High-Performance Routers

H. Jonathan Chao
Dept of Electrical and Computer Engineering
Center of Advanced Technology for Telecommunications
Polytechnic University, NY
chao@poly.edu, http://www.poly.edu/chao
Contents

- Core Network Structure
- Router Architectures
- Network Processors
- IP Route Lookup
- Packet Classification
- Switch Fabrics
- PetaStar: A Petabit Photonic Packet Switch
- Conclusions
Today’s TERA POP Architecture – Why so complex and costly?

Clustering of multiple core routers in POP

WHY?
- Routers lack of port capacity and switching capacity to meet POP to POP demand
- Unreliable routers and lack of network restoration result back to back router configuration
- Lack of connectivity/bandwidth reservation concept in IP networks (tend to over-engineering)

RESULTS
- About 50% of port capacity used for intra POP interconnection – waste customer investment

REAL PROBLEM MOVING FORWARD
- Can this POP Architecture support data traffic growth yet to be realized?
In a few years, POP will look like this

- More Routers thrown into the POP creating serious management nightmare
- More portion of switch ports are used for interconnection
- Service/Network reliability has not been resolved

Need Fundamental Re-thinking
New POP Architecture – Paradigm Shift

- One box solution
  - Carrier-grade reliability
  - Large port counts
  - Every port carries real user traffic
  - 1 – 100 terabit packet switching capacity
What a Router Looks Like

**Cisco GSR 12416**
- Height: 6ft
- Width: 2ft
- Depth: 2ft
- Capacity: 160Gb/s
- Power: 4.2kW

**Juniper M160**
- Height: 3ft
- Width: 2.5ft
- Depth: 19"
- Capacity: 80Gb/s
- Power: 2.6kW
Traffic Manager

- Traffic policing/shaping
  - UPC (Usage Parameter Control) for ATM
  - srTCM (Single Rate Three Color Marker) for IP DiffServ
- Packet scheduling (WRR, WDRR, CBQ, WFQ, WF\textsuperscript{2}Q)
- Buffer management (PPD, EPD, RED, WRED)
- Packet Segmentation and Reassembly
- VOQ (virtual output queue) and VIQ (virtual input queue) management for the switch fabric

Text book: QoS Control in High-Speed Networks by Chao and Guo; John Wiley & Sons, Nov 2001
Network Processor (NP)

- Main functions
  - IP route lookup
  - Packet classification
  - Packet header modification
  - Statistic collections
  - IP Packet fragmentation
- Flexible to new applications and protocols
- Shorten the design cycle and time-to-market
- Reduce the cost by avoiding ASICs
- Two implementation architectures
  - Configurable (multiple special-purpose coprocessors)
  - Programmable (more RISC processors)
- Difficulty to program the NP to support different applications. Some start-up companies specialize in creating machine-codes based on the NP structure.
A Configurable Network Processor

Manager

Classification and table lookup

Packet analysis and modification

Switch fabric forwarding

Classification and table lookup

Packet analysis and modification

Switch fabric forwarding

Classification and table lookup

Packet analysis and modification

Switch fabric forwarding

Input

Output
A Programmable Network Processor

- RISC cluster
- Packet analysis and modification
- Switch fabric
- Switch fabric forwarding
- RISC cluster
- Classification and table lookup
- Manager
- RISC cluster
- Task unit
- Input
- Output
A Forwarding Table Example

<table>
<thead>
<tr>
<th>Name</th>
<th>Prefix</th>
<th>Next Hope</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>0*</td>
<td>H1</td>
</tr>
<tr>
<td>P2</td>
<td>10*</td>
<td>H2</td>
</tr>
<tr>
<td>P3</td>
<td>11*</td>
<td>H3</td>
</tr>
<tr>
<td>P4</td>
<td>1011*</td>
<td>H4</td>
</tr>
<tr>
<td>P5</td>
<td>10110*</td>
<td>H5</td>
</tr>
</tbody>
</table>

Node structure

<table>
<thead>
<tr>
<th>Next-hop-ptr (if prefix)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left-ptr</td>
</tr>
</tbody>
</table>

H. Jonathan Chao
A Multi-bit Trie Example

Node structure

<table>
<thead>
<tr>
<th>Next-hop-ptr (if present)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ptr00</td>
</tr>
</tbody>
</table>

H. Jonathan Chao
# A Packet Classifier Example

<table>
<thead>
<tr>
<th>Rule</th>
<th>IPd</th>
<th>IPs</th>
<th>Prot.</th>
<th>Port#</th>
<th>Appl</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>128.238/16</td>
<td>*</td>
<td>TCP</td>
<td>telnet</td>
<td>*</td>
<td>Deny</td>
</tr>
<tr>
<td>R2</td>
<td>176.110/16</td>
<td>196.27.43/24</td>
<td>UDP</td>
<td>*</td>
<td>RTP</td>
<td>Send to port III</td>
</tr>
<tr>
<td>R3</td>
<td>196.27.43/24</td>
<td>134.65/16</td>
<td>TCP</td>
<td>*</td>
<td>*</td>
<td>Drop if rate &gt; 10 Mb/s</td>
</tr>
</tbody>
</table>

![Diagram showing network topology with routers and IP networks](image)
Function Diagram

H. Jonathan Chao
Header Fields Used For Classification

<table>
<thead>
<tr>
<th>Transport layer header</th>
<th>Network layer header</th>
<th>MAC header</th>
</tr>
</thead>
<tbody>
<tr>
<td>L4-SP</td>
<td>L4-DP</td>
<td>L4-PROT</td>
</tr>
</tbody>
</table>

DA = Destination address
SA = Source address
PROT = Protocol
SP = Source port
DP = Destination port

L2 = layer 2 (e.g., Ethernet)
L3 = layer 3 (e.g., IP)
L4 = layer 4 (e.g., TCP)
## Multi-field Packet Classification

<table>
<thead>
<tr>
<th>Rule</th>
<th>Field 1</th>
<th>Field 2</th>
<th>Field d</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule 1</td>
<td>5.3.40.0/21</td>
<td>2.13.8.11/32</td>
<td>UDP</td>
<td>A₁</td>
</tr>
<tr>
<td>Rule 2</td>
<td>5.168.3.0/24</td>
<td>152.133.0.0/1</td>
<td>TCP</td>
<td>A₂</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rule N</td>
<td>5.168.0.0/16</td>
<td>152.0.0.0/8</td>
<td>ANY</td>
<td>Aₙ</td>
</tr>
</tbody>
</table>

Example: packet \((5.168.3.32, 152.133.171.71, \ldots, TCP)\)

**Packet Classification:** Find the action associated with the highest priority rule matching an incoming packet header.
Classification Schemes

- Linear Search
- Trie-based Schemes
- Geometric Schemes
- Hardware-based Schemes
- Heuristic Schemes
Linear Search

- Keep all rules in a linked list
- Classification (query) requires inspecting rules one by one
- $O(N)$ storage and $O(N)$ query time
Hierarchical Tries

<table>
<thead>
<tr>
<th>Filter</th>
<th>F1</th>
<th>F2</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>00*</td>
<td>11*</td>
</tr>
<tr>
<td>R2</td>
<td>00*</td>
<td>1*</td>
</tr>
<tr>
<td>R3</td>
<td>10*</td>
<td>1*</td>
</tr>
<tr>
<td>R4</td>
<td>0*</td>
<td>01*</td>
</tr>
<tr>
<td>R5</td>
<td>0*</td>
<td>10*</td>
</tr>
<tr>
<td>R6</td>
<td>0*</td>
<td>1*</td>
</tr>
<tr>
<td>R7</td>
<td>*</td>
<td>00*</td>
</tr>
</tbody>
</table>

Arrival packet (001, 110)
A 2-D Set-Pruning Trie

Arrival packet (001, 110)
Ternary CAM (TCAM)

One packet may match multiple entries in the TCAM
Issues of Designing a Terabit Router

- Build a switch fabric that is scalable by adding boards and racks when needed
- Memory speed and size
  - For a OC-768 port with speedup of 2, the memory cycle time < 2.66 ns for 40-byte packets back to back
  - Buffer size: 100 ms worth data – 500 Mbytes
- Packet arbitration for output contention resolution (4ns for above example)
- QoS control (8 ns for packet scheduling and discarding)
- Interconnections and power consumption
  - Chip to chip: 128 SERDES bidirectional I/O @ 20W
  - Rack to rack: VCSEL up to 300 ms with 250mW each bidirectional connection

Text book: *Broadband Packet Switching Technologies*
by Chao, Lam, and Oki; John Wiley & Sons, Aug 2001
Different Structures and Queuing Schemes

- Single stage vs. multiple stages
  - Single stage: less control complexity, but more components
    - Switch elements are proportional to $N^2$ ($N$: switch size)
  - Multiple stage: less components, but more control complexity
    - Require an efficient scheme to reduce internal blocking

- Input Queuing (IQ)
  - Easy to implement
  - Head Of Line (HOL) blocking, throughput 58.6%

- Output Queuing (OQ)
  - Highest throughput
  - Internal speedup $N$

- Virtual Output Queuing (VOQ)
  - Overcome HOL blocking
  - Speedup less than $N$
Virtual Output Queuing (VOQ)

With speedup

Input 1

Input 2

Input 3

Input 4

Output 1

Output 2

Output 3

Output 4
Input-Output Matching for VOQ Switches

- Maximum weight matching
  100% throughput, not practical

- Maximal matching
  100% throughput with speedup 2 or more, not practical for large switches

- Iterative matching
  practical, multiple iterations
  - PIM (parallel iterative matching)
  - iSLIP
  - DRRM (dual round robin matching), less complexity than iSLIP
  - Both iSLIP and DRRM achieve 100% throughput under independent and uniform traffic
Multiple-Plane Single-Stage Switch

- Cisco 12000-series routers (15 x 10Gb/s)
- Agere’s PI40
Multiplane Single-Stage Output Queue Switch

- IBM packet routing switch chip (Q-64G, 32 x 2Gb/s)
- Internet Machine SE200 (64 x 2.5 Gb/s)

H. Jonathan Chao
A Multiplane Multistage Switch

- 5 switch planes, 4 active, 1 standby
- Each links @ 2.5 Gb/s
- Juniper T640, single chassis (or a half rack)
- Supports 32xOC-192
- Forward up to 640 Mp/s
A 3-Dimension Torus Topology (4x3x2)

- Avici terabit switching router (TSR)
- Each node interfaces with 6 others
- Local traffic adds/drops through the ingress and egress lines

H. Jonathan Chao
Switch Fabric Vendors

- Vitesse
- IBM
- Agere
- AMCC
- Internet Machines
- Broadcom
- Crosslayer Networks
- Erlang Technology
- Intel
- Marvell
- Mindspeed
- Paion
- PetaSwitch
- Power X
- SiSilk Networks
- Tau Networks
- TeraBlaze
- TeraCross
- TeraOptic
- TransWarp Networks
- TransSwitch
- ZettaCom

Source: Linley Group
System Architecture of PetaStar

IPC: Input Port Controller
IGM: Input Grooming Module
ODM: Output Demultiplexing Module
OPC: Output Port Controller
PS: Packet Scheduler
VOQ: Virtual Output Queue
r: Cells per Frame
s: Speedup Factor
g: Input Lines per Port
PSF: Photonic Switch Fabric
IM: Input Module
CM: Central Module
OM: Output Module
Photonic Switching Technology

- AWG+ Tunable Laser
- Fast Electro-Optics
- 3-D Fast MEMS
- Liquid Crystal Grating
- 3D Slow MEMS
- Mech. Tuning

Port count

Switching time

H. Jonathan Chao
A 3-Stage Photonic Switching Fabric

**IM** = Input Module, **CM** = Central Module, **OM** = Output Module,
**AWG** = Array Waveguide Grating, **SCU** = Sub-Carrier Unit, **WCU** = Wavelength Conversion Unit

H. Jonathan Chao
Summary

- Terabit routers save the cost by replacing multiple mid-size routers
- Network processors provides flexibility to new applications and protocols, shorten the design cycle and time-to-market, and reduce the cost by avoiding the ASIC approach
- Memory and interconnection technologies are crucial to next generation routers
- Advanced optical technologies can be used in implementing the next generation routers
- Multiple-stage switch fabrics have better scalability, but need an efficient distributed dispatching scheme
References


5. http://www.poly.edu/chao