主題

초고속 라우터 기술

한국전자통신연구원 안병준, 김영선

차 례

Ⅰ. 서론

현재 통신망에서 제공되는 서비스는 그 종류에 따라서 서로 다른 망을 통하여 제공되고 있다. 망 사업자 입장에서 보면 동일하거나 유사한 장비들을 서비스별로 서로 다른 망을 구축하기 위하여 도입하는 것은 중복적인 투자이고 이들을 운용 관리하기 위한 비용 증가도 부담스러운 것이다. 따라서 최근 망 사업자들은 증가하는 IP 트래픽을 수용함과 아울러 지금까지 주 수입원이 되어왔던 전용 회선 및 음성 백본 트래픽들을 계속적으로 지원하기 위한 전략적인 인프라로서 통합 IP 망(converged IP network)의 구축을 생각하고 있다.

이러한 통합 IP 망을 구축을 위하여 고려해야 할 사항들은 많이 있으나 그 중에 첫째는 계속적으로 증가하는 IP 트래픽을 수용하기 위한 확장성이다. AT&T 연구 보고서에 따르면 인터넷 트래픽은 매년 70%~150% 씩 증가하고 있으며[1] ADSL 등과 같은 광대역 서비스 가입자의 급격한 증가로 IP 백본으로 유입되는 트래픽의 양은 더욱 증가하고 있는 추세이다. 두 번째로 중요한 고려 사항은 인터넷 서비스 품질(Quality of Service, 이하 QoS)의 보장이다. 음성과 데이터의 통합, 유선과 무선의 통합 뿐만 아니라 방송 미디어의 통합 까지도 추구하는 통합 IP망에서는 서비스 품질 보장에 대한 인터넷 사용자들의 요구가 더욱 커질 전망이다. 이러한 고려 사항들을 반영하고 기존 서비스의 수용 뿐만 아니라 향후 새로운 서비스를 효율적으로 제공할 수 있는 멀티 서비스 백본망을 구축하기 위해서는 QoS 보장 기능을 갖는 초고속 라우터, 특히 패킷 처리 용량이 수 Tbps에 이르는 테라 라우터의 도입이 필요하다.

이와 같은 초고속 라우터의 도입은 여러 가지 장점을 가지고 있다. 먼저 망 사업자 측면에서 살펴보면 노드 수가 감소됨에 따라서 장비에 대한 투자 비용이 절감되고 망을 운용하고 관리하는 데 필요한 비용도 절감할 수 있는 효과를 기대할 수 있다. 또한 불필요한 노드간 트렁크 감소에 따른 장비의 효율적 운용 및 다양한 서비스 등급(서비스 제공 품질에 따라 서비스 등급을 구분하고 이에 맞게 요금체계를 구성하여 서비스를 제공하는 방법) 개발을 통한 매출 증대

효과를 기대할 수 있을 것이다. 사용자들은 보다 빠른 인터넷 접속이 가능하고 사용 목적에 따라 다양한 서비스 등급을 선택할 수 있어서 인터넷 비용을 절감할 수 있다.

초고속 라우터를 실현하기 위한 기술은 광범위하지만 크게 보면 확장성, QoS 보장, 신뢰성 보장이라는 세가지 목적을 달성하기 위한 것이다. Cisco, Juniper 등을 비롯한 라우터 업체들은 자체 개발한 ASIC을 사용하여 OC-192c/STM-64 라인 카드를 장착한 수백 Gbps 용량의 라우터 제품들을 이미 출시하였고, 수 Tbps 용량의 테라 라우터 시장 확보를 위한 치열한 연구개발 경쟁을 하고 있다. 또한, 최근의 인터넷 장비 시장의 불황에도 불구하고 칩셋 회사들도 1.2Tbps~2.5Tbps 스위칭 용량을 제공하는 스위치 패브릭 칩과 10Gbps wire-speed 패킷 포워딩 칩셋의 출시를 앞 다투어 선전하고 있다. 테라 라우터 제품 및 핵심 칩셋 기술 동향은 참고 문헌 [2]~[5]에서 소개하였고 본 고에서는 확장성과 QoS 보장 관점에서 초고속 라우터의 구조 및 성능에 관한 사항들을 고찰하고자 한다.

Ⅱ. 초고속 라우터 구성 요소

본 장에서는 확장성 및 QoS를 보장하기 위한 초고속 라우터의 구조와 이를 실현하기 위한 기술적인 사항들에 대하여 기술한다. 고속 대용량 라우터는 기본적으로 (그림 1)과 같은 구조를 갖는다[6]~[8]. 즉, Cisco의 IOS, Juniper의 JUNOS 등과 같은 라우터 운영 소프트웨어가 상주하는 라우팅 프로세서(Routing Processor), 망에 접속하여 고속으로 데이터 패킷을 포워딩 하는 라인 카드(또는 blade라고도 함), 라인 카드들간 패킷 전달 채널을 제공하는 스위치 패브릭 등이 고속 백 플레인(back plane)에 장착되어 코어 라우터의 샤시(chassis)를 구성한다.

인터넷 트래픽의 계속적인 증가 추세에 따라서 백본 POP의 패킷 처리 용량을 지속적으로 확장해야 하는 망 사업자들 입장에서는 저용량 라우터들을 끊임 없이 업그레이드 하는 것은 비용이 많이 드는 일이다. 즉, POP 내부의 라우터들을 메쉬(mesh)로 묶기 위해, 망 사업자의 수익을 올리기 위한 트래픽을 주고 받아야 할 라인 인터페이스들을 낭비해야 하는데, 더 많은 라우터들을 POP에 도입할수록 그 숫자가 커지게 되고 궁극적으로는 모든 포트들이 소진되어 라우터를 교체해야 하는 상황이 된다. 실제로 저용량 라우터들의 수명 연한은 18~24 개월이라는 보고도 있다[9].

따라서 백본망을 구성하는 코어 라우터들의 대용량화가 필요한데, 다단 구조로 최대 2.5Tbps까지 단위 스위치 패브릭 용량을 확장할 수 있는 칩셋이 2003년 초에 출시될 예정이다. 지금까지 개발된 코어 라우터들 중 Cisco 12416 GSR, Juniper M160, Lucent NX64000 등은 단일 샤시 구조이지만, Avici TSR, Pluris Teraplex 등은 여러 개의 샤시를 연결할 수 있는 확장 구조이어서 최대 용량을 수 Tbps ~ 십 수 Tbps까지 확장할 수 있다고 주장하고 있다. 그러나 TSR의 경우 각 라인 카드가 스위치 패브릭을 가지고 3-D toroidal mesh topology로 연결되는 분산 스위치 구조를 채택함으로써 스위치 용량 증설은 가능하지만 기본적으로 이 구조는 블록킹 네트워크이기 때문에 실질적인 패킷 처리 용량은 스위치 용량에 미치지 못할 것으로 보인다. Teraplex는 TeraConnect라는 10Gbps 광 링크로 라인 카드와 스위치, 샤시와 샤시를 연결한다. 스위치 카드는 이중화되어서 논리적으로 한 샤시에 8개의 스위치 카드가 있는데 4개는 샤시 내부 연결용이고 나머지 4개가 샤시간 연결용으로 사용된다. 하나의 스위치 카드는 10Gbps 링크들의 9x9 크로스바 스위치인데, 두개의 링크는 내부 연결용 스위치 카드와 접속하고 나머지 7개의 링크로 degree 7의 하이퍼큐브 네트워크를 구성하여 최대 128개의 샤시를 연결할 수 있다. Teraplex 한 샤시는 라인 카드 용량 기준으로 보면 150Gbp의 패킷 처리 용량을 제

공하므로 최대 19.2Tbps 까지 용량을 확장할 수 있다고 주장하였다. 그러나 전체 샤시 용량의 절반을 스위치 네트워크에 할당하면서 확장성을 강조한 구조임에도 불구하고 하이퍼큐브 역시 블록킹 네트워크이므로 실질적인 패킷 처리 용량은 라인 카드 용량보다 적을 수 밖에 없다.

최근 패킷 버스트를 이용한 스위치 용량 확장 시도가 있는데 이 기술은 전기적인 스위치나 광 스위치에 모두 적용될 수 있으나 특히 광 스위치의 잠재적인 확장 가능성에 기대를 걸고 있다. 광 스위치는 전기적인 스위치에 비하여 아직 성능 대비 비용면에서 경쟁력이 떨어지지만, 3차원적인 용량 확장 능력 때문에 저가의 고속 스위칭 능력을 갖는 광 소자가 개발된다면 십 수 Tbps 용량의 광 스위치가 실용화 될 수 있을 것이다. 광 스위치의 3차원적인 용량 확장 능력이란, 다단 구조에 의한 공간적인 확장, 하나의 광 링크를 통한 파장의 다중화, 단일 파장으로 전송할 수 있는 데이터의 대역폭 증가이다.

고속 라우터 기능은 논리적으로 제어 평면의 기능과 데이터 포워딩 평면의 기능으로 크게 분리할 수 있다. 제어 평면의 기능은 라우팅 테이블을 만들고 유지 관리 하는 기능, 장애처리와 시스템 운영 관리 기능 등이다. 라우팅 정보를 교환하기 위하여 RIP, IS-IS, OSPF, BGP와 같은 라우팅 프로토콜이 사용되는 데 통상적으로 전체 네트워크 트래픽 중에서 라우팅 정보를 전달하기 위한 트래픽은 약 5% 정도이다. 라인 카드에서 들어온 패킷을 보고 라우팅 프로토콜 패킷이면 내부 IPC를 통하여 라우팅 프로세서로 전달하는데 전달되는 트래픽이 많지 않으므로 라우팅 프로세서에서는 고속 패킷 처리가 요구되지 않는다. 내부 IPC를 통하여 전달되는 정보의 종류는 라우팅 정보와 시스템 형상, 통계, 장애 등과 같은 운용관리 정보들인데 큰 대역폭을 요구하지 않으며 라우팅 테이블을 초기화 할 경우에도 1Gbps 이하의 대역폭이면 시스템의 성능을 저하시킬 만한 병목이 되지는 않는다. 제어 기능들은 기본적으로 소프트웨어로 구현되며 Cisco IOS와 같은 라우터 운영 소프트웨어에 포함되어 있다. 관리기능에는 하드웨어 장

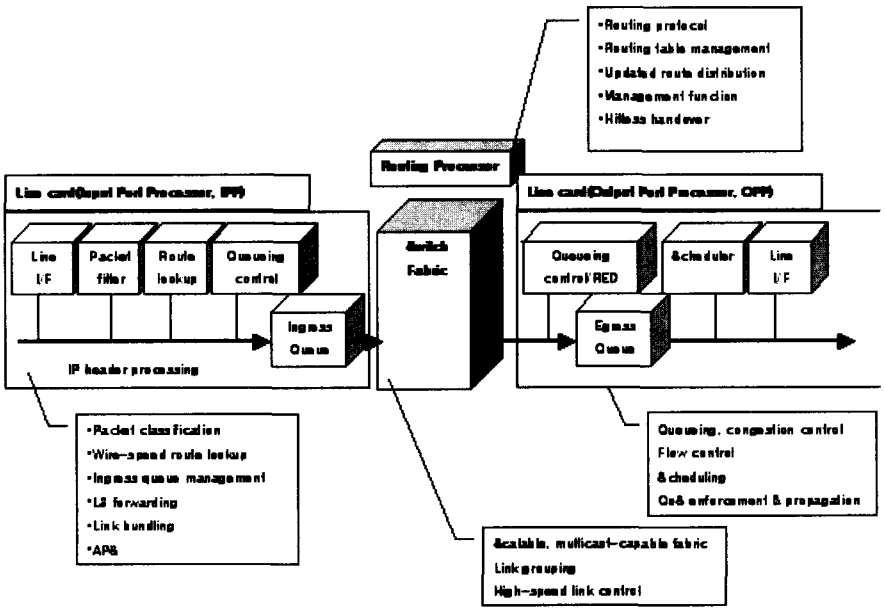

그림 1. 고속 대용량 라우터의 기능 요소들

애 감지와 같은 OAM 기능 뿐만 아니라 과금 및 망 운영관리를 위한 통계정보 수집 등이 포함된다. 이러한 기능들은 IETF RFC1757에 정의되어 있는 RMON과 같은 표준을 사용하여 원격으로 접근될 수 있다.

데이터 포워딩 평면의 기본 기능은 wire-speed 패킷 포워딩이며 이 기능은 라인 카드에서 수행된다. 이밖에 패킷 분류 및 필터링, 트래픽 관리 기능 등이 라인 카드에서 수행되는데 네트워크 프로세서(Network Processor Unit, 이하 NPU)가 핵심적인 역할을 한다. 네트워크 프로세서 기술 동향은 참고문헌 [10], [11]에서 정리하였다.

패킷 분류 기능(classification)은 단순한 L3 라우팅 결정에서부터 복잡한 L4 또는 L7 분류에 이르기 까지 다양하다. 패킷 분류는 네트워크 프로세서가 단독으로 처리하거나 NPU가 외부 검색 엔진 혹은 CAM(Content Addressable Memory)을 이용하여 하기도 한다. 기본적인 IP 라우팅(혹은 포워딩)은 패킷의 IP 헤더에서 추출한 IP 목적지 주소만을 가지고 LPM(Longest Prefix Match) 알고리즘에 따라서 next-hop을 결정하는 것이다. 인터넷에서 QoS를 지원하기 위하여 사용되는 DiffServ 라우팅을 위한 L4 또는 흐름 기반(flow-based) 패킷 분류는 패킷을 흐름(flow)으로 구분하기 위해서 TCP/UDP 헤더에 있는 정보를 사용한다. 이 유형의 패킷 분류는 전형적으로 five-tuple 룩업을 사용하여 수행된다. 즉, IP 발신 및 목적지 주소, TCP 발신 및 목적지 주소, IP 프로토콜에 따라서 단 한번의 검색으로 분류하는 것이다. L4 패킷 분류를 필요로 하는 또 다른 예는 패킷 필터링이다. ACL(Access Control List)에 따라서 IP 발신 및 목적지 주소, TCP 발신 및 목적지 주소 등을 보고 패킷을 필터링 하는데 이에 따른 패킷 포워딩 성능의 저하, 패킷 전달 지연 시간의 증가 등의 문제를 해결하는 것이 코어 라우터의 핵심 기술요소 중 하나이다.

트래픽 관리 기능은 QoS 보장을 위한 결정적인 요소이다. 일단 패킷들이 흐름으로 분류되면, 트래픽 관리 기능은 해당 흐름에 대한 SLA(Service Level Agreement)를 감독하는데 policing & shaping, 폭주 관리, 스케줄링의 3단계로 이루어진다. Policing은 흐름이 SLA에서 할당된 대역폭보다 많은 대역폭을 사용하지 못하게 하는 것이고 트래픽 shaping은 인터넷 트래픽의 버스트를 완화하는 smoothing 기능이다. Policing과 shaping은 dual leaky bucket으로 알려진 메커니즘을 쓴다. 폭주 관리 기능으로 RED(Random Early Discard) 또는 WRED(Weighted RED) 가 사용되는데, RED는 큐가 threshold에 도달하면 무작위로 패킷을 버린다. WRED는 각 트래픽의 클래스별로 다른 threshold와 drop rate을 추가함으로써 RED를 개선한 것인데 현재 대부분의 NPU에서 제공된다. 큐잉 또는 스케줄링은 각 패킷의 전송 시간과 순서를 결정한다. DiffServ 구현에서는 일반적으로 클래스 기반의 큐잉 만을 쓰기 때문에 비교적 적은 수의 큐(수백에서 수천)를 요구한다. 그러나 각 가입자들로부터 들어오는 흐름을 구분하고 우선순위가 다른 트래픽들에 대한 QoS를 보장하는 흐름별 큐잉(per-flow queueing)을 위해서는 수만개의 큐가 요구된다. 스케줄링 알고리즘으로는 WRR(Weighted Round Robin)과 WFQ(Weighted Fair Queueing) 등이 쓰이는데 스케줄러를 계층화 하고 각 계층마다 다른 알고리즘을 쓰기도 한다.

지금부터는 라인 카드내의 여러 인터페이스에 대하여 살펴보기로 한다. 네트워크 인터페이스는 주로 이더넷 종류와 POS(Packet Over SONET)이다. (그림 2)는 이더넷과 POS 인터페이스를 갖는 라인 카드에서 사용되는 표준 인터페이스를 정리한 것이다. GMII는 기가비트 이더넷을 위한 표준이며 8비트, 125MHz에서 동작한다. XGMII는 10기가비트 이더넷을 위한 MAC-to-PHY 인터페이스 표준으로서 32 비트, 312.5MHz에서 동작하며 HSTL

(high-speed transceiver logic) 시그널링을 사용한다. Utopia level 3는 3.2Gbps까지 지원하며 최대 클럭 속도 104MHz에서 32비트 데이터 경로를 제공한다. POS-PHY level 3(PL3)는 Utopia level 3와 유사하며 최대 OC-48까지 지원하는데 PL3는 OIF (Optical Internetworking Forum) SPI-3로 표준화되었다. OIF의 OC-192를 위한 표준인 SPI-4.1(SPI-4 Phase 1)은 200MHz 클럭으로 64비트 single-ended 인터페이스를 제공한다. SPI-4 Phase 2 (SPI-4.2)는 라인 당 622 Mbps를 얻기 위해 DDR(Double Data Rate)을 사용하고 311~322 MHz에서 동작하는 16비트 LVDS(Low Voltage Differential Signaling) 인터페이스이다. SFI-4는 OIF의 OC-192c framer-to-transceiver 인터페이스 표준으로서 16비트 LVDS 데이터 경로를 사용한다.

한편, 기존에 개발된 대부분의 네트워크 프로세서들은 독자적인 스위치 패브릭 인터페이스들을 제공하고 있으나 최근에는 표준 인터페이스를 제공하고자 하는 추세이다. 최근의 고속 스위치 패브릭 인터페이스로는 NPF(Network Processing Forum)의 CSIX-L1, LA-1, SI(Streaming Interface)가 있다. CSIX-L1은 전기적인 인터페이스와 기본적인 프레임 형식을 정의한다. 32, 64, 96, 혹은 128 bit의 데이터 경로 폭을 사용할 수 있고 클럭 속도는 최고 250Mhz까지 사용할 수 있다. OC-48 데이터 전송은 32비트 100MHz로 구현될 수 있고 OC-192는 200MHz에서 동작하는 64비트 인터페이스를 사용하여 구현할 수 있다. CSIX-L1은 각 라인 속도별로 클럭 속도와 데이터 폭의 조합을 규정하지 않기 때문에 호환성 문제가 있다. OC-48 라인 속도에서는 잘 동작하지만, 10Gbps에서는 핀 숫자와 클럭 속도가 문제가 되기 때문에 SI가 나오게 되었다.

NP Forum LA-1은 Look Aside 인터페이스 규격으로서 2002년 7월 첫번째 규격이 만들어 졌다. CAM, classification coprocessor, security processor 들을 NPU에 연결하기 위해서 사용되는데 133~200MHz DDR 동작 속도로 18비트 인터페이스를 제공한다. 144bit 룩업을 위해서 LA-1은 OC-48에서는 패킷 당 4번의 검색을, OC-192에서는 패킷 당 한번의 검색 결과를 전달해 줄 수 있다.

NPF SI(Streaming Interface)는 NPU-to-fabric 인터페이스로서 CSIX-L1을 대체하기 위해서 규격이 제정되고 있다. SI는 flow-through coprocessor들을 NUP와 연결하는데 쓰이거나 하나의 NPU 칩셋 내에 있는 여러 개의 칩을 연결하는데 쓰일 수 있다. CSIX-L2라고도 불리는 SI는 SPI-4.2에 기반을 두고 있으며, CSIX-L1에서 문제가 되었던 인터페이스 핀 숫자를 줄이기 위해서 16비트 LVDS 데이터 경로를 사용한다.

지금까지 개발된 대부분의 네트워크 프로세서들은 외부의 호스트 프로세서들과의 인터페이스로 PCI를 사용하고 있다. 그러나 네트워크 프로세서들의 성능이 향상됨에 따라서 더욱 고속의 호스트 프로세서 인터페이스가 필요하게 되었다. 고속 인터페이스는 네트워킹 장치들을 10Gbps 이상으로 확장하면서도 적절한 핀 숫자를 유지하기 위해서 필요하다. 이것은 NPU와 호스트 프로세서간의 연결에서 특히 문제가 되는 부분이다. HT(HyperTransport), RIO (RapidIO), PCI Express가 PCI와 PCI-X의 대안으로 부상하고 있는데 이들은 모두 고속 점대점 링크를 사용하며 표준 PCI보다 더 넓은 대역폭을 제공한다. RIO가 기술적으로 우수한데, 토폴로지 구성이 자유롭고, 대규모 시스템에서 확장성이 크며, end-to-end 흐름 제어, 오류 복구 기능을 제공할 뿐만 아니라 소프트웨어 오버헤드가 적다. RIO는 full-duplex point-to-point 링크로 구성되는 switched architecture를 정의한다. 물리적인 링크는 8비트 또는 16비트이며 2.5V에서 동작하는 LVDS를 사용한다. 클럭 속도는 125MHz, 250MHz, 500MHz 및 1GHz를 지원하는데 DDR 시그널링을 사용하여 250MB/s (8bit, 125MHz)

~ 4GB/s(16bit, 1GHz)의 대역폭을 제공한다. 1GB/s의 full-duplex 대역폭을 실현하기 위해서는 8비트 500MHz 구성이 일반적인 구현이다. RIO는 패킷 기반의 프로토콜을 사용하고 최대 256 바이트의 데이터 패이로드를 지원한다.

이에 비하여 HT는 제어 평면의 인터페이스에 적당한 사양을 제공한다. HT는 400MHz 클럭과 DDR을 사용하며 한 쌍의 데이터 핀이 800Mbps를 전송한다. 링크는 2, 4, 8, 16 혹은 32비트 폭이다. HT는 IEEE 표준 LVDS가 아닌 고속 diffrential signaling을 사용하는 chip-to-chip 인터페이스이다. 최대 31개의 장치들을 daisy chain으로 지원할 수 있고, 어느 정도의 소프트웨어 오버헤드만으로 적절한 수준의 오류 검출 기능을 제공할 수 있다. HT의 장점은 PCI-like 프로토콜 계층을 제공한다는 것이다. 따라서 이전에 PCI-base 시스템을 위해서 작성되었던 소프트웨어의 포팅이 용이하다는 것이다. PCI Express는 2002년 7월에 발표되었고 HT의 PCI 소프트웨어 호환성과 함께 RIO의 많은 기술적인 장점들을 수용한다. 더욱이 인텔을 비롯한 PCI-SIG업체들이 PCI 이후의 표준 인터페이스로 PCI-Express를 PC에서 사용하기로 결정함에 따라서 PCI Express가 널리 보급될 전망이다.

대부분의 저속 NPU들은 패킷 버퍼 용도로 저가의 SDRAM을 사용하지만 10Gbps NPU가 출현하면서 메모리가 문제가 되고 있다. 비록 최근 133/266MHz 속도의 DDR SDRAM이 보급되고 있지만 simplex 10Gbps NPU 조차도 최소 64bit 뱅크 3개를 요구하고 패킷 버퍼 용도로만 수백개의 핀을 쓰고 있다. 따라서 단순히 DRAM에서 비용을 줄이는 것은 NPU 패키지에서 추가되는 핀의 개수에 따른 비용을 생각하면 의미가 없게 된다. RDRAM(Rambus DRAM), FCRAM(Fast-Cycle DRAM), RLDRAM(Reduced-Latency DRAM)은 핀 당 대역폭을 증가 시켜서 핀 개수를 줄이기 위한 대안이 되고 있다. RDRAM은 표준 장치에는 400/800MHz (400MHz DDR) 속도의 인터페이스를 제공하고, short-channel 장치들에게는 533/1,066MHz를 제공한다. FCRAM은 200/400MHz에서 동작하는 부품은 이미 공급되고 있고 300/600MHz에서 동작하는 부품도 곧 출시될 것으로 보인다. RLDRAM은 FCRAM과 유사한 핀 당 대역폭과 낮은 지연시간을 제공한다. 테이블 룩업 역시 10Gbps에서는 문제가 된다. Full-duplex OC-192 인터페이스에서는 50Mpps(Million packets per second)의 패킷을 wire-speed로

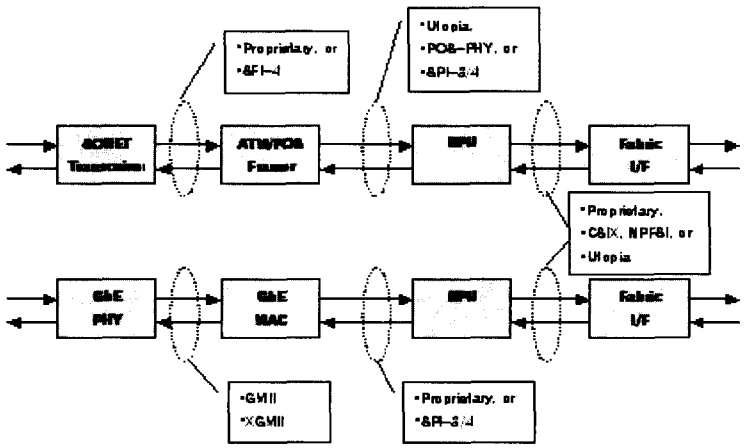

그림 2. 라인카드 표준 인터페이스

처리해야 하는데 패킷 당 두 번의 룩업을 한다 하더라도 100Msps(Million searches per second) 성능을 요구하게 된다. 이것은 현재 출시되어 있는 CAM 중에 일부 빠른 것들이 제공하는 속도이다. 패킷 당 네 번의 룩업을 하려면 각 72비트씩 두 개의 CAM 뱅크를 써야 한다. 패킷 당 더 많은 룩업이 필요하거나 IPv6 처럼 144bit 보다 긴 키(key)가 필요한 경우에는 필요한 CAM 인터페이스의 수를 증가시켜야 하는데 이때는 핀 숫자가 심각한 문제가 된다. 일부 NPU 벤더들은 이러한 이유로 NPF의 새로운 규격인 18비트 LA-1 인터페이스 도입을 시도하고 있다. LA-1은 4개 이상의 CAM 인터페이스를 하나의 NPU에서 제공하는 것을 지원한다.

Ⅲ. 코어 라우터 성능

본 장에서는 차세대 인터넷 백본망 구축을 위한 초고속 라우터의 성능에 대해서 고찰한다. 광범위한 서비스를 제공하는 통합 IP 망을 구축하기 위해서 초고속 라우터에서 제공되어야 할 기능 및 성능 요구사항은 다음과 같다:

- IP VPN과 MPLS L2 및 L3 VPN이 지원되어서 기존에 주 수입원이었던 서비스들을 지속적으로 수용할 수 있어야 한다.
- 또한 전달 지연 시간(latency)은 짧고 일정하게 유지되어야 하며 지터(jitter)는 최소한으로 발생되어야 하고 예측 가능하여야 한다.
- 개별적인 트래픽 흐름(flow)을 관리하고 SLA(Service Level Agreement)를 항상 만족시키는 지 확인하기 위해서는 다양한 QoS 관리 기능을 제공할 수 있어야 하고 이로 인한 성능 저하가 초래되지 않기 위해서는 강력한 큐(queue) 관리 기능도 필요하다.
- 끝으로 차세대 통합 IP 망에서는 IPv6 지원도 하드웨어 기반으로 이루어져야 한다.

이밖에 코어 라우터에 대한 망 사업자의 포괄적인 요구사항은 영국의 BTexact Technologies에서 작성한 문서[12] 등에서 찾을 수 있다. 이와 같은 기능 및 성능 요구사항을 완벽히 만족시키지 못하는 구현은 망 사업자들에게 심각한 손실을 발생시키고 하드웨어의 재설계 및 운영 소프트웨어의 재개발 등을 초래하기 때문에 설계 단계부터 철저한 검증이 필요하다.

이하에서는 Light Reading에서 Charlotte, Cisco, Foundry, Juniper 등의 주요 인터넷 코어 라우터 제품들을 대상으로 실시한 성능 시험 결과 보고서[13]를 바탕으로 코어 라우터의 주요 성능 항목 별 구현 기술 현황을 분석한다.

1. IP 패킷 포워딩 성능

라우터 특히 코어 라우터의 가장 기본적인 기능 요구사항은 OC-48과 OC-192 라인 인터페이스에서의 wire-speed 고속 패킷 포워딩 능력이다. 라우터의 패킷 포워딩 능력을 측정하는 척도로 throughput과 포워딩 속도(forwarding rate)가 사용되는데 throughput은 라우터가 패킷을 하나도 유실하지 않고 전송할 수 있는 최대의 부하이고(RFC1242), 포워딩 속도는 최대 부하가 주어졌을 때 라우터가 정상적으로 포워딩하는 패킷을 Mpps단위로 표시한 것이다. 입력 트래픽의 패턴은 실제 인터넷 트래픽과 유사한 Internet mix와, 40바이트의 짧은 패킷만으로 구성된 패턴이 사용되었다. 시험에 참가한 일부 제품은 약간의 패킷 손실은 있었으나 wire-speed에 근접하는 성능을 보여 주었고, 나머지 제품들은 낮은 수준의 부하에서도 패킷을 손실하거나 과부하에서 라우팅 테이블 엔트리의 일부가 지워지는 등의 문제점을 보였다. 버퍼의 크기를 증가시키면 패킷 포워딩 성능을 어느 정도 향상시킬 수도 있으나 전달 지연 시간 역시 증가하게 되므로 결코 바람직한 방법은 아니다. IP 라우터에서 패킷 포워딩 성능을 말할 때는 throughput, 포워딩 속도, 전달

지연 시간, 패킷 순서 유지 등이 종합적으로 고려하여야 한다.

2. MPLS 패킷 스위칭 성능

L3 기능인 IP 라우팅과는 달리 MPLS는 L2 스위칭 기능이다. 일단 LSP(즉, MPLS 터널)가 설정되면 모든 패킷은 라우팅을 위한 룩업 과정 없이 고속으로 스위칭 된다. MPLS는 모든 패킷에 4바이트의 라벨을 붙이기 때문에 MPLS throughput이 IP 라우팅 throughput보다 약간 떨어질 수도 있으나 룩업 과정이 없으므로 실제 구현에서는 별 차이가 없는 것으로 나타났다. 오히려 시험에 참여한 한 업체 제품의 경우 OC-192 인터페이스에 대한 40바이트 IP 패킷 throughput은 52%였으나 MPLS throughput은 100%에 도달하였다. 일부 업체는 MPLS 라벨 스위칭을 제공하지 않거나 MPLS 코드를 개발 중에 있었다.

3. 패킷 포워딩 지연 시간

패킷 포워딩 지연 시간은 지연에 민감한 서비스에서는 throughput 만큼이나 중요한 성능 요소이다. 라우터는 이상적으로 짧고 일정한 지연 시간을 유지해야 한다. 즉, 지연 시간은 짧을수록 좋으며 최소 지연 시간과 최대 지연 시간의 차이가 최소화되어야 한다. 앞에서도 언급한 바와 같이 패킷 포워딩 지연 시간은 throughput, 패킷 손실율 등과 같은 다른 패킷 포워딩 성능들과 밀접한 상관 관계가 있어서 지연 시간만으로 라우터 전체적인 성능의 우열을 비교하는 것은 무리이다. 아래의 표는 대표적인 코어 라우터 제품들의 패킷 포워딩 지연 시간에 대한 성능 시험 결과를 정리한 것이다.

표 1. OC-48 인터페이스 패킷 포워딩 지연 시간 - 40바이트 IP 패킷 (단위: microsecond)

	최소 시간	평균 시간	최대 시간
Charlotte' s Aranea-1	19.1	8,302.8	32,979.7
Cisco 12416	17.7	1,934.6	15,561.8
Foundry NetIron	10.0	35.7	160.0
Juniper M160	13.2	15.0	21.0

표 2. OC-48 인터페이스 패킷 포워딩 지연 시간 - Internet Mix (단위: microsecond)

	최소 시간	평균 시간	최대 시간
Charlotte' s Aranea-1	23.0	-	110.5
Cisco 12416	15.9	251.9	1,765.6
Foundry NetIron	9.8	130.1	769.5
Juniper M160	13.2	64.9	169.5

표 3. OC-48 인터페이스 패킷 포워딩 지연 시간 - MPLS (단위: microsecond)

	최소 시간	평균 시간	최대 시간
Cisco 12416-40바이트 IP 패킷	16.3	16,034.6	91,215.5
Cisco 12416-Internet Mix	16.0	652.4	6,003.5
Juniper M160-40바이트 IP 패킷	13.5	15.0	20.7
Juniper M160-Internet Mix	14.2	69.7	186.2

표 4. OC-129 인터페이스 패킷 포워딩 지연 시간 - 40바이트 IP 패킷 (단위: microsecond)

	최소 시간	평균 시간	최대 시간
Cisco 12416	14.3	26.4	160.7
Juniper M160	12.7	26.7	58.8

표 5. OC-192 인터페이스 패킷 포워딩 지연 시간 - Internet Mix (단위: microsecond)

	최소 시간	평균 시간	최대 시간
Cisco 12416	14.3	497.0	4,896.9
Juniper M160	13.0	57.8	10,863.1

표 6. OC-192 인터페이스 패킷 포워딩 지연 시간 - MPLS (단위: microsecond)

	최소 시간	평균 시간	최대 시간
Cisco 12416-40바이트 IP 패킷	14.9	13,202.6	58,995.6
Cisco 12416-Internet Mix	14.1	2,324.3	7,098.1
Juniper M160-40바이트 IP 패킷	12.2	26.2	82.3
Juniper M160-Internet Mix	13.5	90.1	39,991.1

4. 패킷 순서 맞춤

라인 인터페이스 링크가 고속화됨에 따라서 패킷 순서 맞춤(packet reordering)이 문제가 되기 시작했다. 특히, Juniper의 OC-192 카드가 입력되는 패킷들을 wire-speed로 처리하기 위하여 4개의 서브 스트림으로 나누어 처리하기 시작하면서 경쟁사와의 뜨거운 논쟁이 계속되고 있다. 문제는 하나의 TCP 연결에 속한 연속한 패킷들이 서로 다른 경로를 통과하면서 패킷의 순서가 맞지 않게 도착할 수도 있다는 것이다. 순서대로 도착하지 않은 패킷들 때문에 발생하는 재전송, 지연, 응용 서비스의 품질 저하 등이 얼마나 심각할 지에 대해서는 쉽게 결론을 내릴 수도 증명을 할 수도 없는 문제이다. 다만, 가능하면 순서 맞춤 문제를 발생시키지 않는 것이 좋겠지만, 패킷 포워딩 엔진의 성능 향상을 위한 기술 개발 속도가 라인 인터페이스 링크의 고속화를 따라가지 못할 경우에는 피할 수 없는 문제이다.

5. LPM

라우팅 테이블을 룩업하는 LPM(Longest Prefix Matching) 성능 역시 라우터의 패킷 포워딩 성능을 결정 짓는 중요한 요소이다. 이론적으로는 라우팅 테이블의 크기의 차이나 유사한 라우팅 엔트리 주소 /서브넷 마스크의 유무 등에 따라서 성능의 차이가 없어야 하지만, 성능 시험 결과를 보면 이에 따른 패킷 포워딩 성능 저하가 많게는 27%까지 발생함을 알 수 있다.

6. 라우팅 테이블 크기 및 MPLS LSP 테이블 크기

인터넷 규모는 최근 경기 불황에도 불구하고 기하급수적으로 증가하고 있다. 통합 IP 망에서 이동통신 가입자를 위한 Mobile IP가 본격적으로 보급되면 향후 몇 년 안에 라우팅 테이블의 크기와 MPLS LSP(Label Switched Path) 테이블의 크기가 2 ~ 3배 커지는 것은 충분히 예측할 수 있는 일이다. 따라서 주소 용량 확장성이 코어 라우터의 수명 연한을 결정하는 한 요인이 될 수 있는데, 앞에서 시험한 라우터 제품들은 대부분 수십만 라우팅 엔트리들을 관리할 수 있음을 보여 주었다. 라우팅 테이블 크기의 확장과 관련하여 고려해야 할 사항은 다음과 같다. 첫째, 라우팅 테이블의 크기는 일반적으로 메모리 크기에 따라 결정된다. 따라서 라우팅 테이블의 크기를 확장하려면 메모리를 증설하면 될 것이다. 일정한 크기가 넘으면 하드 디스크에 스와핑(swapping)하는 방법으로 물리적으로는 얼마든지 라우팅 테이블의 크기를 확장하는 것이 가능할 것이다. 그러나 소프트웨어의 성능이나 기타 제한 조건에 따라서 라우팅 테이블의 크기가 무한히 확장되는 것은 아니다. 더욱이 대부분의 코어 라우터는 고속 패킷 포워딩을 위하여 L3 BGP 테이블과는 별개로 각 라인 카드에 L2 포워딩 테이블을 유지하기 때문에 라우팅 프로세서에 있는 BGP 라우팅 테이블 엔트리 숫자

만큼의 다른 라우터와 패킷을 교환할 수 있는 것은 아니다. 라인 카드에 있는 포워딩 테이블의 크기는 앞 장에서 설명한 바와 같이 메모리, NPU와의 인터페이스 문제 등에 의해서 제한을 받는다.

MPLS LSP 터널 개수는 이미 망 사업자들이 수백만 개를 요구하고 있으나 시험 결과 보고서에 의하면 Juniper는 10,000개, Cisco는 5,000개 까지 제공 가능하였다.

7. 라우트 플래핑과 수렴

실제 인터넷 상에서는 한번에 수 만개의 라우트(route)가 사라지거나 수 천개의 라우트가 추가되는 라우트 플래핑(flapping)이 파상적으로 발생한다. 코어 라우터는 이러한 상황에서도 포워딩 성능을 일정하게 유지하여야 한다. 즉, 라우팅 정보가 바뀌지 않은 안정적인 경로에 대한 포워딩 속도는 일정해야 하고, 라우팅 정보가 바뀐 경로(flapped path)에 대해서도 기본 라우트(primary route) 이외에 2차(secondary), 3차 라우트(tertiary route)가 제공되므로 기본 라우트가 플래핑 되는 도중에도 2차 라우트로 정상적인 패킷 포워딩이 이루어져야 한다. 30초 간격으로 200,000개의 라우트 중에서 50,000개의 라우트를 반복적으로 철회(withdraw)/재고시(readvertise)한 라우트 플래핑 성능 시험 결과는 플래핑이 일어나는 경로에 대한 포워딩 성능이 급격하게 떨어지고 원래의 성능으로 복구하는데 많은 시간이 걸림을 보였다. 일부 제품은 최초의 플래핑이 발생한 후에 포워딩 성능이 급격히 하락하여 회복되지 않은 경우도 있었다.

라우팅 테이블 전체가 철회된 후 재공지 될 경우에는 포워딩 throughput은 이론적으로 0%에서 100%로 수직 상승한 후 트래픽을 최대한 포워딩 함으로써 완벽한 수평선을 보여 주어야 한다. 그러나 시험 결과 보고서에 따르면 수렴(convergence) 시간이 짧은 제품들의 경우에도 20~50초가 걸렸으며 60초 간격으로 반복된 수렴 시험에서 최대 성능까지 도달하지 못하는 경우도 있었다.

8. 필터링

패킷 필터링 규칙의 적용은 라우터 성능 저하를 초래한다는 것은 널리 알려진 사실이다. 실제 시험 결과를 보면 적용된 필터링 규칙에 따라서 패킷 포워딩 성능은 OC-48의 경우에는 50%, OC-192의 경우에는 최대치의 30% 까지 떨어졌다. 전달 지연 시간 역시 필터링의 영향을 받는데 한 업체 제품은 OC-48의 경우 필터링을 적용하지 않을 때보다 4~11배의 전달 지연 시간을 기록했다.

9. QoS

인터넷 서비스 품질 보장은 지금까지의 많은 노력에도 불구하고 아직 해결되지 않고 있는 문제이다. 그럼에도 불구하고 망 사업자들은 코어 라우터에서 흐름별(flow based) QoS는 아니더라도 최소한 CoS(Class of Service)가 제공될 것을 요구하고 있는데, 이것은 각 클래스별로 가입자들에게 서로 다른 요금을 책정함으로써 수익을 낼 수 있는 수단이 되기 때문이다. 라우터에서의 CoS 제공은 패킷 분류에 따른 클래스별 큐를 얼마나 잘 관리하는가의 문제인데, 이에 대한 성능은 두 가지 척도로 측정된다. 첫 번째는 트래픽 폭주 상황에서 각 클래스별 포워딩 속도이다. 즉, 각 클래스에 속한 패킷의 손실율이 어떤지를 보는 것이다. 두번째는 정해진 클래스별 트래픽 비율이 얼마나 잘 지켜지는가 하는 것이다. 성능 시험 결과는 예상대로 낮은 순위의 트래픽을 희생시킴으로써 최우선 순위의 트래픽은 비교적 양호하게 보호됨을 보였다.

Ⅳ. 결론

지금까지 초고속 라우터의 필요성, 구성 요소들의 기술 현황, 현재 인터넷 망에 도입되고 있는 코어 라

우터들의 성능에 대하여 고찰하였다. 살펴 본 바에 의하면 OC-48, OC-192 라인 인터페이스를 갖는 수 Tbps 용량의 라우터 개발은 기술적으로 가능하다. 초고속 라우터 개발에 있어서 기술적인 어려움은 일차적으로 네트워크 프로세서나 스위치 패브릭 칩셋 등 핵심 칩셋 개발에 있지만 구성 요소들간 인터페이스를 위한 고속 신호 처리도 기술적으로 어려운 문제이다. 또한, 본 고에서는 하드웨어 측면의 검토에 치중하였으나 앞장에서 기술한 코어 라우터 성능 시험 결과에서 알 수 있듯이 실제로 라우터 성능은 소프트웨어의 확장성과 안정성에 의해 크게 좌우된다. 향후 인터넷 규모와 트래픽의 기하급수적 증가가 계속되면 100만 라우팅 엔트리, OC-768 라인 인터페이스, 십 수 Tbps 용량의 라우터가 요구될 것이다. OC-768 라인 인터페이스에서 wire-speed로 패킷을 포워딩 할 수 있는 NPU는 가까운 시일에 개발되지 않을 전망이다. 따라서 ingress 트래픽을 몇 개의 sub-stream으로 나누어서 처리하고 egress에서 한데 묶어 주는 방법이 사용될 것이다. 이 경우 패킷 순서 맞춤의 문제가 해결해야 할 과제로 남는다. 전기적인 스위치로 십 수 Tbps 이상 수 십 Tbps 까지 확장 가능한 스위치 패브릭을 개발하는 것은 기대하기 어려운 실정이다. 이러한 용량은 광 스위치로 구현이 가능할 것이다. 아직까지는 광 스위치 소자의 가격 및 스위칭 시간 등이 문제가 되어서 전기적인 스위치에 비하여 경쟁력이 없으나, 십 Tbps 이상의 스위칭 용량이 요구된다면 본격적으로 도입될 것으로 본다. 끝으로, IPv6의 본격적인 도입을 앞두고 하드웨어 기반의 IPv6 패킷 처리를 비롯한 IPv6 기능도 구현되어야 할 것이다.

참고 문헌

[1] Coffman, and Odlyzko, "Growth of the Internet," Optical Fiber Telecommunications IV Journal, July 2001

[2] 이형호, 이규호, 주성순, "초고속 대용량 라우터 기술," 한국통신학회지, 제17권 2호, pp.41~53, 2000년 2월

[3] 변성혁, 이형호, "차세대 IP 스위치 및 라우터 기술," Telecommunications Review, 제10권 1호, pp.23~35, 2000년 2월

[4] 이형호, 김봉완, 안병준, "테라비트 라우터 기술," *Telecommunications Review*, 제11권 2호, pp.237~247, 2001년 4월

[5] 전종암, 변성혁, 안병준, 이형호, "테라비트 라우터 기술 동향," 전자공학회지, 제28권 9호, pp.50~59, 2001년 9월

[6] 이영천, "차세대 초고속 라우터 구조 및 전망," 텔레콤, 제17권 제1호, pp.15~20, 2001년 6월

[7] 남민우, 이현수, "상용 네트웍 프로세서를 이용한 확장 가능형 고속 라우터 구조," 텔레콤, 제17권 제1호, pp.21~33, 2001년 6월

[8] 최명수, 강병창, "Router clustering을 이용한 Tera급 라우팅 성능을 갖는 Scalable router 구조 제안," 텔레콤, 제17권 제1호, pp.34~40, 2001년 6월

[9] Yankee Group, "Core Routers: Challengers & Challenges to the Status Quo," Carrier Convergence Infrastructure Report, Vol.2, No.11, October 2001

[10] 임준서, "네트워크 프로세서의 기술 동향," 전자공학회지, 제28권 10호, pp.82~93, 2001년 10월

[11] 김봉완, 이형호, "네트워크 프로세서의 응용과 표준화 동향," 전자공학회지, 제28권 10호, pp.94~101, 2001년 10월

[12] BTexact Technologies, "Carrier requirements of core IP routers 2002: A technology benchmark from BTexact

Technologies," White Paper, Reference: 42267, Issue 1, Feb. 14, 2002
[13] Light Reading, "Internet Core Routing Test: Complete Results," http://lightreading.com, June 29, 2001

안 병 준

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 주 관심분야 : ATM, 트래픽 제어, QoS, 고속 라우터 기술

김 영 선

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 주 관심분야: ATM, 트래픽 제어, QoS, 고속 라우터 기술, 인터넷, 라우팅 프로토콜, 이동통신