고속 패킷 전송을 위한 멀티캐스트 스위치

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

Multicast Switch for High Speed Packet Transmission

Dong-Wuk Son, Yoo-Ek Son
Department of Computer Eng. Keimyung University

요 약

본 논문은 초고속통신망에서 멀티캐스트 스위칭에서 발생될 수 있는 오버플로우 문제와 블로킹 문제를 보다 효율적으로 해결하고, 공정하게 입력포트에 접근함으로 작은 fanout에 대한 불공정성을 해결하기 위한 멀티캐스트 스위치를 제안하고자 한다. 제안된 스위치는 셀 분할(cell pre-splitting)과 공유된 버퍼, Group Spliiting, 그리고 그룹분할망으로 구성되어지며, 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.

오버플로우는 요구된 복사본의 총 개수가 출력포트의 개수를 넘어설 경우에 발생하며, 패킷을 잃어버리는 원인이 된다. 복사 요구의 총계가 출력포트 수보다 많을 때, 입력 부하가 커지면 커질수록 많은 셀들이 폐기되어진다. 이것은 망 대역폭의 이용을 떨어뜨려 산출량을 저하시키게 되는 원인이 된다.

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. 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 오버플로우의 문제를 해결하기 위해 공정하게 입력포트에 접근하여 복사망으로 들어갈 수 있는 멀티캐스트 스위치를 제안하고자 한다. 공정한 접근과 복사가 가능한 제안된 복사망은 공유버퍼와 cell Pre-Splitting, Group Splitting, 그룹분할망으로 구성되어있다. 제안된 복사망은 큰 fanout에 대한 작은 fanout을 가진 입력포트에 도착한 패킷의 불공정한 대우를 해결하여 시스템 전체 지연 시간을 줄여 산출량을 극대화할 수 있다.

2. 오버플로우와 공정성

오버플로우를 해결하는 방법은 스위치 내 속도를 증가시키거나, 블로킹 시 우회할 수 있는 다중 경로를 제공하거나, 입력 또는 출력 버퍼를 사용하거나 복사망 앞에 정렬망을 사용하는 것이다. fat 반얀망은 다중 경로를 사용해 오버플로우 문제를 해결하였다. 공정성의 문제는 입력포트를 회전하거나 토큰 링을 사용해 입력의 편중성 문제를 해결하고 CDN이 이러한 한 예이다.

$N \times N$ fat 반얀 스위치는 $N/2$개의 2×2 스위치 노드를 각각 가지는 $log_2 N$ 단계로 구성된다. 단계를 연결하는 링크의 확장은 입력으로부터 출력으로 점차 증가한다. L_1=2, L_2=3, L_3=4, L_i는 단계 i에서 fat 스위칭 요소(FSE)의 출력 확장이다. FAB 스위치를 위한 확장 구성은 행렬 벡터$[L_1, L_2, L_3]$로서 명시되어진다.

FAB 스위치에서 FSE의 라우팅 메커니즘은 셀프 라우팅 반얀 구조에 기초를 두고 있다. 주어진 단계에서 FSE는 입력에서 셀의 출력 주소의 대응된 비트를 조사한다. 비트가 "0"이면 입력은 FSE의 상단 출력으로, "1"이면 입력은 FSE의 하단 출력으로 연결된다. FSE의 출력에 대한 셀 경쟁의 수가 FSE의 출력 확장을 넘을 경우에는 중재가 필요하다. 중재는 FSE의 출력 집중기에서 상단에서 하단으로 셀에 우선 순위를 둠으로써 이루어진다. 이런 중재 메커니즘은 FSE 우선 순위가 집중기 입력의 왼쪽에서 오른쪽으로 주는 knockout 스위치 집중기와 유사하다. 대안적으로 셀은 집중기에서 임의적으로 선택되어질 수 있으나 이것은 FSE의 복잡도를 더하는 각 집중기에 대한 난수 생성기를 요구하게 된다.

CDN 복사망은 CDN(Cyclic Distribution Network)과 CRP(Contention Resolution Processor), BBN (Broadcast

Banyan Network), TNT(Trunk Number Translator)의 집합으로 구성되어있다. 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.

3. 멀티캐스트 스위치

3.1 스위치 구조

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 Cell Pre-Splitting과, fanout sum에 의해 한 사이클 내 전송될 수 있는 패킷 분할을 하기 위한 그룹 분할, 블로킹되어질 패킷을 보존하기 위한 공유버퍼와 Fat 반얀망을 개선 및 확장한 그룹 반얀망으로 구성되어있다. 공유 버퍼는 fanout의 요구 수가 출력포트의 수보다 클 경우에 셀들이 폐기되어지는 막기 위해 사용되어진다. 그룹분할 반얀망은 기존의 fat 반얀망을 개선하여 그룹별 다중전송이 가능하도록 공유버퍼와 연결한 모듈적 구조를 가지고 있다.

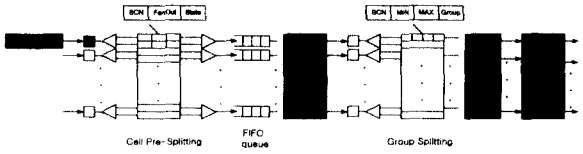

그림 1 제안된 스위치 구조

3.2 Cell Pre-Splitting과 그룹분할

Cell Pre-Splitting은 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. Cell Pre-Splitting 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 우선 순위를 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 포트번호가 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 2에 나와 있다.

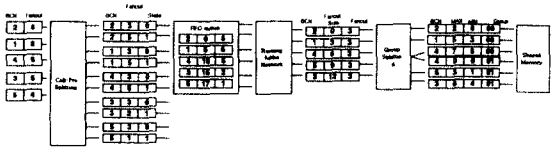

그림 2 Cell Pre-Splitting과 Group Splitting

3.3 그룹 반얀망

그림 3은 두 개의 그룹 반얀망을 사용하여 Fanout sum이 33의 요구에 따른 라우팅을 보여주고 있다. Lee의 스위치에서는 8개만 처리 가능하며, Turner의 방송 반얀망에서는 16개의 확장 링크를 통해 그 요구를 처리한다. 그러나 블로킹의 확률과 전송지연율은 fat 반얀망보다 훨씬 더 높다. 그룹분할은 공유 버퍼와 모듈적 구조로 오버플로우 패킷을 폐기하지 않고 다음 사이클에서 처리한다.

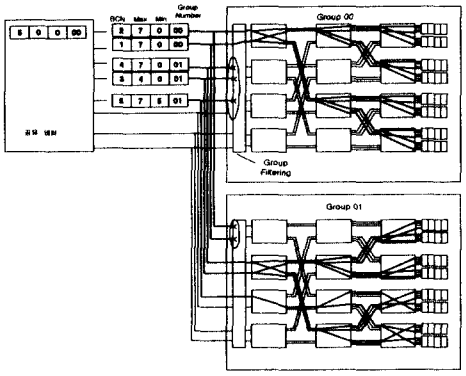

그림 3 두 그룹의 그룹 반얀망

그림 3에서 빗금으로 표시되어진 부분은 전송된 패킷이 그룹 필터링 되어지는 부분이다. 하지만 타 그룹으로 성공적으로 통과함에 의해 패킷은 목적지에 정확히 복사되어 전송된다.

4. Simulation

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

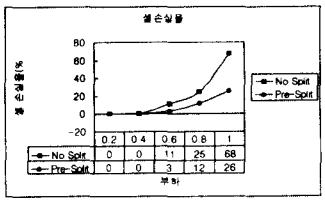

그림 4 셀 손실율

그림 4은 Cell Pre-Splitting을 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.

5. 결론

오버플로우와 공정성 문제를 해결하는 제안된 스위치는 입력 패킷의 요구 수에 따른 적절한 복사와 nonblocking 특성을 가지므로 보다 산출량과 셀 손실에 있어서 상당한 개선을 가져왔다. 특히 pre-splitting과 group splitting, 그룹 분할망은 산출량에서 좋은 결과를 보여주단. 또한 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] 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.

[3] 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.

[4] 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.

[5] 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.

[6] M. Alimuddin, H. M. Alnuweiri and R. W. Donaldson, "Efficient Multicast Copy Network," Broadband Switching Systems Proceedings, 1997. IEEE BSS '97., 1997 2nd IEEE Int'l Workshop on, pp. 169-172, 1997.