ATM 스위치용 스케줄러의 기술 동향

김 형 준* 손 승 일** 홍 성 찬***

◈ 목 차 ◈

1. 서 론

1990년대 중반 이후 인터넷이 대중화되면서 인터넷 트래픽은 폭발적으로 증가하고 있다. 인터넷 트래픽은 1970년대 중반부터 매년 2배 이상씩 증가해 왔고, 최근 몇 년 동안 연간 4~10배씩 증가하고 있다. 인터넷을 통해 전달되는 정보의 종류도 단순한 문자기반에서 비디오나 오디오 데이터와 같은 실시간 대용량 멀티미디어 정보가 점차 증가하고 있다. 이러한 인터넷 환경의 변화에 따라 라우터에도 고속화 대용량화가 요구되어, 기존의 라우터와는 다른 기술적 사항들이 요구되어지고 있다[1]. 수많은 고성능 IP 라우터, LAN 스위치와 ATM 스위치들은 VOQ구조를 갖는 ATM 스위치를 기반으로 한 교환 백플레인에 사용된다. 이 시스템들은 스위칭 패브릭을 통과하기 위한 패킷들을 입력 큐를 사용하여 대기시킨다. 만일 단순한 FIFO 입력 큐가 패킷들을 유지하기 위해 사용된다면, Head-Of-Line (HOL) 블록킹으로 인해 성취 할 수 있는 최대 throughput의 대략 58.6%에 제한된다[2]. 이러한 HOL 블록킹을 제거하기 위해 VOQ (Virtual Output Queueing) 구조를 갖는 ATM 스위치의 입력 버퍼 모듈에서는 수신한 패킷을 각 목적지 별로 구분하여 저장하는 VOQ 구조가 사용된다. 입력 버퍼에 저장된 데이터는 스케줄러에 의해 결정된 입출력 쌍으로 데이터를 스위칭 하여 데이터를 전송하게 된다. VOQ구조를 갖는 ATM 스위치에서 핵심이 되는 것이 입출력 쌍을 결정하는 스케줄러인데, 스케줄러에 사용되는 알고리즘에 따라 스위치의 성능이 결정된다.

본 논문에서는 현재 VOQ구조를 갖는 ATM 스위치에 사용되는 스케줄러 기술 발전에 대해 기술하였다.

2. VOQ 구조의 ATM 스위치

VOQ를 갖는 ATM 스위치의 구조는 그림 1과 같이 나타낼 수 있다.

입력 버퍼는 VOQ구조로 목적지별로 구분되어 큐잉이 이루어진다. 물리적으로는 하나의 메모리이지만 논리적으로 N개의 FIFO처럼 동작하게 운영한다.

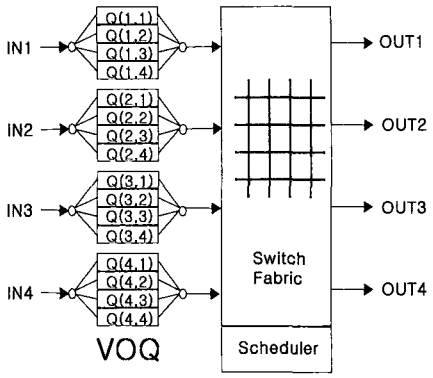
스케줄러에 의해 입출력 쌍을 결정하기 위한 가정은 다음과 같다.

1. 입력단(i)에 도착한 셀은 해당 목적지 출력단(j) 큐 Q(i,j)에 저장된다.
2. 전송할 셀이 있는 입력단은 스케줄러로 경로 설정을 요구한다.
3. 스케줄러는 매 슬럿마다 전송될 입력단 출력단 쌍을 결정하고 해당 VOQ ATM 포인트를 연결한다.

스케줄링에 있어 한 입력단에서 하나의 셀만을 선택해야 하며, 한 출력단으로 하나의 셀만 전송해야한다.

본 연구는 산업자원부 및 과학재단의 시스템 IC 2010의 일환인 "선행핵심 IP 연구 개발" 지원에 의해 이루어졌음.

* 호남대학교 컴퓨터공학과 석사과정
** 한신대학교 정보통신학과 조교수
*** 한신대학교 정보통신학과 부교수



(그림 1) VOQ를 갖는 ATM 스위치의 구조

매 슬럿마다 전송 선택되는 입출력단쌍의 수가 많아야 높은 Throughput을 얻을 수 있다.

3. ATM 스위치의 스케줄러

스케줄링 알고리즘으로는 PIM(Parallel Iterative Matching), 2DRR(Two-Dimensional Round Robin)이 처음 제시되었고 이어 WPIM(Wehgited PIM), iSLIP, MUCS(Matrix Unit Cell Scheduler), APSARA등의 알고리즘들이 차례로 제안되었다. 본 장에서는 이러한 스케줄러 중에서 기본적인 스케줄러 및 현재 주목받고 있는 스케줄러에 대해 설명할 것이다.

3.1 PIM(Parallel Iterative Matching)

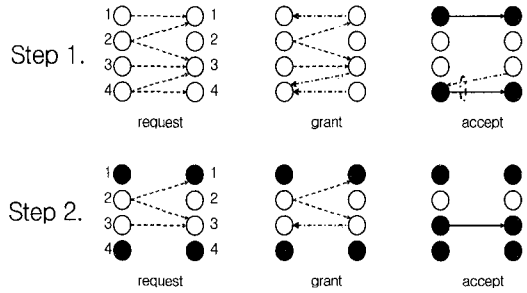
1993년 Anderson이 출력단 마다 서로 독립적인 출력단 중재기를 두고 입력단과 출력단 중재기 간에 request, grant, accept의 3단계 과정을 i번 반복하여 출력단 충돌 없이 전송할 수 있는 셀을 선택해 주는 PIM 알고리즘을 제안하였다. PIM 알고리즘의 각 단계는 다음과 같다.

그림 2는 PIM 알고리즘의 수행과정을 나타낸다.

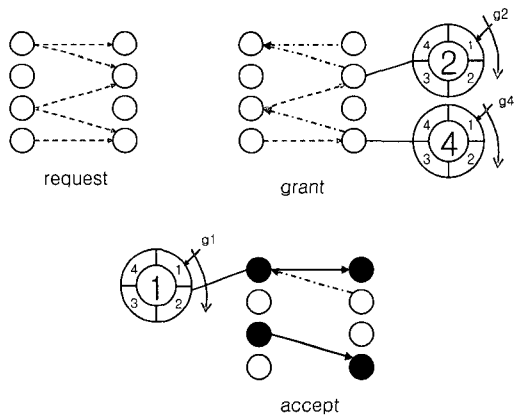
한번의 반복과정이 완료되면, 두 번째 반복과정에서는 매치된 입력단은 request를 보내지 않는다. 이러한 반복 과정을 16x16 교환기의 경우 4번 반복하면 98%정도의 매치를 찾아내는 것으로 알려져 있다[3].

PIM 알고리즘이 제안되고 구현에 성공함에 따라, 이와 유사하면서 PIM의 단점을 보완해주는 많은 알고리

1. request : 매치되지 않은 입력단은 자신의 입력 버퍼에 저장되어 있는 셀들의 목적지 출력단에 request를 보낸다.
2. grant : 매치되지 않은 출력단이 여러 입력단의 request들을 받으면 이들 중 하나를 랜덤하게 선택하고 그 request를 보낸 입력단에게 grant 신호를 보낸다.
3. accept : 입력단은 하나 이상의 grant 신호를 받을 수 있으며, 이들 중 하나를 랜덤하게 선택하고 해당 출력단에게 accept 신호를 보낸다.



(그림 2) PIM 알고리즘



(그림 3) RRM 알고리즘

즘들이 그 뒤를 이어 제안되었으며 대표적인 알고리즘으로 2DRR, WPIM, SLIP, MUCS, APSARA등이 있다.

3.2 RRM(Round Robin Matching)

RRM 방식은 PIM의 복잡성과 불공평성의 단점을 극복하기 위해서 Round-Robin 중재기의 2차원 배열을 구성함으로써 셀들이 각 입력과 출력에 대해 스케줄을 하는 방식이다.

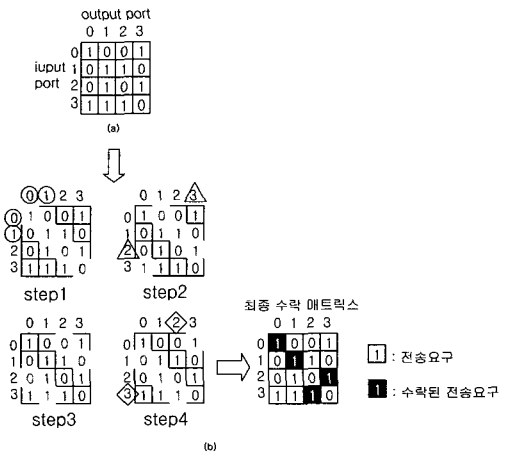
그림 3은 RRM 알고리즘의 수행과정을 나타낸다.

RRM 역시 PIM 알고리즘과 같이 3단계의 과정으로 구성된다. 1단계는 전송 요구 단계로 각 입력단은 자신의 입력 버퍼에 저장되어 있는 셀들의 목적지 출력단에 전송 요구를 보낸다. 2단계는 승인 단계로서 각 출력단은 라운드 로빈 스케줄에 따라 포인터를 증가시키고 한 전송 요구 입력단을 선택한다. 3단계는 마지막 수락단계로 각 입력단은 하나의 출력단을 선택하고, 입력단의 우선순위 포인터를 업데이트 시킨다.

3.3 2DRR(2-Dimensional Round Robin)

다수의 입력 큐를 사용하는 패킷 스위치에서 높은 처리율과 공평한 분배를 위해서 2DRR 스케줄 알고리즘이 제안되었다. 이 알고리즘은 전송 요구 매트릭스(Request Matirx)를 사용한다. 입력포트와 출력포트가 각각 N개인 가상 출력버퍼를 사용하는 스위칭 시스템의 경우 N^2개의 전송요구가 존재하며, 이를 나타낸 전송 요구 매트릭스에서 행과 열은 각각 입력포트와 출력포트를 나타낸다. 행렬에서 1이 명시되어 있는 곳은 행에 해당하는 입력포트와 열에 해당하는 출력포트의 가상 출력 입력큐에 적어도 하나 이상의 해당하는 출력단으로 전송요구를 하는 패킷(셀)이 존재한다는 것을 나타내고, 0은 전송할 패킷(셀)이 없어 비어있는 큐를 뜻한다. 스케줄링 과정은 다양한 제약 조건하에서 스위칭 패브릭의 입출력포트가 겹치지 않는 입출력 쌍의 집합을 찾는 과정이므로, 2차원 라운드 로빈 알고리즘은 입출력포트가 서로 겹치지 않는 N개의 대각선 전송요구들을 N 과정에 거쳐 차례로 전송 요구를 수락하는 과정을 거친다. 전송요구 수락과정은 현재 전송요구와 동일한 행 또는 열에 이미 수락된 전송요구가 없는 경우 현재 전송요구를 수락한다. 전송요구를 수락하는 대각선 패턴이 서비스를 받을 확률이 높으므로, 각 입출력포트의 공평성을 보장하기 위하여 스케줄링 시마다 전송요구 매트릭스의 대각선의 패턴의 순서를 변화시킨다.

그림 4는 2DRR 알고리즘의 수행과정을 나타낸다. a)는 전송요구매트릭스를 나타내며, 실제 수행과정은 b)에 나타나 있다.



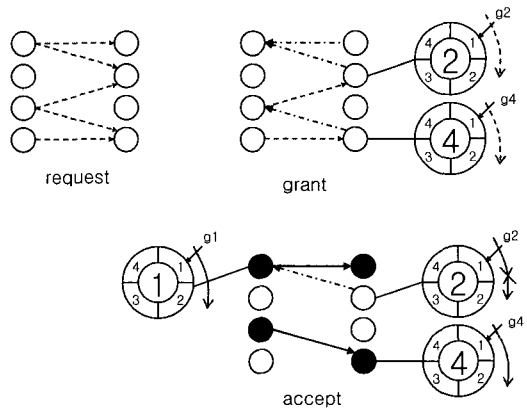
(그림 4) 2DRR 알고리즘

3.4 iSLIP

iSLIP 알고리즘은 RRM의 한 변형으로써 각 출력단 승인 포인터간의 동기화 현상에 의한 성능 저하 문제를 개선하였다. 즉, iSLIP은 승인이 수락되지 않으면 승인 포인터를 움직이지 않는다. iSLIP은 승인 포인터를 업데이트시키는 조건을 제외하면 RRM과 동일한 알고리즘이다.

그림 5는 iSLIP 알고리즘의 수행과정을 나타낸다.

iSLIP 알고리즘은 한 수행 과정은 입출력 쌍의 전송요구를 입력으로 받아 전송 입력 결정 및 전송 출력 결정 과정의 두 과정을 차례로 거쳐, 주어진 입출력 쌍의 집합으로부터 입력과 출력이 겹치지 않는 입출력 쌍의 부분 집합을 나타내는 전송 수락을 결정한다. 전송



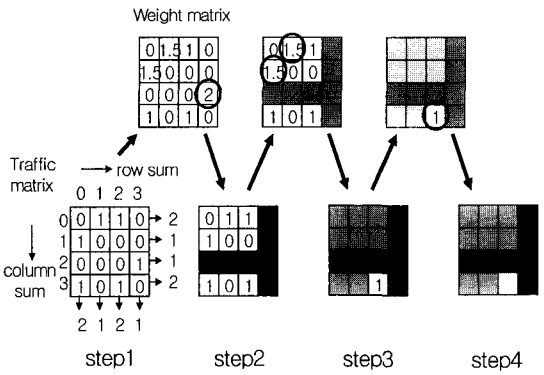
(그림 5) iSLIP 알고리즘

승인 과정에서는 입출력 쌍의 전송 요구 중 동일한 출력포트를 가지는 쌍들 중 입력이 겹치지 않도록 한 쌍을 선택하며, 전송 수락 과정에서는 위 전송 승인 과정의 결과에서 동일한 입력 포트를 가지는 쌍들 중 출력이 겹치지 않도록 한 쌍을 선택한다. 이 때 스케줄링 과정에서 전송할 수 있는 전송 수락의 수를 보다 많이 선택하기 위하여 위 전송 승인 과정 및 전송 수락 과정을 여러 번 반복한다. iSLIP 알고리즘은 한 스케줄링을 마친 후 전송 요구와 전송 수락의 포인터를 수정하며, 승인 신호를 받은 입력간의 수락 신호를 보내지 않으면 승인 포인터를 변형시키지 않는다. 즉, 승인 포인터는 가장 최근에 연결이 이루어진 입력단을 가리키게 된다. 이것은 부하가 높을 때 각 출력단의 승인 포인터를 비동기화 시킴으로써 처리율을 높이는 역할을 한다.

3.5 MUCS(Matrix Unit Cell Scheduler)

MUCS는 트래픽 매트릭스와 웨이트 매트릭스를 사용한다. 웨이트 매트릭스가 가장 큰 입출력단쌍(i,j)를 선택하고 i행과 j열을 트래픽 매트릭스에서 삭제함으로써 입력단 충돌과 출력단 충돌을 해소한다. 그리고 입출력단 충돌이 작은 입출력단 쌍을 선택하여 이후 과정에서 경합에 참여하는 입출력단 쌍의 수가 최대가 되게 하는 알고리즘이다. 그림 6은 MUCS 알고리즘의 수행과정을 나타낸다.

이 알고리즘은 입출력단 충돌이 작은 입출력단쌍을 선택하여 이후 과정에서 경합에 참여하는 입출력단쌍의 수가 최대가 되게 함으로서 최대수 매칭을 구하는 방법



(그림 6) MUCS 알고리즘

으로 높은 Throughput을 얻을 수 있다.

4. 스케줄러의 장단점 분석

PIM 알고리즘의 단점은 랜덤 선택에서 야기된다. 각 중재자가 시간이 변화하는 멤버중에는 랜덤 선택을 해야만 하기 때문에 이를 만족할 만큼 고속으로 구현하려면 비용이 높아지고 설계가 어렵게 된다. 스위치가 자신의 처리용량을 넘어섰을 때 PIM은 연결 알고리즘에서 불공평성을 야기할 수 있다.

RRM은 우선순위 부호기를 구현함으로써 랜덤 중재기보다 훨씬 빠르고 간단하게 스케줄 될 수 있고 우선순위가 순환하기 때문에 연결 요구들 사이에서 보다 공평하게 동등한 대역의 할당이 가능하다. 그러나, 출력단 중재기에 포인터를 업데이트시키는 규칙 때문에 load가 63%가 되면 RRM은 불안정하게 된다는 단점이 있다. RRM은 수락되지 않더라도 출력단의 포인터를 증가시키기 때문에 승인은 했으나 수락되지 않더라도 출력단의 포인터를 증가시키기 때문에 승인은 했으나 수락되지 않은 출력단은 불공평하게 되기 때문이다[2].

iSLIP 알고리즘은 승인 포인터는 가장 최근에 연결이 이루어진 입력단을 가리키게 된다. 이것은 부하가 높을 때 각 출력단의 승인 포인터를 비동기화 시킴으로써 RRM보다 처리율을 높이는 역할을 하고, 단일 매칭 수행시에도 높은 처리율을 보인다. ESLIP, PSLIP은 iSLIP을 개량하여 각 입력별 우선 순위등을 처리할 수 있도록 개량한 것이며, 이 알고리즘은 2DRR 알고리즘들에 비해 공평성과 성능이 좋은 장점이 있으나 구현시 지연 시간 및 면적 복잡도가 커 대용량의 스위칭 시스템에 적용하기 어려운 단점이 있다.

2DRR은 높은 처리율과 공평성을 나타내지만, 포트 수의 크기에 비례하여 반복수행 과정이 증가한다는 단점을 가지고 있다. 따라서 포트 사이즈가 커지면 스케줄링 많은 시간을 소요해야 하는 문제점을 안고 있다.

MUCS 알고리즘의 경우, PIM 알고리즘도 랜덤 선택방식으로 중재하여 입출력 충돌이 적은 입출력단쌍에 전송이 선택될 확률이 높아 입출력단쌍(1,1)과 (2,2)는 출력단 용량의 75%를 사용하고 (2,1)은 25%를 사용한다는 문제가 있었다.

(표 1) 주요 스케줄링 알고리즘의 요약

구분	제안자	특징	응용제품
PIM	Anderson	· 병렬성, 실시간성, 독립성 · 불공평성 존재	AN2(AT&T)
2DRR	Lamaire	· 구현이 용이 · 스케줄링 시간이 포트 크기에 비례	MultiGigabit Router
iSLIP	Mckeown	· 구현이 용이 · 고속 우선권 인코더 필요 · 공평성이 우수 · 높은 throughput	Tiny-Tera, GSR12000
MUCS	H. Duan 강성모	· weight를 적용한 스케줄 · 특정 상황에서 PIM 보다 높은 불공평성	iPOINT

MUCS는 이보다 더 심해 입출력단쌍(2,1)은 아예 선택이 되지 않는 Starvation이 발생한다.

이들 중, iSLIP 방식은 Cisco의 기가비트 라우터 및 Tiny-Tera 시스템에 적용되었고, 2DRR의 아류인 GWFA 스케줄링 방식은 BBN의 Multigigabit Router에 적용되었으며, MUCS는 일리노이 대학에서 iPOINT라는 과제를 통해 구현되었다.

표 1은 본 논문에서 기술한 주요 스케줄링 알고리즘을 요약하여 나타낸 것이다.

5. 결 론

본 논문에서는 고속의 VOQ ATM 스위치의 핵심 소자인 스케줄러의 기술 동향에 대해서 소개하였다. 전 세계적으로 인터넷의 사용이 확산되면서 인터넷 트래픽이 많아지고 있다. 스위치에도 고속화 대용량화가 요구되어, 기존의 스위치와는 다른 기술적 사항들이 요구되어지고 있다. 스위치의 성능을 결정하는 여러 가지 요소가 있겠지만, 그 중에 매 스위칭 슬롯마다 수행되는 스케줄러의 성능이 가장 중요하다. 이에 기존 스케줄러에 비해 빠른 성능을 갖고 비교적 적은 면적을 차지하는 스케줄러 및 새로운 스케줄링 알고리즘에 대한 연구가 필요하다. 뿐만 아니라 스위치 자체적으로도 매 전송 슬롯마다 수행되는 스케줄러의 부담을 줄이기 위한 새로운 방안이 연구되어야 할 것으로 사료된다.

참 고 문 헌

[1] 이형호, 김봉완, 안병준 "테라비트 라우터 기술", Telecommunications Review 제11권 2호, 2001년 3~4월.

[2] N. McKeown, "The iSLIP Scheduling Algorithm for Input Queued Switches", IEEE/ACM Trans. Networking, Vol.7, No.2, pp.188-201, April 1999.

[3] A. Mekkittikul, and N. McKeown, "A Practical Scheduling Algorithm to Achieve 100% Throughput in Input-Queued Switches", In Proc. IEEE Inforcom '98. Vol.2, pp. 792799, San Francisco, Apr. 1998.

[4] R. O. Lamaire and D. N. Serpanos, "Two-Dimensional Round-Robin Schedulers for Packet Switches with Multiple Input Queues" IEEE/ACM Trans. on Networking, vol.2, no.5, pp.471-482, Oct. 1994.

[5] Haoran Duan, John W. Lockwood, and Sung Mo Kang, "Matrix Unit Cell Scheduler (MUCS) for Input-Buffered ATM Switches," IEEE Communications Letters, pp. 20-23, Volume 2, Number 7, July 1998.

[6] B. Prabhakar and N. McKeown, "On The Speedup Required for Combined Input and Output Queue Switching", Automatica, Vol.35, No.12, Dec. 1999.

[7] S. Chuang and N. McKeown, "Matching Output Queuing with a Combined Input Output Queue Switch", in Proc. IEEE INFOCOM'99, pp.1169-1178, April. 1999.

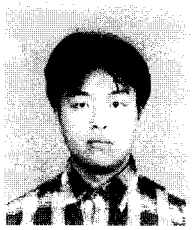
[8] C. Fujihashi and H. Hikita, "Speed-up of Input-Buffered Asynchronous Transfer Mode Switch by Introducing of Parallel Read-Out Structure", in Proc. IEEE Globecom'96, London, pp.819-824, Nov. 1996.

[9] N. McKeown, "Scheduling Algorithm for Input-Queued Cell Switches", PhD Thesis, University of California at Berkeley, May 1995.

[10] P. Gupta, N. Mckeown, "Designing and Implementing a Fast Crossbar Scheduler", IEEE Micro, vol.19, no.1, pp.20-28, Jan. 1999.

[11] Paolo Giaccone, Devavrat Shah, and Balaji Prabhakar, "An Implementable Parallel Scheduler for input-queued Switches", IEEE Micro, pp19-25, Jan. 2002.

◐ 저 자 소 개 ◑



김 형 준

2001년 호남대학교 컴퓨터공학과(학사)
2001년~현재 : 호남대학교 컴퓨터공학과 석사과정
관심분야 : ATM 통신, 네트워크, etc.

손 승 일

1989년 연세대학교 전자공학과(학사)
1991년 연세대학교 대학원 전자공학과(석사)
1998년 연세대학교 대학원 전자공학과(박사)
1998년~2002년 호남대학교 컴퓨터공학과 조교수
2002년~현재 : 한신대학교 정보통신학과 조교수
관심분야 : ATM 통신 및 보안, ASIC 설계, etc.

홍 성 찬

1979년~1982년 고려대학교 통계학과(이학사)
1988년~1990년 일본 게이오대학 이공학부 시스템공학 전공(공학석사)
1990년~1994년 일본 게이오대학 이공학부 시스템공학 전공(공학박사)
1994년~1995년 LG-EDS시스템(주) 컨설팅 부장
1995년~1996년 상명대학교 정보과학과 전임강사
1997년~현재 : 한신대학교 정보통신학과 부교수
관심분야 : XML, 인터넷비즈니스, 정보시스템응용