분산 다중경로를 갖는 오류허용 다단계 상호연결망

이명숙* 손유익**
계명대학교 컴퓨터공학전공
e-mail : mslee@kmu.ac.kr yeson@kmu.ac.kr

A Fault-tolerant MIN with Distributed Multiple Paths

Myung-Suk Lee*, Yoo-Ek Son**
Department of Computer Eng. Keimyung University

요 약

다단계 상호연결망(MIN)은 대규모 병렬처리 능력을 가지는 대표적 연결망 구조이다. 그러나 MIN은 입출력 사이의 단일경로와 블록킹 네트워크를 가지는 특성으로 인해 시스템 성능저하를 가져올 수 있다. 이러한 문제를 개선하기 위해 버퍼를 이용하거나 적은 하드웨어 추가와 스위치 대역폭 확장이 가능한 링크를 N배만큼 확장시키는 방법을 사용한다. 본 논문에서는 기존에 오류를 허용하기 위해 제안된 구조들보다 좀더 많은 오류를 허용하기 위한 방법으로 입력버퍼와 확장경로를 사용하여 HOL 블록킹을 방지하고 오류허용 기능을 향상시키는데 그 목적이 있다. 이에 따라 네트워크에 다중 오류가 발생하더라도 부하분산을 통해 이를 허용할 수 있는 구조를 제안하고 시뮬레이션을 통해 그 성능을 평가한다.

1. 서론

상호연결망의 구성이나 처리방법은 대규모 분산·병렬처리 시스템에서 성능과 밀접한 관계를 가지며, 그 연결 방법에 따라 시스템의 확장성 및 대용량 데이터 처리 시 응답 시간에 많은 영향을 미치게 되므로 많은 연구들이 이루어져 왔다[1].

일반적으로 상호연결망은 연결방법에 따라 공유버스시스템, 크로스바 스위치, 다단계 상호연결망으로 구분된다. 공유 버스 구조는 구현이 쉬운 반면 버스 대역폭의 제한으로 인해 시스템의 확장이 불가능하며, 크로스바 연결은 높은 시스템의 성능을 보장하지만 상호 연결망의 비용이 너무 높아 큰 규모의 시스템에는 적합하지 않다. 이에 반해 MIN구조는 각 단계에서 스위치 비용이 추가되지만 대용량의 시스템으로 확장 시에 하드웨어 복잡도를 낮추면서도 높은 성능을 유지하는데 적합한 구조로 알려져 왔다[2]. 따라서 확장성 있는 병렬 컴퓨터를 구성하기 위한 상호연결 구조에 많이 사용되고 있으며 이를 통하여 병렬 컴퓨터에서 프로세서들 사이의 효율적인 데이터 통신을 가능하게 하며 이는 시스템의 성능향상에 중요한 요인이 된다[3].

네트워크는 여러 개의 스위치들이 모여서 하나의 스테이지를 형성하고 각 스테이지 사이는 특정한 라우팅 알고리즘으로 연결되고 있으며, 상대적으로 높은 처리율, 낮은 지연, 낮은 셀 손실률, 높은 오류허용 차수, 적은 비용 등의 특징을 가지고 있다. 그러나 입력과 출력 사이에 단 하나의 경로만 존재하는 구조적 특성으로 블록킹 문제가 발생할 수 있으며 이는 전체 네트워크의 성능을 감소시키는 원인이 되기도 한다[4]. 이러한 문제점을 해결하기 위해 현재까지 스위치 소자 추가, 네트워크 중복, 스위치 I/O포트 확장, 스테이지 추가, 링크 추가 등 오류를 허용하는 방법들이 많이 연구되어 왔다[5]. 이들 중 링크를 추가시켜 해결하는 방법이 비교적 적은 하드웨어로써 스위치 대역폭 확장이 가능하며 이와 관련된 연구로서는 2-dilated MIN, Augmented network, 2D-ELMIN, 다중경로 버퍼구조 등이 있다.

본 연구에서는 입력버퍼 구조를 갖는 네트워크를 대상으로 이러한 구조에서 발생되는 HOL(Head Of Line) 블록킹을 방지하고, 추가된 링크를 사용하여 경로를 우회시킴으로서 다중의 스위치 소자의 오류를 허용하는 방법에 관하여 언급한다. 성능평가를 위해서 비교대상의 구조들과 비교한 시뮬레이션 결과, 하나의 스테이지에서 두 개 이상 스위치에 오류가 발생하면 full access 기능을 상실하여 처리율이 급격히 떨어지지만 제안된 구조에서는 비교대상의 구조들보다 좀더 많은 오류를 허용하고 분산된 경로와 입력버퍼가 HOL 블록킹 문제를 해결함으로써 처리율이 높음을 알 수 있다. 또한 링크를 추가한 경우 오류허용을 향상시킨 기존의 방식들과 비교함으로써 셀 손실률, 셀 지연, 처리율면에서의 장점을 실험결과를 통해 보인다.

2. 다단계 상호연결망

MIN의 기본구성은 네트워크의 크기가 N인 경우 stage당 m^{n-1}개의 스위치로 구성되며($N=m^n$, $m \times m$ 스위치 소자), 인접한 스테이지에 있는 스위치는 임의의 입력에 임의의 출력으로 모든 셀 전송을 가능하

게 하는 full access 속성을 만족시키는 방법으로 연결되어 있다[6]. 크로스바는 내부적으로 넌블럭킹 네트워크이지만 N^2의 복잡도를 가지므로 네트워크의 크기가 제한적일 수밖에 없다. 네트워크 내에서 스테이지 사이의 상호 연결 라인은 각 입력에서 출력으로 하나의 유일한 경로를 형성하는 방법으로 만들어지며, 이러한 다단계 상호 연결망에서의 임의의 입출력 사이의 경로 설정은 셀프 라우팅에 의해 분산 형태로 이루어진다.

MIN은 여러 가지 장점도 가지고 있지만 내부경쟁으로 인한 블록킹이 발생하는 특성을 가지는데 이 블록킹으로 인한 성능저하를 보완하기 위한 여러 가지 방법 중의 한 방법으로 링크를 추가하여 내부경쟁을 감소시키는 것이다. 이중 대표적인 구조가 병렬반얀 네트워크, 텐덤반얀 네트워크, 확장반얀 네트워크 등이다. 이들 네트워크를 구성하기 위한 단위 스위치 소자의 복잡도는 각각 $O[N(\log N)^2]$, $O[N(\log N)^2]$, $O[N\log N(\log\log N)]$ 이 되며 이들 3가지 네트워크 중에서는 내부링크를 N배 확장하여 구성한 확장 반얀 망이 가장 적은 비용으로 동일한 성능을 나타내 보이고 있다[7].

따라서 본 논문에서는 반얀 구조의 확장 반얀 네트워크를 이용하여 경로를 확장시켜 증가된 라우팅 경로를 분산시킴으로서 HOL 블록킹으로 인한 손실을 줄임으로써 처리율을 향상시키고자 하였다.

3. 제안된 네트워크 구조

기본 베이스라인 네트워크에서의 셀프 라우팅은 입출력간의 단일 경로가 존재하게 된다. 제안된 네트워크에서는 각 스위치 소자에서 두 개의 여분 입력링크와 출력링크 그리고 각각 입력 버퍼모듈로 구성되어 있다. 스위치 소자에 추가된 두 개의 링크는 두 개의 스테이지 사이의 경로를 확장시켜 다음 스테이지의 버퍼에서 블록킹이 발생하거나 링크나 스위치 소자에서 오류가 발생하면 여분의 링크를 사용하게 된다. 특히 첫 단과 마지막 단에서 경로를 분산시켜 줌으로써 더 많은 오류를 허용할 수 있도록 하였다.

본 연구에서는 베이스라인 네트워크를 대상으로 네트워크내의 스위치 구조는 내부 연결 링크가 이중 구조의 형태를 취하고 있다고 가정한다. 베이스라인 네트워크의 링크레벨의 표현은 0에서 $\log_2 N-1$까지 연속적으로 표현되며, 스위치 소자 l은 $(\log_2 N-1)$이 되고 이진표현으로는 $p_l p_{l-1}\cdots p_1$이 된다. 각 레벨 내의 링크는 $p_l p_{l-1}\cdots p_1$까지는 동일하지만 p_0=0이면 링크가 스위치 소자의 상위 출력포트에 연결되고, p_0=1이면 하위 출력포트에 연결된다.

다음의 식(1), 식(2)는 본 논문에서 제안한 구조에서 추가된 여분 링크의 상호연결식이다. 여분 링크는 기본 베이스라인 네트워크의 rule 식에서 상수 값을 보수화 시켰을 때의 패턴과 같다. 식(1)은 상위 링크를 설정하기 위해 i번째 스위치 소자 $(p_l p_{l-1}\cdots p_1)_i$에서 i+1번째 스위치 소자의 연결경로가 $(p_l \cdots p_{l-i+1}0p_{l-i}\cdots p_2)_{i+1}$ 에 의해서 설정되고, 식(2)는 하위 링크를 설정하기 위해 i번째 스위치 소자 $(p_l p_{l-1}\cdots p_1)_i$ 에서 i+1번째 스위치 소자의 연결경로가 $(p_l \cdots p_{l-i+1}1p_{l-i}\cdots p_2)_{i+1}$ 에 의해서 설정된다.

$$\beta'^{0}_{i}[(p_l p_{l-1}\cdots p_1)_i] = (p_l \cdots p_{l-i+1}\,\bar{0}\,p_{l-i}\cdots p_2)_{i+1},$$
$$\text{for link } (p_l p_{l-1}\cdots p_1 0)_{i+1},\ 0 \le i < l \quad \text{식(1)}$$

$$\beta'^{1}_{i}[(p_l p_{l-1}\cdots p_1)_i] = (p_l \cdots p_{l-i+1}\,\bar{1}\,p_{l-i}\cdots p_2)_{i+1},$$
$$\text{for link } (p_l p_{l-1}\cdots p_1 1)_{i+1},\ 0 \le i < l \quad \text{식(2)}$$

스위치 소자 내에서의 경로 선택방법은 모든 스테이지에 있는 스위치 소자는 같은 동작을 하는데 스위치 소자에 도착한 셀 들은 연결된 다음 단의 스위치 소자 혹은 링크에 오류 및 블록킹이 발생하지 않았을 경우 정상경로를 통해서 셀이 전송되고 오류 및 블록킹이 발생했을 경우에는 여분 경로를 통해서 셀 전송이 시도된다. 한번 여분 경로를 통해 전송된 셀은 목적지까지 여분 경로를 통해서만 전송된다. 단, 마지막 스테이지에서 블록킹이 발생할 경우 셀을 폐기시킨다는 가정을 둔다. 4개의 입력포트에서 전달된 셀들은 일반적인 셀프 라우팅에 의해 출력포트로 전달된다.

스위치 소자의 라우팅 알고리즘은 네트워크의 입력은 $S = s_{n-1}s_{n-2}\cdots s_0$, 출력은 $D = d_{n-1}d_{n-2}\cdots d_0$로 가정하고 라우팅 알고리즘은 베이스라인 네트워크와 동일하다. 라우팅은 i가 스테이지 수$(\log_2 N-1)$ 만큼 반복되는데, i번째 스테이지의 스위치 소자에 셀이 도착하면 i+1번째 스테이지의 스위치 소자에게 보낸 요구에 대한 응답에 따라 경로가 설정된다.

라우팅 알고리즘은 표 3-1과 같으며 목적지 주소에 의해 응답이 이루어지고 $0 \le i \le n-1$, 목적지 주소의 i번째 비트가 만약에 d_i=0이면 상위경로로 전송되고, d_i=1이면 하위 경로로 전송이 이루어진다. 여기에서 $[\beta(\beta'), i, s]$는 스위치 소자의 정상(여분) 링크로 출력, 단, 스위치번호]를 나타낸다.

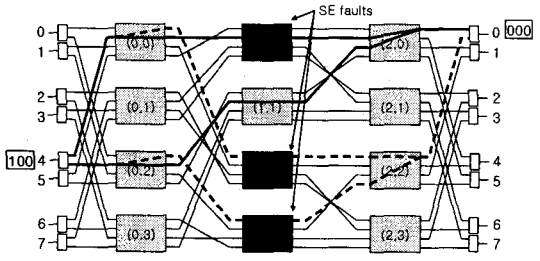
표 3-1. 스위치 소자 라우팅 알고리즘

```
    TR : routing tag
    D : destination output
    n=log2N-1 : no. of stage

begin
TR = D - dn-1dn-2···di···d0

for i=0 to n-1
    case di;
      1 : lower output of SE
      0 : upper output of SE
    end case
    if [β, i+1, next s] is fault then
      send cell to [β', i, next s]
    else
      send cell to [β, i+1, next s]
    if [β', i+1, next s] is fault then
      blocked
end
```

[그림 1]은 기존의 제안된 구조에서 경로를 다른 스위치로 우회시켜 여러 개의 스위치로 부하를 분산시키는 구조를 제안함으로써 이전에 제안된 구조들보다 좀더 많은 스위치 오류를 허용하고 HOL 블록킹을 감소시키는 부하분산 ELMIN(Load Distributing ELMIN : LD-ELMIN)구조이다. 8×8 크기의 제안된 네트워크 구조에서 오류가 발생했을 때 4(100)에서 0(000)으로 가는 라우팅 경로의 예를 나타낸 것이다. 첫 번째 스테이지에서 인접한 세 개의 스테이지에 오류가 발생하거나 두 번째 스테이지에서 세 개의 스위치 소자에 오류가 발생하더라도 MIN의 속성인 full access를 만족하고 목적지까지 여분의 경로를 통해 셀이 라우팅 되는 경로를 보여주고 있다.



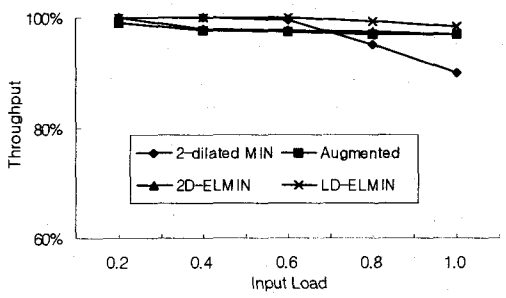
[그림 1] LD-ELMIN의 오류 발생시 라우팅 예

4. 성능평가

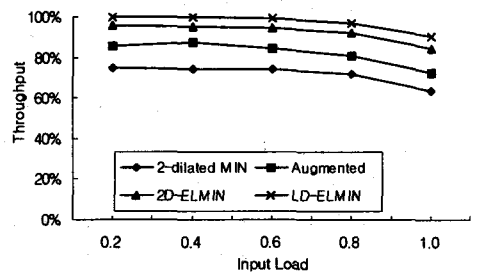
본 논문에서는 네트워크의 성능평가를 위해 이산적인 모델링 방식으로 시뮬레이션 하였으며, 셀 손실률, 처리량, 셀 지연 면에서 성능을 평가하였다. 성능분석은 다음과 같은 가정을 사용한다.

모든 패킷들은 같은 고정크기의 길이를 가지고 있고 동시에 전송된다. 네트워크의 각 소스에 패킷 도착 프로세서는 poisson process를 따르고, 도착 시간은 지수분포에 따라 발생된다. 서비스 원칙은 FCFS이며, 소스에 도착하는 패킷은 모든 목적지로 균등하게 분산된다. 하나의 입력링크는 셀을 저장할 수 있는 내부 버퍼를 가진다. 따라서 하나의 포트 즉, d개의 입력링크 d개의 내부 버퍼를 가진다. 출력링크 속도는 망 내의 내부링크 속도보다 같거나 빨라야 한다. 하나의 스위치에서 여러 패킷이 같은 목적지를 요구할 때 랜덤하게 임의의 하나가 선택되어 전송되고 나머지 하나는 다른 경로를 통해 전송된다. 시뮬레이션 종료는 10,000개의 셀을 생성하여 버퍼에 있는 모든 셀이 처리 될 때까지 시뮬레이션을 동작시킨다.

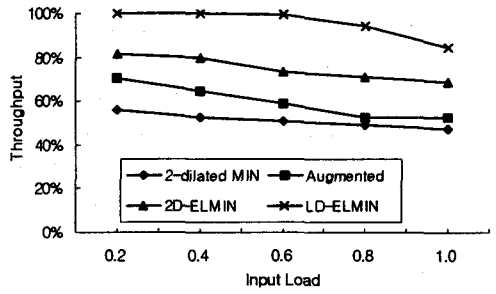
본 논문에서는 10,000개의 고정 셀을 발생시키고 전체 시뮬레이션에서 발생된 총 셀 수중에 총 출력 셀 수의 비율로 처리율을 정의하고 오류 수에 따른 처리율을 측정하였다. [그림 2]는 두 번째 스테이지를 중심으로 스위치 소자에 오류가 하나도 발생하지 않았을 경우, [그림 3]은 한 개의 오류가 발생했을 경우, [그림 4]는 두 개의 스위치 소자에 오류가 발생했을 경우의 처리율을 나타낸다. 스위치 소자의 오류수가 증가할수록 비교대상 구조의 처리율이 급격히 감소한다. 그 이유는 오류가 발생하면 MIN의 특성인 full access를 만족하지 못하기 때문이다. 제안된 구조에서도 두 개의 오류가 발생했을 때는 full access를 만족함에도 불구하고 처리율이 감소하는데 그 이유는 많은 스위치 소자에 오류가 발생하면 블록킹을 해결하기 어렵기 때문이다.



[그림 2] no fault일 때, 부하에 따른 처리율



[그림 3] single fault일 때, 부하에 따른 처리율



[그림 4] double fault일 때, 부하에 따른 처리율

셀 손실은 스위치 소자에 입력된 총 셀의 개수에 대해 출력포트로 출력되지 못하고 오류나 블록킹 등의 원인으로 손실되는 셀로 정의하고 셀 지연은 입력포트로 들어온 셀이 출력포트로 셀이 출력될 때까지 버퍼에서 대기한 평균대기시간으로 정의한다. 제안된 구조는 셀 손실이 낮고 지연도 비교대상의 모델에 비해 작았다. 그 이유는 처음과 마지막 단에서 부하를 분산시켜 줌과 동시에 여분의 경로로 인해 블록킹 발생을 줄여주기 때문이다.

5. 결 론

본 논문에서는 오류를 허용하기 위한 방법으로 기존의 베이스라인 네트워크에서 부가 경로를 확장하고 첫 번째와 마지막 단에 상호연결 패턴을 변형하여 블록킹을 감소시키는 입력버퍼를 이용한 다단계 상호연결망 구조를 제안하였다. 비교대상의 구조와 제안된 구조에서는 하나의 소스에서 목적지로 연결된 경로가 여러 개 있으나 비교 대상 모델에서는 한 스테이지에서 두 개의 스위치 소자를 통해 연결되어 있다. 그러나 제안된 구조에서는 여러 스위치 소자를 통해 분산되어 목적지에 이른다. 특히 두 번째 스테이지에서는 네 개의 모든 스위치 소자로 경로가 분산되어 있고, 처음과 마지막 단에서 경로를 분산시켜 주므로 비교대상의 구조들에 비해 더 많은 스위치 소자의 오류를 허용하는것이 가능하다.

성능분석 방법으로는 AweSim 시뮬레이터를 이용하여 8×8 크기를 가진 네 가지 모델을 입력부하와 오류 수에 따른 시뮬레이션을 수행하여 각 모델에 대한 처리율, 지연시간, 셀 손실율 등을 측정하였다.

시뮬레이션을 통한 분석결과, 제안된 구조는 다른 모델에 비해 높은 처리율과 낮은 셀 손실을 가져왔고, 특히 오류 수에 따른 처리율에서는 오류수가 많아질수록 제안된 구조의 처리율이 더 높은 것으로 나타났다. 비용분석에서는 스위치 소자 내의 버퍼적용은 이중버퍼를 적용하여 모든 구조를 동일한 조건에서 시뮬레이션 하였다. 비용을 줄이기 위한 방법으로 단일버퍼를 적용하는 방법이 있으며 제안된 구조에서 단일 버퍼를 적용하기 위해서는 마지막 스테이지에 다른 알고리즘을 적용해야 한다. 따라서 향후 과제로 제안된 구조에서 싱글 버퍼를 적용할 수 있는 스위치 동작 알고리즘에 대한 연구가 이루어져야 할 것이다.

6. 참고 문헌

[1] 권택근, 초고속 통신망, 홍릉과학출판사, 1996.

[2] Itoh, A., "A fault-tolerant switching architecture for ATM networks," IEEE International Conference on Communications, Vol. 3, pp. 1639-1645, 1992.

[3] Kamiura, N., Kodera, T., Matsui, N., "Design of a Fault Tolerant Multistage Interconnection Network with Parallel Duplicated Switches," IEEE International Symposium on DFT'00, 2000.

[4] Kraimeche, B., "Design and analysis of the Stacked-Banyan ATM switch fabric," Elsevier Science Computer Networks 32, pp. 171-184, 2000.

[5] Awdeh, Ra'ed Y., Mouftah, H. T., "Survey of ATM switch architectures," ELSEVIER Computer Networks and ISDN Systems, Vol. 27, No. 12, pp. 1567-1613, 1995.

[6] Feng, T., "A Survey of Interconnection Networks," IEEE Trans. on Computers, Vol. 14, pp. 12-27, 1981.

[7] 송효정, 권보섭, 윤현수, "내부 버퍼가 있는 확장 베넌 네트워크의 성능분석", 정보과학회 논문지 (A), 26권, 25호, pp. 595-601, 1999.