셀 스케줄러의 설계에 관한 연구

손승일* · 박노식**

*한신대학교

A Study on Design of Cell Scheduler

Seung-il Sonh* · Noh-sik Park**

*Hanshin University

E-mail : saisonh@hanshin.ac.kr

요 약

본 논문에서는 ATM 교환기의 스위치 패브릭을 효과적이고, 빠르게 중재할 수 있는 Cell 스케줄링 알고리즘의 구현에 대해 연구한다. 본 논문에서 설계하는 ATM 셀 스케줄러는 iSLIP 알고리즘을 기본으로 하고 있으며, 이의 고속 구현에 대해 연구한다. 구현되는 셀 스케줄러는 random uniform 트래픽에 대해 100%에 수렴하는 스케줄링 성능을 제공하고 있다. 따라서 본 연구에서는 기본적인 스케줄러의 구조를 제안하고, 이를 HDL로 모델하여 동작 수준 및 타이밍 시뮬레이션을 완료하였다. 그리고 본 논문에서 설계된 셀 스케줄러는 8 포트를 지원하도록 설계하였으며, 이를 기반으로 하여 32 포트로 확장할 수 있다. 8 포트를 지원하는 스케줄러는 grant 및 accept 스테이지를 각각 2단 파이프라인 방식으로 설계하였다.

ABSTRACT

In this paper, we study on an implementation of cell scheduler which arbitrates the ATM exchange efficiently and swiftly. The designed ATM cell scheduler of this paper is based on iSLIP scheduling algorithm. It is aimed at the high-speed implementation. The implemented cell scheduler approximately provides 100% throughput for cell scheduling. We present a basic structure for cell scheduler and describe by using the HDL and perform behavior level and timing simulation. The cell scheduler of this paper is designed to support 8-port switch fabric and can expand in 32-port switch fabric. The cell scheduler for supporting the 8-port switch fabric is designed in 2-stage pipelines for the grant and accept stages respectively. .

키워드

ATM, Switch Fabric, Cell Scheduler, HDL, FPGA, iSLIP

I. 서 론

1990년대 중반 이후 인터넷이 대중화되면서 인터넷 트래픽은 급속도로 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 텍스트 기반에서 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 VOQ(Virtual Output Queueing) 구조를 갖는 ATM 스위치를 기반으로 한 교환 백플레인에 사용된다. 이 시스템들은 스위칭 패브릭을 통과하기 위한 패킷/셀들을 입력 큐를 사용하여 일시적으로 저장한다. 만약에 단일 FIFO(First-in First-Out) 입력 큐가 패킷들을 유지하기 위해 사용된다면, Head-Of- LineHOL) 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 모듈에서

는 수신한 패킷을 각 목적지 별로 구분하여 저장하는 VOQ 구조가 사용된다. 입력 버퍼에 저장된 데이터는 스케줄러에 의해 결정된 입출력 쌍으로 데이터를 스위칭하여 데이터를 전송하게 된다. VOQ구조를 갖는 ATM 스위치에서 핵심이 되는 것이 입출력 쌍을 결정하는 스케줄러인데, 스케줄러에 사용되는 알고리즘에 따라 스위치의 성능에 많은 영향을 미친다.

본 논문에서 VOQ 구조를 갖는 스위치 패브릭의 스케줄러 설계에 대해 연구하며, 이를 HDL을 사용하여 기술하였으며, 최종적으로 이의 동작을 FPGA를 통해 검증하였다.

II. 스케줄링 관련 기술

스케줄링 개념은 컴퓨터나 통신 분야에서 많은 응용 분야가 존재한다. 오늘날 각종 논문을 통해 발표된 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[2]~[6]. 이러한 스케줄링 알고리즘 중에서 몇몇 알고리즘의 장단점을 분석해 보면 다음과 같다.

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 구현 시 지연시간 및 면적 복잡도가 커 대용량의 스위칭 시스템에 적용하기 어려운 단점이 있다.

2DRR은 높은 처리율과 공평성을 나타내지만, 포트수의 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. MUCS는 이보다 더 심해 입출력단쌍(2,1)은 아예 선택이 되지 않는 기근(Starvation)이 발생한다.

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

본 논문에서는 특히 iSLIP 알고리즘을 고속의 하드웨어적으로 구현하고, 이를 직접 검증하고자 한다.

III. 동작 원리

설계되는 스케줄링 알고리즘은 3 단계로 이루어진 반복 매칭을 통해서 입출력을 중재하게 된다. 이러한 3 단계는 다음과 같다.

1) 요청(Request) : 모든 입력단은 그 입력단에 저장되어 있는 셀들의 모든 목적지 출력단으로 request 신호를 보낸다.

2) Grant : 만약 출력단이 1개 이상의 request를 수신하게 되면, 가장 높은 우선권을 가지는 입력단부터 시작하여 정해진 round-robin 스케줄 방식으로 가장 가깝게 있는 입력단에 grant 신호를 보낸다.

3) Accept : 입력단이 1개 이상의 grant 신호를 수신하게 되면, 가장 높은 우선권을 가지는 입력단부터 시작하여 정해진 round-robin 스케줄에 의해 맨 먼저 활성화되어 있는 입력단을 수락하게 된다.

위에서 설명한 1)부터 3)까지의 과정을 반복적으로 수행하게 되며, 반복 회수 많아짐에 따라 중재 성능도 향상된다. 보통의 경우에는 2 - 4회 정도의 반복을 수행한다. 랜덤 uniform 셀 생성 환경에서 시뮬레이션 결과 iSLIP 알고리즘을 3회 반복 수행하였을 때의 매칭 성공률은 99.5%를 보였으며, 4회 반복 수행하였을 때의 매칭 성공률은 99.6%를 보였다.

그림 1은 iSLIP 스케줄 알고리즘을 사용하여 구성된 스케줄러의 블록도를 보여준다. 스케줄러의

3단계인 request, grant, accept 동작은 그림 1의 각 블록에 대응되게 된다.

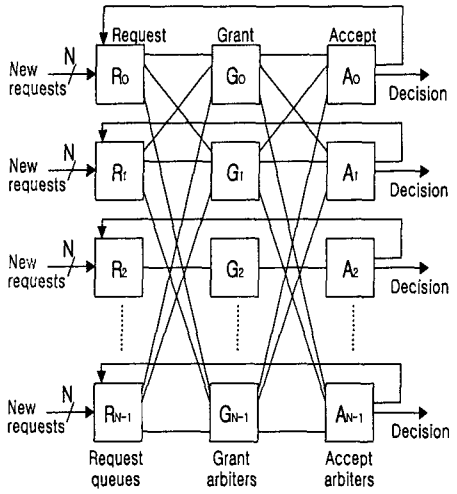

그림 1. iSLIP 알고리즘을 적용한 스케줄러의 블록도

IV. 스케줄러의 설계

우리는 2가지 방식으로 스케줄러를 설계하였다. 하나는 스케줄러의 스테이지를 4개로 분할하여 설계하는 방식이고, 다른 하나는 2개의 스테이지로 분할하여 설계하는 방식을 사용하였다.

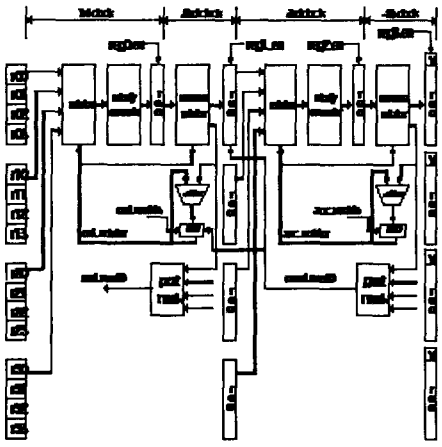

그림 2. 4-스테이지 구조의 스케줄러 블록도

그림 2는 4단 파이프라인으로 설계한 스케줄러의 블록도를 보여주고 있다. 스케줄러의 구성 요소는 rotator, reverse rotator, 및 단순 우선권 인코더를 사용하여 구현이 가능하다. 이 방식은 grant 및 aceept 중재기를 각각 2개의 스테이지로 분할하여 설계하는 경우이다. 이는 고속의 스케줄링을 구현할 수 있는 장점을 제공한다.

그리고 두 번째 구현 방식은 각 중재기마다 단일 클럭 사이클을 할당하여 구현하는 방식이다. 이 방식의 경우에는 PPE(programmable priority encoder)의 고속화 설계를 필요로 한다.

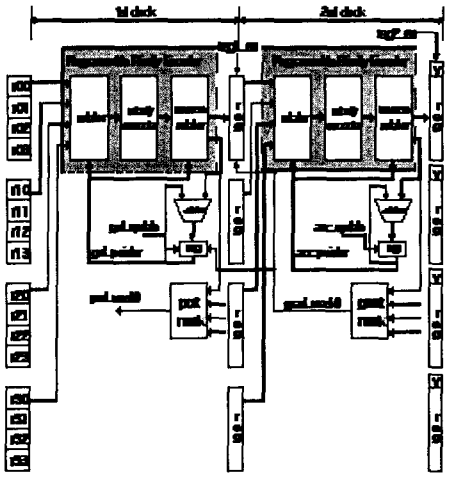

그림 3. 2-스테이지 구조의 스케줄러 블록도

그림 3의 경우 고속의 PPE 구현을 위한 그림 4와 같은 구조의 PPE를 사용하였다[7].

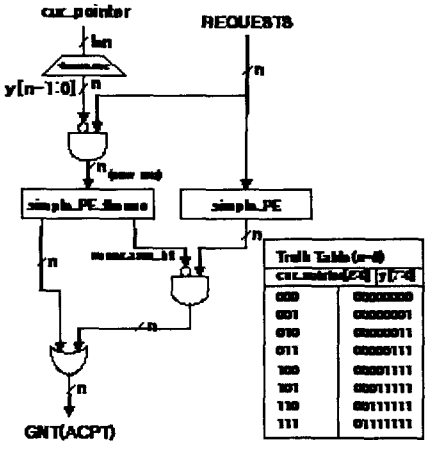

그림 4. 고속의 PPE 구현 블록도

그림 4와 같은 고속의 PPE는 단순한 우선권 인코더만을 병렬로 사용하여 구현이 가능하며, 단일 사이클에 grant 및 accept 중재를 수행할 수 있다. 단, 왼쪽의 단순 PPE는 그림 4에 나타나 있는 진리표의 출력인 y[n-1:0]을 사용하여 현재의

request 입력을 매스크한 신호를 입력으로 받는다. 이는 현재 포인터의 상위 부분에 대해서만 PPE를 수행하기 위함이다.

그림 5는 방식 1을 적용하여 구현한 스케줄러의 시뮬레이션 파형을 보여주고 있다. 본 논문에서 시뮬레이션한 기본 구조는 8 포트의 스위치 패브릭을 지원할 수 있도록 하였다. 그림에서 start 신호는 셀 스케줄이 시작되는 것을 알리는 신호이며, din 벡터는 각 포트의 request 신호를 의미한다. Acc_req 벡터는 accept 단에 수락을 요청하는 신호들의 조합이며, ack_vector 신호는 최종적으로 라우팅이 수락된 포트의 번호를 알려주는 신호이다. 그리고 acc_update_tmp 신호는 3번의 스케줄링을 수행한 후 스케줄러의 포인터를 새롭게 갱신하기 위해 제공되는 신호이다.

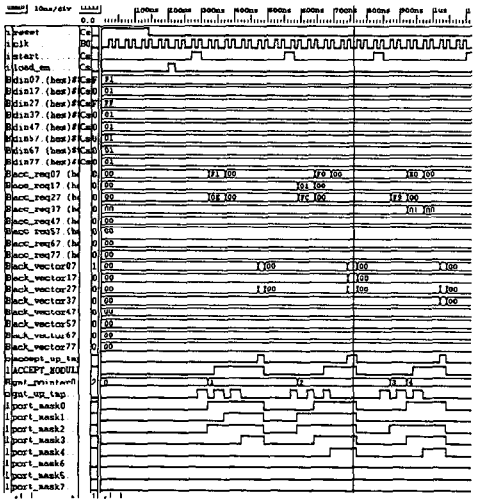

그림 5. 설계된 스케줄러의 시뮬레이션 파형(방식 1)

방식 1은 스케줄링을 위해 3회 반복시 총 9 클록 사이클이 소요된다. 그리고, 방식 2는 스케줄링을 위해 3회 반복시에 총 5 클록 사이클이 소요된다.

V. 결 론

본 논문에서는 iSLIP 알고리즘을 적용한 스케줄러의 구현 방법에 대해 연구하였다. 오늘날 스케줄러는 ATM 스위치의 성능을 좌우하는 중요한 요소로 평가되고 있다. 따라서, 이러한 스케줄러의 속도를 증가시키고, 면적 복잡도를 최소화할 수 있는 설계 방식의 개발 필요성이 부각되고 있는 실정이다. 본 논문에 설계한 스케줄러의 원리를 기본으로 하여 더 많은 포트를 지원할 수 있는 스케줄러로의 확장은 용이하지만, 결과적으로 스케줄러의 복잡도는 $O(N^2)$에 비례하게 된다. 또한 grant와 accept의 구조는 동일한 구조를 가지므로 면적의 최소화를 위해 하나의 PPE만을 사용하여 구현하는 것이 가능하지만, 전체적인 스케줄링 시간을 증가시키는 요인이 된다. 본 논문에서 구현한 스케줄러의 구현 방식을 변형하여 높은 throughput의 달성과 단순한 회로에 의한 구현을 통해 속도의 고속화를 이루는 것이 향후의 연구 과제로 대두되고 있다.

참고문헌

[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.792-799, 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, pp471-482, 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, pp20-23, 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] Pankaj Gpta, Nick McKeown, "Designing and Implementing a fast Crossbar Scheduler", IEEE Micro, pp20-28, Jan. 1999