Performance Analysis of Multistage Interconnection Networks using a Multicast Algorithm

Jaehyung Park and Hyunsoo Yoon
Department of Computer Science & Center for Artificial Intelligence Research
Korea Advanced Institute of Science and Technology
373-1, Kusong-Dong, Yusong-Gu, Taejon 305-701, Korea
{hyeoung,hyoon}@camars.kaist.ac.kr

Abstract

In this paper, we study issues of the multicasting in the multistage interconnection networks (MINs) for large-scale multicomputers. In addition to point-to-point communication among processing nodes, efficient collective communication is critical to the performance of multicomputers. This paper presents a new approach to support multicast communication, on the basis of a restricted address encoding scheme which constructs a short fixed-size multicast header, and a recursive scheme that recycles a multicast packet one or more times through the network to send it to the desired destination nodes. We propose a novel deadlock-free multicast algorithm for multiple multicast packets in MIN-based multicomputers. We also present performance model for the unbuffered MIN using the multicast algorithm and analyze its performance in terms of the network throughput, where several multicast communication are considered.

Keywords: Performance Analysis, Throughput, Multicast Algorithm, Multista ge Interconnection Networks, Deadlock Freedom.

1 Introduction

Multistage interconnection networks (MINs) are a popular network architecture for large-scale multicomputers [1], such as TMC CM-5, IBM SP1/SP2, and NEC Cenju-3. In these systems, processing nodes communicate each other through MINs by passing messages. Each processing node consists of its own processor, local memory, and communication devices. As the number of nodes in the system increases, the total communication bandwidth, memory bandwidth, and processing capability of the system scale up as well. Efficient data communication among processing nodes is critical to the performance of message-based multicomputers [4, 9].

In message-based multicomputer systems, data communication can be classified into point-to-point communication and collective communication [3, 7]. While point-to-point communication deals with the basic operation in which a source node delivers a message to only one destination node, collective communication deals with communication that involves a group of processing nodes. A system-level multicast service, in which the same message is delivered from a source node to an arbitrary number of destination nodes, is fundamental in supporting collective communication primitives. In the computation mode where the same program is executed on different processors with different data, multicast is fundamental to several operations such as replication and barrier synchronization [12]. In a distributed shared-memory system, multicast may be also used to efficiently support invalidation and updating for cache coherence protocol of shared-data [6].

Efficient implementation of multicast communication depends on the particular multicomputer architecture, which includes the network topology as well as underlying switching mechanism. Multicast communication has been extensively studied for multicomputers based on MINs [3, 8, 10, 12]. These works [3, 12] include both hardware and software approaches for wormhole routing technique. Most of such hardware multicast algorithms [3] employ complex header encoding schemes which construct variable-sized headers, in order that a source node directly sends a multicast packet to all destination nodes at a time. In such *header encoding schemes*, switching elements as the basic block of a MIN needs additional hardware to process the variable-sized header. Moreover, it is not easy to implement a deadlock-free multicast algorithm for multiple multicast packets due to irregularity of destination patterns [3, 5]. Multicast algorithms implemented by software approach [3, 10, 12] do not employ a header encoding scheme but a *recursive scheme*, where source and destination nodes which receive a multicast packet send it to the destination

0-8186-7901-8/97 $10.00

nodes which have not received it yet. In implementing multicast communication, such algorithms have the advantage that they need no additional hardware. However, they require at least $\log_2 f$ passes across the MIN to complete given a multicast communication, where f is the number of destination nodes. Furthermore, the proposed algorithms in [3, 10, 12] are not analyzed by numerical model where several multicast communications is considered.

This paper describes a simple header encoding scheme, called a *restricted address encoding scheme*, which constructs a small- and fixed-sized multicast header. The restricted address encoding schemes implement multicast communications by using the recursive scheme. In this paper, we propose a novel deadlock-free hardware multicast algorithm for multiple multicast packets in the wrap-around MIN with the restricted address encoding and recursive schemes. The proposed algorithm exploits the intrinsic nonblocking property of the MIN. We also present performance model for the unbuffered MIN using the multicast algorithm and analyze its performance under several multicast communications in terms of network throughput. The proposed algorithm can be easily applied to MINs supporting a restricted multicast, such as NEC Cenju-3.

The structure of this paper is organized as follows. The next section describes system models including the basic architecture and the intrinsic nonblocking property, and the reigion encoding scheme. In Section 3, a deadlock-free multicast algorithm is proposed which are based on the restricted address encoding scheme and the recursive scheme. We analyze the performance of the unbuffered MIN using the proposed algorithm in Section 4. Section 5 concludes the paper.

2 System Model

This section describes the basic architecture, the nonblocking property of the banyan networks, and the region encoding scheme.

2.1 Basic Architecture and Nonblocking Property

The banyan network under consideration has $n = \log_2 N$ stages. Each stage contains $N/2$ (2×2) switching elements (SEs). The stages are labeled in a sequence from $(n-1)$ to 0 with $(n-1)$ for the first stage. The N input/output links at each stage are labeled using n bits $(a_{n-1}a_{n-2}\cdots a_0)$, within each stage starting from the top. Similarly, the SEs at each stage are labeled using $(n-1)$ bits $(a_{n-1}a_{n-2}\cdots a_1)$ starting from the top. When we refer to one of the two input links of a specific SE, then we use the bit a_0 to represent that input link for simplicity. We consider butterfly interconnection patterns between stages, and a perfect shuffle interconnection pattern between processing nodes and stage $(n-1)$. The formal definitions of these connection patterns are as follows:

Definition 1 The *interconnection function* for the output links at stage i in the banyan network under consideration, for $n-1 \geq i \geq 0$, and that for the outputs of the processing nodes are respectively defined by

$$IC_i[(a_{n-1}\cdots a_{i+1}a_i a_{i-1}\cdots a_0)] = (a_{n-1}\cdots a_{i+1}a_0 a_{i-1}\cdots a_i),$$
$$IC_n[(a_{n-1}a_{n-2}\cdots a_1 a_0)] = (a_{n-2}a_{n-3}\cdots a_1 a_0 a_{n-1}),$$

where the right hand side labels are that of the input links of the next stage [3].

Though we consider a special class of banyan networks with the interconnection patterns as described above, the results we obtain hold for all the networks that are topologically equivalent to this network.

A broadcast banyan network is a banyan network with SEs which are capable of packet replications. A packet arriving at each broadcast SE can be either routed to one of the output links, or it can be replicated and sent out on both links. Figure 1 illustrates the basic wrap-around MIN which is constructed from a $\log_2 N$-stage broadcast banyan network. In a wrap-around MIN, output links at the final stage are connected to processing nodes through external links, hence the network can recycle packets.

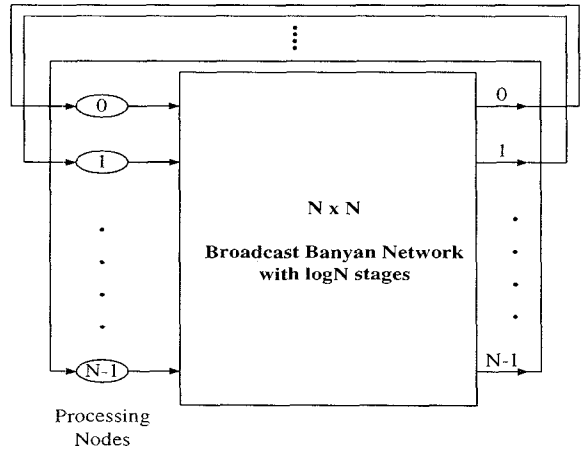


Figure 1. The basic architecture based on the broadcast banyan network

The banyan network itself is a blocking network. However, it is known that if the incoming packets with distinct destination nodes are arranged in an ascending or a descending order and the active sources are all connected to consecutive input links then the banyan network becomes nonblocking [2, 5]. The nonblocking condition is formally described as follows.

Property 1 A broadcast banyan network is *nonblocking* if the active sources $x_1, \ldots, x_k$ and the corresponding destination nodes $y_1, \ldots, y_k$ satisfy the following.
1) Monotone: $y_1 < y_2 < \cdots < y_k$ or $y_k < \cdots < y_2 < y_1$.
2) Concentration: Any source between two active sources is also active. That is, $x_l \leq w \leq x_m$ implies source w is active, where $1 \leq l < m \leq k$.

2.2 Restricted Address Encoding Scheme

The restricted address encoding scheme constructs a multicast routing header from reachable destination nodes which are restricted into a single cube or a single region in the MINs; these are supported in nCUBE-2 and NEC Cenju-3 systems, respectively.

One of restricted address encoding schemes is a region encoding scheme which specifies arbitrary consecutive destination addresses constructing a single region [5]. The multicast routing header for the region encoding scheme indicates the *minimum* and *maximum* addresses of consecutive destination addresses. An SE in the broadcast banyan network has the capability that handles the header with the minimum and maximum addresses. Suppose that an SE at stage i received a packet with the header containing the two addresses: $min_{i+1} = m_{n-1} \cdots m_1 m_0$ and $max_{i+1} = M_{n-1} \cdots M_1 M_0$, where the argument $(i+1)$ denotes an SE at stage $(i+1)$ from where the packet came to stage i. The decision for packet routing and replication is as follows:
• If $m_i = M_i = 0$ or $m_i = M_i = 1$, then send the packet out on link 0 or 1, respectively.
• If $m_i = 0$ and $M_i = 1$, then replicate the packet, modify the headers, and send both packets out on both links.
These rules assume that $m_{i'} = M_{i'}$, $i < i' \leq n-1$ hold for every packet which arrives at stage $i, 0 \leq i \leq n-1$. The modification of a packet header is done according to the following recursion:

- For the packet sent out on port 0,
  $\min_i = \min_{i+1} = m_{n-1} \cdots m_1 m_0$,
  $\max_i = M_{n-1} \cdots M_{i+1} 01 \cdots 1$
- And for the packet sent out on port 1,
  $\min_i = m_{n-1} \cdots m_{i+1} 10 \cdots 0$,
  $\max_i = \max_{i+1} = M_{n-1} \cdots M_1 M_0$

Figure 2 illustrates an example of routing paths through the MIN for a multicast packet with routing tag containing an address interval specified by 0100 and 1000.

3 Recursive Multicast Algorithm in Wrap-Around MINs

In this section, we describes a novel recursive multicast algorithm in wrap-around MINs which is based on the restricted address encoding scheme.

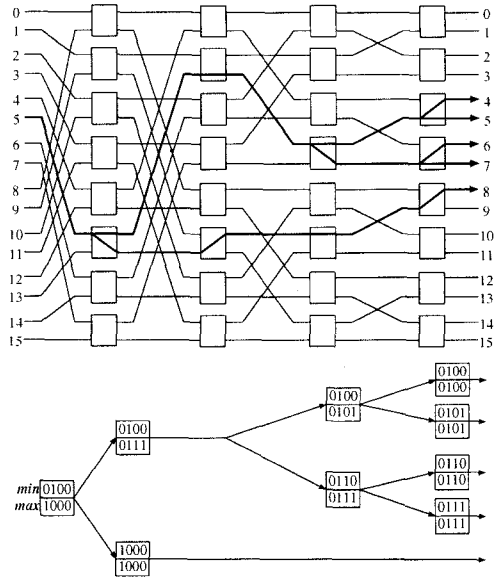


Figure 2. An example of self-routing based on the region encoding scheme

3.1 Two-Phase Multicast Algorithm

The two-phase multicast algorithm with the region encoding is described as follows:

Assume that the f destination nodes for a multicast packet are sorted in the ascending order as $D_0, D_1, \ldots, D_{f-1}$.
Phase 1: Copy from the source node to f consecutive nodes through the MIN. The f consecutive nodes are addressed as $s, s+1, \ldots, s+f-1$ where s is randomly selected. Region encoding scheme is used to copy the packet to f consecutive nodes.
Phase 2: Route the recycled copy from the nodes $(s+l)$ to the destination nodes D_l, where $0 \leq l \leq f-1$. The routing tag of the packet from node $(s+l)$ contains D_l.

In Figure 3, an example of the second phase is shown, where source node 5 sends a multicast packet to destination nodes $\{0, 3, 6, 11, 13\}$. During the copying phase, source node 5 sends a copy to nodes $\{4, 5, 6, 7, 8\}$, where minimum address 4 is selected randomly.

Theorem 1 *In the two-phase multicast algorithm, blocking does not occur among any constituent copies.*

Proof : In the first phase, it is trivial because packets are copied from a single inlet to f consecutive outlets. On the contrary, in the second phase these f copies move from multiple inlets to the corresponding destination nodes through the MIN. However, active nodes which received the copies at the end of the first phase are concentrated on inlets $s, s+1, \ldots, s+f-1$ and the node addresses satisfy the monotonicity property, i.e., for any two destination

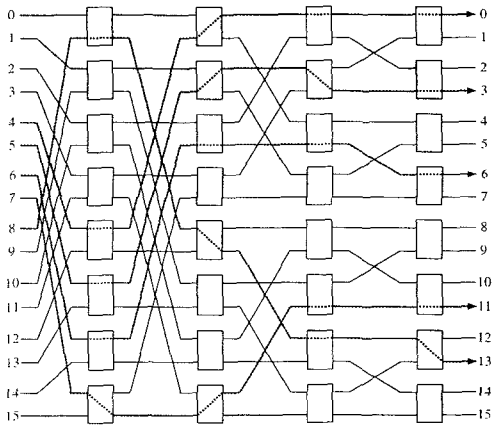


Figure 3. An example of the routing phase in the multicast algorithm

nodes D_l and $D_{l'}$, $D_l < D_{l'}$ where $0 \leq l < l' \leq f-1$. Thus the constituent copies pass without blocking because of Property 1. ■

Theorem 1 guarantees that a multicast packet that is destined for any arbitrary set of destination nodes is sent to the desired destination nodes in two passes across the broadcast banyan network.

3.2 Deadlock Avoidance

These multicast algorithms suffer from potential deadlock due to multiple multicast packets in the unbuffered MIN. Figure 4 shows an example of deadlock situation in a multicast algorithm given two multicast packets $1 \rightarrow \{5,6,7,8\}$ and $15 \rightarrow \{7,8,9,10,11,12\}$. One packet from

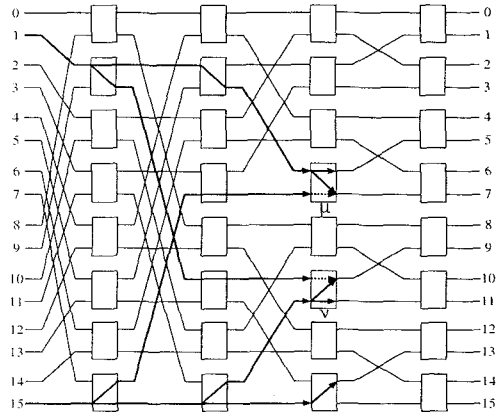


Figure 4. An example of deadlock situation

source node 1 collides with another from source node 15 at SE μ and at SE ν. As shown in Figure 4, at SE μ the packet from source node 1 is randomly selected to be broadcasted to two output links and at SE ν the packet from source node 15 is randomly selected to be broadcasted to the two output links. Subsequently the packet from source node 15 requests the two output links at the SE μ and the packet from source node 1 requests the two output links at the SE ν, under a packet blocking policy of wormhole routing at the SEs. Hence, a deadlock occurs.

In order to avoid deadlock, SEs must use some arbitration mechanism for multiple multicast packets. The *upper input link first*, which prefers the packet incoming from the upper input link, is a distributed scheme.

Theorem 2 *In multicast algorithms, the upper input link first scheme guarantees a deadlock-free transmission through the broadcast banyan network.*

Proof: In the two-phase multicast algorithm, any multicast packet simultaneously requests output link(s) on one or more SEs at the same stage i, $0 \leq i \leq n-1$. Let the source address be $s_{n-1} \cdots s_0$ and the multicast routing header of a multicast packet be $\{m_{n-1} \cdots m_0$ and $M_{n-1} \cdots M_0\}$. The address of the input link at stage $(n-1)$ into which this multicast packet enter is $s_{n-2} \cdots s_0 s_{n-1}$ by Definition 1. Using our notation described in Section 2.1, the label of the particular SE is then $s_{n-2} \cdots s_0$ and the specific input link of this SE is identified by the bit s_{n-1}. Similarly, the address of the input link(s) at stage i, $n-2 \geq i \geq 0$, into which the packet copies enter are $m'_{n-1} \cdots m'_{i+1} s_{i-1} \cdots s_1 s_0 s_i$. Note that $m'_{i'} = M'_{i'}$, for $n-1 \geq i' \geq i+1$, and they are the modified bits of the routing tag of the packets (see Section 2.2). Again using our notation described in Section 2.1, the label(s) of the specific SE(s) are then $m'_{n-1} \cdots m'_{i+1} s_{i-1} \cdots s_1 s_0$ and the specific input link(s) of these SE(s) are identified by the corresponding bit s_i. Thus, the position of the input link(s) at any stage i at which the multicast packet(s) arrive from a particular source node depends only on the source address. Therefore the copies of a multicast packet that request output link(s) at one or more SE(s) at any stage have got the same priority and hence the upper input link first scheme renders deadlock-free multiple multicast communication. ■

4 Performance Analysis of the Wrap-Around MINs

In this section, we model the wrap-around MIN with a multicast capability and analyze the performance of the wrap-around MIN using the proposed recursive multicast algorithm in terms of network throughput, where several multicast communications are considered.

4.1 Assumption and Modelling

Given the rate of packets at processing nodes, network throughput is represented by the number of packets accepted

per time slot. A packet is said to be accepted only if it is successfully transmitted to the requested multiple outputs (nodes). Performance is analyzed on the basis of the following assumptions.

1. Each processing node generates random and independent packets; the packet destination nodes are uniformly distributed over all outputs (nodes).
2. At the beginning of every time slot, each processor generates a new packet with a probability ρ. And $m\rho$ is the average number of multicast packets with fixed fanout f which are generated per time slot by each processing node, where $0 \leq m \leq 1$.
3. The packets which are not accepted are ignored to reenter the switch from the corresponding processing node links later; the packets issued at the next time slot are independent of these blocked packets.

For a multicast packet with fanout f, let c_i be the copy rate, i.e. the probability that the packet is copied at stage i in the wrap-around MIN, where $0 \leq i \leq n-1$. Given the packet rate ρ_i, multicast rate m_i, and copy rate c_i at each of the two inputs of a 2×2 SE at stage i, the expected number of packets both unicast and multicast that it passes to the upper/lower output link in each time slot is given by;

$$\rho_{i-1} = \rho_i(1+m_ic_i) - \frac{1}{4}\rho_i^2(1+m_ic_i)^2 - \frac{1}{2}\rho_i^2 m_i c_i(1-m_ic_i). \quad (1)$$

Because of the symmetry, the same expression holds for lower output link. Note that the expression is equated to ρ_{i-1} since the packet rate from an output link of a SE at stage i is the packet rate at an input link of a SE at stage $(i-1)$.

And the expected number of unicast packets that it passes per time slot is

$$u_{i-1} = \rho_i(1-m_i) - \frac{1}{4}\rho_i^2(1-m_i)(1+m_ic_i). \quad (2)$$

From the relation that $\rho_{i-1} = u_{i-1} + m_{i-1}\rho_{i-1}$, where $0 \leq i \leq n-1$, the probability that a given packet at stage $i-1$ is a multicast packet is given by;

$$m_{i-1} = 1 - u_{i-1}/\rho_{i-1}.$$

For any stage of the wrap-around MIN, the output rate of packets is a function of its input packet rate, multicast rate, and copy rate. Since the output rates of a stage are the input rates of the next stage, one can recursively evaluate the output rates of any stage starting with stage $(n-1)$. We have $\rho_{n-1} = \rho$ and $m_{n-1} = m$. Let ρ' be the expected number of both unicast and multicast packets that a SE at the final stage (i.e., stage 0) passes on to the upper/lower output link and u' be the corresponding number of unicast packets alone. Note that ρ' and u' are obtained from Equations (1) and (2) respectively, for stage 0. Then the throughput per output link of the switch η is given by

$$\eta = \frac{\rho' - u'}{f} + u'.$$

4.2 Numerical Results

To evaluate the throughput for the wrap-around MIN, first the copy rate c_i for a multicast packet with fanout f is computed at stage i, where $0 \leq i \leq n-1$. Where the start address s is randomly selected from N destinations as mentioned in the description of the multicast algorithm, these results are depicted in Figure 5.

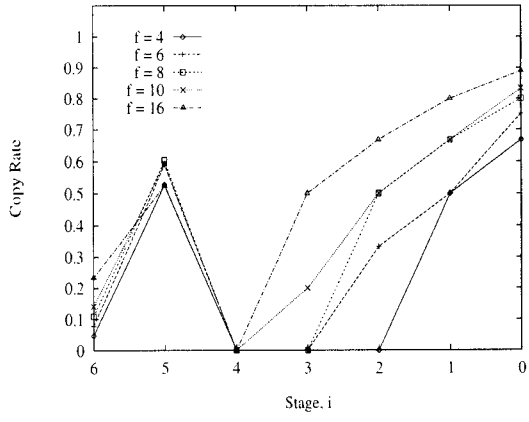
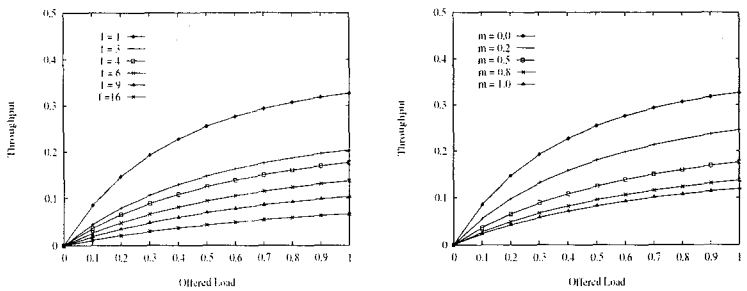


Figure 5. Comparison of copy rates c_i for various fanouts, where n = 7

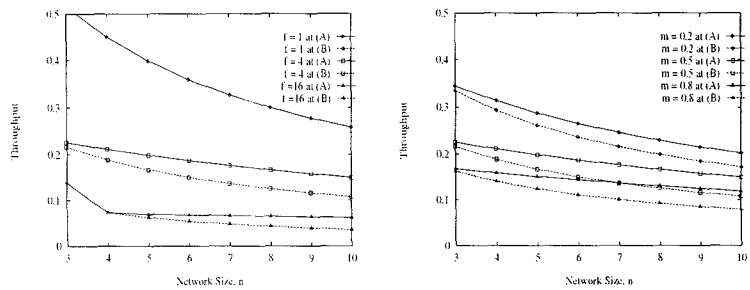
The offered load that is plotted on the x-axis of Figure 6 - 7 is $(1-m)\rho + m\rho f$. Figure 6 shows the throughput η with f or m varying for n equals to 7. The throughput behaviour is identical to that reported in [11] when the fanout f of multicast packets is one.

In Figure 7, the throughput of the MIN using the multicast algorithm is shown varying with network size n. Different curves are shown for various values of multicast rate and fanout. System's offered load is 1.0, that is $(1-m)\rho + m\rho f = 1.0$. In Figure 7, (A) represents the random start address case. In case (B), f copies is generated at the former stages as early as possible. In Figure 7, the throughput of the multicast algorithm using the random start address outperforms that of the algorithm in case (B). This may be attributed to the behaviour of the copy rates in case (A) and (B).



(a) $m = 50\%$ (b) $f = 4$

Figure 6. Throughput η vs. offered load, where $n = 7$



(a) $m = 50\%$ (b) $f = 4$

Figure 7. Network throughput η vs. network size

5 Conclusion

In this paper, we have considered the issues of multicast communication in the MIN for high performance multicomputer systems. To support multicast communication efficiently, the multicast algorithm is proposed on the basis of the restricted address encoding scheme and the recursive scheme which recycles a multicast packet one or more times through the network to send it to the desired destination nodes. The algorithm prevents deadlock among multiple multicast communications in multicomputer systems based on the unbuffered MIN. This paper proposed performance model for the unbuffered MIN using the multicast algorithm and analyzed its performance in terms of the network throughput, where several multicast communications are considered. The proposed algorithm can be also easily applied to wormhole or virtual cut-through MIN-based multicomputers.

References

[1] W. C. Athas and C. L. Seitz, "Multicomputers: Message-Passing Concurrent Computers," *IEEE Computer*, vol. 21, pp. 9–24, Aug. 1988.

[2] K. E. Batcher, "Sorting Networks and Their Applications," in *AFIPS Proc. of the Spring Joint Computer Conference*, pp. 307–314, 1968.

[3] C.-M. Chiang, *Multicasting in Multistage Interconnection Networks*. PhD thesis, Dept. of Comp. Sci., Michigan State Univ., East Lansing, MI, 1995.

[4] K. Hwang, *Advanced Computer Architecture: Parallelism, Scalability, Programmability*. McGraw-Hill Book Co., 1992.

[5] T. T. Lee, "Nonblocking Copy Networks for Multicast Packet Switching," *IEEE Journal on Selected Areas in Communications*, vol. 6, pp. 1455–1467, Dec. 1988.

[6] K. Li and R. Schaefer, "A Hypercube Shared Virtual Memory," in *Proc. of the Int'l Conf. on Parallel Processing*, pp. 125–132, Aug. 1989.

[7] P. K. McKinley, Y. Tsai, and D. F. Robinson, "Collective Communication in Wormhole-Routed Massively Parallel Computers," *IEEE Computer*, vol. 28, pp. 39–50, Dec. 1995.

[8] L. M. Ni, Y. Gui, and S. Moore, "Performance Evaluation of Switch-Based Wormhole Networks," in *Proc. of the Int'l Conf. on Parallel Processing*, pp. 32–40, Aug. 1995.

[9] L. M. Ni and P. K. McKinley, "A Survey of Wormhole Routing Techniques in Direct Networks," *IEEE Computer*, vol. 26, pp. 62–76, Feb. 1993.

[10] J. Park, D. Yoo, H. Yoon, and S. R. Maeng, "Efficient Two-Pass Multicast Algorithms in ATM Switches based on Multistage Networks," in *Proc. of the Int'l Conf. on Systems Engineering*, p. to be appeared, Jul. 1996.

[11] J. H. Patel, "Performance of Processor-Memory Interconnections for Multiprocessors," *IEEE Transactions on Computers*, vol. C-30, pp. 771–780, Oct. 1981.

[12] H. Xu, Y.-D. Gui, and L. M. Ni, "Optimal Software Multicast in Wormhole-Routed Multistage Networks," in *Proc. of Supercomputing*, pp. 1252–1265, Dec. 1994.