State of the art

Routers Architecture

(Juniper, Cisco, Huawei)

October, 15th 2013

Maurizio Ipsale

JNCI, CCSI, HCSI

www.klabs.it

- High-quality training courses
- Real-life experience
- Theoretical knowledge
- Plenty of practical hands-on
- Best learning experience
State of the art Routers Architecture (Juniper, Cisco, Huawei)
Junos Software Architecture

References:

- Management daemon (mgd)
- Routing protocol daemon (rpd)
- Device control daemon (dcd)
- Chassis daemon (chassisd)
Routing sockets are a UNIX mechanism for controlling the routing table. The Junos kernel takes this same mechanism and extends it to include additional information to support additional attributes to create a carrier-class network operating system.
Control Plane

Courtesy of Juniper Networks Education Services
Forwarding Plane

Architecture Details of the Forwarding Plane in Multi-PFE Platforms

Model of PFE Interconnection with Fabric Planes

Courtesy of Juniper Networks Education Services
INTERNAL OVERHEAD ADDED TO THE CELLS/PACKETS NOT SHOWN FOR SIMPLICITY
- Packet Data
- Fabric Queuing
- Revenue Port Queuing
- Delegate Responsibilities:
  - Process Oversubscription (<24x1GbE or <2x10GbE on a MIC)
  - Revenue Port Queuing (H-QoS)
- Route lookups
- MAC lookups
- Class of Service (QoS) Classification
- Firewall filters
- Policers
- Accounting
- Encapsulation
- Statistics

Deep Packet Inspection: 256 bytes into the packet
Each packet is inspected at line rate, and attributes such as Ethernet Type Codes, Protocol, and other Layer 4 information are used to evaluate which buffers to enqueue the packet towards the Buffering Block. Preclassification allows the ability to drop excess packets as close to the source as possible, while allowing critical control plane packets through to the Buffering Block.
Depending on the line card, Trio offers an optional Dense Queuing Block that offers rich Hierarchical QoS that supports up to 512,000 queues with the current generation of hardware. This allows for the creation of schedulers that define drop characteristics, transmission rate, and buffering that can be controlled separately and applied at multiple levels of hierarchy.
1. The packet enters the Buffering Block from the WAN ports and is subject to pre-classification. Depending on the type of packet, it will be marked as high or low priority. The Buffering Block will enqueue the packet as determined by the pre-classification at service the high-priority queue first. A Lookup Block is selected via round-robin and the packet is sent to that particular Lookup Block.

2. The packet enters the Lookup Block. A route lookup is performed and any services such as firewall filters, policing, statistics, and QoS classification are performed. The Lookup Block sends the packet back to the Buffering Block.

3. The packet is sent back to the Fabric Block and is enqueued into the switch fabric where it will be destined to another PFE. If the packet is destined to a WAN port within itself, it will simply be enqueued back to the Interfaces Block.

4. The packet is sent to the switch fabric.
1. The packet is received from the switch fabric and sent to the Fabric Block. The Fabric Block sends the packet to the Buffering Block.

2. The packet enters the Buffering Block. The packet will then be subject to scheduling, shaping, and any other class of service as required. Packets will be enqueued as determined by the class of service configuration. The Buffering Block will then dequeue packets that are ready for transmission and send them to a Lookup Block selected via round-robin.

3. The packet enters the Lookup Block. A route lookup is performed as well as any services such as firewall filters, policing, statistics, and QoS classification. The Lookup Block sends the packet back to the Buffering Block.

4. The Buffering Block receives the packet and sends it to the WAN ports for transmission.
dhanks@R1-RE0> show chassis ethernet-switch
Displaying summary for switch 0
Link is good on GE port 1 connected to device: FPC1
  Speed is 1000Mb
  Duplex is full
Link is good on GE port 2 connected to device: FPC2
  Speed is 1000Mb
  Duplex is full

Link is good on GE port 12 connected to device: Other RE
  Speed is 1000Mb
  Duplex is full

Link is good on GE port 13 connected to device: RE-GigE
  Speed is 1000Mb
  Duplex is full

Courtesy of Juniper Networks Education Services
## MX2020

### Upper Backplane

<table>
<thead>
<tr>
<th>FPC19</th>
<th>FPC18</th>
<th>FPC17</th>
<th>FPC16</th>
<th>FPC15</th>
<th>FPC14</th>
<th>FPC13</th>
<th>FPC12</th>
<th>FPC11</th>
<th>FPC10</th>
</tr>
</thead>
<tbody>
<tr>
<td>CB-RE1</td>
<td>SFB7</td>
<td>SFB6</td>
<td>SFB5</td>
<td>SFB4</td>
<td>SFB3</td>
<td>SFB2</td>
<td>SFB1</td>
<td>SFB0</td>
<td>CB-RE0</td>
</tr>
</tbody>
</table>

### Control Boards, Routing Engines, and Switch Fabric Boards.

### Lower Backplane

<table>
<thead>
<tr>
<th>FPC9</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSM8</td>
</tr>
</tbody>
</table>

### Rear View

#### Power Supply Modules

<table>
<thead>
<tr>
<th>PDM0</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSM7</td>
</tr>
<tr>
<td>PSM6</td>
</tr>
<tr>
<td>PSM5</td>
</tr>
<tr>
<td>PSM4</td>
</tr>
<tr>
<td>PSM3</td>
</tr>
<tr>
<td>PSM2</td>
</tr>
<tr>
<td>PSM1</td>
</tr>
<tr>
<td>PSM0</td>
</tr>
</tbody>
</table>

#### Power Distribution Modules

<table>
<thead>
<tr>
<th>PDM1</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSM14</td>
</tr>
<tr>
<td>PSM13</td>
</tr>
<tr>
<td>PSM12</td>
</tr>
<tr>
<td>PSM11</td>
</tr>
<tr>
<td>PSM10</td>
</tr>
<tr>
<td>PSM9</td>
</tr>
</tbody>
</table>

#### Power Supply Modules

<table>
<thead>
<tr>
<th>PDM2</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSM17</td>
</tr>
<tr>
<td>PSM16</td>
</tr>
<tr>
<td>PSM15</td>
</tr>
<tr>
<td>PSM14</td>
</tr>
<tr>
<td>PSM13</td>
</tr>
<tr>
<td>PSM12</td>
</tr>
<tr>
<td>PSM11</td>
</tr>
<tr>
<td>PSM10</td>
</tr>
<tr>
<td>PSM9</td>
</tr>
</tbody>
</table>

### Power Supply Modules

<table>
<thead>
<tr>
<th>PDM3</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSM17</td>
</tr>
<tr>
<td>PSM16</td>
</tr>
<tr>
<td>PSM15</td>
</tr>
<tr>
<td>PSM14</td>
</tr>
<tr>
<td>PSM13</td>
</tr>
<tr>
<td>PSM12</td>
</tr>
<tr>
<td>PSM11</td>
</tr>
<tr>
<td>PSM10</td>
</tr>
<tr>
<td>PSM9</td>
</tr>
</tbody>
</table>

**Courtesy of Juniper Networks Education Services**
State of the art
Routers Architecture
(Juniper, Cisco, Huawei)
Cisco ASR 1000 Building Blocks

Fully Distributed Architecture for High Performance and High Multi-dimensional Control Plane Scale

- Data forwarding is fully distributed on the line cards
- Control plane split among RSP and LC CPU (same type of CPU as RSP)
- L2 protocols, BFD, CFM, Netflow runs on the LC CPU for high scale

True Modular OS for HA and Operational Simplicity

- Micro-kernel based, true modular OS
- High availability and System stability
- SW patch granularity for operational simplicity

FIA = Fabric Interface ASIC

Active-Active Switch Fabric
Guarantee “0” packet loss during RSP failover

Courtesy of Cisco Education Services
Generic Line Card Architecture

forwarding “slice”

physical interfaces → NP → FIA

replicate “slices” of components to improve performance

FIA = Fabric Interface ASIC

Typhoon LC CPU: Freescale Quad core P4040

Courtesy of Cisco Education Services
References:
1. Cisco Nexus 7000 Hardware Architecture, Cisco Live! 2012
Nexus 7000 decouples control plane and data plane

Forwarding tables built on control plane using routing protocols or static configuration
- OSPF, EIGRP, IS-IS, RIP, BGP for dynamic routing

Tables downloaded to forwarding engine hardware for data plane forwarding
- FIB TCAM contains IP prefixes
- Adjacency table contains next-hop information
FIB TCAM lookup based on destination prefix (longest-match)
FIB “hit” returns adjacency, adjacency contains rewrite information (next-hop)
Pipelined forwarding engine architecture also performs ACL, QoS, and NetFlow lookups, affecting final forwarding result
IPv4 FIB TCAM Lookup

Generate TCAM lookup key (destination IP address)

Ingress unicast IPv4 packet header

Flow Data

Forwarding Engine

FIB TCAM

FIB DRAM

Adjacency Table

Next-hop 1 (IF, MAC)
Next-hop 2 (IF, MAC)
Next-hop 3 (IF, MAC)
Next-hop 4 (IF, MAC)
Next-hop 5 (IF, MAC)
Next-hop 6 (IF, MAC)
Next-hop 7 (IF, MAC)

Load-Sharing Hash

Index, # next-hops

Offset

Index, # next-hops

Modulo function selects exact next hop entry to use

Adjacency index identifies ADJ block to use

Hit in FIB returns result in FIB DRAM

HIT!

Compare lookup key

10.1.1.10

10.1.1.2
10.1.1.3
10.1.1.4
10.10.0.10
10.10.0.100
10.10.0.33
10.1.1.xx
10.1.3.xx
10.10.100.xx
10.1.1.xx
10.100.1.xx
10.10.0.xx
10.100.1.xx

© 2012 Cisco and/or its affiliates. All rights reserved.

Courtesy of Cisco Education Services
State of the art
Routers Architecture
(Juniper, Cisco, Huawei)
System Control and Management Unit
• Route calculation: All routing protocol packets are sent by the forwarding engine to the MPU for processing. In addition, the MPU broadcasts and filters packets, and downloads routing policies from the policy server.

• Outband communication between boards: The LAN switch modules integrated on the MPU provide outband communications between boards. In this manner, messages can be controlled, maintained, and exchanged between SFUs and LPUs.

• Device management and maintenance: Devices can be managed and maintained through the management interfaces (serial interfaces) provided by the MPU.

• Data configuration: The MPU stores configuration data, startup files, charging information, upgrade software, and system logs.

• Data storage: The MPU provides two interfaces for CF cards, which serve as mass storage devices to store data files.
Switching Network

References:
1. HUAWEI NetEngine5000E Core Router, NE5000E Product Description, 2011

Courtesy of Huawei Education Services
The Switch Fabric Unit B(SFUI-200-B) is the new generation switch board. It supports 200G full duplex capacity per slot. The Switch Fabric Unit B(SFUI-200-B) switches data for the entire system at line speed of 3.15 Tbit/s. This ensures a non-blocked switching network.

The NE40E-X16 has four SFUs working in 3+1 load balancing mode. The entire system provides a switching capacity at wire speed of 12.58 Tbit/s.

The four SFUs load balance services at the same time. When one SFU is faulty or being replaced, the other three SFUs automatically take over its tasks to ensure normal delivery of services.
Switching Fabric of NE5000E-X16

Switching Capacity: 3.145Tbps per SFU
786Gbps per slot

SFU

4 pairs of SerDes links to each SFU
4 SFUs = 16 pairs of links from LPU
Each pair of SerDes links max. 6.336 Gbps.

Ingress LPU
0 ~ 15

Segmentation

Reassembly

Egress LPU
0 ~ 15

Cell switching, Fixed length, based on Info in Cell Header

4 SFUs: 3+1 Load Sharing

FIA = Fabric Interface ASIC

Courtesy of Huawei Education Services
Thank you

October, 15th 2013

Maurizio Ipsale
JNCI, CCSI, HCSI

www.klabs.it