다단계 상호연결망에서 영역 부호화 방식을 사용하는 고장 허용 멀티캐스팅 알고리즘

(A Fault-Tolerant Multicasting Algorithm using Region Encoding Scheme in Multistage Interconnection Networks)

김 진 수 † 장 정 환 ††
(Jinsoo Kim) (Jung-Hwan Chang)

요 약 본 논문은 다수의 고장 스위칭 소자를 갖는 다단계 상호연결망(MIN)에서 영역 부호화 방식을 적용하는 고장 허용 멀티캐스팅 알고리즘을 제안한다. 제안된 알고리즘은 MIN의 전체 스위칭 소자들을 같은 크기를 갖는 두 개의 부분집합으로 구분하고, 모든 고장이 동일한 부분집합에 속한 고장 유형을 허용한다. 제안된 알고리즘은 고장을 우회하여 멀티캐스트 메시지를 목적지까지 보내기 위해, MIN을 통해 메시지를 순환시키는 기법을 사용한다. 본 알고리즘은 고장난 MIN에서 임의의 멀티캐스트 메시지를 두 번 순환시켜 라우팅 할 수 있음을 증명한다.

키워드 : 다단계 상호연결망, 멀티캐스트, 영역 부호화, 고장 허용성

Abstract This paper proposes a fault-tolerant multicasting algorithm employing the region encoding scheme in multistage interconnection networks (MIN's) containing multiple faulty switching elements. After classifying all switching elements into two subsets with equal sizes in MIN, the proposed algorithm can tolerate the faulty pattern where every fault is contained in the same subset. In order to send a multicast message to its destinations detouring faults, the proposed algorithm uses the recursive scheme that recirculates it through MIN. We prove that this algorithm can route any multicast message in only two passes through the faulty MIN.

Key words : Multistage interconnection network, multicast, region encoding fault-tolerance

1. 서 론

다단계 상호연결망(Multistage Interconnection Network, MIN)은 다중컴퓨터 시스템에서 프로세싱 노드들 간의 연결을 위한 네트워크로서 널리 사용되는 구조중 하나이다. 또한, MIN은 네트워크 자체가 지닌 셀프라우팅 특성에 의해 고속의 스위칭을 지원하므로, 광대역 종합정보 통신망(B-ISDN)에서 ATM(Asynchronous Transfer Mode) 스위칭 시스템의 내부 연결구조로도 보편적으로 사용되고 있다. MIN은 임의의 입력단에서 모든 출력단으로 네트워크를 한번 통과하여 접근이 가능하며, MIN의 이러한 특성을 완전 접근 능력(full access capability)라고 한다. 또한, MIN의 입력단과 출력단 사이에 유일한 라우팅 경로가 존재한다. MIN의 이러한 특성은 빠르고 효율적인 셀프라우팅 기능을 제공하지만, 이 특성 때문에 다른 네트워크에 비해 고장처리 능력이 저하된다. 즉 MIN에서 하나의 스위칭 소자(Switching Element, SE) 또는 링크가 고장이 발생하더라도, 우회 경로가 없으므로 완전 접근 능력을 상실하게 된다.

MIN에서 고장 허용성(fault-tolerance)의 실현은 크게 하드웨어 접근 방법과 소프트웨어 접근 방법으로 구분된다. 하드웨어 접근 방법은 네트워크에 SE 또는 링크를 추가하여 다중의 경로를 제공함으로써 고장을 극복하는 것이다[1, 2, 3]. 이 방법에서는 고장이 발생하면 네트워크를 재구성하여 완전 접근 능력을 유지하지만, 고장이 발생하지 않으면 추가된 하드웨어 자원은 상당

† 정 회 원 : 건국대학교 컴퓨터응용과학부 교수
jinsoo@kku.ac.kr
†† 종신회원 : 부산외국어대학교 컴퓨터전자공학부 교수
jhchang@pufs.ac.kr
논문접수 : 2001년 2월 19일
심사완료 : 2001년 11월 29일

히 낭비적인 요소가 된다. 또한, 네트워크를 불규칙한 형태로 만들어 MIN 기반 시스템의 모듈화를 저해하게 된다. 소프트웨어 접근 방법은 부수적인 하드웨어 없이 고장 허용 라우팅을 제공하는 것이다[4, 5, 6]. 이 방법에서는 고장을 회피하기 위해 네트워크 상에서 메시지를 여러 번 순환시킨다. 이와 같이 MIN에서 중간 출력단을 통해 메시지를 순환시키는 라우팅 기법을 순환 기법(recursive scheme)이라고 하며, 이 기법을 이용하여 자원의 낭비가 없이 고장을 우회하여 출발지(source)에서 목적지(destination)로 메시지를 라우팅 할 수 있다. 그러나 MIN에서 고장 허용성에 대한 연구들은 대부분 일대일(unicast) 통신을 대상으로 진행되었다.

멀티캐스트 통신(muticast communication)은 집합체 통신(collective communication)[7]의 중요한 형태 중의 하나로서, 하나의 출발지에서 임의의 다수 목적지로 동일한 메시지를 전송하는 일대다 통신을 의미한다. 가장 기본적인 통신 형태인 일대일 통신뿐만 아니라, 멀티캐스트 통신은 병렬 처리와 통신 분야에서 그 중요성이 증가되고 있다. 이 멀티캐스트 통신은 병렬 처리에서 FFT(Fast Fourier Transform)과 경계선 동기화(barrier synchronization)와 같은 연산에 필수적이며, 디렉토리 기반의 캐쉬 일관성 유지 프로토콜을 위한 수정(updating)과 무효화(invalidation)에도 효과적으로 사용된다. 또한, 통신 분야에서 다자간 영상회의와 VOD (video-on-demand)와 같은 응용서비스에서 멀티캐스트 기능이 필수적으로 제공되어야 한다. MIN 기반의 시스템에서 멀티캐스트 기능을 제공하기 위해 순환 기법을 기반으로 한 많은 알고리즘이 제안되었다[8, 9, 10, 11]. 그러나, 이들 연구에서는 고장 허용성이 거의 고려되지 않았다. 즉, 기존의 대부분 연구에서는 순환 기법이 단지 멀티캐스팅 또는 고장을 허용하기 위한 일대일 라우팅에만 활용되었다.

본 논문은 MIN에서 SE들의 고장을 허용하는 멀티캐스트 알고리즘을 제안한다. 제안된 알고리즘은 멀티캐스트 메시지의 연속된 목적지 주소들을 하나의 영역으로 표시하는 영역 부호화(region encoding) 방식을 사용한다. 또한, 고장을 피하고 임의의 멀티캐스트 메시지를 목적지들로 보내기 위해 순환 기법을 기반으로 한다. MIN 상의 SE들을 그 번호에 따라 두 개의 부분집합으로 분할하고, 하나의 부분집합에 속한 다수의 고장 SE들이 고장난 MIN에서 메시지를 두 번 순환시킴으로써 멀티캐스팅을 수행한다. 그러나, 고장 SE들의 일부분이 한 부분집합에 속해 있고, 나머지 고장 SE들이 다른 부분집합에 속해 있는 경우는 허용하지 않는다.

본 논문은 다음과 같이 구성되었다. 2 장에서는 MIN의 기본적인 구조와 상호 무충돌 특성 그리고 영역 부호화 방식을 설명하고, 고장 모델에 대해 기술한다. 3 장에서는 두 단계로 구성되는 고장 허용 멀티캐스트 알고리즘을 제안하고, 제안한 알고리즘의 정확성을 증명한다. 그리고, 4 장에서 결론을 맺는다.

2. 시스템 모델과 용어

본 장에서는 MIN 기반 시스템의 구조와 MIN에서의 메시지간 상호 무충돌 특성에 대해 살펴보고, 임의 목적지를 갖는 멀티캐스트 메시지의 주소를 표현하기 위한 영역 부호화 방식에 대해 설명한다. 또한, 알고리즘에서 사용하는 기본적인 용어를 정의하고, MIN에 대한 고장 모델 등을 기술한다.

2.1 시스템의 구조

네트워크 크기가 N인 MIN은 $n=log_2N$ 개의 단(stage)으로 구성되고, 각 단은 $N/2$개의 2×2스위칭 소자(Switching Element, SE)들로 이루어진다. 각 단에서 N개의 입출력 링크는 위부터 아래로 0에서 $N-1$까지 n-비트로 표현되고, $N/2$개의 SE는 위부터 아래로 0에서 $N/2-1$ 까지 $(n-1)$-비트로 표현된다. 각 단의 번호는 좌측부터 우측으로 $n-1$에서 0로 표현된다.

본 논문에서 고려하는 MIN 기반의 다중컴퓨터는 그림 1과 같은 구조를 갖는다. 그림 1에서 보는 바와 같이, 인접한 두 단간 버터플라이(butterfly) 연결 형태, 노드들과 제일 좌측의 $(n-1)$ 단 사이는 완전 셔플(perfect shuffle) 연결형태를 갖는다. 그리고 메시지의 순환을 위해, 그림 1의 점선과 같이 우측의 출력단에서 노드쪽으로 역방향 링크들이 존재한다. 또한, 멀티캐스트를 지원하기 위해, 모든 SE들은 브로드캐스트 기능을 갖는다. 즉 SE는 입력 메시지의 헤더에 따라 메시지를 한쪽 또는 양쪽 출력 포트로 전송할 수 있어야 한다.

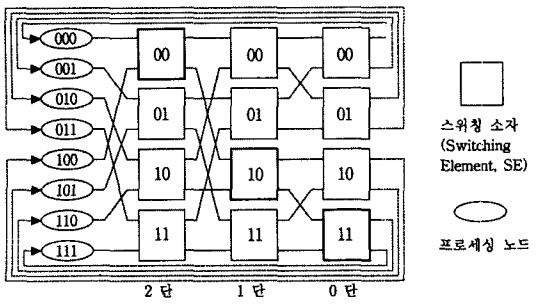

그림 1 MIN 기반의 다중컴퓨터 시스템(N=8)

2.2 MIN의 구조적 특성과 영역 부호화 방식

MIN은 2.1절에서 설명한 연결 구조에 따라 다음과 같은 특성을 갖는다.

[특성-1] 입력이 $s=s_{n-1}\cdots s_1s_0$, 출력이 $d=d_{n-1}\cdots d_1d_0$인 메시지가 i ($n-1\geq i\geq 0$) 단을 통과할 때, 사용되는 SE 번호는 $(d_{n-1}\cdots d_{i+1}s_{i-1}\cdots s_0)$이며 그 SE의 d_i번째 출력 포트가 선택된다.

예를 들어, 그림 1에서 노드 000에서 노드 111로 가는 메시지는 MIN에서 입력과 출력이 각각 000과 111이 된다. 이 메시지가 각 단에서 사용하는 SE와 출력 포트는 2 단에서 00과 1이며, 1 단에서 10과 1이며, 그리고 0 단에서 11과 1임을 알 수 있다.

본 논문에서 고려하는 MIN은 오메가 네트워크와 구조적으로 동등하며 [12], 상이한 두 메시지 간의 충돌이 발생할 수 있다. 두 메시지가 임의의 단에 있는 같은 SE에서 동일한 출력 포트로 전송하려 할 때 충돌이 발생한다. MIN에서의 기존 연구 결과를 활용하여, 다음의 무충돌 특성을 쉽게 유추할 수 있다 [13].

[특성-2] 메시지 m의 입력과 출력을 각각 s^m과 d^m라고 표시할 때, $s^i<s^j$와 $d^i<d^j$인 두 메시지 i와 j가 $s^j-s^i \leq d^j-d^i$를 만족하면, 두 메시지는 MIN에서 충돌을 일으키지 않는다.

일대다 통신 형태를 갖는 멀티캐스팅에서 목적지의 개수는 일정하지 않다. 이러한 멀티캐스트 메시지의 목적지들을 표현하기 위해 여러 가지 방식들이 사용되며, MIN 기반의 시스템에서 널리 사용되는 방식 중 하나가 영역 부호화(region encoding) 방식이다[11, 14, 15]. 영역 부호화 방식은 연속된 주소를 갖는 다수 목적지들을 하나의 영역(region), 즉 최소와 최대 주소의 쌍 [최소주소, 최대주소]로 표현하는 방식이다. 예를 들어 멀티캐스트 목적지가 001, 010, 011, 101 일 때, 목적지들을 두 개의 영역인 [001, 011], [101, 101]로 나타낼 수 있다. 즉 하나의 영역은 1개 이상의 연속된 목적지 주소로 구성되는 집합으로 볼 수 있다. 즉, [001, 011]은 {001, 010, 011}이고 [101, 101]은 {101}을 의미한다.

영역 부호화 방식을 사용하는 MIN에서 SE는 [최소주소, 최대주소]인 멀티캐스트 라우팅 헤더를 처리해야 한다. i ($n-1\geq i\geq 0$) 단의 SE는 헤더가 최소주소 $min = m_{n-1}\cdots m_1m_0$과 최대주소 $MAX = M_{n-1}\cdots M_1M_0$을 갖는 메시지를 받으면, 다음과 같이 메시지를 전송한다.

(1) $m_i=M_i=0$ 이거나 $m_i=M_i=1$ 이면, 헤더의 변경 없이 출력 포트 0이나 1로 메시지를 전송함.

(2) $m_i=0$ 이고 $M_i=1$ 이면, 출력 포트 0과 1로 메시지를 전송하되 각각 다음과 같이 헤더를 변경함.

- 출력 포트 0으로 나가는 메시지의 헤더 : $[min, M_{n-1}\cdots M_{i+1}01\cdots1]$
- 출력 포트 1로 나가는 메시지의 헤더 : $[m_{n-1}\cdots m_{i+1}10\cdots0, MAX]$

그림 2는 출발지가 000이고 목적지 영역이 [001,100]인 멀티캐스트 메시지의 라우팅 예를 보이며, 라우팅 경로상에 있는 SE들의 입력과 출력 포트에서 헤더의 변화를 나타내고 있다. 1 단의 SE 10은 헤더가 [100,100]인 메시지를 받고, $m_1=M_1=0$이므로 헤더의 변경 없이 출력 포트 0으로 메시지를 전송한다. 1 단의 SE 00은 헤더가 [001,011]인 메시지를 받고, $m_1=0$이고 $M_1=1$이므로, 헤더가 [001,001]인 메시지를 출력 포트 0으로, 헤더가 [010,011]인 메시지를 출력 포트 1로 전송한다.

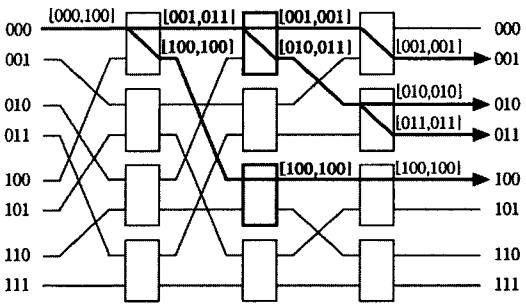

그림 2 영역 부호화 방식의 라우팅 예

2.3 고장 모델과 용어의 정의

본 논문에서는 SE가 고장인 경우를 다루고 있으며, $(n-1)$ 단과 0 단에 있는 SE는 고장이 아님을 가정한다. 이것은 MIN에서 메시지의 입력이 $(n-1)$ 단의 고장난 SE와 연결된 경우 또는 출력이 0 단의 고장난 SE와 연결된 경우에는 근본적으로 라우팅이 불가능하기 때문이다.

[정의-1] $M(0)$과 $M(1)$은 최상위비트가 각각 0과 1인 주소들의 집합으로, $L(0)$과 $L(1)$은 최하위비트가 각각 0과 1인 주소들의 집합으로 정의한다.

예를 들면, 주소 011은 $M(0)$에 속하며, $L(1)$에도 속한다. 당연히, $M(0) \cap M(1) = \phi$ 이고, $L(0) \cap L(1) = \phi$ 이다.

[정의-2] i ($n-2\geq i\geq 1$) 단에 있는 하나의 SE를 $\alpha_{n-1}\cdots\alpha_{i+1}\beta_{i-1}\cdots\beta_0$으로 표시하며, MIN상의 모든 SE를 다음과 같은 두 개의 분할 집합(partitioned sets) F_0와 F_1로 구분한다.

- $F_0 = \{\alpha_{n-1}\cdots\alpha_{i+1}\beta_{i-1}\cdots\beta_0 \mid \alpha_{n-1}\oplus\beta_0 = 0\}$
- $F_1 = \{\alpha_{n-1}\cdots\alpha_{i+1}\beta_{i-1}\cdots\beta_0 \mid \alpha_{n-1}\oplus\beta_0 = 1\}$

또한, 만일 모든 고장 SE들이 집합 F_0에 속한 경우를

F_0 고장, 모든 고장 SE들이 집합 F_1에 속한 경우를 F_1 고장이라고 정의한다.

예를 들면, 4 단으로 구성된 MIN의 모든 단에서 SE 000, 010, 101, 111은 $\alpha_{n-1} \oplus \beta_0 = 0$ 이므로 F_0에 속하며, SE 001, 011, 100, 110은 $\alpha_{n-1} \oplus \beta_0 = 1$ 이므로 F_1에 속한다. 본 논문에서 제안하는 고장 허용 멀티캐스팅 알고리즘은 MIN에서 F_0 고장 또는 F_1 고장이 발생한 경우만을 대상으로 한다.

[정의-3] 두 영역 $R_1 = \{d_1^1, d_1^2, \cdots, d_1^l\}$과 $R_2 = \{d_2^1, d_2^2, \cdots, d_2^m\}$ 간의 이항 관계(binary relation) $<_D$는 다음과 같이 정의된다.

$1 \le j \le l$과 $1 \le k \le m$을 만족하는 모든 j와 k 쌍에 대해 $d_1^j < d_2^k$이면 $R_1 <_D R_2$이고, 그 역도 성립한다.

[정의-4] 두 영역 R_1과 R_2가 $R_1 <_D R_2$를 만족하고, R_1에 속한 최대 주소를 d_1^{max}, R_2에 속한 최소 주소를 d_2^{min}으로 나타낼 때, R_2와 R_1의 차 $R_2 - R_1$은 $d_2^{min} - d_1^{max}$의 값으로 정의한다.

예를 들어, R_1 = [001, 011] = {001, 010, 011} 이고, R_2 = [101, 110] = {101, 110} 이면, $R_1 <_D R_2$ 이다. 또한, $R_2 - R_1$은 2 (=101-011)이다.

[정의-5] $R_1 <_D R_2 <_D \cdots <_D R_m$을 만족하는 정렬된 영역의 집합 $D = \{R_1, R_2, \cdots, R_m\}$이 멀티캐스트 목적지를 나타낼 경우, 이러한 집합 D를 **목적지 영역 집합**이라고 정의하며, 다음과 같이 두 개의 분할 집합 $D0$과 $D1$로 구분하여 표기한다.

- $D0 = \{R_i \mid R_i \subseteq M(0), 1 \le i \le m\}$
- $D1 = \{R_j \mid R_j \subseteq M(1), 1 \le j \le m\}$

예를 들어, 멀티캐스트 목적지가 000, 001, 010, 100, 101, 111 일 때, $D0$은 {[000,010]} 이고, $D1$은 {[100, 101], [111,111]}이다.

[정의-6] $S = \{s^1, \cdots, s^j\}$가 $s^1 < s^2 < \cdots < s^j$ 인 출발지 주소들 집합이고, $D = \{R_1, R_2, \cdots, R_k\}$ $(k \le j)$가 $R_1 <_D R_2 <_D \cdots <_D R_k$ 인 목적지 영역 집합이라고 할 때, $S \Rightarrow^k D$ 는 출발지 s^l $(1 \le l \le k)$에서 목적지 영역 R_l으로 k 개의 메시지들이 동시에 라우팅 되는 것을 표현한다.

예를 들어, S가 {0011,0101,0111}이고 D가 {[1010, 1011], [1110,1110]}일 때, $S \Rightarrow^2 D$는 출발지가 0011이고 목적지 영역이 [1010,1011]인 메시지와 출발지가 0101 이고 목적지 영역이 [1110,1110]인 2 개의 메시지가 동시에 라우팅 되는 것을 나타낸다.

3. 고장 허용 멀티캐스팅 알고리즘

본 장에서는 F_0 고장 또는 F_1 고장인 MIN에서의 멀티캐스팅 알고리즘을 제시한다. 임의의 멀티캐스트 목적지들은 하나의 목적지 영역 집합으로 표현이 가능하며, 알고리즘에서는 목적지 영역 집합을 사용한다. 이 알고리즘의 멀티캐스팅 예를 보이고, MIN의 구조적인 특성을 활용하여 알고리즘의 정확성을 증명한다. 그리고, 다수의 멀티캐스트 메시지가 동시에 라우팅 될 때 발생할 수 있는 문제와 그 해결책을 제시한다.

3.1 멀티캐스팅 알고리즘

멀티캐스트 목적지는 가변적이므로 임의 다수의 목적지 영역으로 표현된다. 고장난 MIN에서 메시지의 헤더가 단일 영역으로 표현되는 영역 부호화 방식을 사용하여, MIN을 한 번 통과하여 하나의 출발지에서 다수의 목적지 영역으로 멀티캐스팅 하는 것은 불가능하다. 이를 처리하기 위해, 고장 허용 멀티캐스팅 알고리즘 (FTM)은 그림 3과 같이 두 단계로 구성된다.

단계 1에서는 목적지 영역의 개수 (=$|D0|+|D1|$) 만큼 메시지를 복사하기 위해, 그 이상의 크기 (= $2 \times max(|D0|,|D1|)$)를 갖는 중간 목적지 영역 IR로 메시지를 라우팅한다. 여기서, IR의 시작 주소는 임의로 선택한다. 단계 1에서 사용되는 메시지는 IR을 나타내는 라우팅 헤더와 데이타 이외에 IR의 시작 주소와 목적지 영역 집합 정보를 추가로 포함한다.

단계 2에서는 IR에 속한 노드들은 이전 단계에서 수신된 메시지를 최종 목적지 영역들로 각각 라우팅을 한다. 각각의 노드는 수신된 메시지에 포함된 IR의 시작 주소와 자신의 노드번호를 이용하여, 자신이 목적지 영역 집합 중 어느 영역으로 라우팅 해야 하는 지를 판단할 수 있으며, 선택된 영역을 메시지의 라우팅 헤더로

가정
1. 모든 고장은 F_0 고장 또는 F_1 고장이다.
2. 멀티캐스트 목적지는 목적지 영역 집합으로 구성된다.
3. $k = max(|D0|, |D1|)$ 이다.

알고리즘

단계 1 : 멀티캐스트 메시지의 복사를 위해, 출발지 $s = s_{n-1} \cdots s_1 s_0$에서 다음의 조건을 만족하는 임의의 중간 목적지 영역 IR ($|IR| = 2k$)로 메시지를 라우팅한다.
- F_0 고장의 경우, $IR \subseteq M(s_0')$
- F_1 고장의 경우, $IR \subseteq M(s_0)$

단계 2 : IR을 두 개의 집합 $S_0 \subseteq L(0)$과 $S_1 \subseteq L(1)$로 구분하고, S_0 과 S_1에 속한 각 노드는 새로운 중간 출발지로서 사용되어, 단계 1에서 수신된 메시지를 다음과 같이 대응되는 최종 목적지 영역으로 라우팅한다.
- F_0 고장의 경우, $S_0 \Rightarrow^{|D1|} D1$ 하고 $S_1 \Rightarrow^{|D0|} D0$ 함
- F_1 고장의 경우, $S_0 \Rightarrow^{|D0|} D0$ 하고 $S_1 \Rightarrow^{|D1|} D1$ 함

IR에서 $(2k - |D0| - |D1|)$ 개의 나머지 노드는 자신의 메시지를 무시한다.

그림 3 고장 허용 멀티캐스팅 알고리즘(FTM)

변경한다. 그리고, F_0 고장의 경우, S_0에서 $|S_0|$-$|D1|$개의 높은 주소를 갖는 노드들과 S_1에서 $|S_1|$-$|D0|$개의 높은 주소를 갖는 노드들은 자신이 수신한 메시지를 무시하며, F_1 고장의 경우도 유사하다.

3.2 알고리즘의 예

그림 4와 5는 출발지 0에서 8개의 목적지 1, 2, 3, 5, 7, 10, 11, 14로 멀티캐스트 메시지를 전송할 때, FTM의 단계별 라우팅 예를 보이고 있다. 이 때 MIN은 2단에서 001, 110과 1 단에서 011, 100 등 4개의 고장 SE를 갖는다. 즉, 모든 고장 SE는 F_1에 속해 있고, F_1 고장이다. 멀티캐스트 목적지는 목적지 영역 집합 D = {[0001,0011], [0101,0101], [0111,0111], [1010,1011], [1110,1110]}으로 표시되며, $D0$ = {[0001,0011], [0101, 0101], [0111,0111]}과 $D1$ = {[1010,1011], [1110,1110]}으로 구분된다. $|D0|$ = 3 이고 $|D1|$ = 2이므로 k는 3이다. 단계 1에서는 그림 4와 같이, 출발지 0에서 크기가 6 (=2×3)인 임의의 중간 목적지 영역 IR = [0010, 0111]로 메시지를 보낸다. 여기서, 출발지가 0 (=0000), 즉 s_0이 0이고 F_1 고장이므로, $IR \subseteq M(0)$인 IR을 임의로 선택한다.

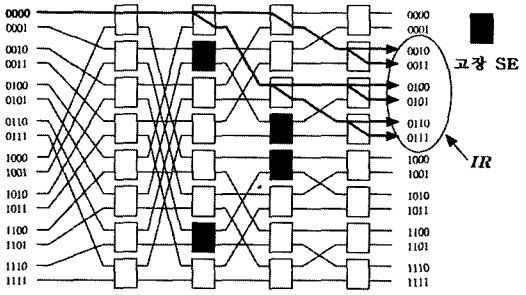

그림 4 FMT 단계 1의 예(메시지의 복사를 위한 라우팅)

그림 5는 단계 2에서의 라우팅 경로를 보여준다. IR은 S_0 = {0010, 0100, 0110}과 S_1 = {0011, 0101, 0111}로 구분된다. 그림 5의 굵은 실선 화살표와 같이, S_0의 중간 출발지 0010, 0100, 0110은 수신된 메시지를 각각 [0001,0011], [0101,0101], [0111,0111]로 전송한다. 또한, 그림 5의 굵은 점선 화살표와 같이, S_1의 중간 출발지 노드 0011과 0101은 수신된 메시지를 각각 [1010, 1011]과 [1110,1110]으로 전송한다. 이 때, S_1의 0111은 자신의 메시지를 버린다. 그림 4와 5에서 보는 바와 같이, 모든 메시지는 고장 SE들을 통과하지 않으며, 또한 2 단의 SE 011, 100과 1 단의 SE 001, 110 등도 통과하지 않음을 알 수 있다.

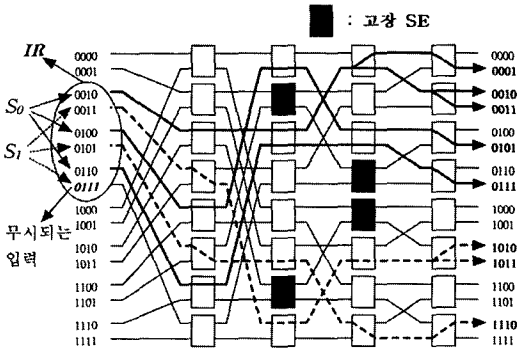

그림 5 FTM 단계 2의 예(최종 목적지 영역으로 라우팅)

3.3 알고리즘의 정확성

FTM은 고장난 MIN에서 문제를 발생시키지 않고 임의의 목적지를 갖는 멀티캐스트 메시지를 라우팅 할 수 있음을 보이고자 한다. 단계 1과 단계 2에서 모두 고장을 회피해야 하며, 단계 1에서 최종 목적지 영역 개수 이상의 크기를 갖는 IR로 메시지를 라우팅 할 수 있어야 한다. 또한 단계 2에서 라우팅되는 다수 메시지들 사이에 충돌이 발생하지 않아야 된다.

[보조정리 1] $NR(m)$를 m-비트 ($m \geq 1$)로 구성할 수 있는 영역의 전체 개수라고 할 때, $NR(m) \leq 2^{m-1}$이다.

증명 : m-비트 ($m \geq 1$)로 구성할 수 있는 영역들을 $R_1, R_2, \cdots, R_{NR(m)}$ ($R_1 <_D R_2 <_D \cdots <_D R_{NR(m)}$) 라고 하자. 그러면, 모든 j ($1 \leq j \leq NR(m)$)에 대해 $|R_j| \geq 1$이므로, 임의의 영역에 속한 주소들의 전체 개수는 $\sum_{j=1}^{NR(m)} |R_j| \geq NR(m)$이다. 또한, 모든 k ($2 \leq k \leq NR(m)$)에 대해 R_k-$R_{k-1} \geq 1$ 이므로, 인접한 영역 사이에 존재하는 주소들의 전체 개수는 $\sum_{k=2}^{NR(m)} (R_k - R_{k-1}) \geq NR(m)-1$이다. 그리고, m-비트 ($m \geq 1$)로 표현 가능한 주소의 전체 개수는 2^m 이다. 따라서, $NR(m)+(NR(m)-1) = 2 \times NR(m)-1 \leq 2^m$ 이며, $NR(m)$은 정수이므로 $NR(m) \leq 2^{m-1}$이다. □

목적지 주소가 모두 홀수이거나 모두 짝수인 경우에 가장 많은 영역을 구성할 수 있다. 3-비트로써 최대 개수로 구성 가능한 영역들은 [000,000], [010,010], [100, 100], [110,110]이며, 그 개수가 4이다.

[보조정리 2] 메시지의 출발지를 $s=s_{n-1}\cdots s_1 s_0$, 목적지를 $d=d_{n-1}\cdots d_1 d_0$ 로 표시할 때, MIN에서 F_0 고장이면, $s_0 \neq d_{n-1}$인 메시지는 고장 SE를 통과하지 않으며,

F_1 고장이면, $s_0=d_{n-1}$인 메시지는 고장 SE를 통과하지 않는다.

증명 : 먼저, F_0 고장인 경우를 살펴본다. MIN의 특성-1에 따라, i ($n-2 \geq i \geq 1$) 단에서 사용되는 SE는, $s_0=0$ 이고 $d_{n-1}=1$ 인 메시지이면 $1\gamma 0$ ($|\gamma|=n-3$) 형태이며, $s_0=1$ 이고 $d_{n-1}=0$ 인 메시지이면 $0\gamma 1$ ($|\gamma|=n-3$) 형태이다. 따라서, $s_0 \neq d_{n-1}$인 메시지는 F_0에 속한 고장 SE를 통과하지 않는다. F_1 고장인 경우도 같은 방식으로 증명할 수 있다. □

예를 들어, 만일 MIN이 F_0 고장이라고 가정할 때, 0000과 같이 $L(0)$에 속한 출발지를 갖는 메시지는 1000과 같이 $M(1)$에 속한 모든 목적지로 고장을 피해 라우팅이 가능하다. 또한 0001과 같이 $L(1)$에 속한 출발지에서 0000과 같이 $M(0)$에 속한 모든 목적지로 라우팅이 가능하다. FTM에서는 두 단계에서 모두 이와 같이 메시지를 라우팅하여 고장을 피할 수 있다.

[보조정리 3] 출발지가 각각 $L(0)$와 $L(1)$에 속한 임의의 두 메시지의 목적지가 서로 다르면, 두 메시지는 서로 충돌하지 않는다.

증명 : 두 메시지의 목적지를 각각 $a_{n-1}\cdots a_1a_0$과 $b_{n-1}\cdots b_1b_0$라고 하자. MIN의 특성-1에 따라, 두 메시지는 j ($n-1 \geq j \geq 1$) 단에서 같은 SE를 통과하지 않는다. 만일, $a_j = b_j$ ($n-1 \geq j \geq 1$)이면, 두 메시지는 0 단에서 같은 SE $a_{n-1}\cdots a_1$에서 만나지만, a_0과 b_0는 같을 수가 없으므로 그 SE에서 다른 출력 포트 a_0과 b_0로 라우팅 된다. 따라서, 충돌이 발생되지 않는다. □

[보조정리 4] 메시지 m의 출발지와 목적지 영역을 각각 s^m과 R_m이라고 표시할 때, $s^j < s^k$와 $R_j <_D R_k$ 인 두 메시지 j와 k가 $s^k - s^j \leq R_k - R_j$이면, 두 메시지는 충돌을 일으키지 않는다.

증명 : 정의-4에 따라, $d^j \in R_j$ 와 $d^k \in R_k$ 인 모든 d^j와 d^k에 대해 $R_k - R_j \leq d^k - d^j$ 이다. 즉, $s^k - s^j \leq R_k - R_j$이면 $s^k - s^j \leq d^k - d^j$ 이므로, 특성-2에 의해 두 메시지 사이에 충돌은 없다. □

예를 들어, 출발지가 0011이고 목적지 영역이 [1010, 1011]인 메시지와 출발지가 0101이고 목적지 영역이 [1110,1110]인 메시지는 서로 충돌이 없다. 왜냐하면, 0011과 0101이 모두 $L(1)$에 속해 있지만, 0101-0011은 2이고, [1110,1110]-[1010,1011]은 3이기 때문이다.

[정리 1] F_0 고장 또는 F_1 고장이 있는 MIN에서, 알고리즘 FTM은 영역 부호화 방식으로 표현된 임의의 멀티캐스트 메시지를 라우팅 할 수 있다.

증명 : F_1 고장은 F_0 고장과 동일하게 유추할 수 있으므로, F_0 고장에 대해 알고리즘의 정확성을 증명한다. 먼저, 임의의 멀티캐스트 목적지 수를 갖는 목적지 영역 집합으로 라우팅이 가능한 지를 살펴본다. 단계 1을 보면, 보조정리 1에 의해 $|D0| \leq 2^{n-2}$ 이고 $|D1| \leq 2^{n-2}$ 이므로, $|IR| \leq 2^{n-1}$이다. 또한 $|M(0)| = |M(1)| = 2^{n-1}$ 이다. 따라서 $|IR| \leq |M(0)| = |M(1)|$ 이다. 단계 2를 보면, $|S_0| = |IC|/2 \geq |D0|$ 이고 $|S_1| = |IC|/2 \geq |D1|$ 이다. 그러므로, 만일 메시지가 고장 SE를 통과하려 하거나 메시지간에 충돌이 없다면 FTM은 임의의 목적지를 갖는 멀티캐스트 메시지를 라우팅 할 수 있다.

메시지가 고장 SE를 통과하려는 문제를 살펴보면, 두 단계 모두 $L(0)$에 속한 출발지는 $M(1)$에 포함된 목적지로 라우팅하고, $L(1)$에 속한 출발지는 $M(0)$에 포함된 목적지로 라우팅한다. 그러므로, 보조정리 2에 의해 모든 메시지는 고장 SE를 통과하지 않는다.

단계 1에서는 하나의 메시지만 라우팅 되므로 충돌 문제를 고려할 필요가 없고, 단계 2에서 라우팅 되는 임의의 두 메시지간에 충돌이 발생할 가능성을 살펴보자. $S_0 \subseteq L(0)$, $S_1 \subseteq L(1)$이므로, 보조정리 3에 의해, $S_0 \Rightarrow^{|D1|} D1$ 의 메시지들과 $S_1 \Rightarrow^{|D0|} D0$ 의 메시지들 사이에는 충돌이 없다. $S_0 \Rightarrow^{|D1|} D1$에 해당되는 메시지들의 실질적인 중간 출발지 집합을 $\{s^1, s^2, \cdots, s^{|D1|}\} \subseteq S_0$, 목적지 영역 집합 $D1$을 $\{R_1, R_2, \cdots, R_{|D1|}\}$라고 하자. $1 \leq j < |D1|$ 인 임의의 j에 대해, $s^{j+1} - s^j = 2$ 이고 $R_{j+1} - R_j \geq 2$ 이며, $|R_j| \geq 1$ 이다. 따라서 $1 \leq j < k \leq |D1|$ 인 모든 j와 k에 대해, $s^k - s^j \leq R_k - R_j$이다. 그러므로, 보조정리 4에 따라, $S_0 \Rightarrow^{|D1|} D1$ 인 메시지들 사이에 충돌이 없다. 또한, 같은 방식으로 $S_1 \Rightarrow^{|D0|} D0$ 의 메시지들 사이에 충돌이 없음을 증명할 수 있다. □

3.4 다수 멀티캐스트 메시지의 처리

FTM 알고리즘은 고장이 있는 MIN에서 단일 멀티캐스트 메시지를 라우팅 할 수 있다. 본 절에서는 다수의 멀티캐스트 메시지가 동시에 라우팅 될 때 발생할 수 있는 교착상태와 이를 해결하는 기술에 대해 설명한다. 내부 버퍼가 없는 스위칭 소자로 구성된 MIN은 여러 개의 메시지에 의해서 교착상태가 발생할 수 있으며[11], 고장난 MIN에 대한 FTM 알고리즘 역시 동일한 현상이 발생 가능하다. 그림 6은 출발지가 1000이고 목적지가 [0001, 0100] 인 메시지 $m1$과 출발지가 1110이고 목적지가 [0011, 0110] 인 메시지 $m2$ 간에 교착상태가 발생한 예를 보여준다.

그림 6에서 보는 바와 같이 메시지 $m1$과 $m2$가 1단에 있는 SE a와 SE b에서 충돌이 발생하였다. SE a에서는 메시지 $m1$이 선택되어 두 출력 포트로 전송되며, 하단 출력 포트로 향하는 메시지 $m2$는 블록킹된다. 또

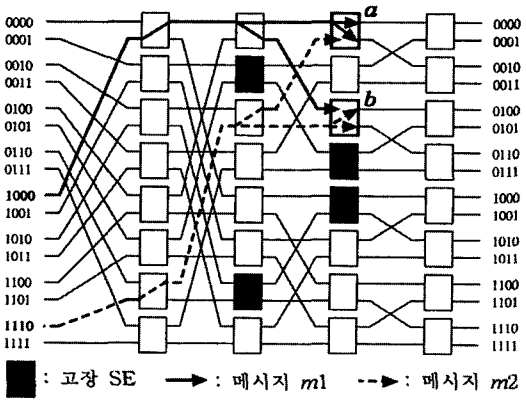

그림 6 교착상태의 발생 예

한, SE b에서는 메시지 $m2$가 선택되어 두 출력 포트로 전송되며, 상단 출력 포트로 향하는 메시지 $m1$은 블록킹된다. 즉, 메시지 $m1$은 SE a의 모든 출력 포트 자원을 점유한 채 SE b의 출력 포트 자원을 요구하며, 메시지 $m2$는 SE b의 모든 출력 포트 자원을 점유한 채 SE a의 출력 포트 자원을 요구한다. 따라서, 두 개의 메시지에 의해 교착상태가 발생한다.

본 FTM 알고리즘은 모든 SE에서 출력 포트에 대한 충돌이 발생할 경우, 상단 입력 포트로 들어온 메시지를 우선적으로 처리하는 상단 입력 포트 우선 기법[11]을 적용하여 교착상태를 방지한다. 그림 6에서와 같이 SE a와 SE b에서 두 메시지간에 충돌이 발생할 경우, 두 SE가 상단 입력 포트로 들어온 메시지인 $m1$을 우선적으로 선택하여 출력 포트로 라우팅 하므로써 교착상태를 피한다. 상단 입력 포트 우선 기법을 사용하면 다수의 멀티캐스트 메시지를 교착상태 없이 전송할 수 있다는 것은 기존의 연구 결과[11]에서 이미 증명되었다.

4. 결 론

본 논문에서는 다수의 스위칭 소자가 고장난 MIN에서의 멀티캐스팅 알고리즘을 제안하였다. 제안된 알고리즘은 연속된 주소를 하나의 영역으로 나타내는 영역 부호화 방식을 사용하여 임의의 목적지를 갖는 멀티캐스트 메시지를 구성하였다. 그리고, 멀티캐스트 메시지가 고장을 통과하는 것을 피하기 위해, MIN 상에서 메시지를 반복하여 순환시키는 기법을 사용하였다. 첫 번째 순환에서 멀티캐스트 목적지 영역 개수보다 큰 중간 목적지 영역으로 메시지를 라우팅 한 후, 두 번째 순환에서 중간 목적지 영역에 속한 노드들은 수신된 메시지를 최종 목적지 영역들로 각각 라우팅을 수행한다. 따라서, 제안된 알고리즘은 고장난 MIN에서 메시지를 두 번 순환시켜 멀티캐스팅을 수행할 수 있다. 또한, 전체 스위칭 소자들을 같은 크기를 갖는 두 개의 부분집합으로 구분하고, 모든 고장 스위칭 소자가 동일한 부분집합에 속한 MIN에서 임의의 멀티캐스트 메시지를 라우팅 할 수 있음을 증명하였다. 그리고, 동시에 다수의 멀티캐스트 메시지가 라우팅 될 때 발생 가능한 교착상태를 회피하는 방법을 제시하였다. 본 알고리즘은 모든 고장 스위칭 소자가 동일한 부분집합에 속한 형태의 고장만을 허용하는 제한점이 있으므로, 보다 일반적인 형태의 고장을 허용할 수 있도록 확장하는 것을 향후 연구 과제로 고려할 수 있다.

참 고 문 헌

[1] G. B. Adams, D. P. Agrawal, and H. J. Siegel, "A Survey and Comparison of Fault-Tolerant Multistage Interconnection Networks", IEEE Computer, Vol. 20, pp.14-27, Jun. 1987.

[2] 박재현, 윤현수, 이홍규, "적응 자기 경로제어 알고리즘을 사용하는 고장 감내 다단 상호연결 네트워크의 설계 및 신뢰성 분석", 한국정보과학회 논문지, Vol. 22, No. 7, pp.1066-1077, Jul. 1995.

[3] S. J. Wang, "Distributed Routing in a Fault-Tolerant Multistage Interconnection Network", Information Processing Letters, Vol. 63, No. 4, pp.205-210, 1997.

[4] A. Varma and C. S. Raghavendra, "Fault-Tolerant Routing in Multistage Interconnection Networks", IEEE Transactions on Computers, Vol. 38, No. 3, pp.385-393, Mar. 1989.

[5] S. Chalasani, C. S. Raghavendra, and A. Varma, "Fault-Tolerant Routing in MIN-Based Supercomputers", Journal of Parallel and Distributed Computing, Vol. 22, No. 2, pp.154-167, Aug. 1994.

[6] N. Das and J. Dattagupta, "Two-Pass Rearrangeability in Faulty Benes Networks", Journal of Parallel and Distributed Computing, Vol. 35, pp.191-198, Jun. 1996.

[7] P. K. McKinley, Y. Tsai, and D. F. Robinson, "Collective Communication in Wormhole-Routed Massively Parallel Computers", IEEE Computer, Vol. 28, No. 12, pp.39-50, Dec. 1995.

[8] R. Cusani and F. Sestini, "A Recursive Multistage Structure for Multicast ATM Switching", Proc. of IEEE Infocom, pp.1289-1295, Apr. 1991.

[9] X. Chen and V. Kumar. "Multicast Routing in Self-Routing Multistage Networks", Proc. of IEEE Infocom, pp.306-314, Apr. 1994.

[10] C. S. Raghavendra, X. Chen, and V. P. Kumar, "A Two Phase Multicast Routing Algorithm in Self-Routing Multistage Networks", Proc. of Int'l Conference on Communications, pp.1612-1618, Jun. 1995.
[11] 박재형, 윤현수, "다단계 상호 연결망에서 제한-주소 부호화를 이용한 재귀적 멀티캐스트 알고리즘", 한국정보과학회 논문지(A), Vol. 24, No. 7, pp.667-674, Jul. 1997.
[12] C. L. Wu and T.-Y. Feng. "On a Class of Multistage Interconnection Networks", IEEE Transactions on Computers, Vol. 29. No. 8, pp.694-702, Aug. 1980.
[13] J. Y. Hui, "Switching and Traffic Theory for Integrated Broadband Networks", Kluwer Academic Publishers, 1990.
[14] C. Chiang and L. M. Ni., "Multi-Address Encoding for Multicast", Proc. of the Parallel Computer Routing and Communication Workshop, pp.146-160, May 1994.
[15] T. T. Lee, "Nonblocking Copy Networks for Multicast Packet Switching", IEEE Journal on Selected Areas in Communications, Vol. 6, pp.1455-1467, Dec. 1988.

김 진 수
1979년 3월 ~ 1983년 2월 서울대학교 컴퓨터공학과(학사). 1983년 3월 ~ 1985년 2월 KAIST 전산학과(석사). 1993년 3월 ~ 1998년 8월 KAIST 전산학과(박사). 1985년 4월 ~ 2000년 2월 한국전기통신공사 선임연구원. 2000년 3월 ~ 현재 건국대학교 컴퓨터·응용과학부 조교수. 관심분야는 상호연결망, 병렬 처리, 초고속 네트워크, 네트워크 보안

장 정 환
1979년 3월 ~ 1983년 2월 경북대학교 전자공학과(학사). 1983년 3월 ~ 1985년 2월 KAIST 전산학과(석사). 1993년 3월 ~ 1998년 8월 KAIST 전산학과(박사). 1985년 4월 ~ 2000년 8월 한국전기통신공사 선임연구원. 2000년 9월 ~ 현재 부산외국어대학교 컴퓨터전자공학부 전임강사. 관심분야는 상호연결망 및 그래프 응용, 초고속통신망, 통신망 보안