2}F\Delta.$$

Expanding this approximate expression using Taylor series and substituting ρ with $(m/d)F\Delta$,

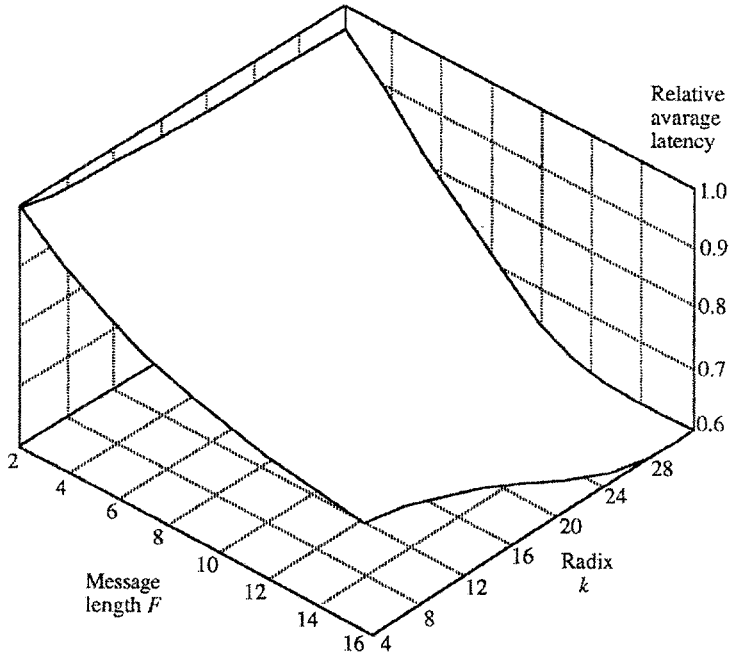


Figure 14. Virtual cut-through switching performance of the pruned 3D k-torus with $m = 0.01$ relative to an unpruned torus.

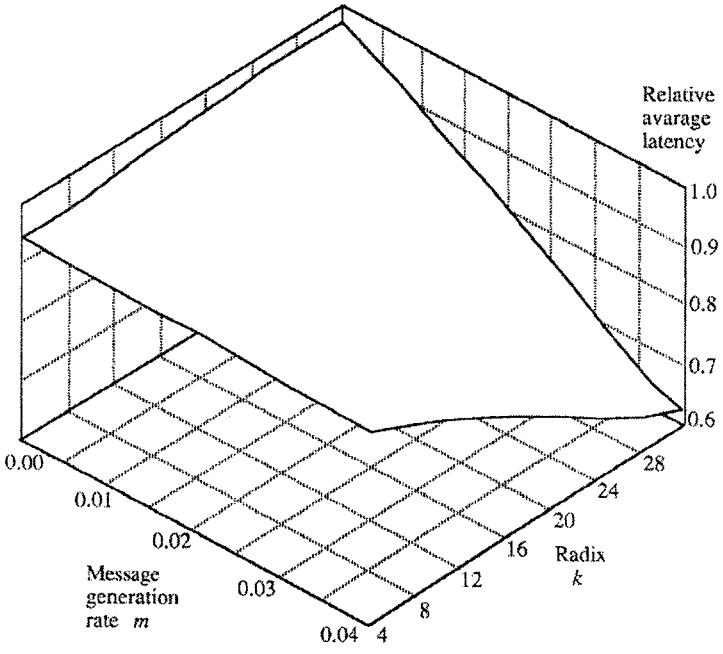


Figure 15. Virtual cut-through switching performance of the pruned 3D k-torus with $F = 4$ relative to an unpruned torus.

we have

$$T = [F + s(\Delta - 1)](1 - \rho) + \rho \frac{d-2}{2d-2} F\Delta (1 + \rho + \ldots)$$
$$= F + s(\Delta - 1) + \frac{m}{d} F\Delta \left(\frac{d-2}{2d-2} F\Delta - s\Delta - F - s + \ldots \right).$$

Figure 15 shows how well the pruned 3D k-torus handles various traffic levels relative to its unpruned counterpart. As expected, pruning offers little or no advantage at low traffic, given the

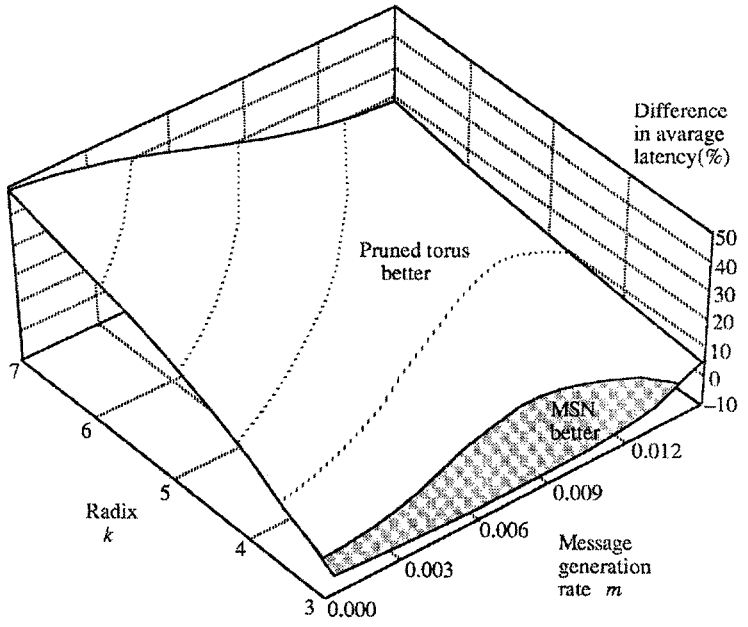


Figure 16. Performance of pruned, undirected torus relative to unpruned, directed torus(MSN), with MSN's more complex router accounted for by increasing its switch delay from three to seven cycles.

assumption of short messages ($F = 4$). The advantage of pruning starts to show for moderate traffic levels and larger network sizes. The 3D k-MSN behaves similarly, except that it reaches saturation more quickly (see Figures 8–11).

Finally, in Figure 16, we show the relative performances of pruned tori and MSNs when a slower switch is assumed to be required for the more complex routing task in MSN.

7. CONCLUSION

We have compared four classes of toroidal networks, corresponding to unpruned/pruned and undirected/directed variations of a torus network. There is more to be done to arrive at definite conclusions, but based on this work, pruned torus networks seem to hold promise for use in designing cost-effective high-performance parallel architectures. Even though the pruned torus network architecture has already been used in the design of Tera MTA [19], our characterization of such networks (using the Cayley graph model) and analytical performance comparisons provide a starting point for more detailed evaluations of such networks for highly parallel processing. Such evaluations must entail both general and application-specific message traffic patterns and be performed with more realistic network models. We strongly expect that comparison results will be fundamentally the same, if not even more in favor of pruned networks. This expectation arises from the fact that random traffic is, in a sense, the worst possible communication pattern for pruned networks with their smaller bisection bandwidths.

Similar advantages can be demonstrated for pruning higher-dimensional tori. However, the comparison will be more involved for $n > 3$. If the layout is carefully planned for expandability, the wire delay of the 3D torus and its pruned derivatives can become virtually independent of network size, while the same cannot be said about higher-dimensional tori (see [35] and the references therein). The dependence of wire delay on the physical dimensions makes it difficult to compare these architectures fairly and realistically. Several attempts based on different assumptions, such as constant bisection width and wire delay, have pointed to widely different conclusions [7,9,30,36].

Our pruned torus network represents but one way to reduce the connectivity of a torus while preserving many of its desirable properties. For example, a 4D torus might be pruned by keeping

links along two dimensions intact, and alternating between links in the other two. This reduces the node degree from eight to six, rather than to four, but offers correspondingly greater performance. Pruning is also applicable to other network topologies; for example, we have previously applied it to chordal rings [37]. In other recent work, we have explored pruned networks as Cayley graphs as well as networks obtained from pruning Cayley graphs [38,39]. We expect the benefits of pruning, demonstrated in this paper for torus networks, to extend to many other networks classes. We will report additional results on such pruned networks in the near future. Combined with hierarchical interconnection networks [40], which also offer reduced connectivities, the design space for cost-effective and scalable networks is quite vast.

REFERENCES

1. Semiconductor Industry Association, *International Technology Roadmap for Semiconductors*, San Jose, CA, USA, `http://public.itrs.net`, (2000).
2. D.J. Kuck, *High-Performance Computing: Challenges for Future Systems*, Oxford University Press, (1996).
3. T.G. Mattson, D. Scott, and S. Wheat, A TeraFLOP supercomputer in 1996: The ASCI TFLOP system, In *Proc. Int'l Parallel Processing Symp.*, April 1996, pp. 84–93.
4. S.L. Scot and G.M. Thorson, The Cray T3E network: Adaptive routing in a high performance 3D torus, In *Proc. Hot Interconnects IV*, Palo Alto, CA, August 1996.
5. B. Parhami, *Introduction to Parallel Processing: Algorithms and Architectures*, Plenum Press, (1999).
6. B. Parhami and D.-M. Kwai, A unified formulation of honeycomb and diamond networks, *IEEE Trans. Parallel and Distributed Systems* **12** (1), 74–80, (2001).
7. A. Agarwal, Limits on interconnection network performance, *IEEE Trans. Parallel and Distributed Systems* **2** (4), 398–412, (1991).
8. J.R. Anderson and S. Abraham, Multidimensional network performance with unidirectional links, In *Proc. Int'l Conf. Parallel Processing*, 1997 pp. 26–33.
9. W.J. Dally, Performance analysis of k-ary n-cube interconnection networks, *IEEE Trans. Computers* **39** (6), 775–785, (1990).
10. D. Banerjee, B. Mukherjee and S. Ramamurthy, The multidimensional torus: Analysis of the average hop distance and application as a multihop lightwave network, In *Proc. IEEE Int'l Conf. Communications*, New Orleans, May 1994, pp. 1675–1680.
11. N.F. Maxemchuk, Routing in the Manhattan street network, *IEEE Trans. Communications* **35**, 503–512, (1987).
12. J. Brassil, A.K. Choudhury and N.F. Maxemchuk, The Manhattan street network: A high performance, highly reliable metropolitan area network, *Computer Networks and ISDN* **26**, 841–858, (1994).
13. T.Y. Chung, N. Sharma and D.P. Agrawal, Cost-performance trade-offs in Manhattan street networks versus 2-D torus, *IEEE Trans. Computers* **43** (2), 240–243, (1994).
14. W.-T. Lee and L.-Y. Kung, Binary addressing and routing schemes in the Manhattan street network, *IEEE/ACM Trans. Networking* **3**, 26–30, (1995).
15. N. Mirfakhraei, Simulation of a Manhattan street network for high-speed ATM applications, In *Proc. IEEE Int'l Conf. Communications*, Seattle,WA, June 1995, pp. 1937–1942.
16. D.-M. Kwai and B. Parhami, A class of fixed-degree Cayley graph interconnection networks derived by pruning k-ary n-cubes, In *Proc. Int'l Conf. Parallel Processing*, August 1997, pp. 92–95.
17. J. Nguyen, J. Pezaris, G. Pratt and S. Ward, Three-dimensional network topologies, In *Proc. Int'l Workshop Parallel Computer Routing and Communication*, Seattle, WA, May 1994, pp. 101–115.
18. B. Parhami and D.-M. Kwai, Incomplete k-ary n-cube and its derivatives, *J. Parallel and Distributed Computing* **64** (2), 183–190, (2004).
19. R. Alverson, D. Callahan, D. Cummings, B. Koblenz, A. Porterfield and B. Smith, The Tera computer system, In *Proc. ACM Int'l Conf. Supercomputing*, Amsterdam, June 1990, pp. 1–6.
20. L. Wittie, Communication structures for large networks of microcomputers, *IEEE Trans. Computers* **30** (4), 264–273, (1981).
21. T.Y. Chung and D.P. Agrawal, Design and analysis of multidimensional Manhattan street networks, *IEEE Trans. Communications* **41**, 295–298, (1993).
22. I. Stojmenovic, Honeycomb networks: Topological properties and communication algorithms, *IEEE Trans. Parallel and Distributed Systems* **8** (10), 1036–1042, (1997).
23. D.-M. Kwai, D.-M. and B. Parhami, Pruned three-dimensional toroidal networks, *Information Processing Letters* **68**, 179–183, (1998).
24. M.S. Krishnamoorthy and B. Krishnamurthy, Fault diameter of interconnection networks, *Computers Math. Applic.* **13** (5/6), 577–582, (1987).
25. K. Day and A.E. Al-Ayyoub, Fault diameter of k-ary n-cube networks, *IEEE Trans. Parallel and Distributed Systems* **8** (9), 903–907, (1997).
26. S. Lakshmivarahan, J.-S. Jwo and S.K. Dahl, Symmetry in interconnection networks based on Cayley graphs of permutation group: A survey, *Parallel Computing* **19**, 361–401, (1993).

27. T.C. Chung,*et al.*, Area array packaging technologies for high-performance workstations and multiprocessors, In *Proc. IEEE Electronic Components and Technology Conf.*, Orlando, FL, May 1996, pp. 902–910.
28. P. Kermani and L. Kleinrock, Virtual cut-through: A new computer communication switching technique, *Computer Networks* **3**, 267–286, (1979).
29. S. Abraham and K. Padmanabhan, Performance of the direct binary n-cube network for multiprocessors, *IEEE Trans. Computers* **38** (7), 1000–1011, (1989).
30. S. Abraham and K. Padmanabhan, Performance of multicomputer networks under pin-out constraints, *J. Parallel and Distributed Computing* **12**, 237–248, (1991).
31. J.W. Dolter, P. Ramanathan and K.G. Shin, Performance analysis of virtual cut-through switching in HARTS: A hexagonal mesh multicomputer, *IEEE Trans. Computers* **40** (6), 669–680, (1991).
32. M.D. Grammatikakis, J.-S. Jwo, M. Kraetzl and S.-H. Wang, Dynamic and static packet routing on symmetric communication networks, In *Proc. IEEE GLOBECOM*, San Francisco,CA, November 1994, pp. 1571–1575.
33. W. Hsu and P.-C. Yew, The impact of wiring constraints on hierarchical network performance, In *Proc. Int'l Parallel Processing Symp.*, March 1992, pp. 580–588.
34. V. Sharma and E.A. Varvarigos, Circuit switching with input queuing: An analysis for the d-dimensional wraparound mesh and hypercube, *IEEE Trans. Parallel and Distributed Systems* **8** (4), 349–366, (1997).
35. G. Bilardi and F.P. Preparata, Horizons of parallel computation, *J. Parallel and Distributed Computing* **27**, 172–182, (1995).
36. R.E. Kessler and J.L. Schwarzmeier, Cray T3D: A new dimension for Cray research, In *Digest of Papers IEEE COMPCON*, San Francisco,CA, February 1993, pp. 176–182.
37. B. Parhami and D.-M. Kwai, Periodically regular chordal rings, *IEEE Trans. Parallel and Distributed Systems* **10** (6), 658–672, (1999).
38. W. Xiao and B. Parhami, Some conclusions on Cayley digraphs and their applications to interconnection networks, In *Lecture Notes in Computer Science Vol. 3033*, Proc. 2^{nd} *International Workshop on Grid and Cooperative Computing* (Edited by M. Li *et al.*) December 2003, pp. 408–412, Springer-Verlag, (2004).
39. W. Xiao and B. Parhami, Hexagonal and pruned torus networks as Cayley graphs, In *Proc. International Conf. Communications in Computing*, Las Vegas, NV, June 2004, pp. 107–112.
40. C.-H. Yeh and B. Parhami, The index-permutation graph model for hierarchical interconnection networks, In *Proc. of the International Conf. on Parallel Processing*, Aizu, Japan, September 1999, pp. 48–55.