IP NGN Backbone Routers for the Next Decade

Josef Ungerman
Consulting SE, CCIE #6167
Agenda

- Motivations for IP NGN
- Trends is IP/MPLS Core Design
- Router Anatomy Trends
- Latest Product Updates
- Switching Fabric Technologies
- Network Processor Technologies
IP traffic will increase 4X (767 exabytes by 2014)

- IPv6 and IPv4 Address Exhaustion
- LTE moving from circuits to packets
- New access technologies – WiFi, FTTX

More Issues:

Source: Cisco Visual Networking Index—Forecast, 2009-2014

© 2010 Cisco and/or its affiliates. All rights reserved.
Challenge of Shifting Environment

Monetization
New revenue streams

Traffic

Profitability

Revenue

Optimization
Efficient delivery
Network Architecture Trends
IP NGN = reducing networks and layers

ACCESS
- PDH MW
- PDH
- FR/ATM
- Metro Ethernet
- DSLAM MSAN
- CPE

AGGREGATION
- SDH/ATM Mobile
- SDH
- FR/ATM
- Metro Ethernet
- CPE

BACKBONE
- SDH
- ATM
- IP/MPLS
- SDH
- ATM

INTERNET
- FMC/Packet Core
- SGSN GGSN IGW
- Internet Peering
- Data Centers

© 2010 Cisco and/or its affiliates. All rights reserved.
IP NGN = reducing networks and layers

**ANY ACCESS**
- 2G
- 3G
- PON
- DSLAM
- LTE, WiFi...

**AGGREGATION**
- CPE
- PDH MW
- Ethernet
- MPLS-TP
- MPLS...

**EDGE**
- CarrierE AGGREGATION
- IP/MPLS

**CORE**
- BACKBONE
- IP/MPLS

**INTERNET**
- Internet Peering
- IGW
- CGv6
- Data Centers

**IP NGN**
- single multiservice IP/MPLS network
- operational hierarchy
- Circuits (Pseudowires) and Clouds (IP)

**1000’s**
**100’s**
**10’s**
**1’s**

© 2010 Cisco and/or its affiliates. All rights reserved.
IP NGN – optimization trends

ANY ACCESS

AGGREGATION

EDGE

CORE

Backbone Optimization
- IP Offload links, 100GE
- Flexible Edge placement
- Consolidation techniques

Existing PE’s

INTERNET

DWDM

1000’s

100’s

10’s

1’s

© 2010 Cisco and/or its affiliates. All rights reserved.
Router Bypass Techniques

- O-E-O regeneration avoided as much as possible
  no need for OTN Switching cross-connects in CEE countries

- Static long lambdas are used
  no need for dynamic G.MPLS in the static Internet backbone

- Importance of OTN interfaces in routers (IPoDWDM)
  STM-256 (OTU3) and 100GE (OTU4)

Real-Life Example:
Warszawa-Poznan, 613km
40G over Siemens 10G WDM
Link Consolidation – 100GE

**Link Bundling (LACP)**
- up to 64x TGE today (32x deployed)
- dynamic adaptable hash (also 3,5,7,11 links)
- 7-tuple hash for equal load-sharing

**100GE**
- throughput (100GE is like a bundle 12-14 TGEs)
- no hashing inefficiencies, easy troubleshooting
- contribution HDTV is 1.24Gbps single stream!

**40GE**
- IEEE 802.3ba, ITU OTU3e
- data center interface
- router to transponder

**100GE**
- IEEE 802.3ba, ITU OTU4
- router interface
- router → router or transponder

---

© 2010 Cisco and/or its affiliates. All rights reserved.
Node Consolidation Techniques

Cluster (ASR9000)

Key motivation is in the Access edge:
**Simpler Access Dual-homing**
- scaling the L2/L3 control plane (not data plane)

Cluster + Satellites
(remote linecards)

Multi-Chassis (CRS)

Key motivation is in the Core:
**Simpler Core PoP**
- scaling the non-blocking data plane
- back-to-back, 2+1, 8+2, etc.

Cluster
(one L2 & IP/MPLS control plane)

Multi-Chassis
(one router)
Optimization: How to move bits cheaper...

*reduce OPEX, CAPEX, and keep reasonable quality?*

1) Reduce the number of networks
   - IP NGN = single multiservice network

2) Reduce the number of layers
   - IP NGN = IP/MPLS + DWDM

3) Reduce the number of nodes
   - Direct Links = huge broadband traffic takes shortest path

4) Reduce the number of links
   - MPLS Technology = statistical multiplex and hierarchy

5) Innovate – make use of modern technologies
   - Moore’s Law = Lower TCO, Price/Gigabit, Watt/Gigabit
Core Trends – Appeal of Innovations

CRS-1 (2005) vs. CRS-3 (2010)

- **Power @ 40G/slot**
- **Power @ 140G/slot**
- **Total Power**
- **W/Gbps @ 40G/slot**
- **W/Gbps @ 140G/slot**

- **CRS Form Factor**
  - 4/S
  - 8/S
  - 16/S

> 60% reduction

© 2010 Cisco and/or its affiliates. All rights reserved.
Silicon has fundamentally followed Moore’s law
Optics is fundamentally an analog problem
Router Anatomy Trends
2004: Cisco CRS-1 – 40G (STM-256) per slot
Focus on Quality (scale, modularity, resiliency)

SPP (Silicon Packet Processor)
- 40 Gbps, 80 Mpps [u-programmable]
- one for Rx, one for Tx processing

4, 8 or 16 Linecard slots + 2 RP slots
2010: Cisco CRS-3 – 140G per slot
Focus on Quality (scale, modularity, resiliency)
2009: Cisco ASR9000 – 8x 10GE per slot
Compact Router/Switch

**Trident Network Processor**
- 30 Gbps, 28 Mpps [u-programmable]
- shared for Rx and Tx processing
- one per 10GE (up to 8 per card)

**RSP (Route/Switch Processor)**
- CPU + Switch Fabric
- active/active fabric elements

Transport LC – 40G
- 4 or 8 Linecard slots
2011: Cisco ASR9000 – 2x 100GE per slot
Compact Router/Switch

Transport LC – 40G
- 4 or 8 Linecard slots
- IOS
- RSP’ (active)

Transport LC – 80G
- IOS
- RSP’ (fab. active)

Transport LC – 2x 100GE
- IOS
- >400G

Transport LC – 24x TGE
- IOS
2003: Cisco 7600 – 4x 10GE per slot

The Switch/Router

Trident Network Processor
- 10Gbps full-duplex
- H-QoS, VPLS, u-programmable

RSP (active)
- CPU + Switch Fabric + Switch ASIC
- active/standby fabric elements

RSP (Route/Switch Processor)

Edge or Core LC – 40G

Edge or Core LC – 20G

ES+

32G local switching

20G

F

F

ES+

EARL Switching ASIC
- 48 Mpps
- Catalyst 6500 compatibility

3, 4, 6, 9 or 13 Linecard/RSP slots
2003: Cisco 7600 – 4x 10GE per slot

The Switch/Router

- NP
- buff.
- F
- IOS
- ES+
- SIP-400
- SIP-600
- SPA

RSP (active)
- SP
- buff.
- RP
- IOS

RSP (standby)
- SP
- buff.
- RP
- IOS

Bus (headers only)

- SIP
- F

Edge or Core LC
- 40G
- 20G

SIP
- 600

IPSEC
- FWSM
- ACE

SPA
- NP
- buff.
How to make a router cheaper...  
...and keep a reasonable quality?

1) Compact Anatomy
   - RSP, Route/Switch Processor (instead of RP and FC)
   - Ethernet-oriented Linecard (non-modular, less memory)

2) Linecard Architecture
   - Multiple smaller NP’s (eg. 4x 10G instead of 1x 40G)
   - One NP is shared for Rx and Tx (not dedicated NP’s per Rx and Tx)
   - Multiple smaller Fabric Ports (eg. 2x 20G instead of 1x 40G)

3) Special Core-facing Linecards
   - 8/16 queues per port (instead of thousands)
   - lower-scale NP (no need for thousands of interfaces)
   - licenses for features that not everybody uses (eg. VPN, OTN, scale)

4) Oversubscribed Cards
   - 2:1 ingress overbooking (eg. PON OLT Aggregation)
IP NGN Routers Update
### Cisco CRS Core

- **~7500 systems**
- **~450 customers**
  (~80 are CRS-3)

<table>
<thead>
<tr>
<th>Chassis</th>
<th>CRS-4/S</th>
<th>CRS-8/S</th>
<th>CRS-16/S</th>
<th>CRS-MC (8+2)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>2004 (CRS-1)</strong></td>
<td># of Slots Height</td>
<td># of Slots Height</td>
<td># of Slots Height</td>
<td># of Slots Height</td>
</tr>
<tr>
<td></td>
<td>4 (+2 RP) 1/3 rack</td>
<td>8 (+2 RP) ½ rack</td>
<td>16 (+2 RP) 1 rack</td>
<td>128 (+2 RP) 10 racks</td>
</tr>
<tr>
<td></td>
<td>40 .320</td>
<td>40 .640</td>
<td>40 1.28</td>
<td>40 10.24</td>
</tr>
<tr>
<td></td>
<td>140/105 1.12</td>
<td>140/105 2.24</td>
<td>140/122 4.48</td>
<td>140/122 35.84</td>
</tr>
</tbody>
</table>

**Maximum BW** / **Protected BW with a failed fabric element**
CRS-3
Interface Modules (PLIMs)

1x 100GBE
- Line-rate performance (100Gbps)
- CFP optics (LR4, 10km)

14x 10GBE-WL-XFP
- Line-rate performance (140Gbps)
- Configurable LAN/WAN PHY

20x 10GBE-WL-XFP
- Oversubscribed (140Gbps)
- Configurable LAN/WAN PHY

Each PLIM requires FP140 or other forwarding card
CRS-3 and CRS-1
Forwarding Cards

**MSC-40** – High-speed edge @ 40Gbps
- H-QoS (8,000 queues), 800 interfaces, WAN

**FP-40** – IP/MPLS Core & Peering @ 40Gbps
- Per-port QoS, IP/MPLS, ACL, Netflow…

**MSC-140** – High-speed edge
- H-QoS (64,000 queues), scale (12,000 vlans)

**FP-140** – IP/MPLS Core & Peering
- Per-port QoS, IP/MPLS, ACL, Netflow…

**LSP-140** – MPLS Core P
- Per-port QoS, MPLS, IP Multicast, limited IP
IPv6 Transition: CGSE
Carrier-Grade Services Engine PLIM

Introducing the new engine for massive Cisco CGv6 deployments (XR 3.9.1)

- 20+ million active translations
- 100s of thousands of subscribers
- 1+ million connections per second
- 20Gb/s of throughput
- XML API (eg. port-forwarding)
- Netflow V9 translation logging
- Security

IPv6 Transition solution feature set

- Carrier-Grade NAT44 (3.9.1)
- NAT64 stateless (3.9.3)
- 6rd BR (3.9.3)
- NAT64 stateful (4.1.2)
- DS-Lite, 4rd, dIVI – planned
### Cisco ASR9000

**Edge and Aggregation**

<table>
<thead>
<tr>
<th>Chassis</th>
<th>ASR-9006</th>
<th>ASR-9010</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong># of Slots</strong></td>
<td>4 (+2 RSP)</td>
<td>8 (+2 RSP)</td>
</tr>
<tr>
<td><strong>Height</strong></td>
<td>¼ rack</td>
<td>½ rack</td>
</tr>
<tr>
<td><strong>Linecard [Gbps]</strong></td>
<td><strong>System [Tbps]</strong></td>
<td><strong>Linecard [Gbps]</strong></td>
</tr>
<tr>
<td>2009</td>
<td>120/80</td>
<td>120/80</td>
</tr>
<tr>
<td>2012</td>
<td>.960</td>
<td>1.92</td>
</tr>
<tr>
<td><strong>Linecard [Gbps]</strong></td>
<td><strong>System [Tbps]</strong></td>
<td><strong>Linecard [Gbps]</strong></td>
</tr>
<tr>
<td><strong>Future</strong></td>
<td>240</td>
<td>240</td>
</tr>
<tr>
<td></td>
<td>1.92</td>
<td>3.84</td>
</tr>
</tbody>
</table>

~3500 systems
~500 customers

Maximum BW / Protected BW with a failed fabric element
ASR 9000 Line Cards

Fixed Ethernet LCs:
- Line Rate: 40xGE, 2x10GE+20xGE, 4x10GE, 8x10GE
- Oversubscribed: 8x10GE (60G), 16x10GE (90G/120G)

Ingress/Egress H-QoS, Netflow, IPoDWDM (G.709, FEC, XFP), Video monitoring, SyncE, E-OAM

L2 Scalability: 1MMACs, 8kBDs, 32kPWs
L3 Scalability: 1M routes, 4kVRFs, 4kL3intfs

3 LC versions (16x10GE OS “Medium Queue” only):

<table>
<thead>
<tr>
<th>Line Card</th>
<th>EFPs</th>
<th>Egress Queues</th>
<th>Policers</th>
<th>Buffering</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Queue</td>
<td>-L</td>
<td>4k</td>
<td>8/portal</td>
<td>8k</td>
</tr>
<tr>
<td>Medium Queue</td>
<td>-B</td>
<td>16k</td>
<td>64k</td>
<td>128k</td>
</tr>
<tr>
<td>High Queue</td>
<td>-E</td>
<td>32k</td>
<td>256k</td>
<td>256k</td>
</tr>
</tbody>
</table>

+licenses: L3VPN (/LC), G.709 (/LC), vidmon (/chassis)

Modular LCs: SIP-700 + max 4x SPA
- QFP based
- 20Gbps Full Duplex
- Ph1: ChOC12

<table>
<thead>
<tr>
<th>ASR 9000</th>
<th>#10GE LR</th>
<th>#10GE OS</th>
</tr>
</thead>
<tbody>
<tr>
<td>6-slot</td>
<td>32</td>
<td>64</td>
</tr>
<tr>
<td>10-slot</td>
<td>64</td>
<td>128</td>
</tr>
</tbody>
</table>
How to make a router cheaper...

...and keep a reasonable quality?

1) Compact Anatomy
   ▪ RSP, Route/Switch Processor (instead of RP and FE)
   ▪ Ethernet-oriented Linecard (non-modular, less...)

2) Linecard Architecture
   ▪ Multiple smaller NP’s (eg. 4x 10G instead of 1x 40G)
   ▪ One NP is shared for Rx and Tx (not dedicated NP’s per Rx and Tx)
   ▪ Multiple smaller Fabric Ports (eg. 2x 20G instead of 1x 40G)

3) Special Core-facing Linecards
   ▪ 8/16 queues per port (instead of thousands)
   ▪ lower scale NP (no need for thousands of interfaces)
   ▪ licenses for features that not everybody uses (eg. VPN, OTN, scale)

4) Oversubscribed Cards
   ▪ 2:1 ingress overbooking (eg. PON OLT Aggregation)
Architecture Matters!
Forwarding Architecture 101

SWITCHING FABRIC
IP/MPLS unaware part
speaks cells/frames

NETWORK PROCESSOR
IP/MPLS aware part
speaks packets

Ingress Linecards

Egress Linecards

1 2 3 4

1 2 3 4

Fabric Port
• addressable entity
• singleduplex pipe

What’s the capacity?
4 fab.ports @10Gbps non-blocking

ENGINEERING:
4 * 10 = 40Gbps fdx

MARKETING:
4 * 10 *2 = 80Gbps
SMP Network Processor Example
2010: CRS QFA (Quantum Flow Array)

- BQS ASIC
  - 64K queues
  - 3L shaping

- 64K queues
- 3L shaping

- 256 Engines

- Processing Pool

- Resources & Memory Interconnect
  - DRAM
  - RLDRAM
  - SRAM
  - TCAM4

- PLU8/TLU5
- Stats/Police
- ACL

- 125 Mpps, 140 Gbps
- 65nm, u-programmable
- 2 per LC (Rx, Tx)

- Distribute & Gather Logic

- Headers only
Bad Network Processor Example (non-Cisco) ACL performance impact

Vendor is Saving on Memory – ACL memory is shared with Route memory

• Effect #1: ACL drastically impacts forwarding performance
• Effect #2: FIB cannot be hierarchical → slow BGP convergence
Switching Fabric and Multicast
Good vs. Bad IPTV Experience

**Good:**
Egress Replication
- Cisco CRS, 12000
- Cisco ASR9K, 7600

10Gbps of multicast eats 10Gbps fabric bw!

**Bad:** Binary Ingress Replication
- dumb switch fabric
- non-Cisco

10Gbps of multicast eats 80Gbps fabric bw!
(10G multicast impossible)
Good Fabric Redundancy

**CRS-1**
- 40G eth. non-blocking with 1 or 2 failed fabric cards

**CRS-3**
- 100G eth. non-blocking with 1 or 2 failed fabric cards

---

**ASR9000**
- 80G eth. non-blocking with failed RSP
- cell dip is not an issue (super-frame based fabric, not cell based)
Bad Fabric Redundancy

Non-Cisco

- 40G system sold as 100G
- if one CMP fails, system turns blocking
- unreliable 100GE

failed SFM/CMP: 105G → 52G
- huge cell dip at >40G
- blocking with 100GE

failed SCB: 126G → 84G
- huge cell dip at >75G
- blocking with m'cast

Non-Cisco

- 70G system sold as 80G
- minimum speedup
- huge cell dip
- multicast impacts unicast
Cell dip and Speedup

<table>
<thead>
<tr>
<th>Percentage of Linerate</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
</tr>
<tr>
<td>20%</td>
</tr>
<tr>
<td>40%</td>
</tr>
<tr>
<td>60%</td>
</tr>
<tr>
<td>80%</td>
</tr>
<tr>
<td>100%</td>
</tr>
</tbody>
</table>

Cell format example:

<table>
<thead>
<tr>
<th>cell hdr [5B]</th>
<th>cell payload [48B]</th>
</tr>
</thead>
</table>

Fixed overhead [cell header, ~10%]
Relative overhead [fabric header]
Variable overhead [padding]

40B IP Packet:

<table>
<thead>
<tr>
<th>cell hdr [5B]</th>
<th>buffer hdr [8B]</th>
<th>IP Packet [40B]</th>
</tr>
</thead>
</table>

41B IP Packet:

|---------------|-----------------|------------------------|

The lower speedup, the more multicast, the more cell dips!

The lower speedup, the more multicast, the more cell dips!

Good efficiency
1Mpps = 1Mcps
1Gb/s → 1.33Gb/s

Cell Dip mitigation:
1. Enough Speedup
2. Packet Packing (CRS)
3. Super-framing (ASR9K)
Quality Differences in 40G Solutions

40 ≠ 40

Good:
40 Gbps per slot
non-blocking

Good-enough:
40 Gbps per slot
non-blocking

Good-enough?:
40 Gbps per slot
non-blocking!

Cisco CRS-1 (MSC-40)

Non-Cisco

Non-Cisco
Impact of too many fabric connections

“How to do 100GE?”

Good 100GE

100GE non-blocking even with failed fabric card
1.4x → 1.2x speedup

Bad 100GE

ECMP/Link-Bundling
No 100G processor available, per-destination IP load-sharing across two old 50G processors with two old 54G fabric ports

More than 54G → Fabric Port gets overloaded!
Head of Line Blocking?

Port-Channel across 2 vlans on the same physical port
(to keep 1 IP and 1 MAC address per port) → no multi-vendor interoperability
Summary

• Motivation for IP NGN = traffic growth
• How to make the Network cheaper
• How to make the Router cheaper
• Quality differences

There are Good, Good-enough or Bad Solutions.
Thank you.