Optimal Constructions of Fault-Tolerant Multistage Interconnection Networks*

Charles Chenggong Fan Jehoshua Bruck
California Institute of Technology
Mail Stop 136-93
Pasadena, CA 91125
{fan,bruck}@paradise.caltech.edu

*Supported in part by the NSF Young Investigator Award CCR-9457811, by the NSF Graduate Fellowship, by an IBM Partnership Award, by the Sloan Research Fellowship and by DARPA and BMDO through an agreement with NASA/OSAT.

Abstract

In this paper we discover the family of Fault-Tolerant Multistage Interconnection Networks (MINs) that tolerates switch faults with a minimal number of redundant switching stages. While previously known constructions handled switch faults by eliminating complete stages, our approach is to bypass faulty switches by utilizing redundant paths. As a result, we are able to construct the first known fault-tolerant MINs that are optimal in the number of redundant stages. Our fault model assumes that a faulty switch can be bypassed and our goal is to guarantee arbitrary point to point and broadcast connectivity. Under this model, we show that to tolerate f switch faults the MIN must have at least f redundant stages. We then present the explicit construction of a MIN that meets this lower-bound. This construction repeatedly uses the singleton basis of the n-dimensional vector space as the mask vectors of the MIN. We generalize this construction and prove that an n-dimensional MIN is optimally fault-tolerant if and only if the mask vectors of every n consecutive stages span the n-dimensional vector space.

1 Introduction

Multistage Interconnection Networks (MINs) enjoyed important applications in fields such as telecommunications and parallel computing in the past decades [1] [3] [9] [11] [12]. They are widely used to construct interconnects in parallel computers as well as various network switches including ATM switches. One of the advantages of MINs is their ability to allow novel ways to tolerate component faults. In this paper, we focus our interest on the fault-tolerance capabilities of Multistage Interconnection Networks over switch faults, and propose a family of constructions that tolerates switch faults with a minimal number of redundant switching stages.

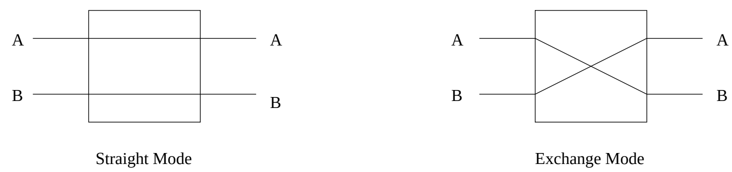


Figure 1: The point to point modes of operation of a 2×2 switch

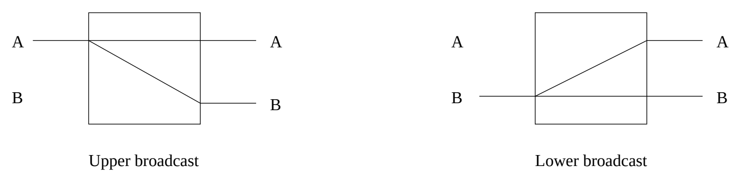


Figure 2: The broadcast modes of operation of a 2×2 switch

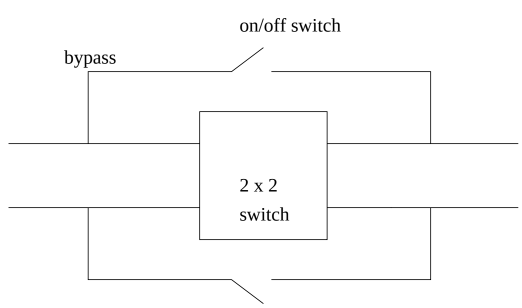


Figure 3: Implementation of the switch fault model for a 2×2 switch

A 2×2 switch is common building block of MINs. For point to point connection, the switch operates in either the *straight* mode or the *exchange* mode, as illustrated by Figure 1. Two additional broadcast modes of operations exist to enable one node to send a message simultaneously to all other nodes (Figure 2). We assume that when a switch is at fault, it is stuck in the *straight* mode. This fault model is accepted by other researchers and can be easily implemented by bypass wires and ON/OFF switches as illustrated in Figure 3. Two fault-tolerance criterions are considered in this paper, namely, the point to point connectivity between any two nodes and the broadcast connectivity. We will prove the results for the point to point model in the early sections, and extend the results to the broadcast model in Section 4.

A Multistage Interconnection Network is shown in Figure 4. This MIN allows point to point connection between any pair of nodes. There are 3 stages of 2×2 switches that interconnect the 8 nodes. Each node is labelled by a binary vector. The length of this vector, n, is the dimension of the MIN. Clearly, $n = log_2 N$, where N is the number of nodes in the MIN. Each switch is also characterized by an n-bit vector, called "mask". The mask indicates the difference between the two input nodes, $B - A$. This difference is obtained by mod-2 vector subtraction. All switches in the same stage have the same masks, therefore we can associate the entire stage with a single stage mask, shown above each stage in the figure.

In this example, The MIN uses the singleton mask set: $\{m_1 = 001, m_2 = 010, m_3 = 100\}$. This mask set forms a basis of the 3-dimensional space, therefore all vectors in this space can be represented as a linear combination of the masks. In other words, this mask set spans the 3-dimensional vector space. Consequently a path exists between any pair of nodes. To route a connection between node A and node B, we decompose the difference between A and B into a linear combination of the masks.

$$B - A = \sum_{i=1}^{n} c_i m_i \tag{1}$$

The switches in stage i go *straight* if $c_i = 0$ and *exchange* if $c_i \neq 0$.

Shown in the bottom half of Figure 4 is the Bar Diagram [7] representation of the same

MIN. Each node in the MIN is represented by a horizontal bar in the Bar Diagram and each switch is represented by a vertical bar. A broken vertical bar in the diagram indicates a faulty switch in the MIN. Connectivity exists between two nodes if and only if a path can be found between these two nodes. Such a path must use at most one switch at each stage and must not change direction inside the MIN, as shown in the figure. Tolerating f switch faults in the MIN is equivalent to tolerating f broken vertical bars in the Bar Diagram.

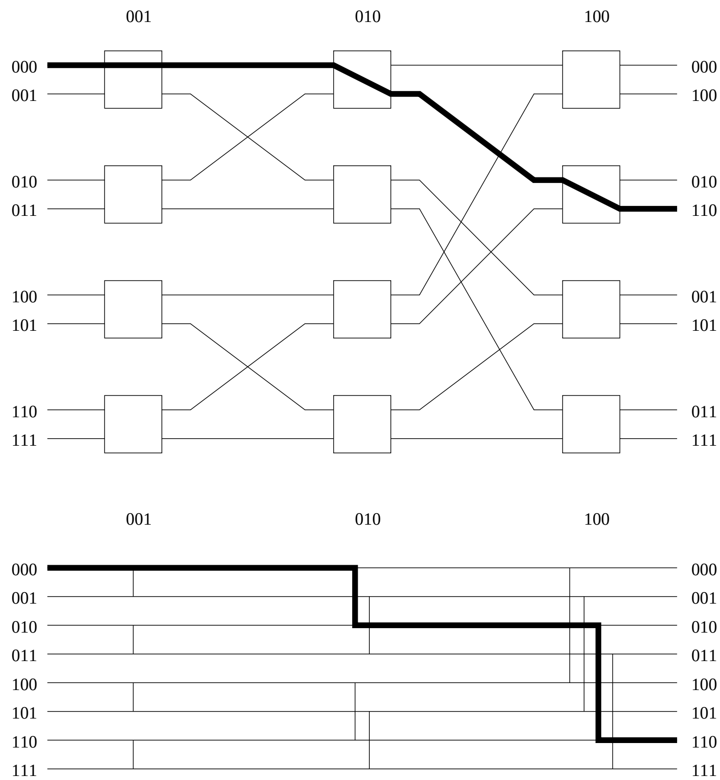
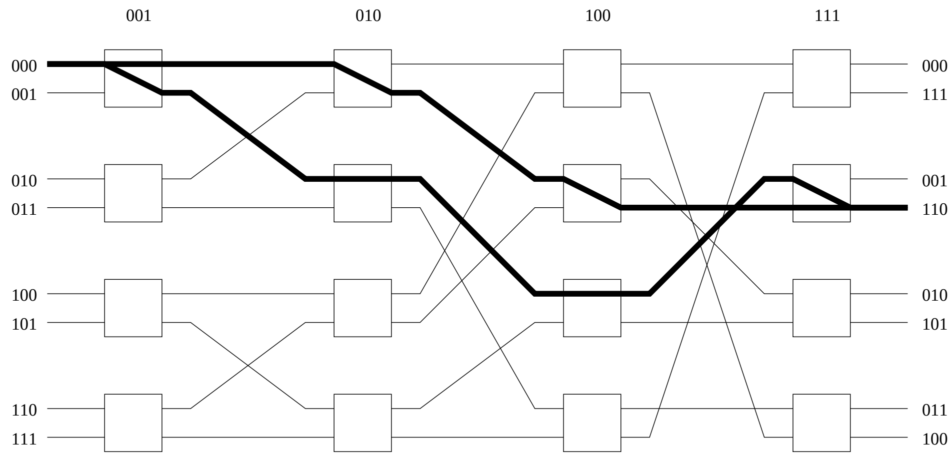


Figure 4: a 3-dimensional Multistage Interconnection Networks using 2×2 switches

To tolerate broken vertical bars in the Bar Diagram, we need to find disjoint paths between any pair of nodes. Two paths are disjoint in a Bar Diagram if they share no vertical bars. To tolerate f broken vertical bars, it is sufficient and necessary to find $f+1$ mutually disjoint paths between all pairs of nodes. It is sufficient because f broken vertical bars can at most break f disjoint paths, and there is at least one paths left between all pair of nodes. It is necessary because if only f disjoint paths can be found between some pair, f broken vertical bars can break all of them, and destroy the connectivity between that pair.

In the MIN shown in Figure 4, one and only one path can be found between all pair of nodes. Therefore it can not tolerate switch faults. To make this MIN single-fault-tolerant, redundant stages need to be added. This problem of tolerating a single switch fault with extra stages has been investigated in the past[2][4]. One solution is to add a stage with an all-1 mask, also known as a “wildcard” stage. The fault-tolerant MIN with the “wildcard” configuration is shown in Figure 5.



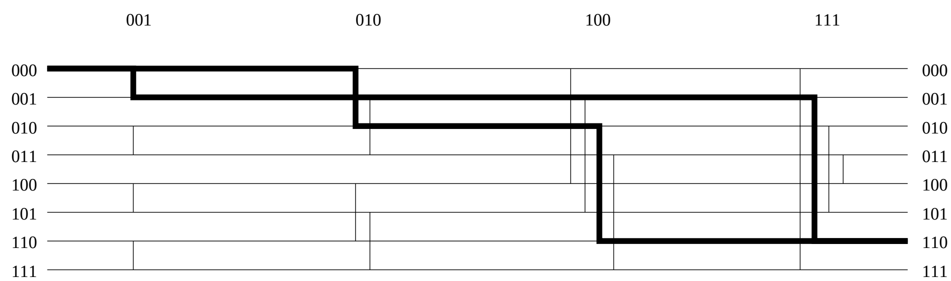


Figure 5: 3-dimensional 1-extra-stage “wildcard” Multistage Interconnection Network

The “wildcard” MIN tolerates one switch fault because we can simply discard the entire stage where the switch fault occurs, and the masks of the remaining three stages still span the space. Therefore the difference between any pair of nodes can still be decomposed into a linear combination of the remaining three masks and a correct routing is therefore available. Two disjoint paths between 000 and 110 are outlined in the figure. This scheme in essense tolerates a stage fault, i.e., it tolerates any number of switch faults if all of them occur in the same stage. The "wildcard" solution does not however tolerate two switch faults since they can occur in different stages.

The "wildcard" construction is not a unique solution to the single-fault-tolerant problem. There exist other solutions that tolerate a single switch fault, which do not necessarily tolerate a single stage fault. We present one of these solutions in Figure 6. This is also a 1-extra-stage construction and the extra stage is masked 001. This MIN does not tolerate a stage fault, since erasing stage 010 or stage 100, the masks of the three remaining stages do not span the space. But this MIN can indeed tolerate a single switch fault. The two disjoint paths between 000 and 110 are outlined in the figure as an example.

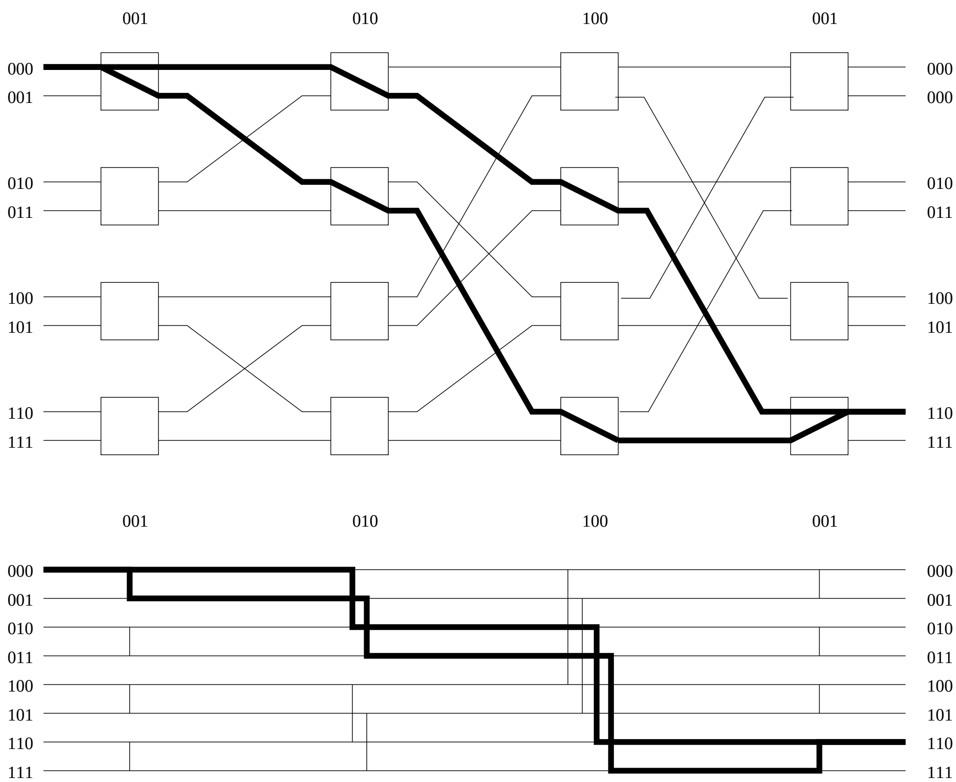


Figure 6: 3-dimensional 1-extra-stage Cyclic Multistage Interconnection Network

The problem of tolerating stage faults have been investigated in previous research works [4] [13]. Bruck and Ho correlated the problem of fault-tolerant MIN to the results in the field of Error-Correcting Codes, and proved that a MIN constructed according to a (n, k, d) code can tolerate d-1 switch faults, as well as stage faults[4]. It showed that a fault-tolerant MIN constructed according to a MDS code uses optimum number of extra stages to tolerate f stage faults. Despite extensive research in the field[5] [6] [8] [10] [11] [14], optimal constructions to tolerate an arbitrary f switch faults, however, have not been proposed.

The two examples we showed led us to consider the following questions: Are the existing solutions the best we can do in tolerating switch faults? If not, what is? Furthermore, if we are able to find optimal constructions, are those constructions the only solutions? The answers to all of these questions are the main contributions of this paper.

In Section 2, we propose a construction of fault-tolerant Multistage Interconnection Networks that uses an optimal number of extra stages to tolerate f switch faults. In that section, we first prove that to tolerate f switch faults, at least f extra stages must be added. None of the previously proposed constructions, however, meets this lower bound. We then propose a new construction that meets this lower bound. A reconfiguration algorithm is also given in that section. In Section 3, we generalize the construction proposed in Section 2 and prove a necessary and sufficient condition for MINs to tolerate an arbitrary number of switch faults with an optimal number of extra stages. While we focus on the MINs that use 2×2 switches under the point to point connection model in Section 2 and Section 3, we extend the results to the Mulstistage Interconnection Networks that use $t \times t$ switches and MINs under the broadcast model in Section 4. In Section 5 we conclude.

2 An Optimal Construction

We first present the following theorem which states the lower bound on the number of extra stages required to tolerate f switch faults for MINs with $t \times t$ switches.

Theorem 1 *To tolerate f switch faults in an n-dimensional Multistage Interconnection Network with $t \times t$ switches, at least f extra stages must be added.*

Proof: (by contradiction) Suppose only $f-1$ extra stages were added. If all the switches in the first f stages that are connected to node 0 failed, only $n-1$ stages can be used to connect node 0 to the other $t^n - 1$ nodes. But it is not possible, since with $n-1$ stages, at most t^{n-1} nodes can be reached. $\Box$

None of the previously proposed MINs meets this lower bound for arbitrary f. For example, the "wildcard" solution only works for $f = 1$ [2]; The number of switch faults that the Error-Correcting Code solutions tolerate is in general less than the redundant stages required [4].

Now we present a new construction of MINs with 2×2 switches that meets this lower bound.

Definition 1 *(Cyclic Multistage Interconnection Networks)*
A (n, k) Cyclic Multistage Interconnection Network is an n-dimensional k-extra-stage MIN which has the singleton basis of the n-dimensional space as the masks of its first n stages and $m_{i+n} = m_i$ for $1 \leq i \leq k$.

A (3, 4) cyclic Multistage Interconnection Network is illustrated in Figure 7. The following theorem implies that this MIN tolerates 4 faults, therefore meets the lower bound stated above. The five mutually disjoint paths can be found between any pair of nodes. In the figure, the paths between node 000 and node 110 are outlined.

Theorem 2 *A (n, f) Cyclic Multistage Interconnection Network with 2×2 switches tolerates f switch faults.*

Proof: We will prove the theorem by explicitly showing that between any two nodes, A and B, there are $f+1$ mutually disjoint paths in the Bar Diagram. Please note that in this proof, the nodes A and B and the masks $\{m_1, m_2, ..., m_{n+f}\}$ are n-bit binary vectors and all arithmatic operations between them are bitwise mod-2. We construct the $f+1$ paths as follows:

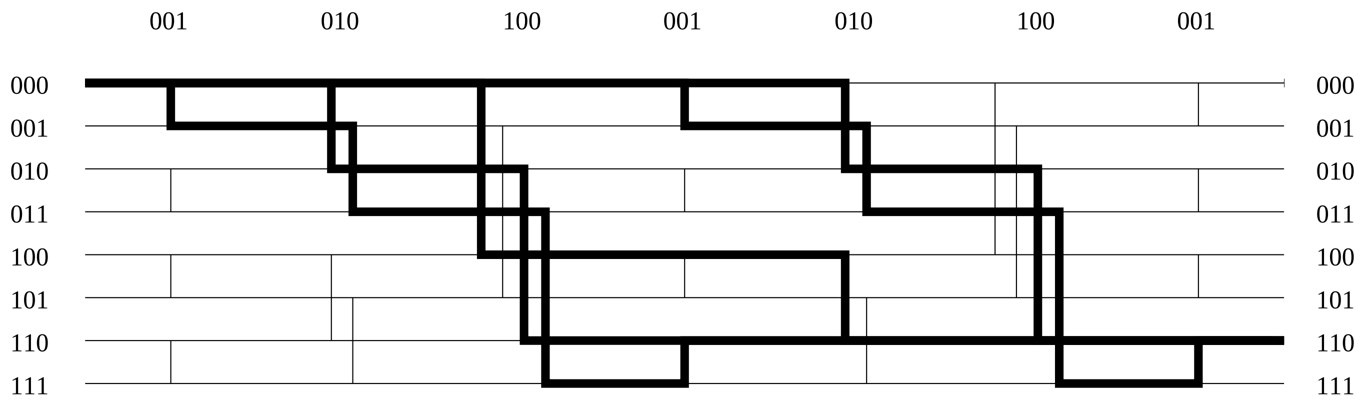


Figure 7: 3-dimensional 4-extra-stage Cyclic Multistage Interconnection Network

For path i, $i \leq f$, the switches at stage i always *exchange*, while stages $i+1$ through $i+n$ are used to perform a regular routing. Since every n consecutive masks span the space, a correct routing from A to B is always achievable. The switches in all other stages go *straight*. The path $f+1$ is constructed by going straight in stages 1 through f and regular routing using stages $f+1$ through $n+f$. Figure 8 illustrates this construction.

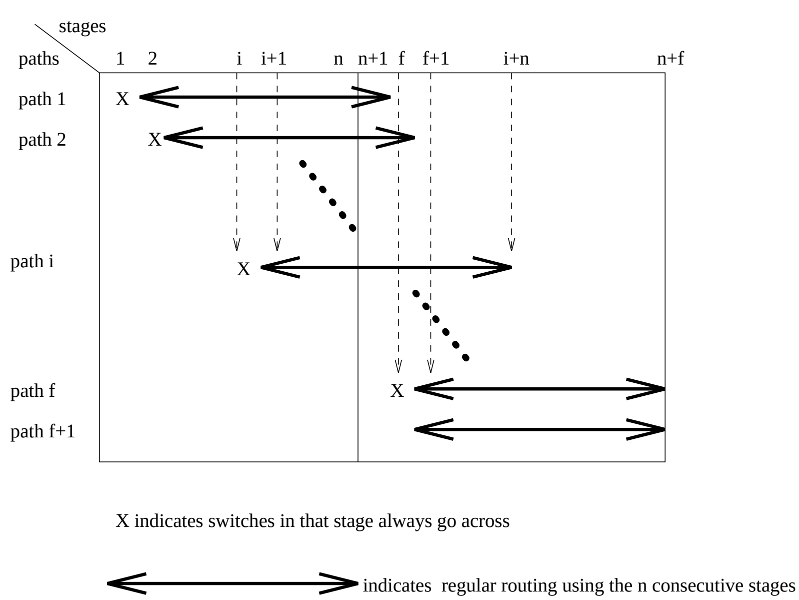


Figure 8: Construction of $f+1$ disjoint paths

We will show that these paths are mutually disjoint from each other by proving that path i, $i \leq f+1$, is disjoint from path j, $j < i$. We consider three cases:

For $j < i - n$, there are no common stages in which the two paths use the switches. Therefore path i and path j are disjoint.

For $j = i - n$, the two paths share stage i. By the construction, path j *exchanges* at stage j while path i goes *straight*. The nodes on the two path are different at bit (j mod n) until stage i. The only case for the two paths to use the same switch is when path j *exchanges* from $A - m_i$ to A while path i *exchanges* from A to $A - m_i$. But it is not possible since path j must reach B after ith stage and $A \neq B$. Therefore path i and path j are disjoint.

For $i - n < j < i$, the two paths are disjoint till stage $j + n$ since path j *exchanges* at stage j while path i goes *straight* and they differ at bit (j mod n). After stage $j + f$ they must agree on bit (j mod n), since they must reach the same destination B. Therefore only one of the two paths will use the switch at stage $j + f$. Consequently path i and path j are disjoint.

Hence the $f + 1$ paths from A to B are mutually disjoint and a (n, f) Cyclic Multistage Interconnection Networks tolerates f switch faults. □

Theorem 2 shows that the performance of Cyclic Multistage Interconnection Networks meets the lower bound stated in Theorem 1. In other words, this construction is optimal in the number of extra stages used to tolerate an arbitrary number of switch faults.

The proof for Theorem 2 explicitly gives the construction of $f + 1$ disjoint paths between any two nodes. This provides a straight-forward way to perform the reconfiguration for Cyclic Multistage Interconnection Networks. When a fault occurs, a node only needs to compare the faulty switch with all the switches in the current routing paths. If a match is found, that path is discarded and the next path according to the construction in the proof will be adopted. Since a routing path uses at most $n + 1$ switches and there are $N - 1$ destinations to reach from a node, the reconfiguration complexity for a node is $O(NlogN)$.

3 A Necessary and Sufficient Condition for Optimal Fault-Tolerance

In Section 2 we introduced the Cyclic Multistage Interconnection Network that demonstrates optimal performance in tolerating any number of switch faults. The construction, however, is not unique. In this section we extend the results to a more general class of fault-tolerant Multistage Interconnection Networks, named Generalized Cyclic Multistage Interconnection Networks.

Definition 2 *(The Generalized Cyclic Multistage Interconnection Network)*
An (n, k) Generalized Cyclic Multistage Interconnection Network is an n-dimensional k-extra-stage Multistage Interconnection Network which has the property that the masks of every n consecutive stages span the n-dimensional vector space.

Figure 9 illustrates a (3, 4) generalized cyclic MIN using a non-singleton and non-repetitive mask set. The 5 disjoint paths between node 000 and node 110 are shown in the illustration.

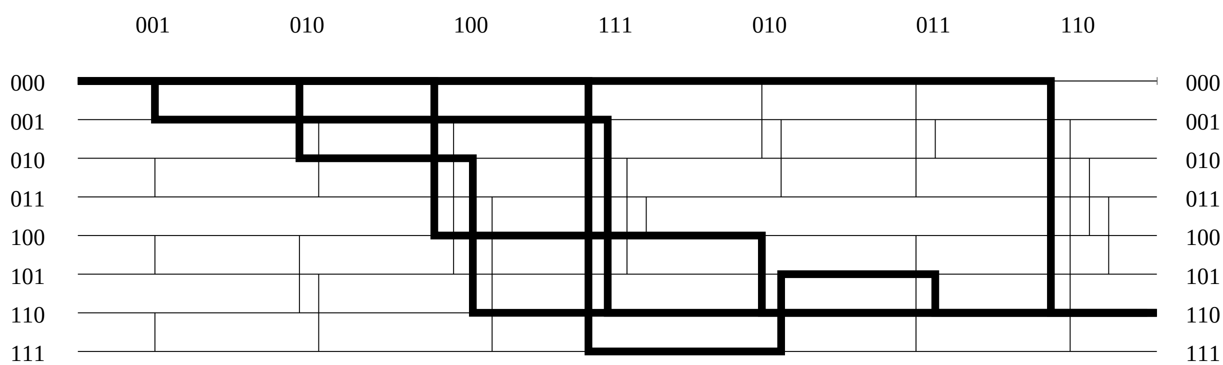


Figure 9: 3-dimensional 4-extra-stage Generalized Cyclic Multistage Interconnection Network

Clearly, the Cyclic MINs is a subclass of the Generalized Cyclic MINs. We will prove that the Generalized Cyclic MINs have the same fault-tolerance capabilities as the Cyclic MINs, namely, they tolerate f faults with f extra stages. In addition, it is the necessary condition for a Multistage Interconnection Networks to demonstrate the optimal fault-tolerance capability.

Theorem 3 *An n-dimensional f-extra-stage Multistage Interconnection Network with 2×2 switches tolerates f switch faults if and only if the masks of every n consecutive stages span the n-dimensional vector space.*

Proof: We prove the forward direction of the theorem by contradiction. Suppose an n-dimensional f-extra-stage Multistage Interconnection Network does not have the property that the masks of any n consecutive stages span the space. There exists n consecutive stages in the MIN whose masks do not span the n dimensional space. When the faults happen in the switches of the remaining f stages at both sides of these n non-spanning stages, and all the faults before the n stages happen at switches connected to node A, and all the faults after the n stages happen at switches connected to node B. The MIN must perform all of the routing in the n nonspanning stages. There exists a B such that a path does not exist between A and B, since the masks of the n stages do not span the n-dimensional vector space while B can take $2^n - 1$ possible values.

The proof of the backward direction is similar to the proof of Theorem 2. The construction of the $f+1$ paths from node A to node B are the same. We need to show that these paths are all mutually disjoint from each other. Again we prove that path i, $i \leq f+1$, is disjoint from path j, $j < i$ by considering three cases:

For $j < i-n$, there are no common stages that the two paths use the switches. Therefore path i and path j are disjoint.

For $j = i - n$, the two paths share stage i. We know that path j *exchanges* at stage j, while path i goes *straight*. Since any n consecutive masks are linearly independent, m_j can not be represented by a linear combination of m_{j+1} through m_{i-1}. Therefore the two paths are disjoint till stage i, and the only way that the two paths are not disjoint is that at stage i, path j *exchanges* from $A - m_i$ to A while path i *exchanges* from A to $A - m_i$. But it is not possible since path j must reach B after ith stage and $A \neq B$. Therefore path i and path j are disjoint.

For $i - n < j < i$, since path j *exchanges* at stage j while path i goes *straight*, the two paths are disjoint till stage $j+n$ with the same reasoning as the previous case. At stage $j+n$, only one of the two paths *exchanges* since they must reach the same destination. Therefore path i and path j are disjoint. $\square$

4 Extensions

In this section, we will make two extensions to the results presented in the previous sections. First, instead of looking at Multistage Interconnection Networks with 2×2 Switching Elements, we will show that the theorems presented in the previous sections also apply to the MINs consisting of $t \times t$ Switching Elements. Following that, we will show that the results are also valid if we are to guarantee the broadcast capabilities of the network.

Let us look at a 9-node 3-extra stage $(2, 3)$ generalized cyclic Multistage Interconnection Network consisting of 3×3 switches. Figure 10 shows the 4 mutually disjoint paths from node 00 to node 20.

Theorem 4 *An n-dimensional f-extra-stage Multistage Interconnection Network with $t \times t$ switches tolerates f switch faults if and only if the masks of every n consecutive stages span the n-dimensional vector space.*

Proof: The proof of the forward direction is the same as the proof for the 2×2 case. To prove the backward direction, we similarly construct $f + 1$ paths from node A to node B and prove that they are disjoint. The difference lies in the construction of the first f paths. The reason for the modification is that a 2×2 switch can only go *straight* or *exchange*, while a $t \times t$ switch has t ways of switching. We say a switch is in mode s if for that switch:

$$output = input + s \times mask \qquad 0 \leq s \leq t - 1 \tag{2}$$

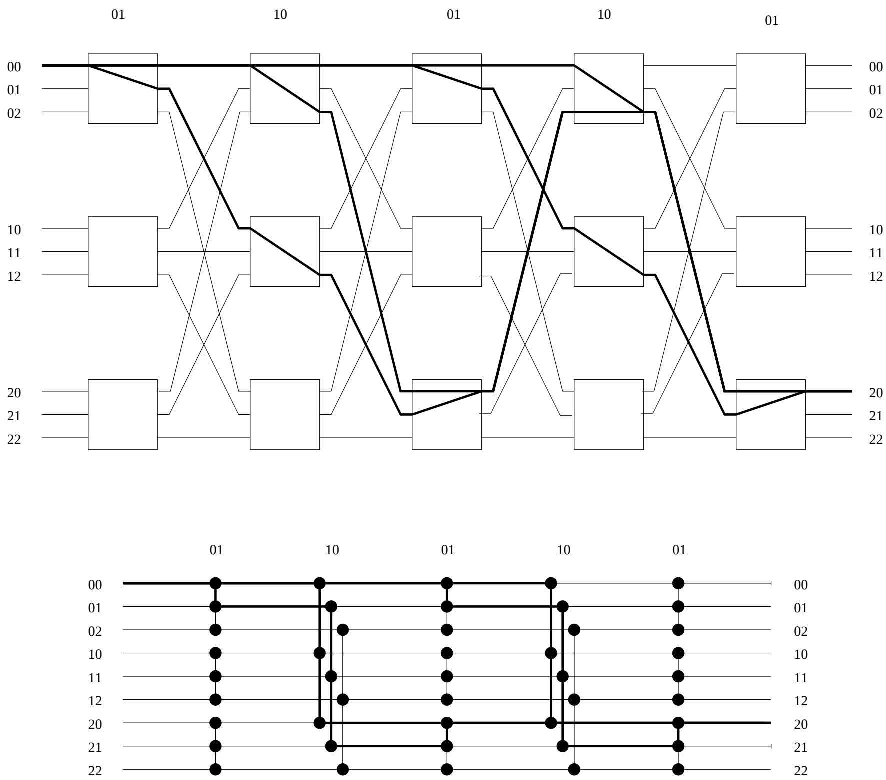


Figure 10: Extension to the MINs with 3x3 Switching Elements

In this proof all vector operations are mod-t. In the construction of path i, $i \leq f$, we decompose the n-dimensional vector $B - A$ into a linear combinations of the masks $\{m_i, m_{i+1}, ..., m_{i+n-1}\}$:

$$B - A = \sum_{j=i}^{i+n-1} c_j m_j \tag{3}$$

Since $\{m_i, m_{i+1}, ..., m_{i+n-1}\}$ span the space, such a decomposition is always possible. If the coefficient $c_i \neq 0$, the switches in stage i is forced to be in mode c_i, i.e., *output* $=$ *input* $+ c_i \times m_i$; If $c_i = 0$, we only need to make sure that the switches in stage i *exchange* to some output, as long as they do not go *straight*. The path i will reach the destination B by a regular routing in the next n stages, i.e., stages $i + 1$ through $i + n$. The construction of path $f + 1$ and the proof that these $f + 1$ paths are disjoint to each other are the same as the proof for the 2×2 case. □

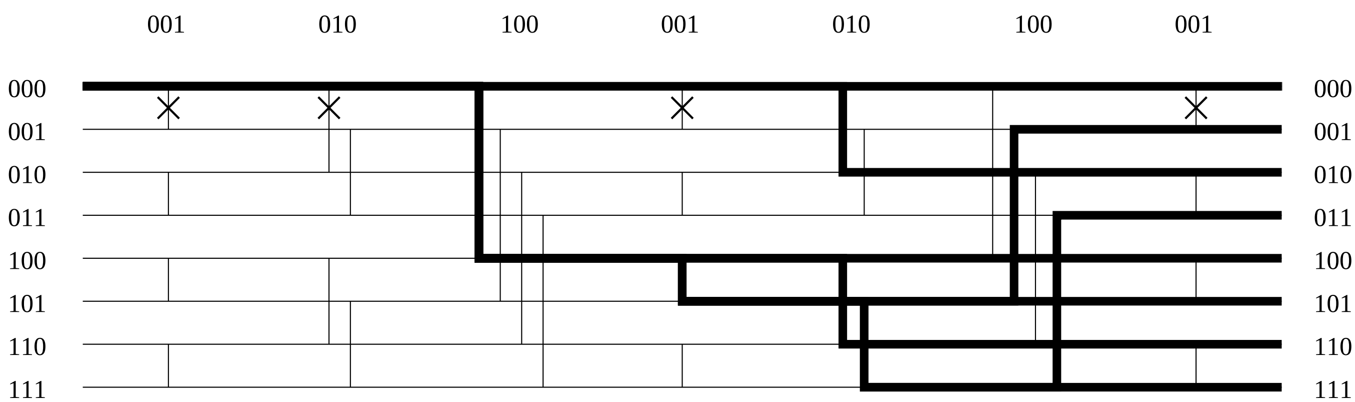


Figure 11: Survived broadcast tree in the presence of 4 faults in a (3,4) Cyclic MIN

We also extend the results to the broadcast connectivity. In the previous sections, we have shown that in an f-extra-stage Cyclic MIN, there exist $f + 1$ disjoint paths between any pair of nodes. It follows that in the presense of f faults, at least 1 path remains intact between any node A and every other nodes. The broadcast from node A is then achieved by picking a survived path from each of these pairs, and construct a broadcast tree from them. Therefore an f-extra-stage Cyclic Multistage Interconnection Network guarantees broadcast connectivity in the presense of f switch faults. As an example, Figure 11 shows the survived broadcast tree in the presence of f switch faults.

5 Conclusion

In this paper we studied the fault-tolerance capabilities of Multistage Interconnection Networks. We constructed the first known extra-stage fault-tolerant Multistage Interconnection Network that tolerates an arbitrary number of switch faults with a minimal number of extra stages. In addition, we proved the general condition that is both sufficient and necessary to achieve this optimal fault-tolerance.

References

[1] G. Adams, D. Agrawal and H. Siefel, "A Survey and Comparison of Fault-Tolerant Multistage Interconnection Networks", *IEEE Computer*, pp. 14-27, June 1987.

[2] G. Adams and H. Siegel, "The Extra Stage Cube: A Fault-Tolerant Interconnection Network for Supersystems", *IEEE Trans. Computers*, pp. 443-454, May 1982.

[3] V.E. Benes, "Optimal Rearrangeable Multistage Connecting Networks", *Bell System Technical Journal*, No. 4, II, July 1964.

[4] J. Bruck and C. Ho, "Fault-Tolerant Cube Graphs and Coding Theory", *Caltech Paradise Technical Report*, ETR007, June 1995. To appear in *IEEE Transaction on Information Theory*, November 1996.

[5] H. Cam and J. Fortes, "Rearrangeability of Shuffle-Exchange Networks", *Proceedings of 3rd Symposium on the Frontiers of Massively Parallel Computation*, pp. 303-314, October 1990.

[6] P.J. Chuang, "CGIN: A Fault Tolerant Modified Gamma Interconnection Network", *IEEE Transactions on Parallel and Distributed Systems*, pp. 1303-1308, December 1996.

[7] D. Knuth, *The Art of Computer Programming: Sorting and Search*, Reading, MA: Addison-Wesley, 1973.

[8] V.P. Kumar and S.M. Reddy, "Augmented Shuffle-Exchange Multistage Interconnection Networks", *IEEE Computer*, pp. 30-40, June 1987.

[9] F. Leighton, *Introduction to Parallel Algorithms and Architectures: Arrays, Trees, Hypercubes*, San Mateo, CA: Morgan Kaufmann, 1992.

[10] Y. Liao, M. Lu and N.F. Tzeng, "The Palindrome Network for Fault-Tolerant Interconnection", *Proceedings of 8th IEEE Symposium on Parallel and Distributed Processing*, pp. 556-561, October 1996.

[11] N. Linial and M. Tarsi, "Interpolation Between Bases and the Shuffle Exchange Network", *European Journal of Combinatorics*, pp. 29-39, October 1989.

[12] M. Pease, "The Indirect Binary n-Cube Microprocessor Array", *IEEE Transactions on Computers*, vol. C-26, No. 5, pp. 458-473, May 1977.

[13] C. Shih and K. Batcher, "Adding Multiple-Fault Tolerance to Generalized Cube Networks", *IEEE Trans. Parallel and Distributed Systems*, pp. 785-792, August 1994.

[14] A. Varma and C.S. Raghavendra, "Fault-Tolerant Routing in Multistage Interconnection networks", *IEEE Transactions on Computers*, pp. 185-393, March 1989.