Improved Algorithms for Computing with Faulty SIMD Hypercubes

C. S. Raghavendra
School of Electrical Engineering Computer Science
Washington State University
Pullman, WA 99164

M. A. Sridhar
Department of Computer Science
University of South Carolina
Columbia, SC 29208

May 17, 1997

Abstract

Computation time for various primitive operations, such as braodcasting and global sum, can significantly increase when there are node failures in a hypercube. In this paper we develop nearly optimal algorithms for computing important basic problems on a faulty SIMD hypercube. In an SIMD hypercube, during a communication step, nodes can exchange information with their neighbors only across a specific dimension. Our parallel machine model is an n-dimensional SIMD hypercube, Q_n, with up to $n-1$ node faults. We use free dimension concept to develop our algorithms, where a free dimension is defined to be a dimension i such that both end nodes on any i-dimension link are not faulty. In an n-cube, with $f < n$ faults, it is known that there exists $n-f+1$ free dimensions. Using free dimensions, we show that broadcasting and global sum can be performed in $n+5$ steps, thereby improving upon the previously known algorithms for these primitives. In this paper, the problem of broadcasting in an n-dimensional SIMD hypercube, Q_n, with up to $n-1$ node faults is studied. In an SIMD hypercube, during a communication step,

nodes can exchange information with their neighbors only across a specific dimension. The broadcasting algorithms must work independent of the location of source node and faulty nodes. In a fault-free hypercube, any source node can broadcast a message to all nodes in n steps, by successive communication along any arbitrary ordering of the n dimensions.

1 Introduction

For massively parallel computers, several processor network architectures have been proposed. Among these meshes and hypercubes are used in many experimental and commercial machines including Intel machines [15], Ncube, and the connection machine [11]. Hypercube is a highly regular, symmetric, and recursive structure, and its properties have been studied extensively [9, 20]. Parallel algorithms for many problems including linear algebra [12], image processing [**?**], and communication problems [**?**]. Also, Ho and Johnsson [13] have been developed for hypercube machines.

Fault tolerance in machines such as hypercubes is important in order to achieve sustained high performance computing. Therefore, it is necessary to compute important primitive functions even in the presence of faults. Hypercube network is quite robust [2, 20]; in fact, at least n faults are needed to disconnect Q_n into 2 components. The symmetry and robustness of hypercube can be exploited to compute many functions efficiently even when there are about n faults in the systems. Several researchers have developed algorithms for solving a wide range of different problems in the presence of processor and/or communication link faults [2, 19, 3].

In this paper, we develop improved algorithms for an important global operation, namely, broadcasting from a specific node in the n-dimensional hypercube to all other non-faulty processors, in the presence of up to $n-1$ faulty processors. We are particularly interested in the SIMD mode of operation, in which processors operate in lock-step and every data transfer that occurs at a given time step is made along the same dimension. Further, the broadcasting algorithms presented in this paper are *source independent*, meaning that the sequence of dimensions used for broadcasting is the same regardless of which source is broadcasting. SIMD operation has the merit that node programs implementing the algorithms are independent of the processor ID, and therefore very simple to develop; thus our results have direct practical application. It is important to point out here that the broadcasting problem in general has been addressed by many researchers in the past. .[Chen Shin injured .] .[Chen Shin depth-first .] .[Lee Hayes .] .[Al-Dhelaan Bose Efficient .] .[Peercy Banerjee distributed .] Some of these techniques utilize only local faulty nodes information and/or

MIMD mode of operation. Others use table based approach and require large amount of preprocessing. However, very few papers have addressed fault-tolerant broadcasting in the SIMD mode of operation. A recent paper studies broadcasting with only link faults and present an optimal broadcasting algorithm with up to $n-1$ link faults. .[Park Bose .] This algorithm works in SIMD mode of operation, however, it is not source independent, which means the successive dimensions used depends on the source node address.

We also make two further simplifying assumptions: one, that there are no more than $n-1$ faulty node processors, and no faulty links, in the network; and two, that we know *a priori* the set of faulty nodes in the network. Our primary intent in making these assumptions is to gain insight into the broadcasting problem in this restricted model of parallel computation, and to lay the groundwork for handling more complex cases. The preclusion of link faults is not a severe restriction, and can easily be relaxed. The assumption of *a priori* knowledge of faulty nodes, on the other hand, allows us to improve upon previously known algorithms for this problem, .[Raghavendra Tien algorithms bounds .] .[Ramanathan Shin .] that require $2n$ time steps. Moreover, since the occurrence of processor faults in practice is infrequent, it is reasonable to assume that fault information is broadcasted to all nodes (using, for example, the algorithms of .[Ramanathan Shin .]) so that the more efficient algorithms proposed here can be used for subsequent application execution that may require frequent broadcast operations.

In the context of these assumptions, we describe new algorithms for broadcasting in the presence of $n-1$ faults and $2n-3$ faults. First we show our technique based on free-dimensions to obtain an optimal $n+2$ steps algortihm for SIMD broadcasting in the presence of at most $\lfloor nover2 \rfloor$ faults. We then develop an algorithm that uses $n+5$ steps for SIMD broadcasting with at most $n-1$ faults. We extend this technique to develop an algorithm that takes $n+10$ steps in the presene of at most $2n-3$ faults. Similar algorithms can be derived for other global operations including global sum and global semi-group logical operations.

2 Background and notation

We will use the term *k-cube* to mean a k-dimensional hypercube. We will use the term *k-subcube* to mean a k-dimensional subcube of some larger cube.

Every subset $S \subseteq \{1, 2, \ldots, n\}$ of the dimensions of Q_n induces a collection of $2^{n-|S|}$ $|S|$-subcubes, each obtained by varying the bits of the dimensions in S over all possible values while fixing the rest of the bits at an arbitrary value. But when given a particular node $x \in Q_n$ and a particular subset $S \subseteq \{1, 2, \ldots, n\}$ of dimensions, there is a unique $|S|$-subcube induced by S containing x. We will denote this subcube by $x\langle S \rangle$. For example, given a node $x \in Q_5$, we denote by $x\langle 2, 4, 5 \rangle$ the 3-subcube induced by the dimensions $2, 4$, and 5 containing x.

It is well known that the hypercube Q_n can be viewed as the Cartesian product of the single edge (Q_1) with itself, taken n times. We will occasionally find it convenient to take a viewpoint related to this, i.e., that Q_n can be thought of as the Cartesian product $Q_k \oplus Q_{n-k}$, for any k in the range $1 \leq k \leq n-1$. This essentially means viewing Q_n as being an $(n-k)$-dimensional cube, each of whose nodes is really a "supernode" that contains a k-dimensional cube.

Given a specific set D of dimensions of Q_n, i.e., $D \subseteq \{1, 2, \ldots, n\}$, we denote by Q_n/D the $(n-|D|)$-dimensional cube that uses the dimensions not in D, whose supernodes use the dimensions in D. We will say that the dimensions in D are the *internal* dimensions (used within the supernodes), and those not in D are the *external* ones.

Definition. A dimension j in Q_n is free if there is no pair of faulty nodes across a dimension j link.

Lemma 1 *In Q_n, with number of faulty nodes $f \leq n$, there exists at least $n - f + 1$ free dimensions. Specifically, when $f \leq n - 1$ there exists at least two free dimensions, say, d_1 and d_2.*

Theorem 1 *Given any $f \leq n$ faulty nodes, Q_n can be partitioned into $(n-f+1)$-dimensional subcubes such that each subcube spanned by the same set of dimensions contains at most one faulty node.*

Proof [?]: From Lemma 1, a free dimension, say i, can be found since $f \leq n$. Squeeze Q_n along dim i to Q_{n-1} whose nodes are 1-dimensional subcubes spanned along i. Since Q_n is collapsed along a free dimension i, where no pair of adjacent faulty nodes lies, Q_{n-1} has the same number of faulty supernodes f. Now, if f is less than or equal to the number of dimensions of this cube, we will be able to find another free dimension.

The above *squeezing process* can be repeated $n - f + 1$ times so that Q_n becomes Q_{f-1}, and in this case, Lemma 1 cannot be applied to find another free dimension, since there are f faulty supernodes. Note in the process of collapsing nodes along free dimensions, the number of faulty supernodes always remains to be f. In Q_{f-1}, each supernode is an $(n - f + 1)$-dimensional subcube. Since, in the squeezing process, no two faulty nodes are collapsed into a supernode by the definition of free dimension, each supernode of size 2^{n-f+1} contains no more than one fault. □

Theorem 2 *In a cube Q_n with one faulty node, the broadcasting problem can be solved in $n+1$ steps.*

Throughout the rest of this paper, we will assume that we are given a particular n-cube Q_n and a particular set F of faulty (or otherwise unavailable) nodes in Q_n. We will consider two situations, one in which $|F| \leq n-1$ and the other $|F| \leq 2n-3$. A doubly faulty edge is defined to be an edge with both end nodes faulty.

3 Improved algorithms for broadcasting

In general, it is possible to completely specify a SIMD algorithm for broadcasting by simply describing the sequence of (not necessarily distinct) dimensions along which the algorithm requires its nodes to transmit data during each time step. To illustrate this point, here is a simple broadcasting algorithm in a fault-free cube:

Algorithm A. Execute n steps, such that during the i-th step, every node sends its message along dimension i. This algorithm is described by the sequence $(1, 2, \ldots, n)$.

This algorithm does not work when one or more faulty nodes exist. But a straightforward modification of this algorithm tolerates $f \leq n-1$ faulty nodes and uses $n+f+1$ time steps:

Algorithm B. Execute $2n$ steps, such that the sequence of dimensions along which the non-faulty nodes send messages is $1, 2, \ldots, n, 1, 2, \ldots, n$.

Algorithm B correctly broadcasts messages from an arbitrary source node to every other node in the presence of up to $n-1$ faulty nodes. .[Raghavendra Tien optimal global .]

We will begin with a preliminary result for the case of fewer faults.

Algorithm C.

Input. A cube Q_n, the source node s for the broadcast and a fault set F containing $\leq \lfloor n/2 \rfloor$ faults.

Output. A dimension sequence of length $n+2$ for SIMD broadcasting from s to all non-faulty nodes.

Method.

[Phase 1.] By Theorem 1, there is a set of dimensions $D = \{d_1, d_2, \ldots, d_{n/2-1}\}$ such that every supernode of Q_n/D contains at most one fault. Consider, in particular, the supernode of Q_n/D that contains the source s. Note that this supernode corresponds to an $(n/2+1)$-dimensional subcube of Q_n, along those dimensions not in D. Because of Theorem 2, we can use $n/2+2$ steps to ensure that all non-faulty nodes in this supernode have been reached. Thus the first $nByTwo + 2$ elements of our output sequence will be the dimensions used by these steps.

[Phase 2.] Next, consider the effect of using the dimensions in D for the next $n/2-1$ steps of the algorithm. After these steps, in any supernode of Q_n/D, the only non-faulty nodes that have not been reached are those nodes x that are images of some faulty node y in some other supernode of Q_n/D along one or more of the dimensions in D. This is because, during phase 2, data propagation is arrested exactly along those paths that encounter a faulty node, and the paths only use dimensions in D.

[Phase 3.] Consider any supernode M of Q_n/D, and map all the faulty nodes into M using the dimensions in D. Pretend for the moment that all the images under this mapping are also faulty. Then M contains no more than $n/2$ faults. Since M is $n/2+1$-dimensional, it contains at least two free dimensions; call these $iandi'$. Note that the *same* dimension i is free in *every* supernode. Therefore, a data transmission across i completes the broadcasting operation, since the fact that i is free guarantees that no two neighbors across i can both be faulty.

We thus have an $(n+2)$-step algorithm for broadcasting in a cube with $n/2$ or fewer faults. □

At this point, it is worth pointing out the following.

Theorem A. ny SIMD broadcasting algorithm in a cube Q_n with 2 or more faults requires at least $n+2$ steps.

This theorem shows that Algorithm C is optimal.

Next, consider the case with upto $n-1$ faults.

Algorithm D.

Input. A cube Q_n, the source node s for the broadcast and a fault set F containing $\leq n-1$ faults.

Output. A dimension sequence of length $n+5$ for SIMD broadcasting from s to all non-faulty nodes.

Method. We will construct the dimension sequence as the concatenation of three subsequences S_1, S_2 and S_3.

(1) Determine two free dimensions d_1 and d_2.

(2) Let $S_1 = (d_2, d_1)$ be the first subsequence.

Suppose we decompose Q_n into subcubes Q_{n-1} and Q'_{n-1} along dimension d_1. Then at least one of these subcubes has $n/2 - 1$ or fewer faults; assume wlog that it is Q_{n-1}. Now regardless of whether or not s is in Q_{n-1}, using the two dimensions d_2 and d_1 (in that order) ensures that a non-faulty node t in Q_{n-1} now contains the data from s.

(3) We now have the problem of broadcasting in the $(n-1)$-dimensional cube Q_{n-1} containing no more than $n/2 - 1$ faults. We can now apply Algorithm C to this case, so that using no more than $n + 1$ steps, we can complete the broadcast within Q_{n-1}.

(4) Finally, we complete the entire broadcast process by using dimensions d_1 and d_2 again, in that order. That this completes the broadcast can be seen by the reverse of the argument in item (2) above.

We thus have a broadcasting algorithm that uses $n + 5$ steps.

Finally, consider the case of upto $2n - 3$ faults, with the additional constraint that the cube is still connected (i.e., no node is isolated). To construct this algorithm, we will need the following lemma.

Lemma 2 *Consider a cube Q_n with $2n-3$ or fewer faults, and suppose that it is partitioned into two subcubes Q_{n-1} and Q'_{n-1} along some dimension i. Then there exists a sequence j_1, j_2, i of dimensions such that given any non-faulty node s, either s is in Q_{n-1} or else the walk from s along dimensions j_1, j_2, i leads to a non-faulty node in Q_{n-1}.*

The proof is a not-so-straightforward counting argument. □

We can now construct the desired broadcasting algorithm as follows. We begin by decomposing Q_n into two subcubes Q_{n-1} and Q'_{n-1}. Without loss of generality suppose that Q_{n-1} has $n-2$ or fewer faults. We can now apply the sequence j_1, j_2, i to get to a non-faulty node in Q_{n-1}, and then use Algorithm D to complete broadcasting within Q_{n-1}. Finally, we use the sequence i, j_2, j_1 to complete the broadcast. Thus we have used a total of $n + 10$ steps.

4 Conclusions

The fundamental question we have tried to address in this paper is the following: In an SIMD hypercube with up to f faulty nodes, the faults being known *a priori,* what is the least number of time steps needed to accomplish broadcasting from a given processor node to all others? We have described improved algorithms for this operation with $\lfloor nover2 \rfloor$ faults, $n-1$ faults, and $2n-3$ faults.

References

[1] B. Aiello, T. Leighton, "Coding theory, Hypercube Embedding, and Fault Tolerance," *Proceeding of the 3^{rd} annual ACM symposium on parallel algorithms and architectures*, 1991.

[2] B. Becker, H. Simon,"How Robust is the n-cube ?" *Information and Computation*, pp.162-178, 1988.

[3] J. Bruck, R. Cypher, and D. Soroker, "Tolerating Faults in Hypercubes Using Subcube Partitioning," *IEEE Transactions on Computers*, May 1992, pp. 599-605.

[4] M. Y. Chan, S. J. Lee," Fault-Tolerant Embeddings of Complete Binary Trees and Rings in Hypercubes," *IEEE Transactions on Parallel and Distributed Systems,* March 1993, pp. 277-288.

[5] Y. Chang, "Fault Tolerant Broadcasting in SIMD Hypercubes," *Proc. of the Fifth Symp. on Parallel and Distributed Processing,* December 1993, pp. 348-351.

[6] M. S. Chen, K. G. Shin," Depth-First Search Approach for Fault-Tolerant Routing in Hypercube Multiprocessors," *IEEE Trans. on Parallel and Distributed Systems*, pp.152-159, April 1990.

[7] B. S. Chlebus, K. Diks, A. Pelc," Optimal Broadcasting in Faulty Hypercubes," *Proc. 21^{th} Inter. Symp. on Fault-Tolerant Computing*, pp.266-271, 1991.

[8] S. Dutt, J. P. Hayes, "Designing Fault-Tolerant Systems Using Automorphisms," *Journal of Parallel and Distributed Computing*, 1991, pp. 249-268.

[9] F. Harary, *et al.*, " A Survey of the Theory of Hypercube Graphs," *Comput. Math. Applic.*, pp.277-289, 1988.

[10] J. Hastad, T. Leighton, M. Newman,"Fast Computation Using Faulty Hypercubes," *Proc. 21st ACM Symp. Theory of Computing*, 1989.

[11] W. D. Hillis, *The Connection Machine*, MIT press, New Haven, CT, 1985.

[12] C. T. Ho, S. L. Johnsson, "Communication efficient basic linear algebra computations on hypercube architectures," J. Parallel and Distributed Computing, Vol. 4, 1987, pp. 133–172.

[13] S. L. Johnsson, C. T. Ho, "Optimal broadcasting and personalized communication in hypercubes," IEEE Trans. on Computers, September 1989.

[14] T. C. Lee, J. P. Hayes,"A Fault Tolerant Communication Scheme for Hypercube Computers," *IEEE Transactions on Computers,* October 1992, pp. 1242-1256.

[15] S. F. Nugent," The iPSC/2 Direct-Connect Communications Technology," *Hypercube Conf.*, 1988.

[16] S. Park, B. Bose, "Broadcasting in Hypercubes with Faulty Links" *Proc. Frontiers of Massively Parallel Computers,* October 1992, 286–290.

[17] M. Peercy, P. Banerjee," Optimal Distributed Deadlock-Free Algorithms for Routing and Broadcasting in Arbitrarily Faulty Hypercubes," *Proc.* 20^{th} *Inter. Symp. on Fault-Tolerant Computing*, 1990.

[18] C. S. Raghavendra, M. A. Sridhar, "Broadcasting Algorithms in Faulty SIMD Hypercubes," *Proceedings of the Fourth IEEE Symposium on Parallel and Distributed Processing,* December 1992.

[19] P. Ramanathan, K. G. Shin," Reliable Broadcast in Hypercube Multicomputers," *IEEE Trans. Comput.*, pp.1654-1657, Dec. 1988.

[20] Y. Saad and M. H. Schultz," Topological Properties of Hypercubes," *IEEE Trans. Comput.*, pp.867-872, July 1988.

[21] Sing-Ban Tien, C. S. Raghavendra, "Algorithms and Bounds for Shortest Paths and Diameter in Faulty Hypercubes," *IEEE Transactions on Parallel and Distributed Systems*, June 1993, pp. 713-718.