일 반 투 고

고속 패킷 스위치 기술 동향

변 성 혁, 안 병 준, 김 영 선

한국전자통신연구원 네트워크연구소 인터넷기술연구부

I. 서 론

최근의 인터넷 트래픽의 급증 및 다양한 QoS 보장에 대한 요구, 그리고 VoIP 기술을 통해 음성트래픽까지 지원하는 공중망으로서의 IP망의 발전은 라우터로 대표되는 네트워크 장비에 대용량화, QoS 지원, 고신뢰성 등을 요구하게 되었다. 이러한 요구사항은 라우터 및 스위치 같은 네트워크 장비의 핵심 요소인 스위치 패브릭(switch fabric)에 그대로 적용되는데, 스위치 패브릭 연구에서의 기본 과제는 높은 스루풋을 내면서 대용량으로 확장 가능하고, QoS까지 보장해 주는 스위치 구조를 찾는 것이라고 할 수 있다. 현재 인터넷 백본에서 사용되는 코어 라우터의 경우 스위치 용량이 160 Gbps 내지 320 Gbps이며 테라비트 라우터의 경우 1 Tbps 이상의 스위치 패브릭을 요구한다. 그러나 일반적으로 높은 스루풋과 대용량화는 동시에 달성하기 힘들며, QoS의 보장을 위해서는 하드웨어가 그만큼 복잡해 지기 때문에 스위치 개발자들은 이러한 목표들을 적절히 만족시키는 구조들을 연구해왔다. 고속 패킷 스위치 기술은 ATM 셀 스위치 기술에 주로 기반하고 있는데, 이는 스위칭을 하드웨어적으로 고속화하기에는 스위칭 단위가 고정길이 패킷인 것이 용이하기 때문이다. 따라서 가변길이 패킷을 다루는 라우터에서도 내부적으로는 입력된 패킷을 고정길이 셀로 잘라서 스위칭한 후 재조립하는 방식을 주로 사용한다. 별도의 언급이 없는 한 본고에서 다루는 스위치는 고정 길이 셀 스위치를 가정하며, 가변 길이 패킷 스위치에 대해서는 별도의 절에서 고정 길이 셀 스위치와 비교해 보고자 한다.(가변길이 패킷 스위치와의 구분을 위해서 스위치 내부의 고정 길이 패킷을 셀로 통칭한다.)

본 고에서는 다음의 순서로 고속 패킷 스위치 기술 동향에 대해 정리한다. II장에서는 스위치 구조에 따른 특성을 정리하면서 현재의 추세를 살펴보고, III장에서 스위치의 대용량화 방법에 대해 알아본다. IV장에서 QoS 지원, 표준 스위치 인터페이스, 가변길이 패킷 스위칭 등의 여러 가지 기술적 이슈를 다루며, V장에서 결론을 맺는다.

II. 단위 스위치 구조

스위치의 구조를 분류함에 있어서 고전적인 기준 중에 하나가 버퍼의 위치에 따른 구분으로서, 스위치의 특성을 잘 표현하기 때문에 많이 사용되고 있다. 패킷 스위치에서는 기본적으로 하나의 출력포트로 향하는 패킷이 동시에 2개 이상 존재할 수 있기 때문에 어딘가에 버퍼를 두어 출력포트 경쟁에 실패한 패킷을 저장해 주어야 한다. 그래서 버퍼의 위치에 따라 입력버퍼 스위치, 출력버퍼 스위치, 공유버퍼 스위치 등으로 크게 구분해 볼 수 있다.

입력버퍼 스위치는 crossbar 등의 공간분할(space-division)형 논블록킹(nonblocking) 스위치 앞에 입력 버퍼를 두며 arbitration을 통해 출력포트로 보낼 패킷을 정한 후에 전송하기

에 arbitration 성능에 따라 스루풋이 결정된다. 공간분할형 스위치가 메모리가 없기 때문에 ASIC 화가 용이하고 포트 속도를 고속화하기 쉬우나 전형적인 입력버퍼 스위치는 단일 FIFO 버퍼를 사용할 경우 HOL(head of line) 블록킹때문에 스루풋이 이론적으로 58%에 불과하다는 단점이 있었다. 그러나 입력 포트마다 출력포트별 별도의 큐(queue)를 두는 VOQ(virtual output queue) 구조와 효율적인 arbitration에 의해 HOL 블록킹을 없애면서 성능이 크게 개선되었고, 고속 포트 속도의 수용이 용이한 구조이기 때문에 Cisco 12000 GSR 등의 상용 시스템의 스위치 구조로 많이 채택되고 있다.

특정 출력 포트가 스위칭 타임슬롯동안 하나의 패킷만 수신이 가능한 입력버퍼 스위치와는 달리, 출력버퍼 스위치는 동일 타임슬롯 내에 모든 입력포트에서 하나의 출력포트로 향하더라도 이를 출력버퍼에서 받아 줄 수 있는 구조로서, 버스나 링 구조가 대표적이다. 출력버퍼 스위치는 성능이 가장 우수하나 출력버퍼의 대역폭이 입력포트의 N+1(N은 스위치 입력포트 수)배가 되어야 하므로 구현상 입력 포트 수에 제한이 있다. 공유버퍼 스위치는 출력버퍼 스위치에서 출력버퍼들을 하나의 공유메모리로 구현한 것으로 공유효과에 의해 출력버퍼 스위치와 동일 성능을 위한 메모리 요구량이 적지만 공유메모리의 대역폭이 입력포트의 2N배여야 하므로 역시 구현 가능한 스위치 크기에 제한이 있다. 동작 원리로 볼 때 공유버퍼 스위치는 논리적으로 출력버퍼 스위치와 동일하기 때문에 이를 출력버퍼 스위치로 분류하기도 한다. 공유버퍼 스위치는 최적의 성능이면서 출력버퍼 스위치보다 적은 메모리 요구량을 갖기 때문에 단일 칩으로 구현 가능한 용량 이하에서 경쟁력을 가지며 Juniper M160 등을 위시한 많은 시스템들이 채택하고 있다. 단일 칩으로 구현가능 한 공유버퍼 스위치의 최대 용량은 반도체 기술의 발전에 따라 높아지는데, 금년에 출시된 Agere의 PI40SAX의 경우 80Gbps (32×32@2.5Gbps)의 스위칭 용량을 갖는 공유버퍼 스위치이다[1].

이상은 이론적인 구분이지만, 실제로 대부분의 상용 시스템에서는 버퍼를 입력이나 출력 측 어느 한 곳에만 두는 것이 아니라 여러 가지 이유로 인해 스위치 구조에 관계없이 입력 버퍼와 출력버퍼는 두고 있으며 공유버퍼 스위치 구조인 경우 스위치 패브릭 내에도 버퍼가 존재하게 된다. 공유버퍼나 출력 버퍼 스위치의 경우 높은 대역폭의 메모리가 필요하기 때문에 ASIC 내부에 메모리를 내장하게 되는데, 하나의 칩 내에 수용 가능 한 버퍼 크기의 제한 때문에 적정 양의 패킷만 칩 내에서 수용하고 공유버퍼나 출력버퍼가 찼을 경우에 입력포트로 백프레셔(backpressure)를 줘서 더 이상의 패킷을 스위치로 보내지 않고 입력포트의 입력 버퍼에 저장하게 한다. 입력 버퍼는 라인 속도로 동작하면 되기 때문에 저가의 대DRAM으로 대용량화가 가능하면서, 공유버퍼나 출력버퍼의 용량을 확장 시킨 효과를 갖도록 한다. 그러므로 입력버퍼는 대부분의 상용 스위치에서 채택하고 있다. 출력 버퍼의 경우도 대부분의 시스템에서 사용되고 있는데, 이는 물리적 포트 속도와 스위치 내부 속도의 차이를 흡수해야 하기 때문이다. 스위치 패브릭 내부 속도는 패킷 분할-재조립(segmentation & reassembly) 및 내부헤더 추가로 인해 입력 포트 속도보다 더 높아야 되며, 하나의 스위치 포트가 여러 개의 저속의 물리 포트를 수용하는 것이 일반적이기 때문에 이러한 속도 차이를 극복하기 위해 출력버퍼가 필요하다. 이 외에 입력버퍼 스위치의 성능을 높이기 위해서 스위치 내부 속도의 speedup을 하는 경우에는 한 타임슬롯에 하나 이상의 패킷을 수신해야 하기 때문에 또한 출력버퍼가 필요한데, 이러한 경우를 특별히 CIOQ(Combined Input and Output Queueing) 스위치라고 분류한다. 입력버퍼 스위치는 스루풋이 출력버퍼 스위치에 비해 낮기 때문에 상용 입력버퍼 스위치에서는 이러한 speedup 방법을 성능 개선을 위해 주로 사용하고 있다.

출력버퍼 스위치가 이론적으로 가장 성능이 우수하나 하드웨어 복잡도가 높아 용량 확대에 어려움이 있다. 그래서 모든 스위치 구조 연구는 하

드웨어 복잡도를 낮추면서 출력버퍼 스위치의 성능에 근접하는 스위치 구조를 찾는 것이라고도 말할 수 있다. 상용 제품에서 채택하고 있는 구조는 공유버퍼 스위치거나 VOQ+crossbar 기반의 입력버퍼 스위치가 일반적이다. 공유버퍼 스위치는 최적의 성능을 가지고 있으나 하드웨어 복잡도 때문에 수십 Gbps 이하의 스위치에서 사용되며, 다음 장에서 다룰 용량 확장 방법을 이용해서 스위치 용량을 확장할 수 있다. IBM의 PRS 시리즈 스위치[2], Agere의 PI40SAX[1], AMCC의 nPX5800, Erlang의 ENET-Se[3], Paion의 GES[4] 등이 상용의 공유버퍼 스위치이며, 라우터 시스템에서는 Juniper의 M시리즈 라우터가 공유버퍼 스위치 구조를 채택하였다.

VOQ 기반의 입력버퍼 스위치는 arbitration의 성능에 따라 100%의 스루풋까지 얻을 수 있으며, 일반적으로 사용되는 crossbar 구조가 ASIC으로 구현하기 용이하기 때문에 많은 스위치에서 적용하고 있다. 이 스위치에서는 각 입력버퍼는 출력 포트별로 논리적 큐를 가지며, 입력 패킷은 자신의 목적지에 대응하는 큐에 저장된다. 스케쥴러(scheduler)는 모든 입력큐에 저장된 패킷의 정보를 바탕으로 충돌 없이 전송할 수 있는 최대한의 패킷을 선택해 전송 시킴으로써 HOL 블록킹을 없앴다. 이러한 구조에서는 스케쥴러의 성능이 스위치의 성능을 결정하는데, 라인속도가 높아짐에 따라 매 패킷마다 수행해야 하는 스케쥴링 시간도 짧아진다. 그리고 한 타임 슬롯에 한번의 스케쥴링으로 최대 매칭을 찾기 힘들기 때문에 여러 번의 매칭을 시도하게 된다. 따라서 스케쥴링 알고리듬은 단순하면서도 높은 성능을 얻을 수 있어야 한다. 그리고 일반적으로 스케쥴링 알고리듬의 복잡도가 O(N) (N : 스위치 입력 포트 수)이기 때문에, 스위치의 크기는 스케쥴링 알고리듬에 의해서도 제한된다. 이러한 스케쥴링 알고리듬 복잡도의 제한 때문에 알고리듬 자체만으로는 출력버퍼 스위치에 근접하는 성능을 얻기 어려워서 스위치 내부 속도를 speedup 하는 것이 일반적이다. 실제 동작속도를 speedup 할 수도 있지만 그 보다는 여러 개의 스위치를 병렬로 두는 다중 스위치플래인 구조를 주로 택한다. 입력 트래픽을 여러 개의 스위치에 분산시켜 처리하기 때문에 개별 스위치의 부하를 낮춰서 물리적 speedup과 유사한 효과를 볼 수 있으며, 특정 스위치가 고장이 날 경우에도 자동적인 load balancing에 의해 그 스위치를 제외한 나머지 스위치들이 입력 트래픽을 처리되게 되어 고 신뢰성을 가질 수 있기 때문이다. 스케쥴링 알고리듬으로 PIM, iSLIP, DRRM 등의 다양한 알고리듬이 연구되어 왔는데[5], Vitesse의 GigaStream[6], AMCC의 Cyclone, PMC-Sierra의 ETT1 등의 스위치 칩셋은 iSLIP을 대표로 하는 round-robin 기반의 스케쥴러를 사용하고 있다. 라우터 중에서는 Cisco의 12000 GSR 라우터와 Juniper의 T640 라우터가 VOQ 기반의 crossbar 스위치 구조를 갖는다.

III. 스위치의 용량 확장

스위치의 용량은 포트 수 x 포트 속도이므로 스위치 용량 확장은 포트 수를 늘이거나 포트 속도를 늘여야 한다. 그러나 스위치 구조와 상관 없이 하나의 칩으로 구현 가능한 스위치 용량은 I/O pin의 개수와 속도, 그리고 단일 칩이 수용가능한 게이트 수의 제한 등의 ASIC 기술의 한계에 의해 제한될 수 밖에 없다. 따라서 대용량 스위치는 멀티 칩으로 구성될 수 밖에 없는데, 어떻게 칩별로 기능을 분리하고, 상호 연결을 하느냐에 따라 다양한 방법이 도출된다. 스위치의 용량 확장 방법은 다음과 같이 정리할 수 있다.

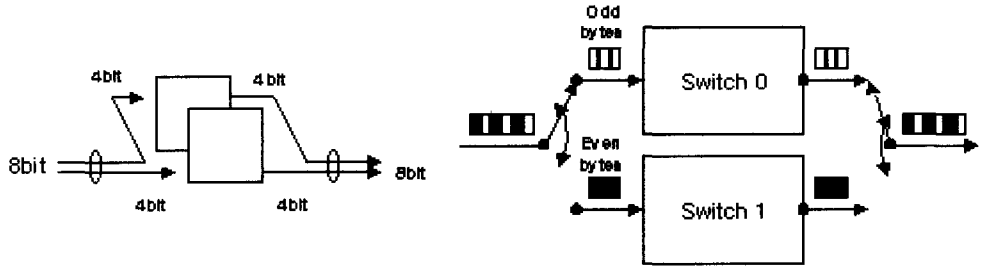
- 포트 속도 고속화
  - 고속 serial 링크(2.5Gbps~3.125Gbps)
  - bit-slicing
  - byte inteleaving
  - multiplane 스위치
  - link bundling
- 포트 수 확장

- 1단 포트 확장 구조
- 다단 스위치 구조
  - ✓Clos 망 구조
  - ✓Memory-Space-Memory(MSM) 구조
  - ✓다단접속망(MIN : MultistageInter-connection Network) 구조

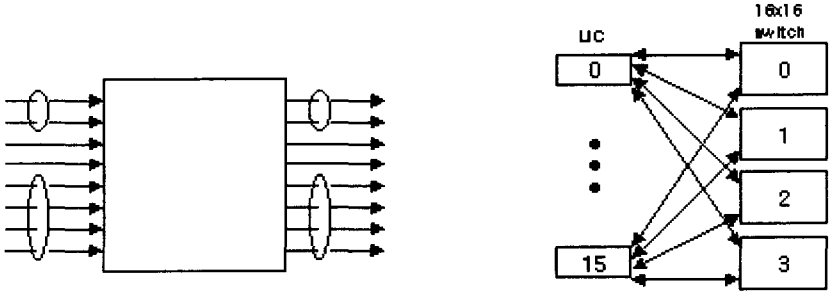
1. 포트 속도 고속화

스위치 포트의 고속화는 우선 신호 자체의 고속화를 통해 이룰 수 있다. 스위치 칩의 I/O 핀의 속도는 수년 전에는 50MHz~100MHz LVTTL 신호가 일반적이었으나, 최근에는 고속 serial 링크 기술인 Serdes 기술의 발전으로 차동신호(differential signal) 채널 당 2.5Gbs내지 3.125 Gbps 속도를 얻을 수 있게 되었다. 이러한 Serdes 링크를 최근에는 스위치 칩에 내장함으로써 데이터 I/O핀 수를 크게 줄이고, 시스템에서 라인카드와 스위치간의 backplane 연결을 보다 단순하게 하게 되었다. Vitesse GigaStream VSC 882 스위치 칩은 16×16@2.5Gbps(2.5Gbps 링크 속도인 16×16 스위치)이며, 가장 최근에 출시된 Agere의 PI40C의 경우 64×64@2.5 Gbps crossbar 스위치이고, Tau의 T64[7] 스위치는 64×64@3.125Gbps이다. 스위치 링크의 최대 속도는 현재 3.125Gbps이며 그 이상의 속도를 얻기 위한 연구도 계속 진행 중이다.

스위치에 Serdes의 내장하게 된 것은 최근 1~2년 사이의 추세로서, 그 이전에 개발된 스위치들은 bit-slicing이나 byte-interleaving과 같은 방법을 사용하여, 스위치 칩의 I/O 대역 한계를 극복하고자 하였다. 〈그림 1(a)〉에 예시되어 있듯이 Bit-slicing 방식은 멀티비트(w bits) 스트림으로 입력되는 패킷을 k bit 단위로 나누어 w/k개의 스위치 칩으로 k bit 단위로 처리하는 방식으로서, 스위치 포트의 대역폭을 최대 w배(k=1일 경우)만큼 확장시킬 수 있는 구조이다. 그러나 각각의 스위치가 동기화되서 동일하게 동작해야 하기 때문에 제어의 어려움은 있으



(a) Bit-slicing (b) Byte interleaving



(c) Link grouping (d) Multiple switchplane

〈그림 1〉 스위치 포트 속도 확장 방법

며, 일반적으로 하나의 스위치가 매스터가 되어 패킷의 제어정보를 바탕으로 나머지 모든 칩을 조정하도록 한다. Byte-interleaving은 패킷을 even byte와 odd byte로 나누어 2개의 동기화된 스위치로 처리하는 방식으로 〈그림 1(b)〉에 도시되어 있다. Byte-interleaving 방식보다는 bit-slicing 방식이 확장성이 더 좋기 때문에 bit-slicing 방식을 많이 사용한다. Cisco 12000 GSR의 스위치는 4개의 bit-slicing crossbar로 구성되며, PMC-Sierra의 ETT1은 14bit-slicing 구조와 Serdes 링크 채택으로 10Gbps 인터페이스를 32개까지 수용가능하다. IBM의 PRS Q-64G도 최대 8bit-slicing 구조와 2Gbps Serdes 링크 기술로 32개의 10Gbps 인터페이스까지 지원한다.

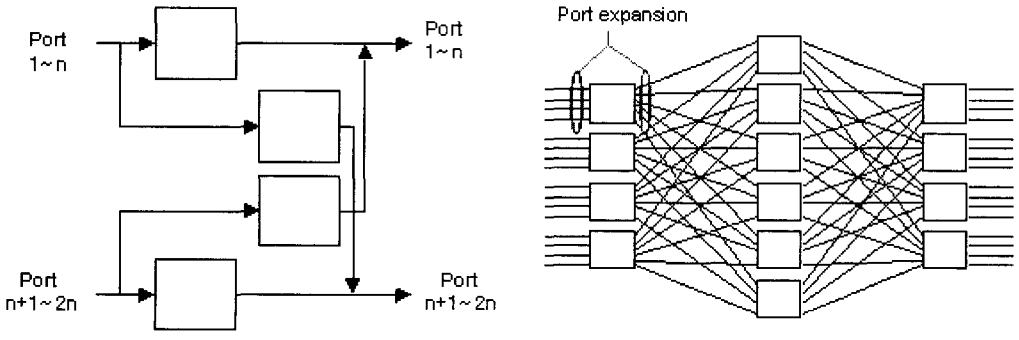
코아 라우터의 경우 OC-192c 인터페이스까지 요구하며, 10GE 시스템이 출시되기 시작하고 있지만 스위치의 물리적 포트 속도는 최대 2~3 Gbps에 불과하기 때문에, bit-slicing 방식을 사용하지 않을 경우 여러 개의 링크를 묶어서 하나의 논리적 링크로 사용하는 링크 그룹핑(link grouping) 방법을 사용한다. 스위치 패브릭에서 초기 설정에 의해 수개의 물리적 링크를 하나의 논리 링크로 설정하면 패킷의 송수신시 링크 그룹내의 링크를 미리 정한 round-robin 순서대로 선택할 경우 패킷 순서가 바뀌지 않고 처리될 수 있으며, 그룹핑에 따른 성능 저하도 유발되지 않는다. IBM의 PRS 스위치나 Agere의 PI40 스위치에서 이런 방법으로 수 Gbps 이하의 물리 포트를 갖는 스위치로 10Gbps 라인 속도를 지원한다.

또 다른 방식은 〈그림 1(d)〉와 같은 다중 스위치 플레인 구조를 사용하는 것이다. 상호 독립적으로 동작하는 스위치 플레인을 여러 개 두고 입력되는 트래픽을 load balancing 알고리듬에 의해 다중 스위치 플레인에 균등하게 분배하여 처리되게 하면 스위치 플레인 수 만큼의 포트 속도 증가 효과를 기대할 수 있다. 단, load balancing 알고리듬이 완벽하지 못할 경우 포트 속도 증가 효과는 그만큼 떨어질 수 밖에 없다. Vitesse의 GigaStream과 Tau의 T64 스위치가 이와 같은 구조이다.

2. 스위치 포트 수 확장

스위치의 포트 수를 단위 스위치가 지원하는 수보다 늘이고 싶으면, 전체 입력 포트를 단위 스위치가 수용할 수 있는 소 그룹으로 나누고 이 그룹간의 연결을 적정 방법으로 구현해야 한다. 일반적으로 스위치를 다단(multistage)으로 구성하는데, 제한적으로 1단 확장 구조도 가능하다.

〈그림 2(a)〉는 IBM의 PRS 스위치에서 사용하는 1단 포트확장 구조를 보여준다. PRS Q-64G 단위 스위치는 내부에 16×16 공유버퍼 스위치 4개의 모듈을 〈그림 2(a)〉와 같이 구성하여 32×32 스위치를 만들었다. 현재 IBM PRS 스위치에서만 사용되는 독특한 구조인데, 포트 수



(a) 1단 포트확장 구조(n×n→2n×2n) (b) Clos망 구조

〈그림 2〉 스위치 포트 수 확장 방법

를 2배 이상으로 확장하기에는 무리가 있으며, 입력과 출력측에서 패킷을 분배하고 선택하는 장치가 추가되는 단점이 있다.

스위치의 대용량화를 위한 가장 일반적인 방법이 〈그림 2(b)〉와 같은 3단 Clos망 구조를 사용하는 것이다. Clos망은 첫번째 단 및 세번째 단의 모듈이 중간 단의 모든 모듈과 연결되는 구조로서, 임의의 입출력 포트에 대해 중간 단의 스위치 모듈 개수 만큼의 path가 존재한다는 것이 특징이다. 다중 path가 존재하기 때문에 path의 수와 path 선택 알고리듬에 따라 전체 스위치의 성능이 결정되는데, 적정 성능을 위해서 그림과 같이 port expansion과 내부 속도 speedup이 요구된다고 알려져 있다. 그리고 하나의 flow내의 패킷은 동일 경로를 가져야 스위칭 도중에 패킷 순서가 뒤바뀌지 않는다. Paion의 GES가 Clos망 확장 방식을 사용한다[4].

Clos망으로 용량 확장할 때는 1, 2, 3단 스위치가 같은 종류인 경우가 일반적인데, Agere의 PI40 스위치는 3단 Clos망 구조이지만 1, 3단 스위치로 공유버퍼 스위치를 사용하고, 2단 스위치에는 버퍼가 없는 crossbar 스위치를 채택한 Memory-Space-Memory(MSM) 3단 구조를 채택하였다[1]. 즉 단위 스위치의 확장방법이 아니라 기본적인 스위치 구성을 3단 Clos 망으로 채택하면서, 2단 스위치를 버퍼가 없는 crossbar을 채택하므로써 동일 flow내의 패킷이 서로 다른 2단 모듈로 스위칭 되더라도 순서 바뀜이 없도록 하였다. 1단 및 3단 공유버퍼 스위치를 aggregation 모듈로 볼 경우 개념적으로는 입출력 포트를 그룹핑한 1단 multiplane crossbar 스위치라고도 볼 수 있다. Agere의 PI40 스위치는 MSM 구조를 사용하여 최대 2.5Tbps까지 확장 가능하다.

이 외에 수 테라비트 이상의 용량을 목표로 하는 테라비트 라우터에서는 확장성이 좋은 MIN 구조를 채택하기도 한다. Avici의 TSR은 3D toroidal mesh 구조를 채택하였는데, 각 라인카드가 3D toroial mesh의 격자점을 구성하는 완전 분산형 격자 스위치 구조로서 라인카드의 증설만으로 스위치 용량이 증가하는 확장성이 좋은 구조이나 3D toroidal mesh 구조 자체가 블록킹 망이기 때문에 높은 스루풋을 내기 어려울 것으로 판단된다.

전통적으로 스위치의 대용량화는 단위 스위치 모듈을 다단 확장 구조로 구성하는 것이 정석이었으니, Clos나 MIN 등의 다단 확장 구조가 단일 스위치의 성능에는 미치지 못하였다. 최근에는 ASIC 기술의 발달로 단일 스위치의 용량 자체가 수십 Gbps까지 커졌으며, bit-slicing 및 다중 스위치 플래인 구조 등의 스위치 대역폭 증가 방법등을 통해 1단 구조에서도 수백기가 급까지 확장이 가능해지게 되어, 다단 구조를 용량 확장 방법으로 채택하지 않는 스위치가 많아 졌다. IBM의 PRS 시리즈는 로드맵에 의하면 계속 1단 확장 구조를 유지하면서 테라비트급 스위치로 발전해 나갈 계획이고, 현재 최고 모델인 PRS Q-64 G는 512Gbps까지 확장 가능하다. Tau의 T64 스위치는 1단 다중 스위치 플레인 구조로서 640Gbps의 최대 용량을 가진다. Agere의 PI40는 기존의 Clos망 개념을 변형하여 논리적으로는 1단 다중 스위치 플레인 구조와 유사한 MSM 구조로 2.5Tbps까지 확장 가능한데, 현재까지 출시된 상용 스위치 중에서 최대 용량을 지원하는 스위치이다.

Ⅳ. 스위치 설계의 기술적 이슈들

1. QoS 지원

Fine grain QoS를 지원하기 위해서는 flow별 큐를 두어야 한다고 얘기되어지는데, 대부분의 QoS 기능은 라인인터페이스의 버퍼, 특히 출력 버퍼에서 구현되어진다. Fine grain QoS를 보장하기 위한 WFQ 등의 기법은 출력 버퍼에서의 동작을 가정하고 있으며, 입력 버퍼의 경우 flow별 대역폭을 보장하려고 해도 스위치 스케쥴링 결과에 의해 출력 여부가 결정되기 때문에 대역폭 보장이 의미를 갖기 어렵다. 스위치의 스케

쥴링에서 QoS를 지원할 경우 지원하는 QoS 클래스 수에 비례해서 스케쥴링이 복잡해지는 것이 일반적인데 스케쥴링은 한 타임슬롯 내에 이루어져야 하기 때문에 결국 지원하는 QoS class의 수가 제한될 수 밖에 없다. 일반적으로 8개 정도의 우선순위를 지원하며 strict priority나 weighted round-robin(WRR) 및 deficit round-robin(DRR) 등으로 각 클래스의 큐를 서비스한다.

2. NP(network processor)와의 인터페이스

라인카드에서의 패킷 처리 기능을 요즘에는 NP로 구현하기 때문에 NP와의 인터페이스는 스위치와 라인카드간의 인터페이스를 말한다. 전통적으로 스위치 칩셋 밴더는 자사의 스위치 칩셋과 연동되는 라인카드용 칩셋을 같이 제공해 왔기 때문에 독자적인 스위치 인터페이스를 사용해 왔다. 그러나 최근에는 라인카드에서의 패킷 처리 기능을 전담하는 NP만을 개발하는 회사, 그리고 스위치 패브릭만을 개발하는 회사가 늘어나면서 표준 스위치 인터페이스 정립이 필요해졌고, 이러한 회사들이 NP Forum을 만들어서 표준 스위치 인터페이스로서 CSIX(common switch interface)-L1 및 NPSI(NP Streaming Interface) v1.0을 제정하였다[8].

CSIX-L1은 NP Forum의 전신인 CSIX Consortium에서 2001년 8월에 제정한 것으로, 스위치와 NP간의 frame format, 플로우 제어 방법, 멀티캐스트 방법등을 명시하며, OC-48급 인터페이스로 32bit 병렬 인터페이스를, OC-192급 인터페이스는 64bit 내지 128bit 인터페이스를 정의하고 있다. OC-48급 CSIX-L1 인터페이스는 스위치 칩셋 중에 Vitesse GigaStream이 채택하였고, NP에서는 Motorola C5-DCP, Intel IXP2400, Vitesse IQ2200 등에서 지원하는 등 널리 사용되고 있다. 그러나 OC-192급 CSIX 인터페이스는 64bit 내지 128bit 인터페이스여서 I/O pin을 많이 요구하고 구현에도 어려움이 예상되어 NP Forum에서는 새로운 OC-192급 표준 인터페이스로 16bit LVDS 규격인 NPSI를 2002년 9월에 제정하였다. NPSI의 물리 규격은 OIF의 SPI4.2에 기반하였다. NPSI는 최근에 스펙이 완성되었기 때문에 아직 지원하는 NP나 스위치는 없지만 내년에 출시 예정인 칩셋들은 NPSI의 지원을 밝히고 있다. 현재까지는 10Gbps급 표준인터페이스로 CSIX-L1 규격을 채택하고 있는데 IBM의 C192 스위치 인터페이스 칩, EZChip의 10G NP인 NP1이 10 Gbps CSIX-L1 을 지원한다.

AMCC나 Agere, IBM 등과 같이 라인카드용 칩셋과 스위치 칩셋을 같이 공급해 온 밴더들은 표준 인터페이스 채택에 미온적이었으나 시스템 업체로부터의 지원 요구가 높아지고, 또한 경기 침체로 자사가 모든 칩셋을 동시에 제공하기 힘들게 되면서 표준 인터페이스를 칩셋에 적용하거나 표준 인터페이스를 자사의 고유 인터페이스로 변환시켜주는 브릿지 칩을 제공하는 등 점차 수용하고 있는 추세이다.

3. 고정 길이 셀 vs. 가변 길이 패킷 스위치

IP 패킷이 가변길이 임에도 이를 시스템 내에서 스위칭하기 위해서 고정길이 셀로 잘라서 스위칭하는 이유는 가변길이 패킷 스위칭이 복잡하다는 일반적인 믿음 때문이다. VOQ를 채택한 크로스바 스위치에서 가변길이 패킷 스위칭을 한다고 가정해 보자. 전송중인 가변길이 패킷은 언제 전송이 끝날지를 모르기 때문에 스케쥴러는 계속 전송 상태를 감시해야 한다. 한 패킷의 전송이 끝났을 때 스케쥴러는 전송을 끝낸 입력 포트의 여러 출력 큐 중에서 하나를 선택해서 다음 전송을 시도해야 하고, 출력 포트에선 여러 개의 입력포트에서 전송을 시도하는 패킷들 중 하나를 공정하게 선택해야 한다. 그런데 패킷의 길이가 다 달라서 공정한 선택 알고리듬을 찾기 힘들며, 입력 트래픽 패턴에 따라서는 특정 포트의 패킷이 계속 전송 기회를 잃는 기근(starvation) 현상이 발생할 수도 있다[9].

그러나 셀 스위칭을 함으로써 생기는 문제점도 존재한다. 스위치 내부 셀의 크기를 N 바이트로 정했다고 가정할 때, N+1 바이트의 패킷이 입

력되면 이는 2개의 N 바이트 셀로 만들어 져야 한다. 따라서 worst case로 N+1 바이트의 패킷이 계속 입력될 경우는 스위치의 내부 속도가 입력의 2배가 되어야 정상 처리될 수 있다. 이를 "N+1 문제"라고 부르는데, 이러한 최악의 상황이 일어날 수 있기 때문에 셀 스위칭을 하는 상용 스위치는 2배 이상의 내부 speedup을 하고 있다. 예로 Agere의 10Gbps NP인 APP750은 스위치와의 인터페이스 대역폭이 25Gbps이다. 또한 출력측 인터페이스에서는 패킷을 재조립(reassembly)해야 하기 때문에 재조립 버퍼가 필요하고 재조립 지연이 유발된다.

기존의 스위치 연구에서는 셀 스위칭 자체의 지연시간만을 주로 고려해 왔으나 IP 패킷의 측면에서는 패킷 분할 및 재조립 지연까지를 포함해서 스위치에서의 지연을 나타내는 것이 타당하다. 셀 스위치 스케쥴링 알고리듬인 iSLIP을 가변 길이 패킷 스케쥴링용으로 수정하여 셀 스위칭과 가변길이 패킷 스위칭을 비교한 결과를 보면, 랜덤 균일 트래픽에 대해서는 셀 지연시간과 패킷 지연시간 모두에 있어서 가변길이 패킷 스위칭이 우수한 결과를 보여주며 스루풋도 차이가 나지 않고 있다[10]. 앞에서 언급한 기근 문제와 공정성 문제는 특정 비균일 트래픽 패턴에서 발생하기 때문에, 가변길이 패킷 스위치에 대한 연구는 보다 진전되어 그러한 worst case를 예방할 수 있는 알고리듬이 개발되어야 할 것이다. 그러나 라인 인터페이스의 속도가 10Gbps에 이르렀고 추후 40Gbps까지 높아진다고 볼 때, 셀 스위칭에서의 N+1 문제를 해결하기 위해 2배의 speedup을 하는 것은 스위치 설계에 있어서 큰 부담이다. 따라서 고속 인터페이스를 갖는 스위치에서는 가변길이 패킷 스위치의 장점이 충분히 있다고 판단된다.

V. 결 론

지금까지 많은 스위치 연구들이 높은 스루풋과 낮은 지연특성, 대용량으로의 확장성, QoS 보장 등을 목표로 이루어져 왔다. 현재의 추세는 단위 스위치 구조측면에서 공유버퍼 스위치와 VOQ 기반의 crossbar 스위치 구조가 각광을 받고 있다. 공유버퍼 스위치는 최적의 성능이, crossbar 스위치는 우수한 확장성이 각각 장점이라고 할 수 있다. 용량 확장을 위해서는 다단 구조보다는 1단 구조에서의 확장 방법이 주로 채택되고 있는데, bit-slicing, 다중 스위치 플레인구조 등이 그것이다. 이것은 요구되는 스위치 용량이 현재의 기술로는 1단 스위치 구조로도 가능하기 때문이나, 수 Tb/s 이상의 스위치 패브릭은 다단 구조가 불가피할 것이다. 현재 MSM 구조로 최대 2.5Tbps까지 확장 가능한 스위치가 개발되었다. 스위치에서의 QoS 지원은 8개 정도의 CoS(class of service)를 지원하는 수준이며, CSIX나 NPSI와 같은 표준 스위치 인터페이스가 정착해 가고 있다. 그리고 현재는 고정 길이 셀 스위칭이 일반적인데, 라인 속도가 고속화와 셀 스위칭의 N+1 문제들을 고려할 때, 가변길이 패킷 스위칭이 새로 조명받을 것으로 보인다.

참 고 문 헌

〔1〕 http://www.agere.com/enterprise_metro_access/switch_fabric.html

〔2〕 http://www_3.ibm.com/chips/products/wired/products/switch_fabric.html

〔3〕 http://www.erlangtech.com/products/products.cfm

〔4〕 http://paion.com/

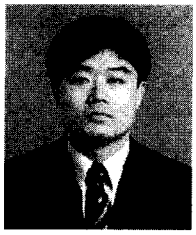
〔5〕 H. Jonathan chao, cheuk H. Lam, and Eiji Oki, Broadband Packet Switching technonogies, John Wiley & Sons, 2001.

〔6〕 http://www.vitesse.com/products/index.cfm

〔7〕 http://www.taunetworks.com/products.html

〔8〕 http://www.npforum.org/techinfo/approved.shtml
〔9〕 Nick McKweon, "Fast switched backplane for a gigabit switched router," Business Communication Review, Dec. 1997
〔10〕 Sung-Ho Moon and Dan Keun Sung, "High-Performance Variable-Length Packet Scheduling Algorithm for IP Traffic," GLOBECOM'01, pp.2666~2670, 2001.

저 자 소 개

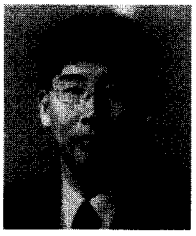


邊 性 赫

1991년 2월 한국과학기술원 전기 및전자공학과 졸업 (공학사), 1986년 2월 한국과학기술원 전기 및전자공학과 졸업 (공학석사), 1999년 2월 한국과학기술원 전기 및전자공학과 졸업 (공학박사), 1999년 2월~현재 : 한국전자통신연구원 인터넷 기술연구부 선임연구원, 〈주관심 분야 : 고속 라우터, QoS, 스위치 구조, IPv6〉

安 炳 俊

1984년 2월 한양대학교 전자통신공학과 졸업 (공학사), 1986년 2월 한양대학교 전자통신공학과 졸업 (공학석사), 1999년 5월 Iowa State University 졸업 (Computer Engineering) (공학박사), 1986년 2월~현재 : 한국전자통신연구원 책임연구원, 라우터구조팀장, 〈주관심 분야 : ATM, 트래픽 제어, QoS, 고속 라우터 기술〉



金 煐 善

1980년 2월 고려대학교 전자공학과 졸업 (공학사), 1982년 2월 고려대학교 전자공학과 졸업 (공학석사), 1991년 8월 고려대학교 전자공학과 졸업 (공학박사), 1982년 3월~현재 : 한국전자통신연구원 책임연구원, 인터넷기술연구부장, 1994~1998 : 전북대학교 컴퓨터공학과 겸임교수, 1980~현재 : 대한전자공학회 스위칭 및 라우팅 연구회 전문위원장, 논문지 편집위원, 상임이사 (회지편집 위원장), 기획위원회 위원, 평의원, 1989~현재 : 한국통신학회 교환 및 라우팅 연구회 전문위원장, 학회지 편집위원, 대전. 충남지 부 지부장, 평의원, 2000년~2001년 : 과학기술부 국가연구개발사업 평가위원, 1993년~1997년, 2001년 : 정보통신연구진흥원 정보통신연구개발기금사업 심사위원, 2000년 : 특허청 특허기술상 심사협의회 위원, 〈주관심 분야 : ATM, 트래픽 제어, QoS, 고속 라우터 기술, 인터넷, 라우팅 프로토콜, 이동통신〉