The Chordal Spoke ATM Interconnection Network †

Mark J. Clement, Bryan S. Morse, J. Kelly Flanagan and Wei Wei
Department of Computer Science
Brigham Young University 3372 TMCB
Provo, Utah 84602-6576
{clement, morse, kelly, weiwei}@cs.byu.edu

Phyllis E. Crandall
Computer Science & Engineering
University of Connecticut
191 Auditorium Rd., U-155
Storrs, CT 06269-3155
(860) 486-2672
crandall@cse.uconn.edu

Contact Author: Mark Clement clement@cs.byu.edu

January 15, 1997

Abstract

Networks of workstations have become a popular platform for parallel computing. The use of Asynchronous Transfer Mode (ATM) interconnection networks can improve the communication characteristics of these parallel clusters. This paper presents Chordal Spoke (CS) networks which are a class of topologies built from small ATM switches. These networks are cost effective and scalable to include a large number of processors. The CS network topology is presented along with an addressing scheme which provides efficient communication between processors. The low diameter and scalable bisection width of CS networks give them a performance advantage over traditional workstation cluster configurations. We show that CS networks compare favorably with other multicomputer topologies in terms of several important metrics.

Index Terms: Multi-hop ATM networks, interconnection networks, switching networks, embedding, routing, fault tolerance, multicomputers.

† This research was supported by Sprint Corporation.

1 Introduction

Many interconnection topologies have been proposed to provide communications on distributed memory parallel processing platforms [3, 8, 12, 20, 21, 25]. Although several of these networks have exhibited acceptable performance, their excessive cost has made them impractical for many parallel applications. As a result, many applications are run on clusters of workstations that communicate through a shared media Local Area Network (LAN) such as Ethernet. Because of the fixed bandwidth available in these configurations, shared media communication networks are not practical for connecting large numbers of workstations.

Recently, high-performance communication networks have been developed that overcome many of the problems inherent in Ethernet LANs. ATM networks have been proposed as a flexible interconnection media for cluster computing platforms [4, 17]. ATM switching technology is currently being used for telephony, and the hardware promises to become less expensive as telephony deployment continues [5, 7, 9, 14, 22]. Using ATM technology, Gigabit optical communication links are available for each node in a multicomputer, overcoming the bandwidth limitations of Ethernet-style LAN connections [16]. The cost of ATM switches, however, increases exponentially as more inputs and outputs are added to the switch. For this reason, Small ATM Switch-based Interconnection (SASI) networks [23] may be more feasible for cost-sensitive cluster computing solutions. Networks that use a large number of small ATM switches (with four to sixteen ports on each switch) can be implemented so that the cost of a high-performance network increases linearly with the number of processors. This research will investigate network technologies that enable cluster solutions to provide acceptable performance at a fraction of the cost of traditional parallel platforms.

This paper presents the Chordal Spoke (CS) optical interconnection network as a viable option for workstation cluster communications. Previous research into SASI networks has shown that Generalized Fat Trees (GFT) [15] Meshes, Pyramids, and many other topologies [18, 19] can be efficiently constructed with small ATM switches. The Round Table (RT) SASI network [23] maximizes the usability of switch ports, but does not provide scalable bisection width. The chordal spoke network was developed to overcome this bisection width limitation and to provide increased fault tolerance for workstation clusters, thus providing a scalable and high-performance technology for implementing clusters of workstations.

2 ATM Properties

The ATM protocol specifies a fix-sized, cell-based, connection-oriented protocol. This protocol simplifies the hardware implementation of ATM switches and allows them to run at higher speeds. Standard ATM data rates range from 51.84 Mbps to 2,488.32 Mbps [16]. Although the use of ATM by the telephony industry decreases the cost of network connections, the latency induced by adherence to standards will generally be greater than that found in custom MPP networks. Nevertheless, the increased bandwidth and projected decreasing cost of ATM networks makes them attractive for many parallel applications.

An ATM switch can be viewed as a connection unit with w pairs of input and output ports. Each pair of input port and output ports has a unique address labeled from 0 to $w-1$. A switch with width w can also be thought of as a w-node completely connected graph. If processing elements are connected to each port of the switch, there is a direct, contention-free connection between each pair of nodes [4, 5]. However, if several source nodes send a message to a common destination node at the same time congestion can occur, which can cause increased delays or lost data in an ATM switch.

3 Chordal Spoke Networks

In order to connect large numbers of workstations through smaller ATM switches, some of the ports will be connected to processing elements and some ports will be connected to other switches. For simplification, we assume that the input and output ports are connected to the same endpoint. Since the ATM protocol is connection-oriented, a connection must be established between two ports on a switch before data can be transmitted. A connection of two switches through their ports is a bidirectional edge. The chordal spoke network is a collection of ATM switches with internal connections between switches and leaf ports connected to workstations or leaf nodes. A path is a sequence of edges from one leaf node to another leaf node. The distance of a path is the number of switches that must be traversed in the path.

A chordal spoke network $\mathrm{CS}(w, r, i)$ is a hierarchical network with switch width w, radius r, and inside width i. It can be characterized in the following way:

- A $\mathrm{CS}(w, 1, i)$ is a network that consists of $w + 1$ ATM switches with width w: one switch (atomic center) connects to w perimeter switches in a star topology [1]. The perimeter switches (leaf switches) have chordal ring connections [2]. The radius is defined as the number of edges from the atomic center to the leaf switches. The inside width is defined as the number of ports on each leaf switch that are connected to other leaf switches and the center. Inside width values are restricted to odd values since one port will be connected to the atomic center and pairs of ports are used to create the chordal connections ($i = 3, 5, 7, ..., w - 1$). Larger values for the inside width are used to create networks with higher bisection width. We define the outside width as the number of ports for each leaf switch that can be connected to workstations or to outer layers of the hierarchical switch (i.e., the difference between switch width and inside width). The addresses of leaf ports in a $\mathrm{CS}(w, 1, i)$ network are the leaf switch's outside port number appended to the corresponding center's port number.

- A $\mathrm{CS}(w, r, i)$ with $r > 1$ is a network that uses a $\mathrm{CS}(w, r - 1, i)$ as its center, with a perimeter switch connected at each leaf port of the $\mathrm{CS}(w, r - 1, i)$. These perimeter switches are connected to each other as a chordal ring (Fig. 1).

  When the chordal ring connections of a $\mathrm{CS}(w, r, i)$ are removed, the network becomes an acyclic graph. We call this the natural spanning tree of the chordal spoke network (Fig. 2).

For expository purposes we assume a network where all switches have the same number of ports.

3.1 Topological Properties

The chordal spoke network $\mathrm{CS}(w, r, i)$ is symmetrical about the atomic center. This symmetry simplifies routing and node addressing and makes the CS networks easy to scale up to large numbers

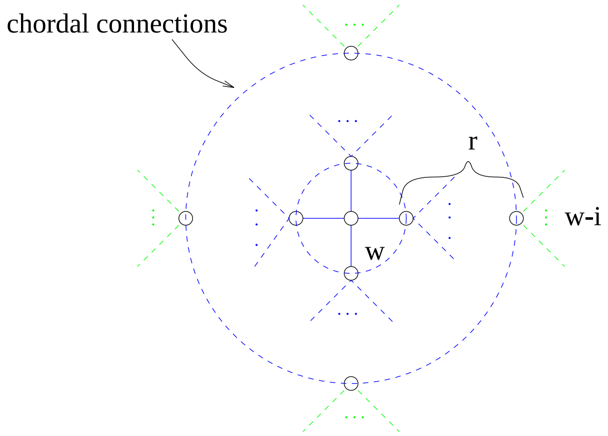


Figure 1: An example of CS(w, r, i).

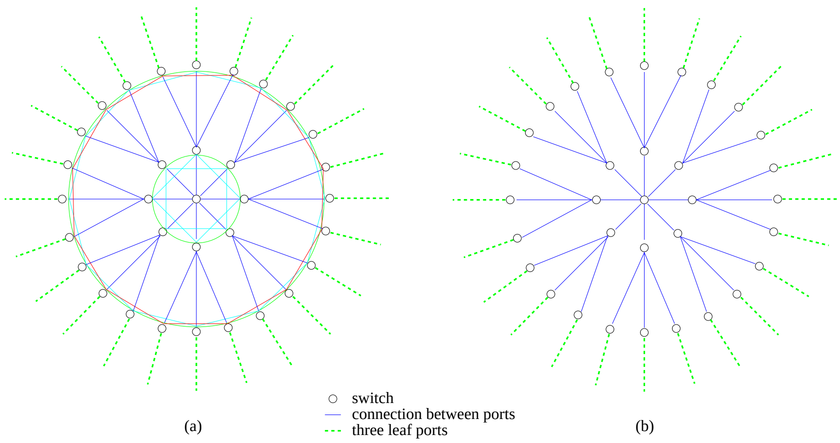


Figure 2: A CS$(8, 2, 5)$ (a) and its spanning tree (b)

of processors. The distance from each leaf node to the atomic center is the same, resulting in predictable communication latency between nodes. The diameter (the longest distance between two leaf nodes) of CS(w, r, i) is $2r + 1$. The distance from a leaf node to the atomic center is r and the distance from the atomic center to another leaf node is r, so there are $2r + 1$ switches in the path.

3.1.1 Number of Leaf Ports

The atomic center for CS(w, r, i) has w leaf ports connected to w perimeter switches. These w perimeter switches have both inside width ports connected to each other and outside width leaf ports connected to next-level perimeter switches. Thus, if a level j has internal width i_j, each of its switches has $w - i_j$ leaf ports connected to the next level. At the outermost layer, the number of leaf ports (nlp) for CS($w, r, \{i_j\}$) is

$$\begin{aligned} \mathrm{nlp}(w, r, i) &= w(w - i_1)(w - i_2)\ldots(w - i_r) \\ &= w \prod_{j=1}^{r} (w - i_j) \end{aligned} \tag{1}$$

where $i_j \in \{1, 3, 5, \ldots, w - 1\}$ is the internal width of each level j.

From Eq. 1 we see that the number of leaf ports is dependent on the switch width w and inside width i_j at each level. The wider the switch width w is, the greater the number of leaf ports is. The smaller the inside width i at each level, the greater the number of leaf ports.

3.1.2 Number of Switches

A CS(w,1,i) consists of a single ATM switch connected to its w perimeter switches. As the network grows, a next level perimeter switch will be added to each leaf port. The generalized number of switches (ns) in a CS($w, r, \{i_j\}$) is thus the sum of the leaf ports across all levels plus one for the atomic center:

$$\begin{aligned} \mathrm{ns}(w, r, i) &= 1 + w + \sum_{k=1}^{r-1} \mathrm{nlp}(w, k, \{i_j\}) \\ &= 1 + w + \sum_{k=1}^{r-1} w \prod_{j=1}^{k} (w - i_j) \end{aligned} \tag{2}$$

where $i_j \in \{1, 3, 5, \ldots, w - 1\}$ is again the internal width of each level j.

3.1.3 Bisection Width

For a given network, the bisection width (bw) is the minimum number of connections that must be removed in order to divide the network into two equal halves. Since the atomic center of a CS network must be cut in order to divide the network into two equal halves, the bisection width of the CS(w, r, i) is equal to the bisection width of the atomic center added to the number of chordal ring connections that must be removed to bisect the network.

The center ATM switch can logically be viewed as a w-node fully-connected graph, the number of edges for which is $w(w - 1)/2$. The bisection of a w-node fully connected graph will result in two ($w/2$)-node fully-connected subgraphs. The number of edges that must be removed is equal to the difference between the number of edges in a w-node fully connected graph and the number of edges in the two $(w/2)$-node fully-connected graphs:

$$\begin{aligned} \frac{w(w-1)}{2} - 2\left[\frac{\frac{w}{2}\left(\frac{w}{2}-1\right)}{2}\right] &= \frac{w(w-1)}{2} - \frac{w}{2}\left(\frac{w}{2}-1\right) \\ &= \frac{w}{2}\left(w - 1 - \frac{w}{2} + 1\right) = \frac{w}{2}\left(\frac{w}{2}\right) \\ &= \left(\frac{w}{2}\right)^2 \end{aligned} \tag{3}$$

To bisect the chordal rings, each ring must be cut in two places, and all of the internal connections that cross the cut must be removed. The number of such internal connections that cross each cut at level j with internal width i_j is

$$\sum_{k=1}^{(i_j-1)/2} k$$

The number of chordal ring connections that must be removed to bisect the entire network is the sum of these across these levels:

$$2\sum_{j=1}^{r}\sum_{k=1}^{(i_j-1)/2} k \tag{4}$$

Hence, the bisection width (the bisection width of the atomic center (Eq. 3) plus the number of chordal ring connections that must be removed to bisect the network (Eq. 4)) is

$$\mathrm{bw}(w, r, \{i_j\}) \quad = \quad \left(\frac{w}{2}\right)^2 + 2\sum_{j=1}^{r}\sum_{k=1}^{(i_j-1)/2} k$$

Higher bandwidth ATM connections at the atomic center will result in a larger bisectional bandwidth for the CS networks. Larger values for inside width at each level will also result in greater bisection width. But these large values will limit the number of processors that can be connected to the network. In network structure design, high bisection width and a high number of leaf ports are both important goals. Therefore, decisions must be made to resolve this unavoidable conflict depending on a number of considerations, including performance requirements and cost.

3.1.4 Leaf Port and Bandwidth Utilization Efficiency

The leaf ports efficiency in a CS network can be determined by dividing the number of leaf nodes by the total number of ports available on switches in the network.

$$\mathrm{effp}(w, r, i) = \frac{\mathrm{nlp}(w, r, i)}{w \;\; \mathrm{ns}(w, r, i)}$$

The efficiency of switch bandwidth utilization in a CS network can be determined by dividing bisection width by half the number of leaf ports in the network. This metric shows the bandwidth available to each processor when half of the network tries to communicate with the other half. It represents the worst-case bandwidth availability for the network.

$$\mathrm{effb}(w, i, r) = \frac{\mathrm{bw}(w, i, r)}{\mathrm{nlp}(w, i, r)/2} = \frac{2\mathrm{bw}(w, i, r)}{\mathrm{nlp}(w, i, r)}$$

3.1.5 Optimization of Inside Width

To balance the bisection width with the reduction [13] in leaf ports caused by varying values for the inside width, our analysis will begin with a single switch. Figure 3 shows a switch with width 8 and its different inside and outside width configurations.

In Fig. 3a the inside width is five and the outside width is three. If all the inside links communicate in the (o) direction at the same time, congestion occurs. On the other hand, when all the outside ports want to communicate in the (i) direction at the same time, the excess inside paths are available (congestion can occur only when two or more outside ports communicate through the same inside link). Figure 3a shows that communication in the (o) direction has a greater chance of experiencing congestion compared with the (i) direction. Assuming randomly distributed communications, the level of congestion is dependent on the ratio of inside and outside widths and the amount of communications in both the (i) and (o) directions. Figure 3b shows the opposite results. When both (i) and (o) directions have the same communication demands, the minimum congestion will occur when the inside width is equal to the outside width (Fig. 3c).

Congestion can occur even when excess ports are available if a single link is chosen for multiple communications. But since the ATM switches use leaky bucket algorithms to smooth the communication stream, low levels of congestion will merely result in increased delay. The amount of effort that can be justified in eliminating congestion will depend on the communication patterns for applications that will be run on the cluster.

The inside width in the CS networks must be an odd number, so the inside and outside widths can be equal only when the switch width is $w = 2(2k+1)$ for some integer k. When the inside and outside widths are not equal, the ratio of (i) and (o) direction communication demands will decide the inside and outside widths. Broadcasts are an (o) type communication since a single message injected into the network will result in nlp(w,i,r) (o) type communications. Reductions are an (i) type communication, and point-to-point communications are neutral. If the ratio of broadcasts to reductions is known, then the network can be configured with the same ratio for outside width over inside width. Otherwise we will assume that a near-equal ratio between inside and outside widths will be best on average. In this case the inside width will be

$$i = \begin{cases} \frac{w}{2} & \text{when } \frac{w}{2} \text{ is odd} \\ \frac{w}{2} \pm 1 & \text{otherwise} \end{cases}$$

When $\frac{w}{2}$ is even, the inside width $i = \frac{w}{2} + 1$ will be larger and the resulting network will have a greater bisection width. If we assume a constant inside width for all levels of the network, the parameters of the network can be simplified as follows:

$$\begin{aligned} \mathrm{nlp}(w, r, i) &= w(w-i)^r \\ \mathrm{ns}(w, r, i) &= 1 + \frac{w(w-i)^r - 1}{w-i-1} \\ \mathrm{bw}(w, r, i) &= \left(\frac{w}{2}\right)^2 + \frac{r}{4}(i^2 - 1) \\ \mathrm{effp}(w, r, i) &= \frac{(w-i)^r}{1 + \frac{w(w-i)^r - 1}{w-i-1}} \end{aligned}$$

$$\mathrm{effb}(w,r,i) \quad = \quad \frac{2\left(\left(\frac{w}{2}\right)^2 + \frac{r}{4}(i^2-1)\right)}{w(w-i)^r}$$

The leaf port efficiency is affected only slightly by the radius, but it's affected greatly by switch width, as shown in Fig. 4a. Switch width values of sixteen or larger result in near-optimal port efficiency. The bandwidth efficiency is affected by switch width and the radius of the network. Figure 4(b) shows that switch width values less than sixteen and radius values less than six result in networks with acceptable bandwidth efficiency.

Tables 1 and 2 show the topological properties of two CS networks with inside width (w/2)-1 and varying radius values. The leaf port efficiency remains constant at approximately 0.5 and the bisection width is large enough to maintain an acceptable bandwidth efficiency for up to 10,000 processors.

4 Routing

In order to perform the routing in ATM networks, the source processor must specify the path through all the switches to the destination. Since ATM Networks are connection-oriented, the path through intermediate switches must be established before data can be sent [6, 24]. Signaling will be much easier if an appropriate addressing scheme is adopted.

4.1 Addressing

The identical address connection method (IACM) has been developed to simplify connection setup in the network. Using this scheme, connections between two switches utilize identical port numbers. Port 5 on the first switch will be connected only to port 5 on the second switch. This addressing method enables the address of an entity to specify its exact position in the network.

For instance, in Fig. 5a, port number 0 in the atomic center has a spoke connection to port 0 on the perimeter switch with address 0. Similarly, the additional perimeter switches are given an address that corresponds to the port on the center switch to which they are connected. In CS(8,1,1) network, the 8 perimeter switches are labeled from 0 to 7.

In addition to the spoke connections, the IACM will also be used for the chordal links. One way of making these connections is to choose the first pair of ports that have the same address between neighboring perimeter switches. The first port following the switch address that is shared between switches 3 and 4 in Fig. 5a is 5, so port 5 is used as the chordal connection between these switches in Fig. 5b. Figure 5 shows the $\mathrm{CS}(8,1,i)$ networks with varying inside widths.

Using the addressing scheme described above, the three leaf ports at the top of Fig. 5c are $6 \cdot 2, 0 \cdot 2$, and $1 \cdot 2$. The three leaf ports at the bottom of Fig. 5c are $2 \cdot 6, 4 \cdot 6$, and $5 \cdot 6$.

Addressing for multilevel networks begins with the single-level $\mathrm{CS}(w,1,\frac{w}{2})$. We can think of it as a high level center that has $\mathrm{nlp}(w,1,\frac{w}{2})$ ports. Each port is connected to a perimeter switch. The only difference between an atomic center and a high-level center is that a high-level center has

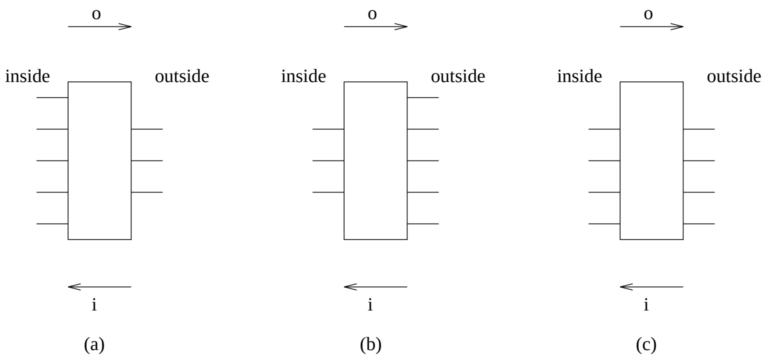


Figure 3: Three configurations of inside and outside widths for a switch with width 8

Radius	Diameter	Bisection Width	Switches	Leaf Ports	Effp	Effb
1	1	18	9	40	0.555556	0.9
2	3	20	49	200	0.510204	0.2
3	5	22	249	1000	0.502008	0.044
4	7	24	1249	5000	0.5004	0.0096
5	9	26	6249	25000	0.50008	0.00208
6	11	28	31249	125000	0.500016	0.000448

Table 1: Topological Properties for CS(8,r,3)

Radius	Diameter	Bisection Width	Switches	Leaf Ports	Effp	Effb
1	1	76	17	144	0.529412	1.055556
2	3	88	161	1296	0.503106	0.135802
3	5	100	1457	11664	0.500343	0.017147
4	7	112	13121	104976	0.500038	0.002134

Table 2: Topological Properties for CS(16,r,7)

redundant port addresses (nlp$(w, 1, \frac{w}{2}) > w$) and is made up of multiple switches. Figure 6 gives an example of perimeter switch addressing in a multilevel network.

A leaf port p in CS(w, r, i) has address $\{a_0, a_1, ..., a_r\}$ where $a_i \epsilon \{0, 1, ..., w-1\}$.

For example, in Fig. 7, leaf ports A, B, and C have the following addresses:

$A : 1 \cdot 4 \cdot 0$

$B : 1 \cdot 6 \cdot 0$

$C : 1 \cdot 2 \cdot 6$

4.2 Path Finding

Two techniques have been developed to determine the path between the source and destination node. These methods are as follows:

1. Spanning Tree Path

   The path from source leaf node to destination leaf node uses the natural spanning tree. This path is unique.

   Suppose there exist source leaf node s and destination leaf node d in CS(w, r, i) with the following addresses:

   $s : \{s_0, s_1, ..., s_{r-1}, s_r\}$

   $d : \{d_0, d_1, ..., d_{r-1}, d_r\}$

   Then the spanning tree path is:

   $\{s_0, s_1, ..., s_{k-1}, s_k, d_k, d_{k-1}, ..., d_1, d_0\}$ with $s_k \neq d_k$ and $k \leq r$

   For example, in Fig. 7, the path from leaf node A to leaf node C is $\{1, 4, 0, 6, 2, 1\}$. The spanning tree path from leaf node A to leaf node B is $\{1, 4, 6, 1\}$. The spanning tree path is shortened since $s_2 = d_2$.

2. Shortcut

   Chordal links offer potentially shorter paths and can decrease the congestion on spanning tree connections. Instinctively, we should choose chordal paths only if the length is less than the spanning tree path. For this reason these paths will be called shortcuts. A shortcut will exist between two nodes only if at some level L the number of positions in the ring between s_L and d_L is less than $\frac{i-1}{2}$.

   A position table is used to determine the number of ring positions between switches in each level. The following algorithm is used to find the shortcut path.

   Shortcut Algorithm:

   Input: Spanning tree path: $\{s_0, s_1, ..., s_{k-1}, s_k, d_k, d_{k-1}, ..., d_1, d_0\}$ and position table

   Output: Shortcut path

For $j = 1$ to k {

If s_j and d_j are neighbors within $\frac{i-1}{2}$ then {

- Read chordal link number c from the position table.
- Output shortcut:

  $\{s_0, s_1, ..., s_{j-1}, c, d_{j-1}, ..., d_1, d_0\}$ where $j \leq k$
- Terminate the algorithm.

}

}

If not terminated, output = Input. □

If we use Fig. 7 for a communication between leaf node A and leaf node B, the spanning tree path is $\{1, 4, 6, 1\}$. Since switch $4 \cdot 0$ and $6 \cdot 0$ differ by one position in the ring, a direct chordal link exists. The corresponding shortcut is $\{1, 7, 1\}$.

5 Comparison

A major difference between the CS networks and traditional interconnection networks is that the network is constructed from ATM switches. The processing nodes are connected to leaf ports of the CS network. Processors are involved in communications only if they are the source or destination of the communications. These properties reduce communication overhead and allow the network to be constructed with lower-cost commodity ATM switches.

Table 3 compares metrics of interest for Hypercubes, Generalized Fat Trees, Extended Generalized Fat Trees (XGFT), and chordal spoke networks. These topologies were chosen for the following reasons: The Hypercube has been used extensively in previous research and many algorithms can be mapped to this topology. Although generalized fat trees have not yet been implemented, in theory they compare favorably with many popular interconnection networks [10, 11, 15]. The topology used in the Connection Machine CM-5 is a special case of the extended generalized fat tree [10, 15].

Table 3 shows that CS(w, r, i) has nearly double the number of leaf nodes when compared to the GFT(r,w,p) and more than that found in a Hypercube with the same diameter. Both the Hypercube and GFT have greater bisection width than CS networks, but the CS(w,r,i) bisection width scales up with the number of nodes in the network. Table 4 gives a specific example of how the metrics compare for the three networks.

5.1 Embedding

Embedding can be used to show that a host graph is at least as powerful as a given guest graph. The dilation of the embedding is the maximum distance on the host between adjacent nodes in the guest topology. The load is the maximum number of guest nodes mapped onto a single host node. The edge congestion is the maximum number of edges in the guest that are mapped to a single edge in the host.

5.1.1 Embedding a Square Mesh into the Chordal Spoke network

Suppose we have a Mesh(k, q) with wrap-around connections between processors in adjacent rows or columns. The parameter k is the number of nodes in one row of Mesh and q is the number of dimensions of Mesh. Each node has degree $2q$ and the number of nodes is k^q (when $q = 1$, the Mesh becomes a Ring).

We use a customized chordal spoke network in order to embed a Mesh. The smaller switches can be used at the perimeter of the network to reduce the number of leaf ports. At the perimeter level the switch width is $2q + 2$ instead of the w ports switch used in the interior of the network (assuming $2q + 2 < w$). $2q$ ports correspond to the node degree on the Mesh, one port connects to the center, and the last port connects to the processing element. The inside width for the perimeter switches is $2q + 1$ and the outside width is 1. The center is a switch or a chordal spoke network, depending on the size of k^q. At the center, the number of leaf ports is equal to the number of nodes in Mesh ($k^q = w(\frac{w}{2})^r$). The relation between the radius (r) of the chordal spoke network and the dimension (q) of the Mesh is

$$r = \frac{q \log(k) - \log(w)}{\log(w/2)}$$

Theorem 1 In CS$(w, r, \frac{w}{2})$, a Mesh M(k, q) can be embedded with optimal load 1, edge congestion 1, and dilation $\frac{2q \log(k) - \log(w)}{\log(w/2)}$.

Proof:

Figure 8a shows M(4,2). In Fig. 8b the mesh is redrawn in a chordal ring shape. Figure 8c shows the embedding into the CS$(16, 1, 5)$ network.

For a given Mesh with k^q nodes, we can always find a chordal spoke network with nlp$(w, r, i) \geq k^q$. Each node in the Mesh is mapped to a leaf port with load = 1.

Since there is at most one edge between any two nodes in the Mesh (except between the first node and the last node, which have q edges), the edge congestion is one (for the first and last nodes in Mesh, the edge congestion is q).

Since the center will provide a path when there is no shorter path between the source and the destination nodes, the dilation will be equal to the diameter $2\frac{qlog(k) - log(w)}{log(\frac{w}{2})}$.

□

5.1.2 Embedding a Fat Tree into the Chordal Spoke Network

The example in Fig. 9 is used to illustrate the general procedure. Suppose we have a fat tree FT$(4, 2, 2)$ with four child connections on each non-leaf node, height of two and two connections to parent nodes. Processors are connected to the leaf nodes [11]. There are sixteen leaf nodes and twelve non-leaf nodes [18] (Fig. 9a). Figure 9b shows the fat tree redrawn in a chordal spoke shape.

We can build a customized chordal spoke CS$(16, 1, 3)$ network that has small switches with width four (instead of sixteen) at the perimeter. Figure 9c illustrates one of these networks and Fig. 9d shows the CS$(16, 1, 3)$ redrawn in a fat tree shape.

The sixteen leaf nodes in the fat tree are mapped into the sixteen perimeter switches in the chordal spoke network. The twelve non-leaf nodes in the fat tree are mapped into the center of the CS. Processors are connected with a single edge to the leaf nodes in each topology. The load is twelve and the dilation is one. The edge congestion is two because the two edges in the fat tree that connect to the parents are mapped to a single edge in the CS.

Another customized CS$(7, 2, 3)$ could be built that uses switches with width four in the atomic center and perimeter. Switches with a width of seven are used for the interior of the network. See figures 9e and 9f. In this case, the sixteen leaf nodes in the fat tree are mapped into sixteen perimeter switches in the CS. The eight second-level nodes in the fat tree are mapped into four interior switches in chordal spoke, and the four highest-level nodes in fat tree are mapped into the atomic center of the chordal spoke. The load is four, the dilation is one and edge congestion is two.

Although these embeddings are not perfect, the CS host graph has several properties that are superior to the guest graphs. Table 5 shows that the CS(16,1,3) has a greater bisection width and smaller diameter, and that it uses fewer switches and inside edges than the fat tree with the same number of processors.

6 Conclusions

Parallel-processing platforms based on networks of workstations are being utilized extensively to solve challenging computational problems. ATM network technology promises to provide an inexpensive and flexible interconnection medium for applications on workstation clusters that require high bandwidth and congestion-free communications. The chordal spoke SASI network can be used to connect large numbers of processors with inexpensive ATM switches. This network topology has scalable bisection width, acceptable port and bandwidth efficiency, and it compares well with other interconnection networks. Because of its regular topology, routing can be performed in a structured manner.

Current research is focused on developing simulators and analyzing embeddings from additional topologies into chordal spoke networks. Additional research will explore three-dimensional SASI network topologies and extend the definition of the CS networks to include the custom networks used in embeddings. When the information highway is deployed, parallel processing can be performed using machines that are geographically distant. Future work will analyze the impact of CS networks in Wide Area Networks (WANs) and suggest topologies that will provide the resources necessary for parallel processing. Chordal spoke networks provide a flexible and economical alternative to existing MPP networks and promise to improve the functionality of clustered computing.

Comparison of topological properties				
Network	Leaf Ports	Diameter	Degree	Bisection Width
Hypercube Q(2r)	4^r	$2r-1$	$2r$	2^{2r-1}
Fat trees GFT($\frac{w}{2}, r, p$)	$(\frac{w}{2})^r$	$2r-1$	p	$\frac{w \cdot p^r}{4}$
XGFT(r;$w_1, ..., w_r; p_1, ..., p_r$)	$\prod_{i=1}^r w_i$	$2r-1$	p_1	$\frac{w_r \cdot \prod_{i=1}^r p_i}{2}$
chordal spoke CS($w, r-1, \frac{w}{2}$)	$2(\frac{w}{2})^r$	$2r-1$	$\frac{w}{2}$	$(\frac{w}{2})^2 + \frac{r-1}{4}(\frac{w^2}{4} - 1)$

Table 3: CS($w, r-1, \frac{w}{2}$) compared with other topologies

Comparison of topological properties				
Network	Leaf Nodes	Diameter	Degree	Bisection Width
Hypercube $Q(2 \cdot 4)$	256	7	8	128
Fat tree $GFT(9,4,3)$	6561	7	3	324
Chordal spoke $CS(16,3,7)$	11664	7	7	100

Table 4: CS(16,3,7) compared with other topologies

Comparison of topological properties				
Network	Bisection Width	Diameter	Switches	Inside Edges
Fat tree $FT(4,2,2)$	8	5	28	48
Chordal spoke $CS(7,2,3)$	8	5	21	40
Chordal spoke $CS(16,1,3)$	66	3	17	32

Table 5: FT(4,2,2) compared with chordal spoke networks

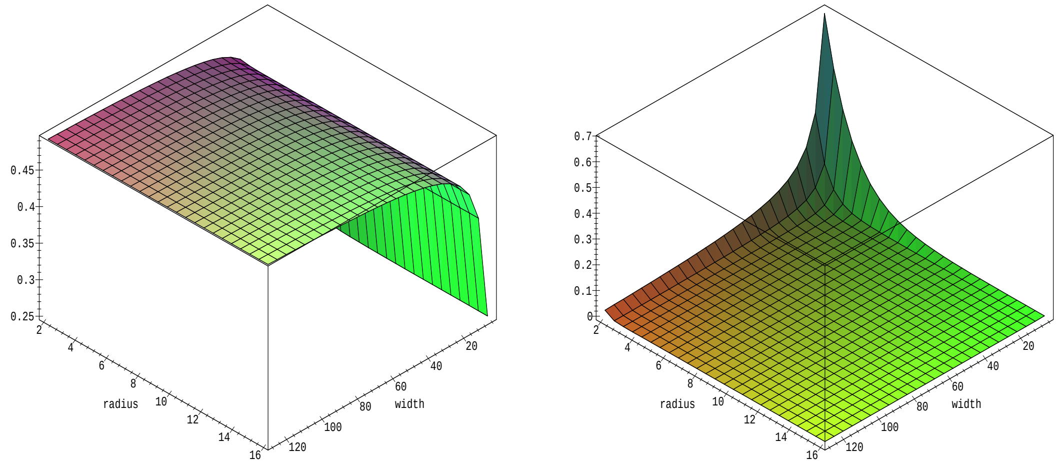


Figure 4: Leaf Ports Efficiency (a) and Bandwidth Efficiency (b)

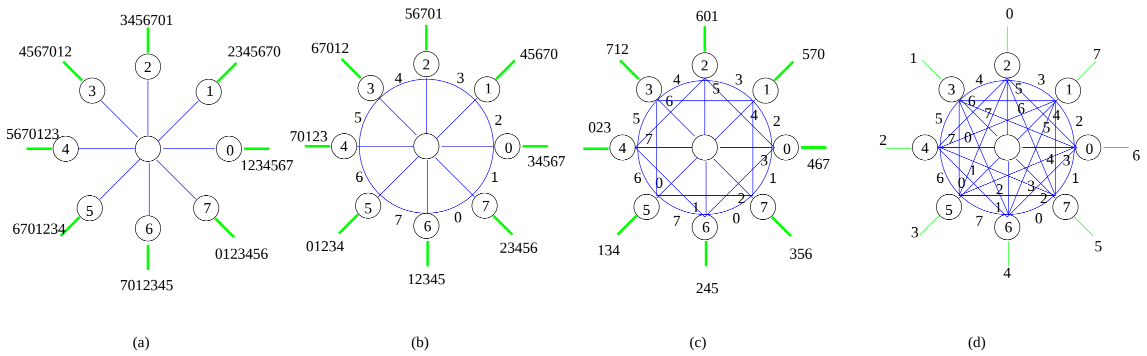


Figure 5: CS$(8,1,i)$ networks with (a) CS$(8,1,1)$, (b) CS$(8,1,3)$, (c) CS$(8,1,5)$, and (d) CS$(8,1,7)$

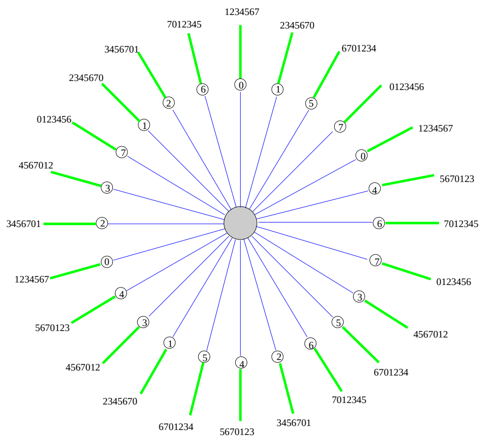


Figure 6: Addressing scheme for perimeter switches in CS$(8,2,1)$

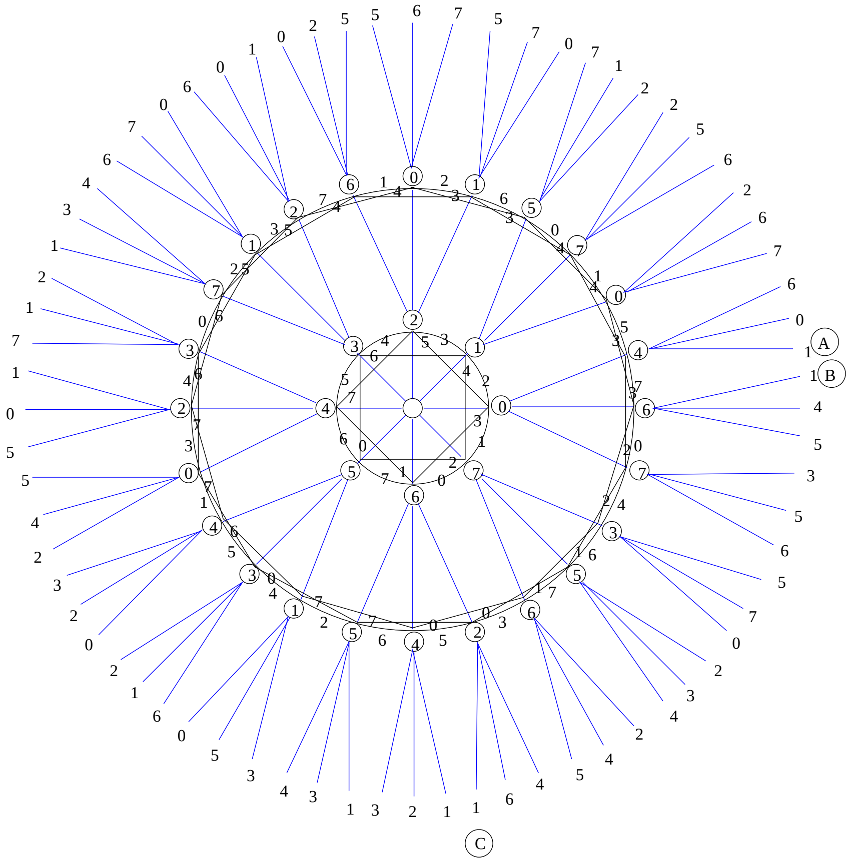
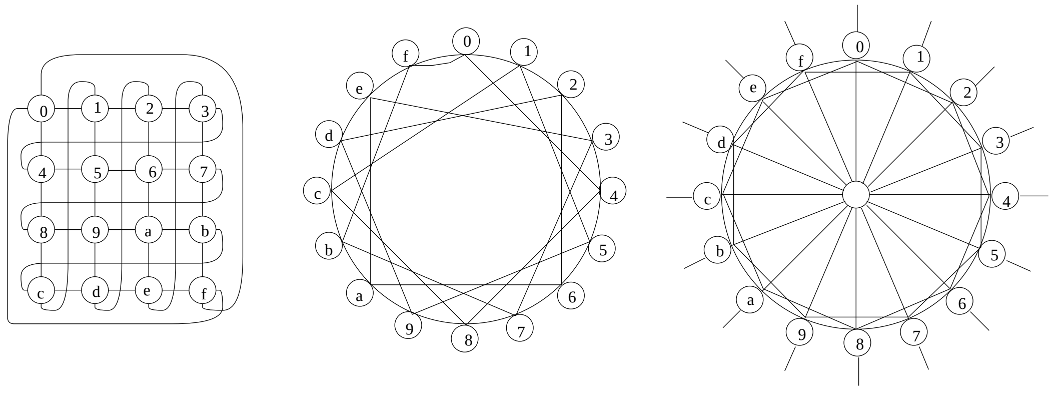


Figure 7: Addressing scheme for leaf ports in CS$(8, 2, 5)$

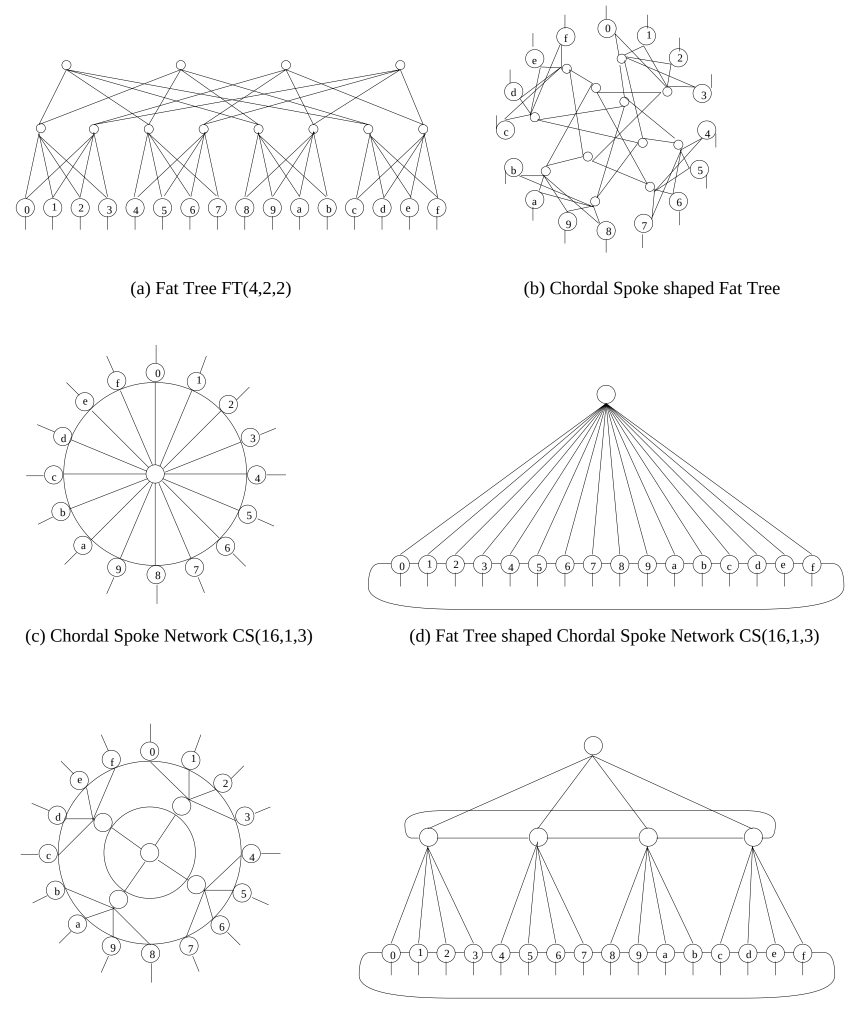


(a) 4x4 Mesh with wrap-around connections

(b) Chordal Ring Shaped Mesh

(c) Chordal Spoke Network CS(16, 1, 5)

Figure 8: Embedding a Mesh into a chordal spoke network



(a) Fat Tree FT(4,2,2)

(b) Chordal Spoke shaped Fat Tree

(c) Chordal Spoke Network CS(16,1,3)

(d) Fat Tree shaped Chordal Spoke Network CS(16,1,3)

(e) Chordal Spoke Network CS(7,2,3)

(f) Fat Tree shaped Chordal Spoke Network CS(7,2,3)

Figure 9: Embedding a fat tree FT$(4,2,2)$ into a chordal spoke network

References

[1] Akers, S. B. , Harel, D. , and Krishnamurthy, B. . "The Star Graph: An Attractive Alternative to the n-Cube". In Proceedings International Conference on Parallel Processing, pages 393–400. IEEE, 1987.

[2] Browne, R. F. and Hodgson, R. M. . "Symmetric degree-four chordal ring networks". IEE Proceedings, 137 Pt.E(4):310–318, July 1990.

[3] D. Duh, G. C. and Fang, J. . "Algorithms and Properties of a New Two-Level Network with Folded Hypercubes as Basic Modules". IEEE Transactions on Parallel and Distributed Systems, 6(7):714–723, July 1995.

[4] Dowd, P. W. , Srinidhi, S. M. , Blade, E. , and Claus, R. . "Issues in ATM Support of High Performance Geographically Distributed Computing". In First International Workshop on High-Speed Network Computing, pages 19–28. IEEE, 1995.

[5] FORE Systems, Inc. "ForeRunner ASX-200 ATM Switch User's Manual". FORE Systems, Inc, 1994, ch.1 pp. 1-12.

[6] Huang, C. , Devetsikiotis, M. , Lambadaris, I. , and Kaye, R. . "Fast Simulation for Self-Similar Traffic in ATM Networks". In 1995 IEEE International Conference on Communications, pages 438–443. IEEE, 1995.

[7] Kalampoukas, L. and Varma, A. . "Performance of TCP over Multi-Hop ATM Networks: A Comparative Study of ATM-Layer Congestion Control Schemes". In 1995 IEEE International Conference on Communications, pages 1472–1477. IEEE, 1995.

[8] Kruskal, C. P. and Snir, M. . "A Unified Theory of Interconnection Network Structure". Theoretical Computer Science, 48:75–94, 1986.

[9] Kwon, T. and Jeong, C. . "A Simple, Extendible ATM Switch with Load-balanced Rounding Copy Network". In 1995 IEEE International Conference on Communications, pages 1122–1126. IEEE, 1995.

[10] Leigserson, C. E. , Abuhamdeh, Z. S. , Douglas, D. C. , Feynman, C. R. , Ganmukhi, M. N. , Hill, J. V. , Hillis, W. D. , Kuszmaul, B. C. , Pierre, M. A. S. , Wells, D. S. , Wong, M. C. , Yang, S. , and Zak, R. . "The Network Architecture of the Connection Machine CM-5". In Proc. Fourth Ann. ACM Symp. Parallel Algorithms and Architectures, pages 272–285. ACM, 1992.

[11] Leiserson, C. E. . "Fat-Trees: Universal Networks for Hardware-Efficient Supercomputing". IEEE Trans. Computers, C-34(10):892–901, Oct. 1985.

[12] Leung, K. and Hamdi, M. . "Evaluating PVM and Express on Various Network Clusters". In First International Workshop on High-Speed Network Computing, pages 57–66. IEEE, 1995.

[13] Miguet, S. and Robert, Y. . "Reduction Operations on a Distributed Memory Machine with a Reconfigurable Interconnection Network". IEEE Transactions on Parallel and Distributed Systems, 3(4):500–505, July 1992.

[14] Murakami, K. and Kim, H. S. . "Joint Optimization of Capacity and Flow Assignment for Self-Healing ATM Networks". In 1995 IEEE International Conference on Communications, pages 216–220. IEEE, 1995.

[15] Ohring, S. R. , Ibel, M. , Das, S. K. , and Kumar, M. J. . "On Generalized Fat Trees". In 9th International Parallel Processing Symposium, pages 37–44. IEEE, 1995.

[16] Partridge, C. . "Gigabit Networking". Addison-Wesley Publishing Company, 1994, ch.1, pp. 1-13, ch.2, pp. 28-31, and ch.3, pp. 51-53.

[17] Prycker, M. D. . "Asynchronous Transfer Mode". Ellis Horwood, 1993, ch.3, pp. 103-146.

[18] Quinn, M. J. . "Parallel Computing: Theory and Practice". McGraw-Hill, Inc., 1994, ch.3, pp. 52-89.

[19] Ramanathan, G. , Clement, M. , and Crandall, P. . "Hyperweave: A Fault-tolerant Expandable Interconnection Network". In Proceedings of the Fourth IEEE Symposium on Parallel and Distributed Processing, pages 479–482. IEEE, 1992.

[20] Scherson, I. D. and Youssef, A. S. . "Interconnection Networks for High-Performance Parallel Computers". IEEE Computer Society Press, 1994, ch.1, pp. 1-2.

[21] Seitz, C. L. . "The Cosmic Cube". comm. ACM, 28(1):22–33, Jan. 1985.

[22] Sole-Pareta, J. , Sarkar, D. , Liebeherr, J. , and Akyildiz, I. F. . "Adaptive Multipath Routing of Connectionless Traffic in an ATM Network". In 1995 IEEE International Conference on Communications, pages 1626–1630. IEEE, 1995.

[23] Wei, W. , Clement, M. J. , and Flanagan, J. K. . "The Round Table ATM Interconnection Network". In Proceedings of the 1995 International Conference on Parallel and Distributed Processing Techniques and Applications, pages 445–454. CSREA, 1995.

[24] Winkler, R. . "A Routing Framework for Multimedia Multipoint Communications in ATM Networks". In 1995 IEEE International Conference on Communications, pages 815–820. IEEE, 1995.

[25] Zhou, H. and Geist, A. . "Faster Message Passing in PVM". In First International Workshop on High-Speed Network Computing, pages 67–73. IEEE, 1995.