Appeared in "Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA'95)"

The Round Table ATM Interconnection Network †

Wei Wei, Mark J. Clement and J. Kelly Flanagan

Department of Computer Science Brigham Young University 3372 TMCB
Provo, Utah 84602-6576 {weiwei, clement, kelly}@cs.byu.edu

Abstract

Networks of workstations are becoming a popular platform for parallel computing. Asynchronous Transfer Mode (ATM) network connections can improve the communication characteristics of these parallel clusters. This paper introduces a new class of topologies for constructing multicomputers called Small ATM Switch based Interconnection Networks (SASIN). The Round Table Network (one kind of SASIN) is introduced and analyzed. It is hierarchical, symmetrical, has a low diameter and is easy to implement. The bisection width, diameter, number of switching elements and efficiency of this network are compared to other interconnection topologies.
Keywords: ATM, Interconnection Networks, routing, multicomputers.

1 Introduction

Recent advances in the power of parallel computers have made them a cost effective platform for solving large computational problems [5, 11, 12]. Networks of workstations have been explored for use as a parallel processing platform for the following reasons:

- Each node of a workstation cluster is generally less expensive than the corresponding node in a custom parallel machine.
- A single node in a workstation cluster generally exhibits higher performance than a single processor in a Massively Parallel Processing (MPP) machine. This occurs because of the long lead time in developing the interconnection network for large parallel machines.

† This research was supported by Sprint Corporation.

- The software development environment is more mature on workstation clusters. The large volume market for workstations enables vendors to develop high performance compilers and debugging tools.

Workstation clusters have traditionally communicated through a shared media Local Area Network (LAN) such as Ethernet. Shared media communication networks are not practical for connecting large numbers of workstations because of the fixed bandwidth available in these configurations. This research will investigate network technologies which can enable cluster solutions to be used as low cost, high performance parallel processing platforms.

In the past decade, general purpose communication networks have achieved significant increases in performance. ATM networks have been proposed as a flexible interconnection media for cluster computing platforms[1, 8]. ATM switching technology[2] is currently being used for telephone networks and the hardware promises to become less expensive as telephony deployment continues. Using ATM technology, Gigabit [7] communication links are available for each node in a multicomputer, overcoming the bandwidth limitations of Ethernet style LAN connections. The cost of ATM switches increases exponentially as more inputs and outputs are added to the switch [7]. For this reason networks consisting of many small switches may be more feasible for cost sensitive cluster computing solutions. We will call the class of networks constructed in this way Small ATM Switch based Interconnection Networks (SASIN).

ATM networks are different from MPP networks in several ways. Traditional MPP networks utilize electrical media and generally operate at a fixed bandwidth [9]. Since optical fiber can be used as a medium for ATM networks, much higher bandwidth connections are available. The ATM protocol specifies a fixed-size, cell based, connection-oriented network which results in switching fabrics which are simpler than many MPP machines. Standard ATM data rates range from 51.84 Mb/s to 2,488.32 Mb/s [7]. 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. The improvement in bandwidth and the low cost of ATM networks makes them attractive for many parallel applications that are not highly sensitive to latency.

SASIN is very flexible. It can be used to construct Generalized Fat Trees (GFT)[6] Meshes, Hypercubes, Hyperweave, Pyramids and many other topologies [9, 10]. In this paper, we introduce one class of SASIN – Round Table Network. This network uses a minimum number of ATM switches to connect the maximum number of leaf nodes.

2 Definitions and Notations

A *switch* is a connection unit with w pairs of input and output ports. We will denote the *width* of a switch as the number of input and output ports. Each input port and output port has a unique *address* from $0, 1, ..., w-1$. A switch with *width* w can also be thought of as a w-node completely connected graph. If processing elements were connected to each port of the switch, there would be a direct, contention free connection between each pair of nodes. But if several source nodes send a message to a common destination node at the same time congestion can occur. Congestion can cause increased delay or lost data in an ATM switch.

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 connect switches. For simplification, we say a switch has w ports instead of w pairs of input and output ports. 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 called an *edge*. It is the shortest or unit *distance* between two switches.

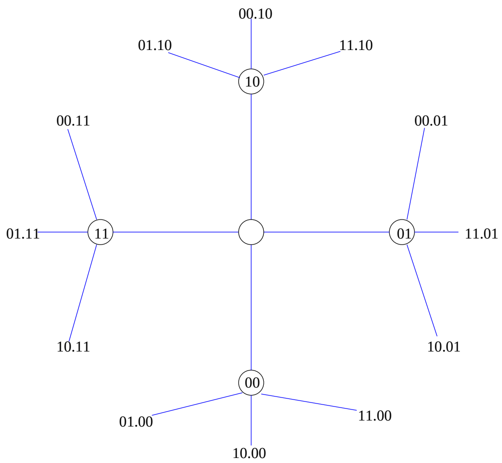

Figure 1: An example of RTN (4, 2)

A SASIN 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.

3 Round Table Networks

A *Round Table Network* RTN(w, r) is a hierarchical network that has many appealing properties. It can be characterized in the following way:

- A RTN$(w, 1)$ is an ATM switch. It is the *atomic center* of its w ports. The *addresses* of leaf ports in RTN$(w, 1)$ are 0, 1, ..., w-1.

- A RTN(w, r) with $r > 1$ is a network which uses an RTN$(w, r-1)$ as its *center* with a *perimeter* switch connected at each leaf port of RTN$(w, r-1)$. The *address* of each leaf port of RTN(w, r) is the *perimeter* switch's port address followed by the leaf port address in the RTN$(w, r-1)$ where the *perimeter* switch is connected.

  For instance, a RTN(4, 1) has leaf ports 00, 01, 10 and 11 and a RTN(4, 2) can be built by adding four *perimeter* switches to RTN(4, 1)'s four ports. The RTN(4,2) has 12 leaf ports shown in Figure 1.

The *radius* r of RTN(w, r) is defined as the number of switches between a leaf port and the RTN(w, r) network's *atomic center*. For RTN$(w, 1)$, the radius is 1. The number of leaf ports in RTN(w, r) is denoted by nlp(w, r) and the number of switches is denoted by ns(w, r).

4 Topological Properties

The Round Table Network RTN(w, r) is symmetrical about the atomic center. This symmetry simplifies routing and node addressing and makes the RTN easy to scale up to large numbers of processors. The distance from each leaf node to the atomic center is the same, resulting in predictable communication latency between nodes. There exists only one path between any two leaf nodes in a RTN. Because of this property, and the fact that there is a path between every pair of leaf nodes, RTN inherits the desirable properties of Banyan networks[3]. The *diameter* – the longest distance between two leaf nodes – of RTN(w, r) is $2r - 1$ since 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 exist $2r - 1$ switches in the path.

The RTN(w, r)'s atomic center has w leaf ports which are connected to w perimeter switches. These w perimeter switches each have $w - 1$ leaf ports which are connected to $w(w - 1)$ perimeter switches. The number of leaf ports for RTN(w, r) is

Theorem 4.1

$$nlp(w, r) = (w - 1)nlp(w, r - 1) = w(w - 1)^{r-1}$$

Proof:

- The formula obviously holds for $r = 1$:
  $nlp(w, 1) = w$
- We make the induction hypothesis that $nlp(w, r) = w(w - 1)^{r-1}$ for some positive integer r.
  This implies
  $$nlp(w, r + 1) = (w - 1)nlp(w, r) = (w - 1)w(w - 1)^{r-1} = w(w - 1)^r$$

$\square$

A RTN(w,1) consists of a single ATM switch connected to leaf nodes (ns(w, 1) = 1). As the network grows, a perimeter switch will be added to each leaf port. The generalized number of switches in a RTN(w, r) is

Theorem 4.2

$$\begin{aligned} ns(w, r) &= ns(w, r - 1) + nlp(w, r - 1) \\ &= 1 + \sum_{i=0}^{r-2} w(w - 1)^i \\ &= 1 + w\left(\frac{(w - 1)^{r-1} - 1}{w - 2}\right) \end{aligned}$$

Proof:

- The formula obviously holds for $r = 1$:
  $ns(w, 1) = 1$
- We make the induction hypothesis that $ns(w, r) = 1 + \sum_{i=0}^{r-2} w(w - 1)^i$ for some positive integer r.
  This implies
  $$\begin{aligned} ns(w, r + 1) &= ns(w, r) + nlp(w, r) \\ &= 1 + \sum_{i=0}^{r-2} w(w - 1)^i + w(w - 1)^{r-1} \\ &= 1 + \sum_{i=0}^{r-1} w(w - 1)^i \\ &= 1 + w\left(\frac{(w - 1)^r - 1}{w - 2}\right) \end{aligned}$$

□

The average distance between nodes in a RTN(w, r) is shorter than the diameter since some paths do not go through the atomic center. One of the advantages of RTN is that $1/w$ of the paths are shorter than the diameter.

Theorem 4.3

$$al(w,r) = \frac{(2r-1)(w-1)^r + (w-2)\sum_{i=0}^{r-2}(2i+1)(w-1)^i}{w(w-1)^{r-1}-1} \quad \textit{for } r \geq 2 \textit{ and } w \geq 3$$

Proof:

- The formula holds for $r = 2$:

$$al(w,2) = \frac{1(w-2)+3(w-1)(w-1)}{w-2+(w-1)(w-1)} = \frac{3(w-1)^2+w-2}{w(w-1)-1}$$

- We make the induction hypothesis that

$$al(w,r) = \frac{(2r-1)(w-1)^r + (w-2)\sum_{i=0}^{r-2}(2i+1)(w-1)^i}{w(w-1)^{r-1}-1}$$

  for some positive integer r.
  This implies

$$\begin{aligned} al(w,r+1) &= \frac{1}{nlp(w,r+1)-1} \cdot \{(2r-1+2)(w-1)^r(w-1) \\ &\quad +(w-2)\sum_{i=0}^{r-2}(2i+1)(w-1)^i + (w-2)(2r-1)(w-1)^{r-1}\} \\ &= \frac{(2(r+1)-1)(w-1)^{r+1} + (w-2)\sum_{i=0}^{(r+1)-2}(2i+1)(w-1)^i}{w(w-1)^{(r+1)-1}-1} \end{aligned}$$

□

For a given network, the *bisection width* is the minimum number of connections removed in order to divide the network into two equal halves. Since the atomic center of a RTN must be cut in order to divide the network into two equal halves, the bisection width of the RTN(w, r) is equal to the bisection width of the atomic center. For an ATM switch with w ports, the bisection width is

Theorem 4.4

$$bw(w,r) = \left(\frac{w}{2}\right)^2$$

Proof: An ATM switch can logically be viewed as a w-node fully connected graph. The number of edges for a w-node fully connected graph 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

Table 1: Topological properties for RTN($4,r$)

radius	diameter	bw	ns	nlp	effp	effb	al
1	1	4	1	4	1.0	1.0	1.0
2	3	4	5	12	0.6	0.333333	2.545455
3	5	4	17	36	0.529412	0.111111	4.142857
4	7	4	53	108	0.509434	0.037037	5.813084
5	9	4	161	324	0.503106	0.012346	7.526316
6	11	4	485	972	0.501031	0.004115	9.260556

Table 2: Topological Properties for RTN(16,r)

radius	diameter	bw	ns	nlp	effp	effb	al
1	1	64	1	16	1.0	4.0	1.0
2	3	64	17	240	0.882353	0.266667	2.828452
3	5	64	257	3600	0.875486	0.017778	4.701584
4	7	64	3857	54000	0.875032	0.001185	6.584307

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

$$\begin{aligned} bw(w,r) &= \frac{w(w-1)}{2} - 2\frac{\frac{w}{2}(\frac{w}{2}-1)}{2} = \frac{w(w-1)}{2} - \frac{w}{2}(\frac{w}{2}-1) \\ &= \frac{w}{2}(w-1-\frac{w}{2}+1) = \frac{w}{2}(\frac{w}{2}) = (\frac{w}{2})^2 \end{aligned}$$

□

Higher bandwidth ATM connections at the atomic center will result in larger bisectional bandwidth for the RTN. Large radius RTN may require higher bandwidth ATM switches for applications which are sensitive to bisectional bandwidth.

The *leaf port efficiency* in a RTN can be determined by dividing the number of leaf nodes by the total number of ports available on switches in the network. RTN achieves maximum efficiency because if any connection between switches is removed, the graph is disconnected.

$$effp(w,r) = \frac{nlp(w,r)}{w \cdot ns(w,r)} = \frac{(w-1)^{r-1}}{1 + w(\frac{(w-1)^{r-1}-1}{w-2})}$$

The *bandwidth efficiency* in a RTN can be determined by dividing bisection width by the number of leaf ports in the network. It shows the worst case bandwidth sharing.

$$effb(w,r) = \frac{bw(w,r)}{nlp(w,r)} = \frac{w}{4(w-1)^{r-1}}$$

Table 1 and Table 2 show the topological properties of two RTNs. The efficiency for RTN($4,r$) is lower than RTN($16,r$), but the network scales up more gradually. With 16 port ATM switches, a radius of 3 is big enough to connect thousands of processors.

The leaf port efficiency is affected little by radius but strongly by switch width. The wider the switch width is, the higher the efficiency is. The results are shown in the left part of Figure 2.

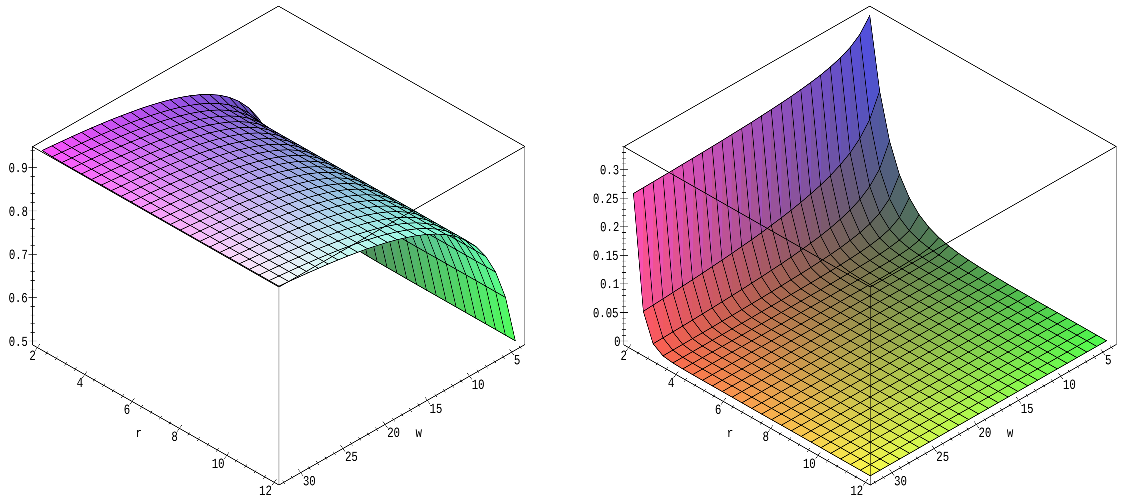

Figure 2: Leaf Ports Efficiency and Bandwidth efficiency

The bandwidth efficiency is affected by switch width and strongly by radius. The bigger the radius is, the lower the efficiency is. The results are shown in the right part of Figure 2.

The low bandwidth efficiency means that in the leaf level of RTN, the $w-1$ leaf ports will share the bandwidth of one port of the leaf level switch. Correspondingly, the leaf level switches will share the bandwidth of one port on the center switch. So the leaf or near leaf level switches will not effectively utilize their bandwidth. One way to overcome this problem is to use high bandwidth switches (OC-12 or higher) in centers and use low bandwidth switches (OC-3 or lower) in leaf or near leaf regions. Figure 3 is an example of how to balance the bandwidth efficiency.

Table 3 compares metrics of interest for Hypercubes, Generalized Fat Trees (GFT), Extended Generalized Fat Trees (XGFT) and Round Table 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 been implemented, in theory, they compare favorably with many popular interconnection networks. The topology used in the Connection Machine CM-5 is a special case of the Extended Generalized Fat Tree[4, 6].

Table 3 shows that RTN(w, r) has less leaf nodes than GFT(r,w,p), but far more than

Table 3: RTN(w,r) compared with other topologies

Comparison of topological properties					
Network	nlp	diameter	Degree	bw	Cost
Hypercube Q(2r)	4^r	$2r-1$	$2r$	2^{2r-1}	$2r(2r-1)$
Fat Trees GFT(w,r,p)	w^r	$2r-1$	p	$\frac{w \cdot p^r}{2}$	$(2r-1)p$
XGFT(r;$w_1, ..., w_r; p_1, ..., p_r$)	$\prod_{i=1}^{r} w_i$	$2r-1$	p_1	$\frac{w_r \cdot \prod_{i=1}^{r} p_i}{2}$	$(2r-1)p_1$
Round Table RTN(w,r)	$w(w-1)^{r-1}$	$2r-1$	1	$(\frac{w}{2})^2$	$2r-1$

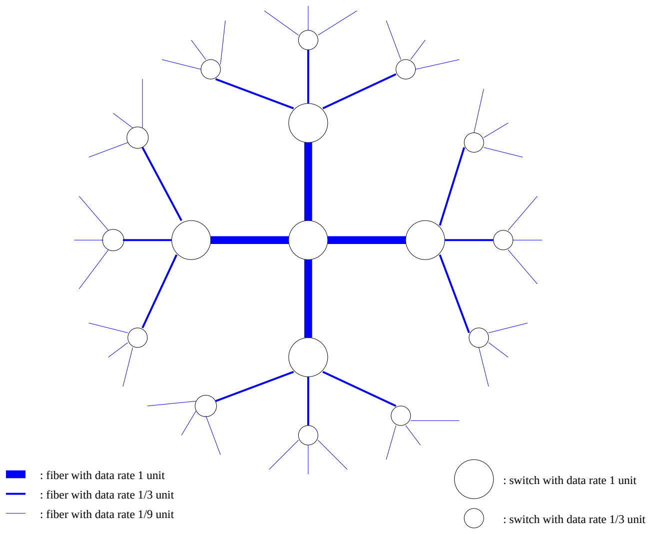

Figure 3: Switches with different bandwidth to achieve bandwidth efficiency

that found in a Hypercube with the same diameter. The RTN has degree of one while Hypercube has degree 2r and GFT has degree p (number of parent edges). This allows for economical connection of processing elements on RTN, but limits the fault tolerance and bisection width of RTN. The *cost*, which is the product of the diameter and degree [6] is lower for RTN than Hypercubes or GFTs. Both the Hypercube and GFT have higher bisection width and leaf nodes than RTN, but the RTN(w,r) is equivalent or better than GFT(r,w,1) in diameter, degree and cost and has a higher bisection width ($(\frac{w}{2})^2$ instead of $\frac{w}{2}$). The RTN(16,4) compares favorably with other topologies. See Table 4.

5 Communications

The routing scheme for RTN follows from the Identical Address Connection Method. Port number 00 in the atomic center is connected to port 00 in one of the perimeter switches. Similarly, additional ports on interior switches are connected to identical port numbers in next level switches. Node addresses are a concatenation of the leaf port address followed by the intermediate switch port numbers between the leaf and the atomic center. Using this connection method and addressing scheme for nodes, routing from one leaf node to another is simple. Figure 4 is an example of an RTN(4,3) network.

A path from leaf node A (which has address $10 \cdot 01 \cdot 11$) to leaf node B (which has address $01 \cdot 11 \cdot 00$) can be obtained by combining the source node's address with the reverse of the destination node's address. The path would be $10 \cdot 01 \cdot 11 \cdot 00 \cdot 11 \cdot 01$.

Table 4: RTN(16,4) compared with other topologies

Comparison of topological properties					
Network	nlp	diameter	Degree	bw	Cost
Hypercube $Q(2 \cdot 4)$	256	7	8	128	56
Fat Trees $GFT(16,4,p)$	65536	7	p	$8 \cdot p^4$	$7 \cdot p$
Round Table $RTN(16,4)$	54000	7	1	64	7

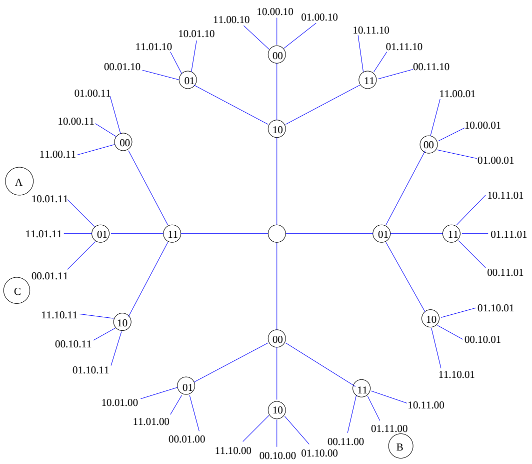

Figure 4: An example of RTN (4, 3)

The length of the path is 5. This path can be used by the signaling protocol to set up a connection through the ATM network.

The path from leaf node A (which has address $10 \cdot 01 \cdot 11$) to leaf node C (which has address $00 \cdot 01 \cdot 11$) can be shortened to $10 \cdot 00$ since the two leaf node have the same RTN(4, 2) *center* address $01 \cdot 11$. The length of the path is 1.

Broadcasting can be performed easily in RTN using the multicast facilities provided by many ATM switch vendors. Since a RTN(w,r) is a spanning tree, a multicast message sent through the atomic center will propagate to all leaf nodes without inducing congestion.

6 Conclusion

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 media for applications on workstation clusters that require high bandwidth, congestion free communication services. The SASIN class of network topologies can be used to connect a large number of workstations with inexpensive ATM switches. The Round Table SASIN maximize switch utilization efficiency and have many other desirable properties. RTN is hierarchical, symmetrical, has a low diameter and is easy to implement.

Compared with Generalized Fat Tree, RTN has the same diameter, average distance and lower cost metric. Due to the high bandwidth and multicasting ability of ATM switches, reasonable bisection bandwidth and broadcasting capabilities can be achieved using RTN. The structure of RTN also provides for simple and extensible routing schemes.

Current research is focused on developing other SASIN with scalable bisection width and increased fault tolerance. SASIN provide a flexible and economical alternative to existing MPP networks and promise to improve the functionality of clustered computing.

References

[1] P. W. Dowd, S. M. Srinidhi, E. Blade, and R. Claus. Issues in ATM Support of High Performance Geographically Distributed Computing. In *First International*

Workshop on High-Speed Network Computing, pages 19–28. IEEE, 1995.

[2] FORE Systems, Inc. *ForeRunner ASX-200 ATM Switch User's Manual.* FORE Systems, Inc, 1994.

[3] Clyde P. Kruskal and Marc Snir. A Unified Theory of Interconnection Network Structure. *Theoretical Computer Science*, 48:75–94, 1986.

[4] Charles E. Leigserson, Zahi S. Abuhamdeh, David C. Douglas, Carl R. Feynman, Mahesh N. Ganmukhi, Jeffrey V. Hill, W. Daniel Hillis, Bradley C.Kuszmaul, Margaret A. St. Pierre, David S. Wells, Monica C. Wong, Shaw-Wen Yang, and Robert Zak. The Network Architecture of the Connection Machine CM-5. In *Proc. Fourth Ann. ACM Symp. Parallel Algorithms and Architectures*, pages 272–285. ACM, 1992.

[5] Ka-Cheong Leung and Mounir Hamdi. Evaluating PVM and Express on Various Network Clusters. In *First International Workshop on High-Speed Network Computing*, pages 57–66. IEEE, 1995.

[6] Sabine R. Ohring, Maximilian Ibel, Sajal K. Das, and Mohan J. Kumar. On Generalized Fat Trees. In *9th International Parallel Processing Symposium*, pages 37–44. IEEE, 1995.

[7] Craig Partridge. *Gigabit Networking.* Addison-Wesley Publishing Company, 1994.

[8] Martin De Prycker. *Asynchronous Transfer Mode.* Ellis Horwood, 1993.

[9] Michael J. Quinn. *Parallel Computing: Theory and Practice.* McGraw-Hill, Inc., 1994.

[10] Gowri Ramanathan, Mark Clement, and Phyllis Crandall. Hyperweave: A Fault-tolerant Expandable Interconnection Network. In *Proceedings of the Fourth IEEE Symposium on Parallel and Distributed Processing*, pages 479–482. IEEE, 1992.

[11] Isaac D. Scherson and Abdou S. Youssef. *Interconnection Networks for High-Performance Parallel Computers.* IEEE Computer Society Press, 1994.

[12] Honbo Zhou and Al Geist. Faster Message Passing in PVM. In *First International Workshop on High-Speed Network Computing*, pages 67–73. IEEE, 1995.