테라비트 라우터 기술 동향

KRnet2001

2001. 6. 27

안병준
bjahn@etri.re.kr
Contents

- IP router overview
- Design issues on fast routers
- Switching fabrics
- Case study
- Summary
Functions of IP Router

- **IP Packet Processing**
  - Lookup packet DA in forwarding table
    - If known, forward to correct port
    - If unknown, drop packet
  - Decrement TTL, update header checksum
  - Forward packet to outgoing interface
  - Transmit packet onto link

- **Maintain routing information**
  - Listen for updates/update neighbors
  - Construct forwarding table
Functional Elements of Fast Router

Routing protocol
Routing table management
Updated route distribute management function

Network Processor

Line card(input port processor, IPP)
- Line I/F
- Packet filter
- Route lookup
- Queueing control
- IP pkt
- Ingress Queue
- Packet classification
  - Line-speed route lookup
  - ingress queue management
  - L3 forwarding to Gbits/second

Switch Fabric
- IP Header processing

Switching Fabric (SF controller)
- Highly scalable, multicast-capable fabric for inter-card communications

Line card(output port processor), OPP
- Queueing control / RED
- Egress Queue
- Scheduler
- Line I/F
- Queuing, congestion handling flow control scheduling
- QoS enforcement & propagation

- Line-speed route lookup
- Ingress queue management
- L3 forwarding to Gbits/second
First Generation Routers

Typically <0.5Gb/s aggregate capacity

Source: http://www.stanford.edu/~nickm
Second Generation Routers

Typically <5Gb/s aggregate capacity

Source: http://www.stanford.edu/~nickm
Third Generation Routers

Switched Backplane

Typically <50Gb/s aggregate capacity

Source: http://www.stanford.edu/~nickm
Classification of Routers

- Access Router
- Enterprise Router
- Core Router
Access Routers

- Link home or small business office with an ISP
- Various media technologies
  - POTS modem, ADSL, cable modem
- Various access protocols
  - SLIP, PPP, VPN
- Capacity: Several Mb/s (ADSL)
- Low cost
Enterprise Routers

- Interconnect many subnets of a campus or an enterprise network
- Large number of ports
- Multiple protocols (IP, IPX, ATM..)
- Quality of Service (QoS)
- Multicast
- Security (firewall)
Core Routers

- High speed ports (OC-12, OC-48, OC-192)
- Large routing table (~ 200K)
- Reliability and speed
- Gigabit/Terabit router
테라급 라우터의 필요성

인터넷 이용자의 급증, 서비스의 다양화 및 고급화, 속도의 고속화 요구
 백분당에서 처리해야 할 트래픽 급증
 ◆ 전세계적으로 매 3~6개월마다 2배씩 (미국의 경우 2001년 4.5Tbps, 2005년 4,500Tbps)
 ◆ 국내의 경우 매 4~6개월마다 2배씩 (2000년 250Gbps, 2005년 250Tbps)
 ◆ 국내의 경우 250Tbps의 트래픽을 1Tbps 라우터로 처리하기 위해 소요되는 백분 라우터의 수는 약 850 대임
IP 백분당의 트래픽 증가 고려, OC-192, OC-768 WDM 링크의 망 적용이 필요

Growth of Internet POPs
Source : Nortel, assuming 5X/year growth

Terabit Routing: Simplifying the Core
테라급 라우터 설계시 고려 사항

<table>
<thead>
<tr>
<th>Requirement</th>
<th>Definition</th>
<th>Benefit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scalable Capacity</td>
<td>• Large capacity range (10’s Gigabits to 10’s Terabits)</td>
<td>• 3+ year product life</td>
</tr>
<tr>
<td></td>
<td>• Incremental scalability</td>
<td>• Commonality across product lines</td>
</tr>
<tr>
<td></td>
<td>• On-line expansion</td>
<td>• Pay as you grow economics</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Flexibility</td>
</tr>
<tr>
<td>Multi-Service</td>
<td>• Integrated packet and circuit switching</td>
<td>• Lower system cost</td>
</tr>
<tr>
<td></td>
<td>• Adaptive usage of packet and circuit</td>
<td>• Rapid provisioning of services</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Better utilization of bandwidth</td>
</tr>
<tr>
<td>High Port Density</td>
<td>• 1 Tbps non-blocking switching capacity in half-rack</td>
<td>• Small footprint</td>
</tr>
<tr>
<td></td>
<td>• ≤2 switching fabric chips per 10G port across capacity range</td>
<td>• Low chip count</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Lower cost, total power</td>
</tr>
<tr>
<td>Quality of Service (QoS)</td>
<td>• Multiple classes of service</td>
<td>• Service differentiation</td>
</tr>
<tr>
<td></td>
<td>• WFQ or strict priorities among classes</td>
<td>• Enables revenue generation via SLA-based services</td>
</tr>
<tr>
<td>Reliability, Availability &amp;</td>
<td>• Fault tolerance</td>
<td>• 99.999% uptime</td>
</tr>
<tr>
<td>Serviceability</td>
<td>• “N+1” redundancy</td>
<td>• Lower total cost of ownership</td>
</tr>
<tr>
<td></td>
<td>• In-service maintenance</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Common FRUs</td>
<td></td>
</tr>
</tbody>
</table>
Design Issues on Fast Routers

- Wire-speed IP packet forwarding
  - Flow classification
  - Longest Prefix Matching (LPM) IP address Lookup

- Fast switching fabric
  - Large Capacity, Scalability
  - High performance

- QoS support
  - Scheduling (WFQ, WRR, DRR)
  - Per-flow/per-class queueing
  - Congestion control (RED)
Wire-Speed Packet Forwarding (1)

- Wire-speed means
  - 26M pkt/sec at OC192(10Gb/s) interface (40byte pkt)
  - In full-duplex, each port must handle 52 Million PPS
  - OC768(40Gb/s) interface will be required soon.

- Forwarding Engine (FE) for each interface is not sufficient  
  Parallel/pipelined approach
3 parallel/pipeline flows allow 3x the throughput over normal serial processing

.. per packet processor
Complex line cards

Typical IP Router Linecard

Optics

Physical Layer

Framing & Maintenance

Packet Processing

Lookup Tables

Buffer & State Memory

Buffer Mgmt & Scheduling

Switch Fabric

Arbitration

OC192c linecard:
~10-30M gates
~2Gbits of memory
~2 square feet
>$10k cost
100’s of Watts

Source: http://www.stanford.edu/~nickm
Scalability Limitations on Switch Fabrics

- Output queuing/shared queueing switch
  - Internal speedup of N or 2N limits the maximum switching capacity
- Single Chip
  - Clock speed ~ 100MHz (I/O)
  - Pin Limitation: 400 ~ 600 pin (package)

- Input queueing switch for large packet switch
- Bit Slice or Multistage configuration to increase the switching capacity
Trends in Fast Switching Fabric Design

- Crossbar with VOQ + scheduler architecture is preferred
- Most switch fabrics are fixed length cell switches. (variable length packet should be segmented before switching)
- Single stage architecture with 16 or 32 I/O ports
- Very fast line speed such as 2.5Gb/s or 10Gb/s
Scalability Limits of Crossbar with VOQ

- Most scheduler assumes single crossbar
- Scheduling complexity limits the maximum number of I/O ports (generally up to 16 or 32)
- Switching capacity can be expanded with bit-slicing method, but also limited.

⇒ Multistage architecture is inevitable in terabit router
Multistage Architecture : Clos Network

- Multiple path between any input/output pairs
- Complex path allocation control needed
- Port expansion factor and internal speedup are required for nonblocking
Objectives

- Provide a standard interface for connecting traffic managers to switch fabric
- Support board level connections of 6~8 inches
- Support operation at up to OC192

Non-objectives

- Support a connector between traffic managers and fabric
- Provide interoperability between different traffic manager products
- Provide interoperability between different fabric products

Members: Power X, XaQti(now Vitesse), AMCC, Broadcom, C-Port(now Motorolar), Maker Communicaitons(now conexant), Vitesse, Agere, MMC Networks(now AMCC), NeoCore, IBM, etc.
CSI-X-Based Switch Design
Case Study

- Juniper M160
- CISCO 12416 GSR
- Avici TSR
- Lucent NX64000
- Pluris Teraplex 20
Juniper M160 Overview

- Switching capacity: 160Gbps (full duplex 80G)
- 8 slots (10Gb/s per slot)
- 4 switch fabric/IP forwarding module (SFM)
- Distributed shared memory switch
- 4 centralized FEs, each with 40Mpps performance (total 160Mpps)
- Designed with custom ASICs including FE, Buffer manager, and even SONET framers
- Mid-plane architecture
- BSD-based JUNOS OS
- QoS
  - 4 Class of Queue per interface
  - Policing, WRR, RED
  - Packet Classification
M160 Router Shelf

M160 Router Front and Back Views

Front View
- Craft Interface with LEDs
- FPCs with PICs
- OC-192c/STM-64 PIC
- Cable Management System
- Air Filter

Back View
- Upper Impeller
- SFMs
- MCS
- Routing Engines
- Lower Impeller
- Circuit Breaker Box
- Dual Power Supplies
M160 Packet Forwarding Architecture

- **SFM**
  - Route lookup: 40Mpps
  - Filtering
  - Packet distribution

- **FPC**: 4 I/O manager + 2 packet director

- **Processor**
  - SFM / FPC: PowerPC 603e
  - Routing Engine: Intel mobile pentium 333MHz
M160 Flexible PIC Concentrator

- 4 PIC slots
- 4 I/O Manager ASICs
  - SAR
  - Queueing
- 12.8 Gbps of throughput
- 2 FPC types
  - FPC1: DS3 ~ OC12, GE
  - FPC2: OC48c, Tunnel
- PIC and FPC of OC192c interface are integrated in a single board
Packet Flow in M160

Logical View of M160 Architecture

- Incoming Traffic
- Outgoing Traffic
## M160 Port Densities

<table>
<thead>
<tr>
<th>PIC Type</th>
<th>Ports per PIC</th>
<th>PICs per Chassis</th>
<th>Ports per Chassis</th>
<th>Ports Per Rack</th>
</tr>
</thead>
<tbody>
<tr>
<td>OC-3/STM-1* SONET/SDH</td>
<td>4</td>
<td>32</td>
<td>128</td>
<td>256</td>
</tr>
<tr>
<td>OC-3/STM-1* ATM</td>
<td>2</td>
<td>32</td>
<td>64</td>
<td>128</td>
</tr>
<tr>
<td>OC-12/STM-4 SONET/SDH</td>
<td>4</td>
<td>32</td>
<td>128</td>
<td>256</td>
</tr>
<tr>
<td>OC-12/STM-4 ATM</td>
<td>1</td>
<td>32</td>
<td>32</td>
<td>64</td>
</tr>
<tr>
<td>ChOC12 /STM4* to DS3</td>
<td>1</td>
<td>32</td>
<td>32 (384 DS3s)</td>
<td>64 (768 DS3s)</td>
</tr>
<tr>
<td>(12 DS3 Ports)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GE (SX and LX)</td>
<td>2</td>
<td>32</td>
<td>64</td>
<td>128</td>
</tr>
<tr>
<td>OC-48/STM-16 SONET/SDH</td>
<td>1</td>
<td>32</td>
<td>32</td>
<td>64</td>
</tr>
<tr>
<td>OC-192c/STM-64 SONET/SDH</td>
<td>1</td>
<td>8</td>
<td>8</td>
<td>16</td>
</tr>
</tbody>
</table>
Cisco 12416 GSR

- 16 slots of 10Gb/s capacity
- Max. 15 line cards (one for routing engine GRP)
- Crossbar switch fabric with virtual output queues
- 4 bit-slicing crossbar (2 CSC + 3 SFC, 4:1 redundancy)
- 25Mpps IP forwarding performance at each line card
- WFQ, RED, WRED
- IOS
- IDT R5000 for main processor
Switch Fabric

- 2.5Gbps switch fabric (12016) vs. 10Gbps switch fabric (12416)
- Minimum 1 CSC + 3 SFCs
- Clock and Scheduler Card (CSC)
  - Switch fabric + scheduler + system clock
- Switch Fabric Card (SFC)
- Second CSC for 4:1 redundancy
Maintenance Bus (MBus)

- Links all line cards, GRP, switch fabric cards, alarm cards, power shelf, and blower modules
- Each router module has MBus module
- Two redundant buses
- Functions
  - Power-up and power-down control
  - Device discovery
  - Code download
  - Environmental monitoring and alarms
  - Diagnostics
Avici TSR (Terabit Switch Router)

- 40 slots per Bay, up to 14 Bays
- 5.6 Tbps claimed (if 10Gbps per slot)
- Currently 2-Bay configuration available
- 1-port dual slot OC-192c module set
- Distributed 3-d toroidal mesh interconnection (each line card has switch fabric part)
- Composite Links: logical OC768 link with 16 OC links
- Software: iPriori
TSR Bay Configuration

Single Bay

14 Bays: Full Configuration
TSR Line Card Modules

- Midplane design (I/O + router/switch module)
- Input/output framing ASIC for L2 protocols
- Forwarding ASIC for IP lookup, routing, classification
- Packet scheduler ASIC for multicast, composite trunks, fabric scheduling and queueing
- Fabric switching ASIC and packet memory
- QoS ASIC for output queueing, WRED and WFQ
- 200MHz RISC processor for control plane processing
1-port OC48c Line Card of TSR

From NGN’98 presentation
Server Module

- Two module for 1:1 redundancy
- 576MHz CPU, 320MB memory, 32KB NVRAM
- Warm stand-by server
Composite Links Concept
Composite Links Technology

- Combine multiple physical links into a single logical pipe with single IP prefix
- Combine up to 16 links (16 OC-48 = OC-768)
- 4:1 speed mismatch between composite link members are supported. (mix of OC12c and OC48c)
- Packet ordering is accomplished by a hash value from the IP source/destination address
- Traffic is balanced across all link members
- Link failures are rerouted over other member links in 45 ms
Lucent NX64000

- 1.6 Tbps per switch fabric, 6.4Tbps with 4 switch fabric (claimed)
- Single chassis
- 16 line interfaces
- 1xOC48c available, 1xOC192c expected in Q2/2001
- Low latency (40ms)
- IP, ATM, FR, MPLS support
- Software: NX-IS
Pluris Teraplex 20 Architecture

- **Line Cards**
  - Up to 15 cards per shelf
  - 10Gbps per card

- **Fabric Cards**
  - 90 Gbps Each (9x9 @ 10G)
  - 16 card per shelf
  - 2(redundancy) x (4 internal fabrics + 4 external fabrics) ?

- **Control Card**: main processor module
- 10Gb/s fiber optic interconnection links between cards and shelves: TeraConnect
- Expands to 128 shelves (19.2 Tb/s)
- Degree 7 hypercube network (128 node) ==> Blocking Network!!
QoS Functions implemented with 6 ASICs

Classifer: Up to 128k rules
Policer: Per flow, wide rate range: 64Kbps to 2.488 Gbps
Queuing: WFQ, priority-based, weighted round-robin, 256 q/line card, 64 q/OC-48port
eWRED: Enhanced WRED for congestion mgmt.
Shaper: Three methods
Scheduler: Egress
IP Bond: Link Trunking Technology

- Support different-speed interfaces (OC48c and OC12c ports)
- Distribute and load-share traffic over multiple ports
- Enable redundant links

IP Bond across two ports
Summary

- Most “Terabit” router has less than 320Gbps switching capacity (full duplex 160Gbps)
- OC-192c interface is available now.
- Most terabit routers are designed with custom ASICs
- MPLS is the common feature of Terabit routers
- Link trunking technologies are proposed by two vendors (Avici, Pluris)
국내 라우터 기술 개발 동향(1)

수백 Mbps 이하급 소형 라우터
- Ethernet 허브로부터 시작하여 Ethernet 스위치로 발전
- 쌍용정보통신, 한아시스템 등이 80%의 시장을 점유로 경쟁력 확보
  - 핵심 부품 및 소프트웨어를 국외 상용기술에 의존
  - 시장 적시성과 가격 경쟁력 확보가 어려운 실정임

캠퍼스 및 기업망용 중형 라우터
- 시스코, 노텔, 모토롤라 등의 외산 제품 위주
- 국내 회사로 쌍용정보통신, 다산인터넷 등에서 최근 일부 제품 출시

ISP 백분망을 구성하는 에지 및 코어급 중대형 라우터
- 연구개발 투자 미흡 : 연구개발 투자보다는 외산 제품의 대행 판매에 의존(시스코, 주니퍼 등의 외산 제품이 100% 시장 점유)
- 라우터 기술 기반 미확보 : 국산 제품은 없음
국내 라우터 기술 개발 동향(2)

■ 최근 MIC과제 수행을 통해 라우터 기반 기술을 부분적으로 확보
  ◆ 고속 라우터 기술 개발
    ▪ 80Gbps급 edge 라우터(2001년 상용화 목표)
    ▪ 2.5G POS I/F 및 10Mpps IP Forwarding Engine
  ◆ Gigabit Ethernet 기술 개발
    ▪ 1.2 Gigabit Ethernet Interface 기술
    ▪ 32 Gbps 급 Gigabit Ethernet 스위칭 기술
    ▪ 3 Mpps 급 IP forwarding engine ASIC 기술
  ◆ 20 Gbps급 MPOA 기반 Switched Router 기술 개발
    ▪ 기본 라우팅 프로토콜 포팅 기술
    ▪ STM-1, STM-4c 접속 기술
    ▪ 10 Gbps 급 셀 스위칭 ASIC 기술
  ◆ HAN/B-ISDN 사업
    ▪ OC-4, OC-12 interface 기술
    ▪ 40 Giga 급 switching fabric 기술
    ▪ 160 Giga 급 switching fabric 기술 개발 중
테라급 라우터 기술 과제

주관연구기관 및 연구책임자:
◈ 한국전자통신연구원 라우터기술연구부장 이형호

총 연구기간 및 연구비:
◈ 2001. 3. 1 ~ 2005. 2. 28 (4년)
◈ 정부출연금 260억원, 기업체 부담금 260억원

단계별 연구목표
◈ 1단계 (2001. 3. 1 ~ 2003. 2. 28) : 핵심 기술 및 640Gbps급 시제품 개발
◈ 2단계 (2003. 3. 1 ~ 2005. 2. 28) : 테라급 실장 기술 및 5.2Tbps 라우터 개발

공동 연구 기관
◈ 시스템 개발 업체 : LG전자, 삼성전자, 다산인터넷
◈ 부품 개발 업체 : 크로스반도체, Engedi Networks

최종목표 : 차세대 인터넷 백분명용 테라급 라우터 기반 기술 및 시스템 개발
◈ 5Tbps급 스위치 네트워크 기술
◈ 10/40Gbps WDM 인터페이스 및 10Gigabit Ethernet 인터페이스 기술
◈ 30Mpps 이상의 패킷 포워딩 처리 기술
◈ IPv6 구현 기술, 멀티캐스트 라우팅 프로토콜, QoS 라우팅 프로토콜
◈ 지능형 라우터 형상 제어 및 트래픽 제어 기술
◈ 백분용 라우터 시험 기술 및 상용화 기술
1단계 640Gbps 라우터 시제품
- Fault-tolerant 시스템 구조
- 640Gbps 스위칭 용량
- OC-48/OC-192 POS 인터페이스
- 30Mpps IP 패킷 포워딩 라인 인터페이스 카드
- 2.5 Gbps IP 패킷 포워딩 엔진 chip set
  - 100만 루트 엔트리, L3/L4, QoS(8 service classes), 2.5Gbps POS, CSIX-L1
  - Unicast/Multicast/Broadcast
  - Traffic shaping, RED, Policing
- 모듈화된 소프트웨어 구조
  - IP 라우팅 프로토콜, 정책기반 QoS 관리, IP 트래픽 관리
  - 운영관리 기능
    - ROM
    - VPN
2단계 5.2Tbps 라우터 시스템

- Fault-tolerant 구조
- 수백기가급 스위칭 용량의 단위 스위치
- Optical Interconnection Network
  - 5.2Tbps 이상의 스위칭 용량
  - 2.5/10Gbps WDM ports
  - 표준 VSR 스위치 링크 I/F
- OC-48/OC-192/OC-768 POS 및 10GbE I/F
- 50Mpps IPv6 패킷 포워딩 라인 인터페이스 카드
- 10Gbps IPv6 패킷 포워딩 엔진 chip set
- 개방형 소프트웨어 구조
- 대규모 converged network 라우팅
  - 50,000 IPv6 flow 식별 및 상태 관리, flow별 QoS scheduling
  - IPv6 기반 라우팅 프로토콜, 멀티캐스팅
- 차세대 인터넷 백분망 운용보전
  - 다중 링크 부하 분배 및 제어
  - 지능형 형상 제어 및 트래픽 제어