2001년 한국정보처리학회 춘계 학술발표논문집 제8권 제1호

비균일 트래픽하의 공정한 멀티캐스트 스위치

손동욱*, 손유익
계명대학교 컴퓨터공학전공
psalm8@hcc.ac.kr, yeson@kmucc.kmu.ac.kr

A Fair Multicast Switch under Nonuniform Traffic

Dong-Wuk Son*, Yoo-Ek Son
Department of Computer Science Keimyung University

요약

본 논문은 작은 fanout에 대한 불공정성과 hot-spot의 문제를 해결하기 위해 공정하게 입력포트에 접근하여 복사망으로 들어갈 수 있는 멀티캐스트 스위치를 제안하고자 한다. 제안된 스위치는 큰 fanout에 대한 작은 fanout을 가진 입력포트에 도착한 패킷의 불공정한 대우를 해결하여 시스템 전체 지연 시간을 줄여 산출량을 극대화할 수 있다.

1. 서론

화상회의, 오락용 비디오, 파일 분배와 같은 멀티캐스트는 광대역 ISDN에 필수적으로 요구되어진다. 이러한 다지점 통신을 제공하기 위한 필수적 요소는 멀티캐스트 패킷 스위치이다. 멀티캐스트 스위치의 대부분은 복사망과 점대점 라우팅망을 직렬 연결함에 의해 멀티캐스트 기능을 제공한다. Lee에 의해 제안된 멀티캐스트 패킷 스위치가 전형적인 예이다[1]. Lee의 멀티캐스트 스위치의 복사망은 RAN(Running Adder Network)과 DAE (Dummy Address Encoder)의 집합, 방송 반얀망으로 구성되며, self-routing, nonblocking 특성과 일정 지연을 가지고 있다.

그러나 Lee의 복사망의 문제점은 오버플로우 문제와 입력에 있어서 불공정성이다. 입력에 있어서 불공정성은, 입력포트에 도착하는 패킷은 상위 포트가 하위 포트보다 우선 하므로 상위 포트의 fanout이 최대가 되어지면 맨 하위 포트는 계속해서 다음 사이클로 미루어지는 점이다. 결과적으로 폭주를 피하기 위해 스위치 구조는 전송에 보다 많은 clock rate이 운영되어져 왔다. 이런 단점은 복사망의 사용 능력을 제한한다[2]. 아울러 hot spot일 경우에도 그와 같은 불공정성이 발생한다. hot spot 비 균일 트래픽은 특정의 출력포트에 대해 동시에 많은 요구가 발생하는 것으로 일정 균일 트래픽에 부과된 접근률 보다 훨씬 높게 단일 출력포트에 집중되어지는 것을 말한다. 그러므로 hot spot의 시스템 지연은 들어오는 셀의 비 균일 분산의 결과로서 증가될 수 있다.

본 논문은 이러한 불공정성과 hot spot의 문제를 해결하기 위해 공정하게 입력포트에 접근하여 복사망으로 들어갈 수 있는 멀티캐스트 스위치를 제안하고자 한다. 균일 트래픽뿐만 아니라 비 균일 트래픽인 hot spot에도 공정한 접근과 복사가 가능한 제안된 복사망은 공유버퍼와 fanout 분할, 그룹분할을 하는 방법을 사용한다. 제안된 복사망은 큰 fanout에 대한 작은 fanout을 가진 입력포트에 도착한 패킷의 불공정한 대우를 해결하여 시스템 전체 지연 시간을 줄여 산출량을 극대화할 수 있다.

2. 불공정성에 대한 기존 연구

2.1 SSP 망

Lee의 복사망과 Turner의 방송패킷 스위치는 요구된 패킷의 복사 전체 개수가 출력포트의 수를 초과하는 오버플로우의 문제를 가지고 있다. 그러므로 복사망 내에 오버플로우가 발생하지 않도록 하기 위해서는 흐름제어가 필수적이다. 흐름제어는 패킷 주소 체계에 기초를 두고 있어, 오버플로우를 발견할 때 보다 높은 번호를 가진 입력에서의 패킷은 블록 및 재전송 되어진다. 그러나 이것은 패킷이 복사망에 접근하는 데 있어서 불공정하다. 이런 불공정의 문제를 해

결하기 위해, complementary RAN이 추가되어지고 두 개 망은 전진과 후진 모드로 교대로 작동되게 하는 방법이 제안되었다. 후진 모드에서 running adder는 포트 $N-1$에서 i까지 복사본 수의 running sum i를 생성하며, 전진모드에서는 0에서부터 i까지 running sum을 계산한다. 결과적으로 주소 체계는 높은 번호의 포트 쪽으로 편중되어진다[3].

불공정의 문제는 복사망의 앞 끝에 SSP(Shift Sequence Permutation)을 cascading함에 의해 해결되어질 수 있다. SSP 망은 한 타임슬롯에서 입력포트를 회전한다. 기본적으로 타임슬롯 t에서 입력포트 j에서 $1 \le j \le N$에 대해 출력포트 $(j+t) \bmod N$에 map되어진다. 그것을 SSP index라 한다. 이런 방법에서 SSP망의 각 입력포트는 타임-평균 기초 위에서 복사망 접근에서 동일한 기회를 가진다. SSP 망은 $N \times N$ 반얀망에 의해 구현되어질 수 있다. $0 \le i \le N-1$에 대해 $(i,i+1,....N-1,0,....,.i-1)$로 정수$(0,1,.....N-1)$의 i shift sequence permutation (i-SSP)를 정의한다. $N \times N$ 반얀망은 목적지 집합이 $0 \le i \le N-1$에 대해 i-SSP를 형성할 때 블로킹 없이 복사 및 라우팅을 완성한다.

그림 1은 SSP를 보여준다. 한 슬롯 타임에 trigger 되어지는 n-비트 이진 카운터(CNT)는 AM(Address Mapper)에 cyclic sequence $\{0,1....N-1\}$을 제공한다. AM은 modulo N adder의 집합이다. AM의 기능은 CNT로부터 입력포트 번호의 각각을 더함에 의해 SSP를 생성한다. SSP망은 우선 순위 복사를 이루기 위해 batcher sorting망과 같은 sorter 네트워크와 결합되어질 수 있다.

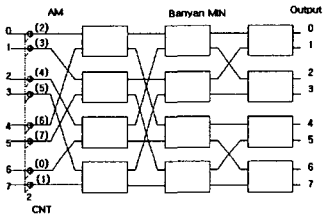

그림 1 SSP Network

2.2 CDN 복사망

CDN 복사망은 CDN(Cyclic Distribution Network)과 CRP(Contention Resolution Processor), BBN (Broadcast Banyan Network), TNT(Trunk Number Translator)의 집합으로 구성되어있다[4].

CDN의 기능은 master 셀을 CRP로 cyclic하게 분배하는 것이며, 모든 CRP가 균일하게 공유되어지도록 해준다. CDN은 running adder 망과 역 반얀망으로 구성되어있다. CDN의 구축은 $K \times K$(K : 네트워크의 크기) reverse 반얀망이 입력 셀이 연속적인 출력 주소 modulo K를 가진다면 nonblocking이다라는 특성에 기초를 둔다. $K \times K$ running adder망은 들어오는 active 셀의 수의 합을 modulo K로 처리한다. 합의 결과에 따라 셀은 역 반얀망에 의해 적절한 CRP로 라우팅 되어진다. 또한 RAN의 최종 출구에서 합의 결과는 최종 출구를 RAN의 첫 입력으로 연결함에 의해 다음 타임슬롯에서 사용되어진다. 그러므로 각 타임슬롯에서 CDN에 의해 받아들여진 셀은 cyclic 방법으로 CRP로 분배되어진다.

CRP는 토큰링의 제어 하에 서로간 조정되어진다. 확장 BBN을 수용함에 의해 복사 생성 원칙이 동일하게 유지되는 한 출력의 수는 입력의 수보다 훨씬 커질 수 있다. CRP의 기능은 CDN에 의해 분배된 master 셀을 저장하고 FIFO 방식으로 셀을 처리하며, 요구된 복사본의 수만큼 BBN의 연속 출력을 예약하기 위한 master 셀의 헤드를 갱신하여 갱신된 master 셀을 BBN으로 전송하는 것이다.

BBN은 앞에 있는 스위치 요소가 셀 복사 외에 셀 라우팅을 추가한 것을 제외하고는 반얀망이다. 입력은 $0,1,2....(K-1)$로 번호가 붙여진 $K \times K$ BBN을 가정한다. BBN은 Lee의 복사망에서 nonblocking 특성을 가진다고 증명하였다[1].

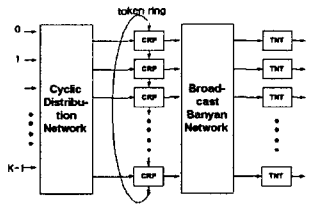

그림 2 CDN Network

3. 제안 구조

3.1 스위치 구조

앞에서 각 입력포트의 접근에 대한 공정성을 기하기 위한 방법으로 입력포트를 회전하거나 토큰 링을 사용해 입력의 편중성 문제를 해결하였다. 하지만 입력포트의 접근에 있어서 공정성을 가지지만 큰 fanout에 대한 비해 작은 fanout을 가지고 입력포트에 도착한 패킷의 희생을 해결할 수 없다. 최악의 경우 높은 번호를 가진 포트들은 적어도 몇 사이클 후에야 복사를 위해 반얀망에 들어갈 수 있다. 공유버퍼가 존재하는 경우에는 버퍼에 저장되어지지만, 버퍼가 없는

경우에는 fanout sum을 계산한 후 공정하게 입력단에 도착하였더라도 패킷은 복사되어지지 못하고 폐기되어 차후 사이클에 다시 재 전송되어진다.

제안 스위치의 구조는 큰 fanout을 나누기 위한 fanout 분할과, fanout sum에 의해 한 사이클 내 전송될 수 있는 패킷 분할을 하기 위한 그룹 분할[6], 블로킹되어질 패킷을 보존하기 위한 공유버퍼[7]과 복사를 위한 반얀망으로 구성되어있다.

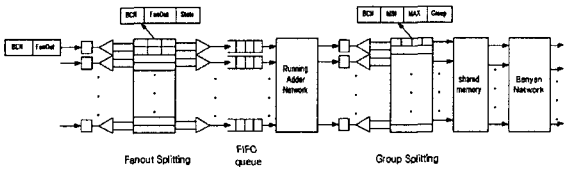

그림 3 제안된 스위치 구조

3.2 Fanout 분할과 그룹분할

Fanout 분할은 RAN(Running Adder Network) 앞에서 큰 fanout을 작은 fanout으로 나누는 역할을 하게 된다. 이렇게 하는 이유는 큰 fanout에 의해 작은 fanout의 희생을 막기 위해서이다. 큰 fanout은 알고리즘에 따라 나뉘어져 먼저 전송되어지는 패킷과 다음 사이클에 전송되어지는 패킷으로 나누어진다. 먼저 전송되어지는 패킷에 대한 패킷 내 표시는 state 필드에 의해 표시되어지며 state 필드는 *"0"*, *"1"*의 값을 가진다. *"0"*은 분할된 상위 패킷을 의미하며, 우선 순위를 가지고 FIFO queue 먼저 진입한 다음, RAN으로 들어가서 fanout sum을 계산하게 된다. fanout 분할 알고리즘이 아래에 나와있다.

```
if Fanout_i > log2N
    Fanout_i0 = log2N
    Fanout_i1 = N- log2N
else
    Fanout_i0 = Fanout_i
```

$Fanout_{i0}$에서 i는 입력포트, *"0"*은 상위 패킷 및 우선순위를 의미한다. log_2N은 fanout을 나누기 위한 threshold 값으로 적정 fanout 분할 값을 의미한다. N은 네트워크의 크기이다.

그룹분할은 RAN에서 계산된 fanout sum에 따라 한 사이클 내에 전송되어질 수 있는 그룹을 분할하기 위해 사용되어진다. 그룹분할 알고리듬은 아래와 같다. G_i는 입력포트 i의 그룹번호이며, fanout sum에 의해 그룹 값이 결정되어진다. 예를 들면 입력포트 i의 fanout sum이 6이면 *000 110*$_{(2)}$으로 표시하여 앞의 세 자리가 그룹 번호(G_i)가 되어지고, 뒤의 세 자리는 포트 번호가 되어 진다.

```
i가 0이면 G_i = 0
if G_i = G_i-1  no packet 분할
  //G_i : 입력 포트 i의 그룹 번호
else if G_i > G_i-1 // packet 분할
      Fanout_i0 = N - Fanout_i
      Fanout_i1 = Fanout_i- Fanout_i0
      G_i0   = G_i-1
      G_i1   = G_i
else  overflow , i'th packet discard
min_i = Fanout_i-1
max_i = Fanout_i -1     (if i=0, min_i = Fanout_N-1)
```

위의 알고리듬에 따라 패킷 분할한 그림이 그림 4에 나와 있다.

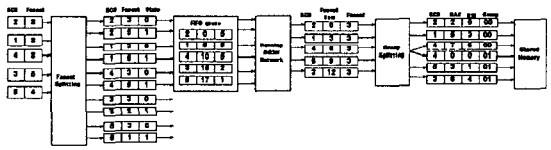

그림 4 Packet 분할

3.3 복사 과정

그림 5는 공유메모리와[5] 반얀망과의 관계를 나타내고 있다. 공유 메모리에는 다음 사이클에서 전송되어질 그룹번호 *"01"*이 저장되어있고 그룹번호 *"00"*은 복사를 위해 반얀망으로 들어간다. 입력포트에서 셀은 목적지 주소에 따라 self-routing에 의해 출력포트로 나가게 된다.

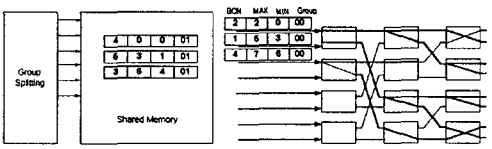

그림 5 패킷 복사

하지만 반얀망 내부에서 블로킹의 문제를 피할 수 없다. 여러 개의 입력 셀이 같은 출력포트를 요구하는 경우, 오직 한 셀만이 통과 및 진행 가능하기 때문에 하나의 출력포트에 대하여 경쟁을 벌이게 되는 내부 블로킹이 발생하고 스위치의 수행 능력이 떨어지게 된다. 이러한 경우 내부 링크의 처리율을 높이거나

한번에 하나의 셀만 처리하여야 한다. 반얀망이 nonblocking 특성을 가지는 위해서는 반얀망 앞에 sorting 네트워크를 붙이거나 목적지 주소로 sorting 하는 방법을 사용한다.

3.4 결과 분석

시뮬레이션 수행 시 fanout과 제공된 입력 부하를 변수로 설정하고 변수의 변화가 미치는 영향에 따른 성능의 변화를 관찰하였다. 시뮬레이션의 결과의 분석과 비교를 위하여 사용된 각 용어와 성능 평가의 정의는 다음과 같다.

ⓐ 입력 부하 : 스위치의 각 입력 단에 대하여 매 주기마다 새로운 셀이 도착할 확률.

ⓑ 산출량 : 매 주기마다 출력되는 셀의 개수. 본 논문에서는 임의의 제한 시간 내 네트워크의 출력링크를 통과한 셀의 합으로 정의한다.

ⓒ 셀 손실률 : 스위치에 입력된 총 셀의 개수에 대해 출력단으로 출력되지 못하고 손실되는 셀의 비율

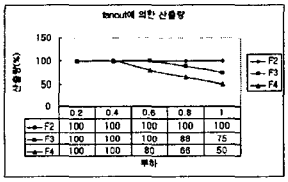

그림 6 Fanout에 의한 산출량

그림 6은 fanout 분할에 있어서 log_2N의 그 적정성을 실험하였다. 실험 복사망은 망의 크기가 N이 8인 8×8 반얀망으로 log_2N으로 분할이 산출량에 있어서 우수함을 보여준다. $F2$의 표기는 fanout을 2로 반얀망에 적용한 모델을 의미한다. fanout이 4 이상부터는 산출량이 급격히 떨어진다.

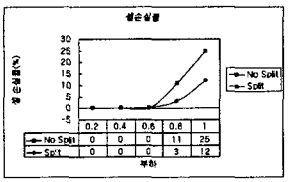

그림 7 셀 손실율

그림 7은 fanout 분할을 적용한 모델과 적용하지 않은 모델의 셀 손실율의 비교이다. 부하가 0.6인 위치까지는 거의 같은 셀 손실율을 기록하지만 그 이상에서는 확연히 구분이 되어지고 있다. 그러므로 적은 fanout을 가진 패킷을 먼저 처리하고 큰 fanout을 다음 사이클에 처리하는 것이 전체 셀 손실률을 줄일 수 있다.

4. 결론

Lee의 복사망의 문제점은 오버플로우 문제와 큰 fanout의 처리, 그리고 입력에 있어서 불공정성이다. 입력에 있어서 불공정성은, 입력포트에 도착하는 패킷은 상위 포트가 하위 포트보다 우선 하므로 상위 포트의 fanout이 최대가 되어지면 하위 포트는 계속해서 다음 사이클로 미루어지는 점이다.

이런 문제점을 해결하기 위한 제안된 스위치는 공유버퍼와 fanout 분할, 그룹분할을 하는 방법을 사용한다. 그러므로 본 논문은 큰 fanout에 대한 작은 fanout을 가진 입력포트에 도착한 패킷의 불공정한 대우를 해결하여 시스템 전체 시간을 줄여 산출량을 증가시켰다.

참고문헌

[1] Tony T. Lee, "Nonblocking Copy Networks for Multicast packet Switching", IEEE Journal on Selected Areas in Comm., Vol. 6, No. 9, pp. 1455-1467, Dec. 1988.

[2] J.S. Turner, "A Practical Version of Lee's Multicast Switch Architecture," IEEE Trans. on Communications, Vol. 41, No.8, pp. 1166-1169, August 1993.

[3] C.L.Tarng, J.S.Meditch, A.K.Somani, "Fairness and Priority Implementation in Non-Blocking Copy Network", International Conference on Communications, pp. 1002-1006, 1991.

[4] Wen De Zhong, Yoshikuni Onozato, Jaidev Kaniyil, "A Copy Network with Shared Buffers for Large-Scale Multicast ATM Switching", IEEE/ACM Trans. Networking, Vol. 1, No. 2,, pp. 157-165. 1993.

[5] Feihong Chen, Bülent Yener, ALi N. Akansu, Sirin Tekinary, "A Novel Performance Analysis for the Copy Network in a Multicast ATM Switch," Proceedings of the Int'l Conference on Computer Communications and Networks , pp. 99-106, 1998.

[6] Xinyi Lju and H. T. Mouftah, "A Dynamic Cell -Splitting Copy Network Design for ATM Multicast Switching," Global Telecommunications Conf., 1994. GLOBECOM '94. Comm.: The Global Bridge., IEEE , pp. 458 -462, Vol.1, 1994.

[7] Xinyi Liu, H.T Mouftah, "Overflow Control In Multicast Networks", Proc. of Canadian Conf. on Electrical and Computer Engineering, Vancouver, B.C., pp. 542-545, 1993.