Multicast Switch for Internet

Paul Hartke, Dean Liu, Hossam Fahmy, David Lie

1.0 Abstract

This paper describes and evaluates, various proposed methods for multicast support in an interconnection network. This topic has been extensively studied for parallel computers. Our focus is on the applicability of these results to a core Internet Router.

2.0 Problem Definition

There is long standing and widespread interest in using the Internet to replace other forms of broadcast communication methods. The current distribution forms are Television, Cable Television, Radio, and US Postal Mail Delivery (for bulk mailings). For such systems, unicast transmission is very inefficient since multiple copies of the same information are sent to several destinations. The common links among destinations could be utilized to save link bandwidth. The trade-off is more complex routing and flow-control mechanisms. The communication model that sends from one source to many receivers is called "Multicast."

This problem of multicast has already been addressed by the parallel computing community. Parallel computers must support various forms of collective communication services that require an efficient multicasting capability. Our objective is to determine which of these techniques are appropriate for use in an Internet switch.

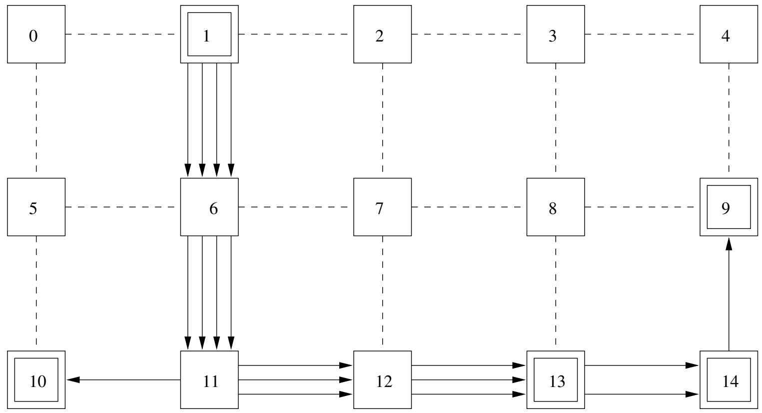
There are several key differences between an Internet Router and a parallel computer interconnection. The first is that a parallel computer is a closed system that is essentially self-throttling. For example, a node will stop transmitting once it has generated a certain number of memory references or it has reached a synchronization point. Also, there is an opportunity to a priori determine traffic flow patterns as a result of both program profiling and compiler placement of data and code throughout the nodes of the machine. In contrast, an Internet Router does not have self-throttling traffic--the presence of congestion will not substantially decrease the arrival rate of traffic. This occurs for several reasons. First, the end-to-end flow control mechanisms currently used in the Internet are slow to react to such congestion. Second, a large fraction of internet traffic is constant bit rate and has no flow control. Third, the highest-capacity Internet Routers are used in the network core where there is significant aggregation of independent flows. These independent flows make it very difficult to characterize the expected workload.

It is important to note that we are not addressing the end-to-end Internet multicasting problem. There are ample proposals in that area. Instead, we focus on a single Internet Router box. As a result, the operation of the end-to-end multicast routing protocols can be logically separated. An incoming packet is received and the packet is processed by the end-to-end multicast routing protocol to determine what port or ports the packet should leave on. After the set of output ports is determined, this set and the packet itself is inserted into the interconnection network. What happens to the packet after it is inserted into the network is the problem we are addressing. Note that since the focus is on a single Internet Router, the interconnection network within the Router will have a regular structure.

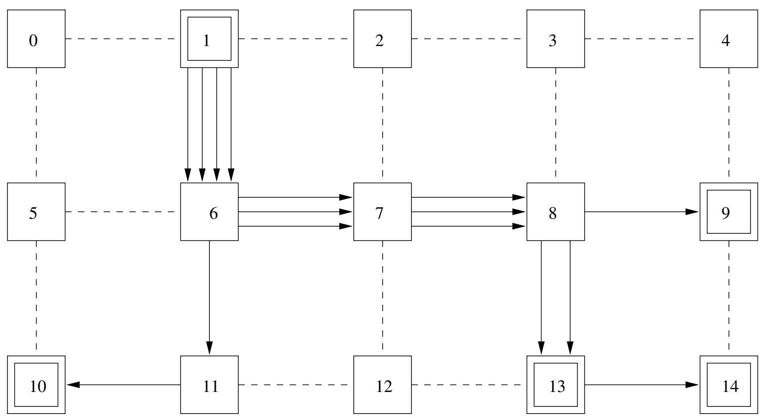
3.0 Internet Router Requirements

We have established that the Internet Router architecture problem is similar to the parallel computer interconnect problem. It is therefore useful to apply the lessons learned in this area to Internet Routers. However, an Internet Router does have different requirements.

There are several key figures of merit for differentiating different multicast methods. The first is the total number of channels required to support a single multicast. The second is the multicast latency measured from when the packet is first injected into the network to when the last multicast destination receives the packet. A third is the computational complexity required to perform the multicast routing algorithm before the packet is injected into the network. Figure 1 shows how different routing algorithms can affect the latency and resource utilization.



Latency Between 1 and 9: 6 Hops
Total Traffic (Many One-to-One): 18
Total Traffic (Multicast): 7



Latency Between 1 and 14: 5 Hops
Total Traffic (Many One-to-One): 16
Total Traffic (Multicast): 8

FIGURE 1. Multicast Routing Figures of Merit

For an Internet Router, the multicast latency is relatively less important since the timing for an Internet packet is much different than for a parallel computer packet. End-to-End delays of tens of milliseconds through the network are acceptable for real-time interactive communication. In contrast, a barrier synchronization stops the participating threads until it is resolved. As a result, this latency is of paramount importance for a parallel computer. Also, since an Internet Router is in a non self-throttling system, the full offered load is presented to the system. The internal link bandwidth is the resource that needs to be optimized for, not latency.

In a parallel computer no packets may be dropped by the interconnection network. The self-throttling aspect of the system makes this tenable. However, in an Internet Router, an output link may be indefinitely oversubscribed. Since there can only be finite buffering in the Router, packets must be dropped in order to prevent queues from growing without bound.

Finally, the expected fanout of a multicast is a key parameter. In a parallel computer, synchronization mechanisms often involve a large fraction of the nodes. Indeed, broadcast messages are not uncommon. However, in a core Internet Router, the expected multicast fanout is much less. Broadcasts would be almost nonexistent.

4.0 Implementation Issues

4.1 Address Encoding

The encoding of the multicast destinations for each packet is important. It is desirable to use an encoding of the smallest size for the expected number of destinations. Several methods are possible. The first, All-Destination Encoding, is to simply list the destinations as in Figure 2a. A second method involves representing the N nodes as an N bit vector and using this vector as the destination address, Figure 2b. A compromise solution, Multiple Region Bit String Encoding, allows many Beginning and Ending addresses. A bit vector is used to indicate which destinations between the Beginning and Ending nodes are used.

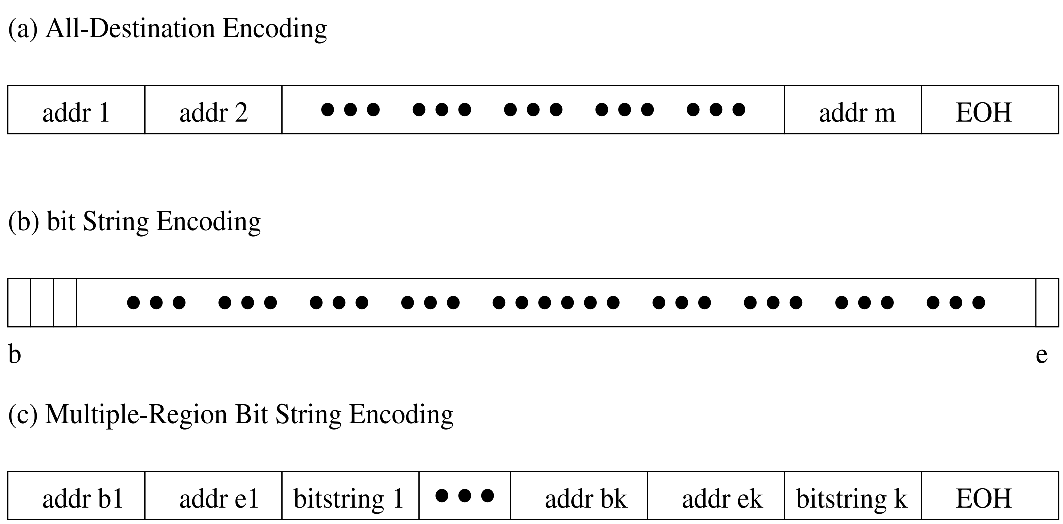


FIGURE 2. Multicast Address Encoding Schemes

The expected fanout of the multicast address largely determines the method used. For a low fanout, the All-Destination Encoding is more appropriate. For large N, on the order of a thousand, a Bit String become prohibitively large. The Multiple-Region Bit String Encoding is attractive for parallel computers where to some extent, multicast messages can be consolidated to adjacent nodes. As discussed below, the actual implementation of the multicast system affects the multicast fanout expectation.

4.2 Multicast Communication Models

The simplest multicast communication model is to convert a multicast message for N destinations into N unicast messages. This scheme allows existing unicast interconnection networks to be used for multicast communication. Any explicit support for multicast naturally increases the implementation complexity which would slow down the overall network. However, it is clear that resources, particularly channel bandwidth, are not efficiently used. This inefficient utilization might not be too detrimental in small diameter networks, but it is severe for large networks.

More sophisticated multicast communication models for interconnection networks can be grouped into two main categories: Path-Based and Tree-Based. Path-Based multicast schemes deliver the packet to each multicast destination in a sequential manner. This path must be computed before the packet is injected into the network. This technique tends to use more channels per multicast than a Tree-Based approach. This is especially true if the multicast destinations are not topologically close together.

In a core Internet Router, multicast destinations will be spread throughout all destinations and not concentrated. For this situation, a Tree-Based scheme is useful. In a Tree-Based scheme, the source is the root of a tree that extends to cover all the multicast destinations. A tree can be used to minimize the number of channels that the multicast packet uses. If the number of channels used is to be minimized, then this problem is analogous to the Steiner Tree problem. Alternatively, the multicast latency can be minimized at the expense of more channels used.

Path-Based designs are needed in cases where packet replication is not possible at intermediate hops. Such replication does complicate the router design at each intermediate hop. However, our objective is to minimize the channel traffic. As a result, Tree-Based designs are particularly suitable for our criteria.

4.3 Deadlock in Multicast Routing

As with unicast routing, deadlock in multicast routing can occur. Deadlock is a more complicated problem for multicast since each multicast requests many more resources than a unicast message. Of course, the multicast needs to request more resources since it is delivering to many destinations compared to the one destination of a unicast message.

Wormhole routing further complicates the problem as shown in Figure 3. Before the B flits can proceed across the output link, both A flits must advance. However, the bottom branch of the multicast is blocked and this prevents the top branch from preceding even though it is not blocked. The top branch cannot proceed since there are only a few flits of buffering at each node. If Virtual Cut Through or Store and Forward switching had been used, then the packet would have been entirely buffered at Node N2 and sent out each branch as it became available. The flow control scheme must keep states for each output that the packet is destined to. As a result, wormhole routing is preferred since it combines both low latency packet transmission and requires only small buffers.

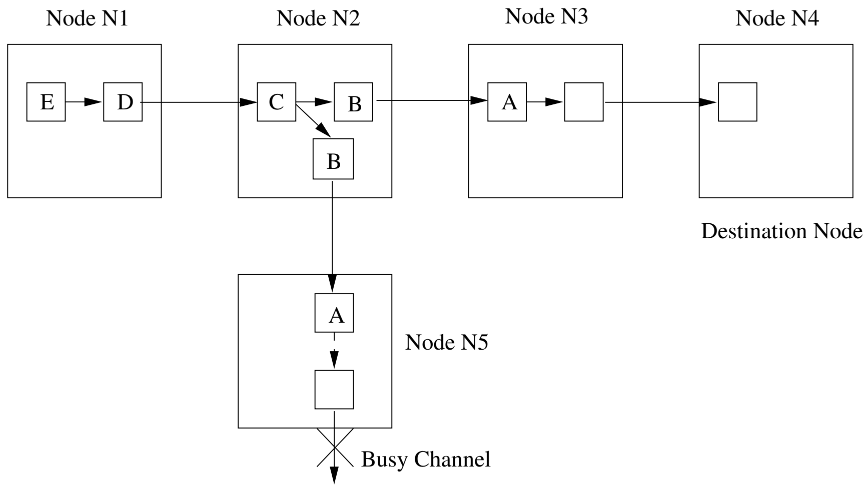


FIGURE 3. Example of blocking for Multicast routing

4.3.1 Double-Channel XY Multicast Wormhole Routing

The general solution to the deadlock problem is to make sure that the routing algorithms have no cyclic dependencies. For torus and mesh based topologies, one way to achieve this is to split the multicast into several independent parts.[1] Subnetworks are created to allow deadlock free multicasts for all nodes in a particular direction for each dimension. Figure 4 shows example subnetworks for a 3x4 mesh. For a multicast originating at a particular source, the destinations are partitioned into groups that correspond to which multicast subnetwork they lie on.

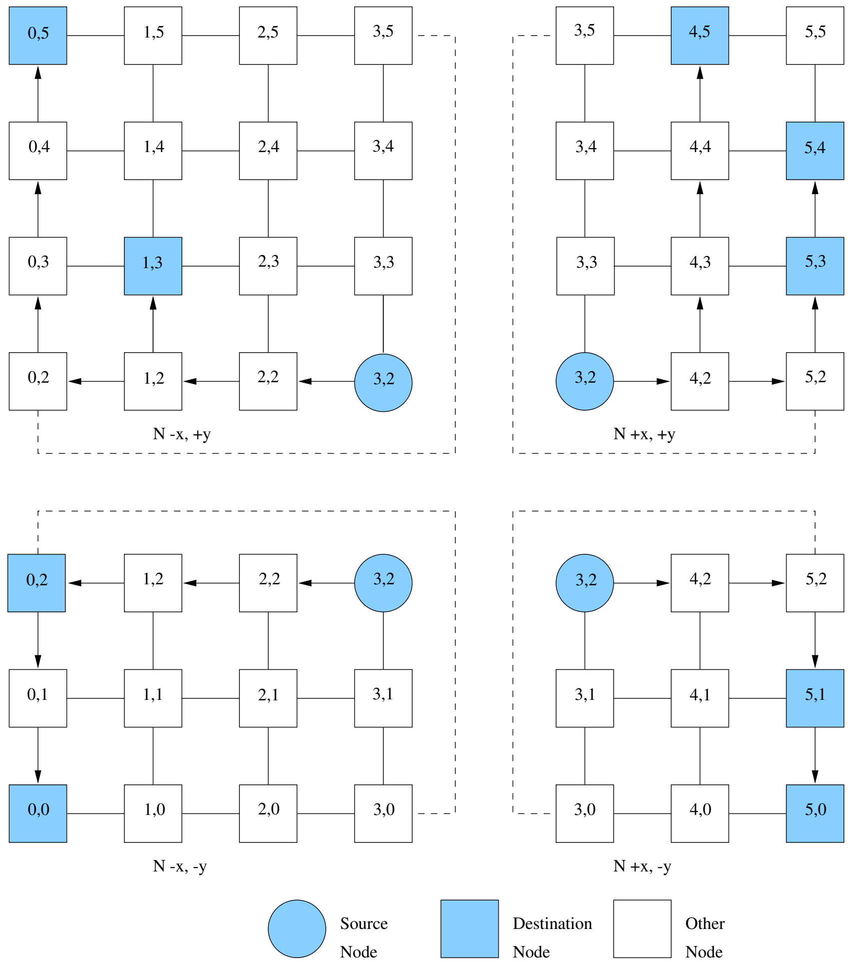


FIGURE 4. Double-Channel XY Multicast Wormhole Routing

This scheme is simple but does have a few drawbacks. The number of multicast subnetworks grows exponentially with the number of dimensions. For a 3-cube, eight multicast subnetworks are required. However, virtual channels are used to create the subnetworks so this increase is manageable. The partitioning also means that the number of channels used will be greater than the minimum. Nodes laying near a subnetwork boundary relative to the source node will each receive separate multicasts--one for each side of the boundary.

4.3.2 Tree-Based Multicast with Pruning

Replication of messages at each node can be done synchronously or asynchronously. With synchronous replication, the message needs to be sent out to all the ports at the same time. In contrast, with asynchronous replication, the message can be sent out at different times. In a tree-based multicast routing, branch pruning can be used with asynchronous message replication to recover from deadlock. As shown in figure 5a, one message from A is blocked at G_2, and the flits are stopped at stages G_2, G_1, and G_0. The branch of message A that is not blocked at stage G_2 is pruned after it goes through the stage. Figure 5b shows that once the pruned branches release the channels they were occupying at stage G_2, the blocked branches are able to advance.

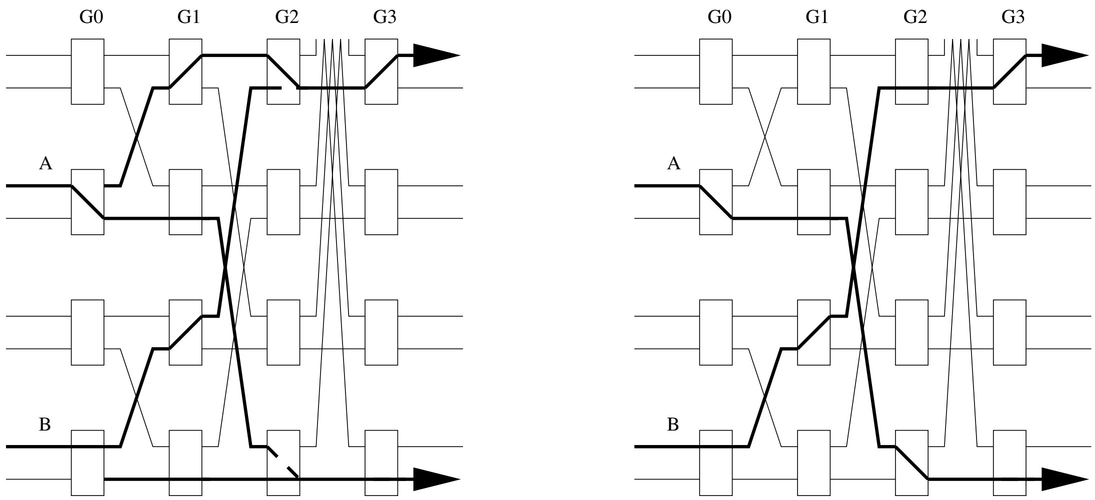


FIGURE 5. Mulitcast Pruning as Demonstrated on a Butterfly Network

The branch pruning technique is applicable only to asynchronous message replication; therefore, it does not work with pure wormhole switching. One alternative is to have a hybrid switching strategy that combines Virtual Cut Through with Wormhole. With this strategy, the router can buffer the packet as in VCT when the channels are blocked. However, there are only a few of these buffers available at each node. This strategy is not a very good technique because it complicates hardware implementation and when the channels are not blocked, the extra buffers needed to support VCT are wasted.

4.3.3 Path-Based Multicast Routing

Path-Based multicast routing is attractive since only one resource (the next hop) is requested at a time. The Tree-Based schemes request many resources and thus have a higher chance of being blocked. As the expected utilization and duty factor of the network as a whole increases, Path-Based approaches become attractive. This occurs because individual branches of the tree have a higher probability of becoming blocked.

The packet should only be delivered to each node once with a minimal overall path length. The requirements of the path are the same as a Hamiltonian path covering a graph. Since an optimal or minimum Hamiltonian path computation is NP-complete, simpler heuristics have been developed.

These heuristics involve labeling the nodes in ascending order. The source node splits the multicast destinations into two sets. Two multicast packets are generated: one with destinations labeled higher than the source and one with destinations lower. Then the packet is sent through the two subnetworks. Figure 6 shows an example of dual-path multicast routing.

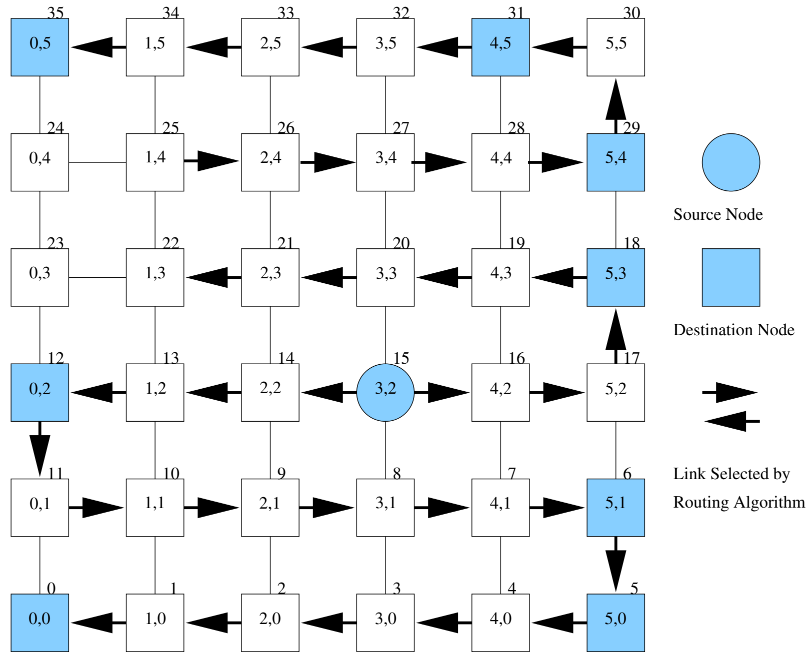


FIGURE 6. Dual-Path Routing Algorithm

Two modifications can be made to this scheme. First, the average length of a multicast is large since the packet must follow the ascending/descending labeling instead of a minimum path connecting the nodes. If each node has a degree greater than two, multi-path multicast routing can be used, whereby more than two paths are generated at the source node. This scheme lowers the average path of a multicast transmission but increases utilization of the channels around a source node. This increase is tolerable if the network is lightly loaded but is detrimental at high loads.

Second, some non-minimal adaptive routing can be allowed. All of the above multicast routing algorithms are deterministic and are thus susceptible to the same problems as unicast deterministic routing algorithms. Thus non-minimal routing can be used as long as nodes are visited in ascending (or descending) order.

Another alternative for adaptive multicast routing can be viewed as an extension of Duato's algorithm for unicast routing.[2] Consider a deadlock free routing algorithm such as Dual-Path. Virtual channels can be used to add other minimal routes to the routing function. As in Duato's unicast Algorithm, these less restrictive routes are used to support adaptivity but are subject to the rules of the underlying deadlock free routing algorithm. Figure 7 shows the extensions to Duato's Algorithm.

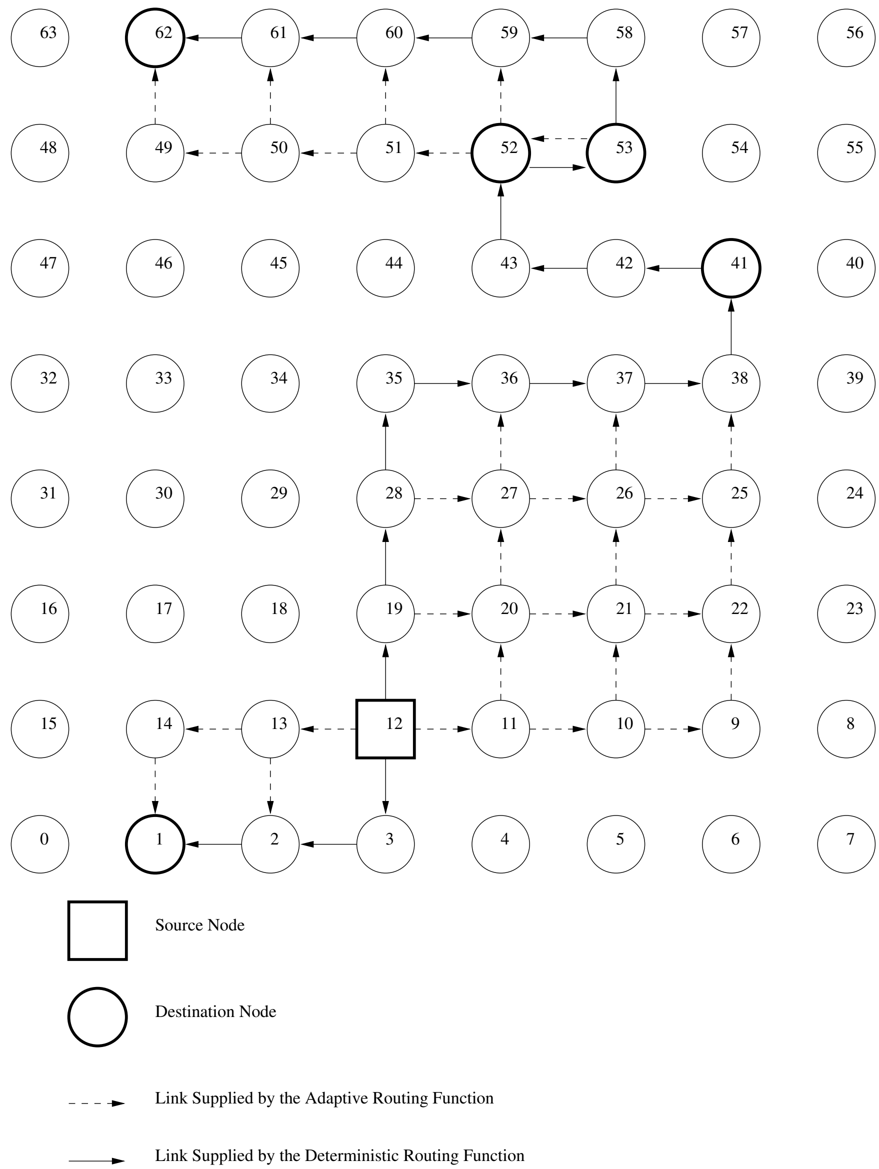


FIGURE 7. Example of Multicast Extensions to Duato's Algorithm

4.3.4 Base Routing Conformed Path

The above Path-Based schemes all treat unicast and multicast traffic differently. This leads to more complicated implementations that affect the overall speed of the system. In many applications, unicast packets are the dominant traffic type. In these cases, the extra overhead of the previous multicast support schemes are not justified. An alternative is to use the existing unicast routing to support multicast Path-Based routing.

Routing algorithms can support fully adaptive routing by providing escape channels that can always be used to drop into a deadlock free routing subnetwork. As long as at each node a valid escape channel is available, fully adaptive routing is supported.[3] In a multicast transmission, it is possible that two sequential destination nodes could not be reached by the escape channels. The destination nodes need to be ordered so that escape channels exist.

There may not exist any ordering that allows the destinations to be reached entirely via escape channels. In this case, the nodes are partitioned into groups that can be reached solely via escape channels. A single node in each group is selected as a leader and the source node unicasts the message to each leader and the leaders relay the packet via multicast to the other members in their respective groups.

5.0 Recommendations

The above multicast routing schemes have been proposed, and in many cases implemented, for parallel computers. However, as discussed above, the requirements for an Internet Router affect the applicability of each scheme. For the encoding scheme, All-Destination encoding is preferred for two reasons. First, the target number of system nodes is large, on the order of a thousand, and a Bit String encoding would be prohibitively wasteful. Second, the expected fanout is small compared to the total number of nodes in the system. This is especially true if each multicast is split into several sub-multicasts which is the cast for the schemes below.

The tree model allows the most utilization of the internal link bandwidth since it optimizes for minimum channel utilization. The Double-Channel XY Multicast Routing Algorithm allows the advantage of a tree structure to be leveraged in a deadlock free manner. The downside is that extra virtual channels are required. However, these extra virtual channels are manageable. The Double-Channel XY is deterministic which is a drawback. Deterministic algorithms is not preferred since it cannot adaptively route around areas of high congestion.

The Multicast extension to Duato’s Algorithm can be used to provide adaptability for a path based model. This adaptability comes at the cost of more channel utilization but the gains of adaptability may be a worthwhile trade-off. The expected multicast loading and the fanout determines the crossover point.

Multicast traffic is only going to grow in the Internet. In addition, the expected applications are high bandwidth requiring efficient delivery of multicast packets. Multicast Interconnection schemes originally developed for parallel computing applications can be used to address this new requirement.

References

[1] X. Lin, P.K. McKinley, and L.M. Ni, "Performance evaluation of multicast wormhole routing in 2D-mesh multicomputers," *Proceedings of the 1991 International Conference on Parallel Processing*, vol. I, pp.294-297, August 1993.

[2] J. Duato, "A theory of deadlock-free adaptive multicast routing in wormhole networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 6, no.9, pp. 976-987, September 1995.

[3] D. K. Panda, S. Singal, and P. Prabhakaran, "Multidestination message passing mechanism conforming to base wormhole routing scheme," *Proceeding of the Workshop on Parallel Computer Routing and Communication*, pp. 131-145, May 1994.

[4] J Duato, S. Yalamanchili, and L. Ni, *Interconnection Networks: an Engineering Approach*, pp. 175-238, IEEE Computer Society Press, Piscataway, NJ., 1997.

Figures in this paper are adapted from the text.