Fault-Tolerant Multicasting in Multistage Interconnection Networks

Jinsoo Kim*
Seoul Telecommunication O & M Research Center,
Korea Telecom,
17, Woomyeon-Dong, Seocho-Gu, Seoul 137-792, Korea
jinsoo@dambi.kotel.co.kr

Jaehyung Park
Department of Computer Science,
Purdue University, West Lafayette, IN, 47907
hyeoung@cs.purdue.edu

Jung Wan Cho and Hyunsoo Yoon
Department of Computer Science,
Korea Advanced Institute of Science and Technology
373-1, Kusong-Dong, Yusung-Gu, Taejon 305-701, Korea
{jwcho,hyoon}@camars.kaist.ac.kr

Abstract

In this paper, we study fault-tolerant multicasting in multistage interconnection networks (MINs) for constructing large-scale multicomputers. In addition to point-to-point routing among processor nodes, efficient multicasting is critical to the performance of multicomputers. This paper presents a new approach to provide fault-tolerance multicasting, which employs the restricted header encoding schemes. The proposed approach is based on a recursive scheme in order to send a multicast packet to the desired destinations detouring faulty element(s). In the proposed fault-tolerant multicasting, a multicast packet is routed to its own destinations in only two passes through the MIN having a number of faulty elements by exploiting its nonblocking property.

1. Introduction

Multistage interconnection networks (MINs) are popular and efficient interconnection for large-scale multicomputers, such as IBM SP1/SP2 [12] and NEC Cenju-3 [5]. Many of them are a class of networks which consist of $\log_2 N$ stages of 2×2 switching elements connecting N input ports to N output ports. These networks have the property of full access capability that any output can be reachable from any input in a single pass through the network. In addition, there exist a unique path between any pair of input and output in these networks. The unique path property helps the use of a simple and efficient routing algorithm for setting up connections.

However, any single fault on a link or a switching element (SE) of these networks may cause to destroy the full access property. Interconnection networks have the feature of *fault-tolerance* if they can sustain to provide connection in spite of having faulty components. Fault-tolerance criterion for networks in this paper is preserving full access capability [2, 13].

To achieve fault-tolerance in MIN-based multicomputers, there are two alternative approaches. The first is to add SEs and/or links in the network [1, 10], which provide multiple paths to detour faulty elements. In this scheme, the failure of SE(s) and/or link(s) in the network causes reconfiguration of the network in order to preserve full access capability. The reconfigured network by such scheme has the same communication capability as the original network. However, this scheme renders an enormous waste of resources [13] or the modification of its original routing algorithm. In addition, extra logics to tolerate faults may cause irregularity in designing the internal structure; this results in decreasing the modularity of its structure for

*This author is also currently working in Department of Computer Science, Korea Advanced Institute of Science and Technology.

multicomputers.

Instead of augmenting additional elements, the second is to expense routing overhead in order to minimize the loss of resources [6, 13]. Thus, the influence of the faulty SE(s) and/or link(s) can be decreased by allowing multiple passes through the network. The network is known to possess *dynamic full access capability* if every output can be reachable from every input in a finite number of passes, as routing the packet through intermediate outputs if necessary [13]. Even though a single fault destroys the full access capability, some faults do not destroy the dynamic full access capability. A routing algorithm is known as *recursive scheme*, which allow routing through intermediate destinations and recycling through the network. Without loss of resources, in this scheme, a destination can be reachable from its source detouring faulty element(s).

In this paper, we propose fault-tolerant multicasting in MIN-based multicomputers. While unicasting means that a source node delivers a packet to only one destination node. multicasting means that the same packet is delivered from a source to an arbitrary number of destinations. In many multicomputer systems, it is important to provide multicasting as well as unicasting [8, 9]. The proposed multicasting employs the restricted header encoding schemes in order to specify packet's destinations and is based on the recursive scheme which allows a packet recycle at the output to its input in order to send its own destination. The proposed fault-tolerant multicasting exploits the intrinsic nonblocking property of the MIN. Hence, a multicast packet is routed to its own destinations in only two passes in the MIN having certain fault sets which satisfy some conditions.

The structure of this paper is organized as follows. The next section describes the MIN topology, its intrinsic properties, and the restricted header encoding schemes as a system model. Section 3 describes the fault model and terminologies used in this paper. In Section 4 a fault-tolerant multicasting is proposed under certain fault-set environment, which is based on the recursive scheme. Section 5 concludes the paper.

2. System Model

This section describes the MIN topology of multicomputers, its intrinsic properties, and the restricted header encoding scheme.

2.1. Basic Architecture

The MIN is an $N \times N$ interconnection network with $n = \log_2 N$ 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 ports at each stage are labeled using n binary digits $(a_{n-1}a_{n-2}\cdots a_0)$, within each stage starting from the top. And the SEs at each stage are labeled using $(n-1)$ binary digits $(a_{n-1}a_{n-2}\cdots a_1)$ starting from the top.

The MIN that we consider in this paper has butterfly interconnection patterns between stages, and a perfect shuffle interconnection pattern between input controllers and stage $(n-1)$, as shown in Figure 1. In MIN-based multicomputers, output links at the final stage are connected to processing nodes through external links, hence packets can be recycled through the MIN.

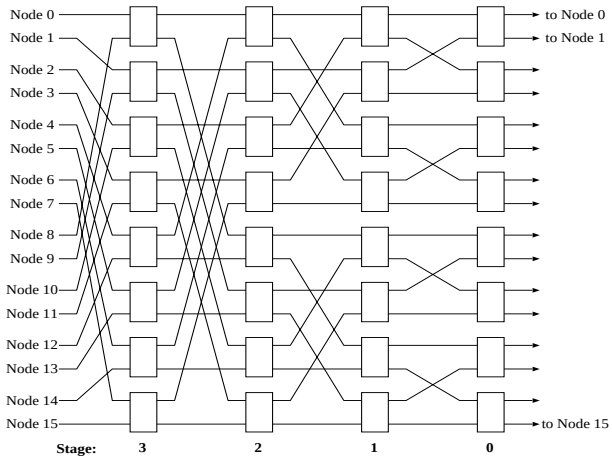

Figure 1. A MIN-based multicomputers

The MIN has the following property on account of its interconnection patterns.

Property 1 L et the source and destination of a packet be $a = a_{n-1}a_{n-2}\cdots a_0$ and $b = b_{n-1}b_{n-2}\cdots b_0$, respectively. In stage i, the SE $b_{n-1}\cdots b_{i+1}a_{i-1}\cdots a_1a_0$ is used for the packet, the b_ith output port of the SE is selected.

2.2. Nonblocking Properties

The MIN considered in this paper is topologically equivalent to the *omega network* [14]. We can easily analogize the following two nonblocking properties from the results of omega network [3]. We represent the source and destination of a packet p_i as s_i and d_i, respectively.

Property 2 A ssume that $s_i > s_j$ and $d_i > d_j$. If the difference between two source addresses is not greater than the difference between two destination addresses, that is, $s_i - s_j \leq d_i - d_j$, then two packets cannot induce any blocking in the MIN.

Definition 1 $S_{msb}(a,b)$ is the number of bits which are identical from $n-1$ to 0 in the binary expansions of a and b, and $S_{lsb}(a,b)$ is the number of bits which are identical from 0 to $n-1$.

Property 3 T wo packets p_i and p_j cannot induce any blocking in the MIN if and only if $S_{lsb}(s_i, s_j)+S_{msb}(d_i, d_j) < n$.

2.3. Restricted Header Encoding Scheme

The restricted header encoding scheme constructs a multicast routing header from reachable destinations which are restricted into a single cube or a single region in the MINs.

As one of the restricted header encoding scheme, a cube encoding specifies arbitrary destination addresses forming a single cube C. The multicast routing header for the cube encoding scheme is specified by $\{R, M\}$, where $R = r_{n-1} \cdots r_1 r_0$ contains the routing information and $M = m_{n-1} \cdots m_1 m_0$ contains the multicast information [4]. To handle the multicast header $\{R, M\}$, an SE at stage i $(0 \leq i \leq n-1)$ examines r_i and m_i. If m_i is 0, the normal unicast routing is performed according to r_i. If m_i is 1, r_i is ignored and the broadcast is performed.

Other restricted header encoding scheme is a region encoding scheme which specifies arbitrary consecutive destination addresses forming a single region [7, 11]. The multicast routing header by the region encoding scheme indicates the *minimum* and *maximum* addresses of consecutive destination addresses. An SE in the MIN 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 in stage $(i+1)$ from where the packet came to stage i. The decision for packet routing and replication is described as follows:

1. If $m_i = M_i = 0$ or $m_i = M_i = 1$, then send the packet out on port 0 or 1, respectively.
2. If $m_i = 0$ and $M_i = 1$, then replicate the packet, modify the headers, and send both packets out on both ports.

These rules 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 :

- $\min_i = \min_{i+1} = m_{n-1} \cdots m_1 m_0,$
  $\max_i = M_{n-1} \cdots M_{i+1} 01 \cdots 1$

for the packet sent out on port 0, and

- $\min_i = m_{n-1} \cdots m_{i+1} 10 \cdots 0,$
  $\max_i = \max_{i+1} = M_{n-1} \cdots M_1 M_0$

for the packet sent out on port 1, at stage i.

3. The Faulty Model and Terminologies

We assume that SEs in stage $(n-1)$ or 0 cannot be faulty, otherwise packets with some sources and destinations alway cannot be routed. We also assume that the mean time to repair faults is quite large.

The *destination group* is a set of destinations such as a region or a cube. The *group packet* is the packet routed to a destination group from a source.

Definition 2 The *binary relation* $<_D$ is defined between two destination groups D_1 and D_2 as follows:
$D_1 <_D D_2$ if and only if $d_1 < d_2$ for all d_1 and d_2 such that $d_1 \in D_1$ and $d_2 \in D_2$.

$M(\alpha)$ and $L(\beta)$ are a set of addresses whose most significant bits are α, and a set of addresses whose least significant bits are β, respectively. Thus, $M(0)$ and $M(1)$ are disjoint groups. Similarly, $L(0)$ and $L(1)$ are also disjoint groups.

Definition 3 D^{α} is defined as a set of destination groups Ds such that at least one destination in D is an element of the set $M(\alpha_{n-1})$. $D^{\bar{\alpha}}$ is defined as a set of destination groups Ds such that any destinations in D are not in $M(\alpha_{n-1})$.

Let $A = \{s_1, \ldots, s_m\}$ be a set of source addresses satisfying that $s_1 < s_2 < \cdots < s_m$, and $B = \{D_1, \ldots, D_n\}$ be a set of destination groups, satisfying that $D_1 <_D D_2 <_D \cdots <_D D_n$. The notation $A \Rightarrow^k B$ means that each packet is routed from a source s_i to a destination group D_i, for all i, $1 \leq i \leq k$, where $1 \leq k \leq m$ and $1 \leq k \leq n$.

A faulty SE at stage i is represented by $f = \alpha\beta$ or $\alpha_{n-1} \cdots \alpha_{i+1} \beta_{i-1} \cdots \beta_0$. Therefore, each packet whose source is in $L(\beta)$ and destination is in $M(\alpha)$ always passes the faulty SE $\alpha\beta$ in banyan network, according to Property 1. Consequently, if such packets are excluded, the faulty SE cannot induce any problem in routing.

4. Fault-Tolerant Multicasting in MINs

In this section, we propose fault-tolerant multicasting in MIN-based multicomputers with certain fault set.

4.1. On Region Encoding Scheme

Definition 4 Let R_1 and R_2 be two regions satisfying that $R_1 <_D R_2$. $R_2 - R_1$, is defined the value of $d_{min2} - d_{max1}$ where d_{min2} and d_{max1} are the minimum destination in R_2 and the maximum destination in R_1.

For example, if $R_1 = [0000, 0010]$, $R_2 = [0100, 0100]$, then $R_2 - R_1 = 0100 - 0010 = 2$.

Fault-Tolerant Multicasting I (FTM-I)

Phase 1: Copy from the source to $2k$ consecutive intermediate destinations SR through the MIN, where $k = max(|D^{\alpha}|, |D^{\bar{\alpha}}|)$. The start address is randomly selected within the restriction that all the consecutive intermediate destinations are in $M(\alpha_{n-1}^{-})$ if the source is in $L(\beta)$.

Phase 2: Route the recycled copies from SR to the regions as follows :

- Case 1 : $A_1 \Rightarrow^{|D^\alpha|} D^\alpha$ if $A_1 = \{a | a \in SR$ and $a \in L(\beta_0)\}$.
- Case 2 : $A_2 \Rightarrow^{|D^{\bar{\alpha}}|} D^{\bar{\alpha}}$ if $A_2 = \{a | a \in SR$ and $a \in L(\beta_0)\}$.

In Figure 2, an example of the second phase is shown, where source 0 sends a multicast packet to destinations $\{1, 3, 4, 7, 8, 10, 11, 14\}$ in a banyan network with a faulty SE 000 at stage 2. Thus, α is 0 and β is 00. Therefore, $D^\alpha = \{[0001, 0001], [0011, 0100], [0111, 1000]\}$ and $D^{\bar{\alpha}} = \{[1010, 1011], [1110, 1110]\}$. $|D^\alpha| = 3$, $|D^{\bar{\alpha}}| = 2$, and $k = 3$. During the first phase, source 0 sends a copy to 6 intermediate destinations $SR = \{8, 9, 10, 11, 12, 13\}$. Although the source 0 is in $L(\beta)$ or $L(00)$, destinations are not in $M(0)$ and then the packet does not pass the faulty SE. In this case, the start address 8 is randomly selected. In the second phase, $A_1 = \{9, 11, 13\}$ and $A_2 = \{8, 10, 12\}$. The thick solid lines of Figure 2 shows the case 1 in which the recycled sources $9, 11$, and 13 in A_1 send their own copies of the multicast packet to destination groups $[0001, 0001]$, $[0011, 0100]$, and $[0111, 1000]$ in $|D^\alpha|$, respectively. The dotted lines shows the case 2 in which the recycled sources 8 and 10 in A_2 send copies to destination groups $[1010, 1011]$ and $[1110, 1110]$ in $|D^{\bar{\alpha}}|$, respectively. Note that the recycled source 12 of the first phase discards its packet.

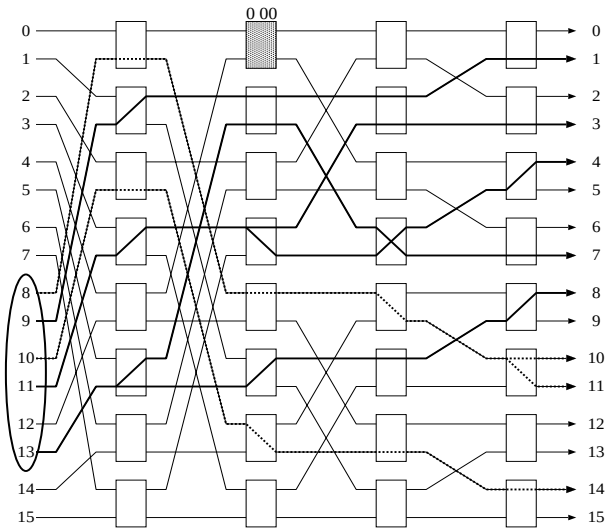

Figure 2. An example of the routing phase in FTM-I

Lemma 1 Let s_i and s_j be the source addresses of two group packets, and R_i and R_j be their destination regions, respectively. Assume that $s(i) < s(j)$ and $R_i <_D R_j$. If $s_j - s_i \leq R_j - R_i$, then two group packets cannot induce any blocking.

Proof : By Definition 4, it is clear that $s_j - s_i \leq R_j - R_i \leq d_j - d_i$ for all d_1 and d_2 such that $d_i \in R_i$ and $d_j \in R_j$. Therefore, two group packets cannot induce any blocking according to Property 2, if $s_j - s_i \leq R_j - R_i$. ■

Lemma 2 Blocking cannot occur between any two group packets whose sources are in $L(0)$ and $L(1)$ respectively, if their destination groups D_1 and D_2 are disjoint.

Proof : We consider it for any two packets whose destinations are $d_1 = d^1_{n-1} \cdots d^1_0 \in D_1$ and $d_2 = d^2_{n-1} \cdots d^2_0 \in D_2$. Based on Property 1, two packets cannot pass the same SE in stage j, $n-1 \geq j \geq 1$. If $d^1_j = d^2_j$, $n-1 \geq j \geq 1$, they encounter the same SE $d^1_{n-1} \cdots d^1_1$ in stage 0. Since D_1 and D_2 are disjoint, $d_1 \neq d_2$. Hence, two packets are routed to different output ports d^1_0 and d^2_0 at the SE. Therefore, two packets cannot induce any blocking in all stages. ■

Lemma 3 The total number of regions that can be composed of k bit destinations is less than or equal to 2^{k-1}.

Proof of Lemma 3 is trivial.

Theorem 1 Algorithm FTM-I with the region encoding scheme can route a multicast packet with any arbitrary set of destination groups in two phases across the MIN having a single faulty SE at stage i, $n-2 \geq i \geq 1$.

Proof : We first consider the routing possibility of any arbitrary set of destination groups. The maximum number of consecutive destinations in the first phase are guaranteed to be 2^{n-1} at the worst case that the source is in $L(\beta)$. Besides, both $|D^\alpha|$ and $|D^{\bar{\alpha}}|$ are less than or equal to 2^{n-2} respectively, according to Lemma 3. Thus, $|A_1| \geq |D^\alpha|$ in the case 1 and $|A_2| \geq |D^{\bar{\alpha}}|$ in the case 2. Therefore, FTM-I can route any arbitrary set of destination groups if routing problems do not occur.

Routing problems that may occur in the faulty MIN are the blocking and the packet passing the fault. We prove that such problems cannot occur in two phases. In the first phase, sources in $L(\beta)$ cannot allowed to send a copy to any destination in $M(\alpha_{n-1})$. Thus, packets that such sources send do not pass any SE $\alpha_{n-1}\gamma\beta$,including the faulty SE, at stage i where $|\gamma| = n - i - 2$. Obviously, packets whose sources are not in $L(\beta)$ don't pass the faulty SE at stage i. Moreover, it is clear that a single group packet does not induce any blocking.

In the second phase, all the packets cannot pass the faulty SE, since the SE that can be used in routing at any stage is $\alpha_{n-1}\gamma\bar{\beta}_0$, $\bar{\alpha}_{n-1}\gamma\bar{\beta}_0$, or $\bar{\alpha}_{n-1}\gamma\beta_0$, where $|\gamma| = n - 3$. There is no blocking between any two group packets in the case 1 and the case 2 respectively according to Lemma 2. Let the active sources and the regions of group packets in the case 1 be two ordered sets as $\{s_1, s_2, \ldots, s_x\}$ and $\{R_1, R_2, \ldots, R_x\}$, respectively. It is clear that $s_{j+1} - s_j = 2$ and $R_{j+1} - R_j \geq 2$ for any j such that $1 \leq j < x$. Consequently, $s_k - s_j \leq R_k - R_j$, $1 \leq j < k \leq x$, since the number of destinations in each region is greater than or

equal to 1. Therefore, blocking cannot occur among the group packets in the case 1. It can be analogized that there is no blocking in the case 2 by similar arguments to the case 1. ■

4.2. On Cube Encoding Scheme

A set of regions that represent all the destinations of a multicast packet satisfies the relation $<_D$ between adjacent regions. However, any arbitrary set of cubes does not so and may cause routing to be more complex. The cube $C = c_{n-1}c_{n-2}\cdots c_0$, where $c_j \in \{0, 1, *\}, n-1 \geq j \geq 0$ satisfying the following condition is called the *least significant bit ordered (LSBO) cube*.

- If $c_j = *$, then $c_k = *$, for all k such that $j > k \geq 0$

For example, $00**$ is LSBO cube, but $00*0*$ is not. Let $S = \{C_1, C_2, \ldots C_m\}$ be an ordered set of LSBO cubes that represent any arbitrary multicast destinations. It is clear that $C_j <_D C_k$ for j, k such that $1 \leq j < k \leq m$.

Fault-Tolerant Multicasting II (FTM-II)
Assume that a set of LSBO cubes represents multicast destinations.
Phase 1: Copy from the source to a single cube SC with $2k$ consecutive intermediate destinations through the MIN, where k is the maximum number of $2^{\lceil \log_2 |D^\alpha| \rceil}$ and $2^{\lceil \log_2 |D^{\bar{\alpha}}| \rceil}$. The cube SC is randomly selected within the same restriction as that of FTM-I.
Phase 2: Route the recycled copy from SC to the LSBO cubes with the appropriate routing headers as follows :

- Case 1 : $A_1 \Rightarrow^{|D^\alpha|} D^\alpha$ if $A_1 = \{a | a \in SC$ and $a \in L(\beta_0)\}$.
- Case 2 : $A_2 \Rightarrow^{|D^{\bar{\alpha}}|} D^{\bar{\alpha}}$ if $A_2 = \{a | a \in SC$ and $a \in L(\beta_0)\}$.

The remaining $(2k - |D^\alpha| - |D^{\bar{\alpha}}|)$ destinations in SC discard their packets.

4.3. Under Certain Fault Set

We consider fault-tolerant multicasting in the MIN with certain fault set.

Definition 5 Let f^1 and f^2 be any two faulty SEs, which they are represented as $\alpha^1_{n-1}\cdots\alpha^1_{i+1}\beta^1_{i-1}\cdots\beta^1_0$ and $\alpha^2_{n-1}\cdots\alpha^2_{i+1}\beta^2_{i-1}\cdots\beta^2_0$. α-match is defined if $S_{msb} > 0$ and α-mismatch is otherwise. Also, β-match is defined if $S_{lsb} > 0$ and β-mismatch is otherwise.

Using the previous approaches FTM-I and FTM-II, any multicast packet is sent to its own destinations without blocking through the MIN having a certain fault set.

4.3.1. In case of α-match and β-match

In Figure 3, an example of the second phase is shown, where source 0 sends a multicast packet to destinations $\{1, 2, 3, 7, 8, 10, 11, 14\}$ in a MIN with a faulty SE 000 at stage 1 and 010 at stage 2. Thus, α_{n-1} is 0 and β_{n-1} is 0. Therefore, $D^\alpha = \{0001, 001*, 0111\}$ and $D^{\bar{\alpha}} = \{1000, 101*, 1110\}$. $|D^\alpha| = 3$, $|D^{\bar{\alpha}}| = 3$, and $k = max(2^{\lceil \log_2 3 \rceil}, 2^{\lceil \log_2 3 \rceil}) = 4$. During the first phase, source 0 sends a copy to a single cube $1 * **$ with 8 destinations. In the second phase, the recycled source $9, 11$, and 13 in A_1 send their own recycled copies of the multicast packet to destination groups $0001, 001*$, and 0111 in D^α, respectively. the recycled source $8, 10$, and 12 in A_2 send their own recycled copies to destination groups $1000, 101*$, and 1110 in $D^{\bar{\alpha}}$, respectively. While the recycled sources 14 and 15 discard their packets.

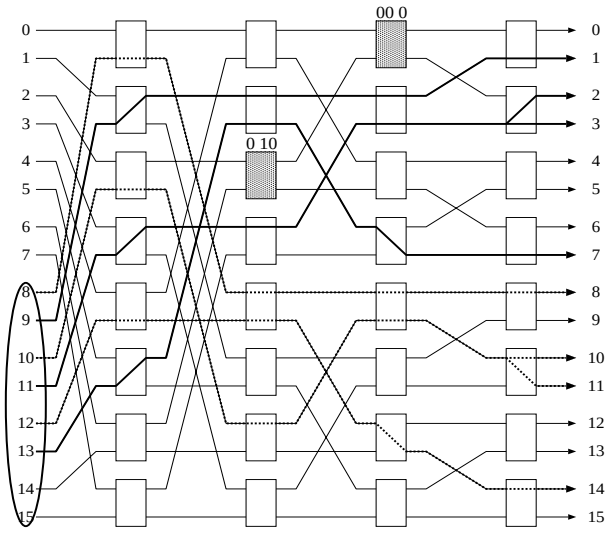

Figure 3. An example of α-match and β-match

In case of α-match and β-match, any multicast packet is sent to its own destinations without blocking through the MIN having two or more faulty SEs. It is because α_{n-1} is always same and β_{n-1} is, too.

4.3.2. In case of α-mismatch and β-mismatch

In the first phase, the source s sends a multicast packet to $M(f^k_{n-1})$ such that $s_0 \neq f^k_0$, where k is 1 or 2. In Figure 4, an example of the first phase is shown, where source 1 sends a multicast packet to destinations $\{1, 2, 3, 7, 8, 10, 11, 14\}$ in a MIN with a faulty SEs 001 and 100 at stage 2.

In the second phase, the destinations $8, 10, 12, 9, 11, 13$ send their own recycled copies of the multicast packet to $0001, 001*, 0111, 1000, 101*, 1110$, respectively.

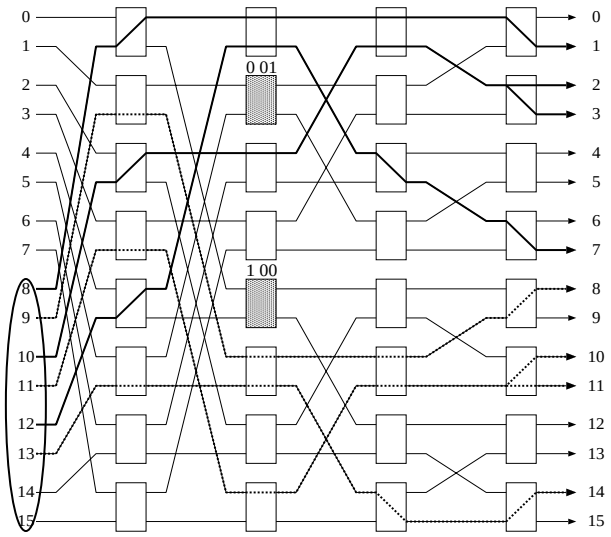

Figure 4. An example of α-mismatch and β-mismatch

Lemma 4 Let MC_k be the number of cubes that can be composed of k bit destinations. Then, $MC_k \leq 2^{k-1}$.

Proof : It can be proved by induction on k. It is clear that the property is true for $k = 1$, since $MC_1 = 1$. Assume that it is true for $k-1$. Let C_k^0 and C_k^1 be two sets of cubes of which k bit destinations are in $M(0)$ and $M(1)$ respectively. By the assumption, $|C_k^0| \leq 2^{k-2}$ and $|C_k^1| \leq 2^{k-2}$. If there are two cubes $C_a = 0a_1a_2\cdots a_{k-1} \in C_k^0$ and $C_b = 1b_1b_2\cdots b_{k-1} \in C_k^1$ where $a_j, b_j \in \{0, 1, *\}$ and $a_j = b_j$, for all j such that $1 \leq j \leq k-1$, then such cubes are merged into a single cube $*a_1a_2\cdots a_{k-1}$. Therefore, $MC_k \leq |C_k^0| + |C_k^1| \leq 2^{k-1}$. ■

Lemma 5 Let a and b be two n-bit numbers such that $a < b$. Then, $S_{lsb}(a, b) \leq \log(b-a)$.

Proof : Let $k = S_{lsb}(a, b)$. Then, a and b can be represented by $a_{n-1}a_{n-2}\cdots a_k c_{k-1}\cdots c_0$ and $b_{n-1}b_{n-2}\cdots b_k c_{k-1}\cdots c_0$ respectively. Thus, $b - a \geq 2^k$, and accordingly $S_{lsb}(a, b) = k \leq \log(b-a)$ ■

Lemma 6 Let $S = \{C_1, C_2, \ldots, C_m\}$ be an ordered set of LSBO cubes. $S_{msb}(C_a, C_b) \leq n - \log(b-a+1) - 1$, if $b > a$.

Proof : Let $k = S_{msb}(C_a, C_b)$. Since $C_a <_D C_j <_D C_b$, $S_{msb}(C_a, C_j) \geq k$, for all j such that $a < j < b$. Thus, the most significant k bits of the cubes $C_j, a \leq j \leq b$ are identical. Considering the remaining $n-k$ bits, the number of such cubes, that is $b-a+1$, must be less than or equal to the maximum number of cubes that composed of addresses with $n-k$ bits. By Lemma 4, $b - a + 1 \leq 2^{(n-k)-1}$. Consequently, $S_{msb}(C_a, C_b) = k \leq n - \log(b-a+1) - 1$. ■

Theorem 2 FTM-I and FTM-II can route a multicast packet with any arbitrary set of destination groups in two phases across the MIN having some faulty SEs.

Proof : By the similar argument to Theorem 1, the following facts can be proved.

- $2^{n-2} \geq |A_1| \geq |D^\alpha|$ in the case 1 and $2^{n-2} \geq |A_2| \geq |D^{\bar{\alpha}}|$ in the case 2.
- Any packet cannot pass the faulty SE in the first and the second phases.
- There is no blocking in the first phase.
- Any two group packets in the case 1 and the case 2 respectively, do not induce any blocking.

We consider blocking problems among the group packets in the case 1 of the second phase, those in case 2 are similar. Let the active sources and the cubes of group packets in the case 1 be two ordered sets as $\{s_1, s_2, \ldots, s_{|D^\alpha|}\}$ and $SC = \{C_1, C_2, \ldots, C_{|D^\alpha|}\}$, respectively. Assume that $1 \leq j < k \leq |D^\alpha|$. By Lemma 5, $S_{lsb}(s_j, s_k) \leq \log(s_k - s_j) = \log(2(k-j))$. $S_{msb}(C_j, C_k) \leq n - \log(k-j+1) - 1$ by Lemma 6. Consequently, $S_{lsb}(s_j, s_k) + S_{msb}(C_j, C_k) < n$. According to Property 3, blocking cannot occur among the group packets in the case 1. ■

5. Conclusions

In this paper, we proposed fault-tolerant multicasting in the wrap-around MIN for large-scale multicomputers. The proposed algorithms can employs both region and cube encoding schemes as the header encoding scheme. They are based on a recursive scheme in order to send a multicast packet to the desired destinations. A multicast packet is routed to its own destinations in only two passes on the MIN having a certain fault set. It has been proved that these algorithms can route any arbitrary multicast destinations without any blocking, by exploiting well-known nonblocking properties of MIN. The proposed approach can be easily applied to wormhole or virtual cut-through routed MINs for multicomputers.

Acknowledgement

This work is supported in part by KOSEF(Korea Science and Engineering Foundation) through *Center for Artificial Intelligence Research* at Korea Advanced Institute of Science and Technology.

References

[1] G. B. Adams III, D. P. Agrawal, and H. J. Siegel. A Survey and Comparison of Fault-Tolerant Multistage

Interconnection Networks. *IEEE Computer*, 20:14–27, 1987.

[2] S. Chalasani, C. S. Raghavendra, and A. Varma. Fault-Tolerant Routing in MIN-Based Supercomputers. *Journal of Parallel and Distributed Computing*, 22(2):154–167, Aug. 1994.

[3] V. Chandramouli and C. S. Raghavendra. Nonblocking Properties of Interconnection Switching Networks. *IEEE Transactions on Communications*, 43(2/3/4):1793–1799, Feb./Mar./Arp. 1995.

[4] X. Chen and V. Kumar. Multicast Routing in Self-Routing Multistage Networks. In *Proc. of IEEE Infocom*, pages 306–314, Apr. 1994.

[5] N. Koike. NEC Cenju-3: A Microprocessor-Based Parallel Computer. In *Proc. of the Int'l Parallel Processing Symposium*, pages 396–401, Apr. 1994.

[6] V. P. Kumar and S. J. Wang. Dynamic Full Access in Fault Tolerant Multistage Interconnection Network. In *Proc. of the Int'l Conf. on Parallel Processing*, pages 621–630, 1990.

[7] T. T. Lee. Nonblocking Copy Networks for Multicast Packet Switching. *IEEE Journal on Selected Areas in Communications*, 6(9):1455–1467, Dec. 1988.

[8] P. K. McKinley, Y. Tsai, and D. F. Robinson. Collective Communication in Wormhole-Routed Massively Parallel Computers. *IEEE Computer*, 28(12):39–50, Dec. 1995.

[9] D. K. Panda. Issues in Designing Efficient and Practical Algorithms for Collective Communication Wormhole-Routed Systems. In *Proc. of the Int'l Conf. on Parallel Processing*, pages 8–15, Aug. 1995.

[10] J. Park and H. Lee. Ring Banyan Network: A Fault-Tolerant Multistage Interconnection Network and Its Fault Diagnosis. In *Lecture Notes in Computer Science*, volume 852, pages 511–528. Springer-Verlag, 1994.

[11] C. S. Raghavendra, X. Chen, and V. P. Kumar. A Two Phase Multicast Routing Algorithm in Self-Routing Multistage Networks. In *Proc. of Int'l Conference on Communications*, pages 1612–1618, Jun. 1995.

[12] C. B. Stunkel and *et al.* The SP2 Communication Subsystem. Technical report, IBM Thomas J. Watson Research Center, Aug. 1994.

[13] A. Varma and C. S. Raghavendra. Fault-Tolerant Routing in Multistage Interconnection Networks. *IEEE Transactions on Computers*, 38(3):385–393, Mar. 1989.

[14] C. L. Wu and T.-Y. Feng. On a Class of Multistage Interconnection Networks. *IEEE Transactions on Computers*, C-29(8):694–702, Aug. 1980.