Ponoření do architektury ASR9000
T-SP3
Jiří Chaloupka – Cisco
Prosíme, ptejte se nás

- Twitter  [www.twitter.com/CiscoCZ](www.twitter.com/CiscoCZ)
- Talk2cisco  [www.talk2cisco.cz/dotazy](www.talk2cisco.cz/dotazy)
- SMS  721 994 600
Program

- ASR9000 family
- RSP2/Trident LC
- New RSP440/Typhoon LC
- New ASR9001
- Nv Cluster
- Nv Satellite
## ASR 9K Chassis Overview

<table>
<thead>
<tr>
<th>Max Capacity (bidirectional)</th>
<th>ASR 9001 (Ironman)</th>
<th>ASR 9006</th>
<th>ASR 9010</th>
<th>ASR 9922 (Megatron)</th>
</tr>
</thead>
<tbody>
<tr>
<td>240 Gbps</td>
<td>120Gbps</td>
<td>440G/slot 4 I/O slots</td>
<td>440G/slot 8 I/O slots</td>
<td>1.2T/slot 20 I/O slot</td>
</tr>
<tr>
<td>3.5Tbps</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 Tbps</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>48 Tbps</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Max Power</td>
<td>2RU</td>
<td>10RU</td>
<td>21RU</td>
<td>44RU</td>
</tr>
<tr>
<td>750W</td>
<td>6KW</td>
<td>9KW</td>
<td>24KW</td>
<td></td>
</tr>
<tr>
<td>Air Flow</td>
<td>Side to side</td>
<td>Side to back</td>
<td>Front to back</td>
<td>Front to back</td>
</tr>
<tr>
<td>FCS</td>
<td>4.2.1 release</td>
<td>Shipping</td>
<td>Shipping</td>
<td>4.2.2 release</td>
</tr>
</tbody>
</table>

*Air Flow Options: Side to side, Side to back, Front to back*
ASR 9010 and ASR 9006 Chassis
Identical HW components across two chassis*

Integrated cable management with cover

System fan trays

Line Card (0-3)

RSP (0-1)

cable management

Three Modular Power Supplies

Front-to-back airflow

Side-to-back airflow

RSP (0-1)

Line Card (0-3, 4-7)

System fan trays

Air draw

Six Modular Power Supplies
Power and Cooling
Existing Power Supply and Fan are ready for 400G/slot

- Fans unique to chassis
- Variable speed for ambient temperature variation
- Redundant fan-tray
- Low noise, NEBS and OSHA compliant

DC Supplies
- 2.1/1.5 kW

AC Supplies
- 2.1 kW
- 3 kW

- 6 & 10 slot use same power supplies
- Single power zone
- All power supplies run in active mode
- Power draw shared evenly
- 50 Amp DC Input or 16 Amp AC for Easy CO Install
RSP Engine

- Performs control plane and management functions
- Dual Core CPU processor with 4GB or 8GB (in 4.0) DRAM
- 2MB NVRAM, 4GB internal bootdisk, 2 external compact flash slots
- Dual Out-of-band 10/100/1000 management interface
- Console & auxiliary serial ports
- Hard Drive: 70G HDD
RSP Engine Architecture

RP/0/RSP0/CPU0:viking#dir ?
/all List all files
/ena Recognize sub directories
/recurse Recursively list subdirectories encountered
WORD file name
bootflash: bootflash: file system \rightarrow boot image
compactflash: compactflash: file system
compactflasha: compactflasha: file system
disk0: disk0: file system \rightarrow XR image, configurations
disk0a: disk0a: file system
disk1: disk1: file system
disk1a: disk1a: file system
harddisk: harddisk: file system \rightarrow SYSLOG, core dump
harddiska: harddiska: file system
harddiskb: harddiskb: file system
location Fully qualified location specification
nvram: nvram: file system \rightarrow ROMMON variables

© 2010 Cisco and/or its affiliates. All rights reserved.
RSP Operations Impact Fabric?
Guarantee “0” packet loss for RSP failover or OIR

- Switch fabric ASIC reside on the RSP blade physically
- Switch fabric ASIC is controlled by low level hardware, it operates separately from RSP function
- All fabric ASIC run in active mode regardless of the RSP status
- RSP SW switch over, reload, crash including kernel crash have NO impact on fabric operation
- RSP OIR has no traffic impact due to long/short pin backplane design and instant fabric switch over
  - Short pin trig the control signaling for fabric switchover in hardware
  - Long pin is used for data packet. It can continue draining the in-flight packets from the fabric during the extended short period of time
ASR 9K Ethernet Line Card Overview

First-generation LC (Trident NP)
- L, B, E

- A9K-40G
- A9K-4T
- A9K-8T/4
- A9K-2T20G
- A9K-8T
- A9K-16T/8

Second-generation LC (Typhoon NP)
- TR, SE

- A9K-24x10GE
- A9K-2x100GE
- A9K-MOD80
- A9K-MOD160

MPAs
- 20x1G
- 2x10G
- 4x10G
- 1x40G
- 2x40G
NP: Network Processor
Main forwarding ASIC
L2 & L3 forwarding, features (QoS, ACL, etc), control plane policing, mcast replication, etc

10Gbps bi-directional with features applied

CPU (same as RSP)
Program HW forwarding tables
Distributed Control planes
SW switched packets
Inline Netflow

FIA: Fabric Interface ASIC
Provide non-blocking data connection to switch fabric
Internal system queues/VoQ
Intelligent mcast replication

B: Bridge FPGA
Provide non-blocking data connection between NP and FIA
Internal System queues
Intelligent mcast replication

10G PHY for one 10G port, or 10x1G port

Example: A9K-8T

Note, Bridge FPGA provide non-blocking connection between NP and the FIA. Functionally it does the HW conversion due to different interface format on NP and FIA. It’s part of the switch fabric connection. To make it logically simple, it will be removed from the remaining slides.
Line Card HW Components – Counters

- RP/0/RSP1/CPU0:SJC#show controllers fabric ?
  - Arbiter  Arbitration ASIC show screens.
  - Crossbar  XBAR ASIC show screens.
  - fia  Show command for fabric interface asic

- RP/0/RSP1/CPU0:SJC#show controllers fabric fia bridge stats location 0/0/cpu0

- RP/0/RSP1/CPU0:SJC#show controllers fabric fia stats location 0/0/cpu0

- RP/0/RSP1/CPU0:SJC#show controllers np ?
  - counters  Display contents of global stats counters
  - crashinfo  Display NP Crash info
  - drvlog  Display Driver Logging
  - fabric-counters  XAUI counters dump
  - interrupts  Show NP interrupt data
  - memory  NP Raw Memory Dump
  - portMap  Show port mapping on NP
  - ports  Shows physical ports associated with each np

<snip>
4xNPs Line Card Family

- **A9K-4T-E/B/L**
  - PHY 3 → NP0
  - PHY 2 → NP1
  - PHY 1 → NP2
  - PHY 0 → NP3

- **A9K-2T20G-E/B/L**
  - PHY → NP0
  - PHY → NP1
  - PHY → NP2
  - PHY → NP3

- **A9K-8T/4-E/B/L**
  - PHY 3 → NP0
  - PHY 7 → NP1
  - PHY 6 → NP2
  - PHY 5 → NP3

- **A9K-40G-E/B/L**
  - PHY → NP0
  - PHY → NP1
  - PHY → NP2
  - PHY → NP3

Oversubscribed line card
Up to 60Gbps bandwidth
8xNPs Line Card Family

Oversubscribed line card
Up to 120Gbps (~117Gbps) bandwidth
Each FIA has one fabric channel which is 23 Gbps bi-directional, to each of the switch fabric ASIC.

The diagram shows the following:

- **NP0** to **FIA0**
- **NP1** to **B0**
- **NP2** to **B1**
- **NP3** to **FIA0**
- **NP4** to **B0**
- **NP5** to **B1**
- **NP6** to **FIA0**
- **NP7** to **FIA1**

- **FIA0** has 60Gbps bi-directional
- **FIA1** has 60Gbps bi-directional

- **PHY** connections to **NP0** to **NP7**

**Bandwidth Specifications**:

- **30Gbps and 25M pps (combined ingress and egress)**
- **15Gbps bidirectional**
- **30Gbps bidirectional**
Line Card Memory Options – Queue Scale

- 3 memory options for each line card: Extended (or high queue), Base (medium queue), Low (low queue)*

- Different memory option has different QoS queue scale and L2 sub-interface scale. All other system wide scale is the same across different type of the line cards, including FIB, MAC address, Bridge-domain, L3 sub-interface, VRF, etc

- All line cards have the same HW → Identical features

- Mixed different type of line cards are supported on the same chassis with same system wide scale and identical features
Each NPU has Four Main Associated memories TCAM, Search/Lookup memory, Frame/buffer memory and statistics memory
- TCAM is used for VLAN tag, QoS and ACL classification
- Lookup Memory is used for storing FIB tables, Mac address table and Adjacencies
- Stats memory is used for all interface statistics, forwarding statistics etc
- Frame memory is buffer memory for Queues

- E/B/L line card have different TCAM, Stats and Frame Memory size, which give different scale number of the QoS queues and L2 sub-interfaces per line card
- Lookup Memory is the same across line card s→ why?
  - To support mix of the line cards without impacting the system wide scale including routing, multicast, MAC address, L3 interface, MPLS label space scale
Switch Fabric Overview

- Active-active load balancing: Unicast: per-packet load balancing, Multicast: per (S,G) load balancing
- Arbiter for fabric access control. Arbiter is in active/standby mode, which is controlled by low level hardware signalling
- Frame format over fabric: super-frame, it can aggregate multiple small packet into a big sup-frame to improve the fabric throughput
Switch Fabric Bandwidth Access Overview
Intelligent Fabric and Internal System QoS

1: Fabric Request
2: Arbitration
3: Fabric Grant
4: load-balanced transmission across fabric links
5: credit return
Backpressure and VoQ Mechanism

Egress NP congestion $\rightarrow$ backpressure to ingress FIA $\rightarrow$
Packet is enqueued in the dedicated VoQ $\rightarrow$
No impact of the packet going to different egress NP $\rightarrow$
No head-of-line-block issue

VoQ Scale: Each FIA has P1/P2/BE queue set for every NP and RSPs in the entire system

Backpressure: egress NP $\rightarrow$ egress FIA $\rightarrow$ fabric Arbiter $\rightarrow$ ingress FIA $\rightarrow$ VoQ

Ingress side of LC1

Packet going to different egress NP put into different VoQ set $\rightarrow$
Congestion on one NP won’t block the packet going to different NP

Egress side of LC2
Two-Stage Packet Forwarding
Fully Distributed Forwarding on Line Cards

Packet is forwarded to the egress NP based on the information in the NP/fabric header

- Each line card has independent AIB only for local interfaces
- Each line card has independent Interface DB for local interfaces
- Both Ingress and Egress FIB – allows forwarding features to be independently applied on LCs

Ingress NP look up → Get egress NP information, add those information into fabric/NP header

Egress NP look up → Get egress logical port, VLAN, MAC, ADJ information, etc for packet rewrite
Multicast Packet Replication (1)
Switch Fabric and Egress LC Replication

1. **Fabric Replication** → replicate single copy to LCs which receive IGMP join, based on FGID table in switch fabric

2. **FIA Replication** → replicate single copy to Bridge which has IGMP join, based on MGID table in FIA

3. **Bridge Replication** → similar as FIA replication, single copy to NP

4. **NP Replication** → replicate copy per receiver based on multicast FIB table

---

**FGID** – Fabric Group ID

**MGID** – Multicast Group ID

**MFIB** – Multicast Forwarding Information Base
ASR 9000 Flexible Ethernet SW Infrastructure
(“EVC” SW Infrastructure)

- EFP (Ethernet Flow Point) or sub-interface
- Flexible VLAN tag classification
- Flexible VLAN tag rewrite
- Flexible Ethertype (.1Q, QinQ, .1ad)

1. P2P local connect

2. Flexible service mapping and multiplexing
   L2 and L3, P2P and MP services concurrently on the same port
# ASR 9000 RSP2 VS RSP440

<table>
<thead>
<tr>
<th></th>
<th>Current RSP2</th>
<th>RSP440</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Processors</strong></td>
<td>2 x 1.5GHz Freescale 8641D CPU</td>
<td>Intel x86 Jasper Forest 4 Core 2.27 GHz</td>
</tr>
<tr>
<td><strong>RAM (user expandable)</strong></td>
<td>4GB @133MHz SDR 8GB</td>
<td>6GB (RSP440-TR) and 12GB (RSP440-SE) version @1066MHz DDR3</td>
</tr>
<tr>
<td><strong>Cache</strong></td>
<td>L1: 32KB L2: 1MB</td>
<td>L1: 32KB per Core L2: 8MB shared</td>
</tr>
<tr>
<td><strong>Primary persistent storage</strong></td>
<td>4GB</td>
<td>16GB - SDD</td>
</tr>
<tr>
<td><strong>Secondary persistent storage (HD/SSD)</strong></td>
<td>30GB - HDD</td>
<td>16GB - SDD</td>
</tr>
<tr>
<td><strong>USB 2.0 port</strong></td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Acceleration / Security</strong></td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>HW assisted CPU queues</strong></td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>nV Cluster – EOBC ports</strong></td>
<td>No</td>
<td>Yes, 2 x 1G/10G SFP+</td>
</tr>
<tr>
<td><strong>Switch fabric bandwidth</strong></td>
<td>184G/slot (with dual RSP)</td>
<td>440G/slot (with dual RSP)</td>
</tr>
</tbody>
</table>
RSP440 – Front Ports

1G/10G SFP+
EOBC ports for nV Cluster

IEEE 1588, GPS
SyncE, IEEE1588 master and slave
10/100M Copper Ethernet

BITS/J.211
Sync 0, Sync 1
RJ45

USB Type A

LEDs
Status, Alarm

Aux

Management Ethernet

Console

Note, red color is the new front ports, which is supported on RSP440 only, not RSP2
NG Switch Fabric Overview
3-Stage Fabric

8x55Gbps = 440Gbps with dual RSP
4x55Gbps = 220Gbps with single RSP

8x7.5G = 60G raw bandwidth
55G available user bandwidth per fabric channel
Back-compatible: NG Switch Fabric
Mixed New Linecard and Existing Linecard

FIA0
Dual-FIA 8xNPs Linecard

FIA1
8x23G bi-directional

FIA
Single-FIA 4xNPs Linecard

4x23G bi-directional

RSP0
8x55G bi-directional

RSP1

NG Line Card

8x55Gbps = 440Gbps with dual RSP
4x55Gbps = 220Gbps with single RSP
Line Card Architecture Overview

Trident Line card
- PHY
- NP0
- NP1
- NP2
- NP3
- CPU
- 3x 10G
- 3x10GE
- SFP +
- 4x23G

Typhoon Line card
- 3x10GE SFP +
- 8x55G
- 4x23G
- A9K24x10G
- A9K-4T

Switch Fabric
- RSP0
- RSP1
- Trident Fabric
- Typhoon Fabric
LC Architecture – 36x10G

36x10G line card

6x10GE Hex PHY

Typhoon

FIA

Switch Fabric

ASIC

CPU

8x55 G

Switch Fabric

RSP0

RSP1
LC Architecture – 2x100G

100GE MAC/PHY

100G

100G

100G

100G

Ingress Typhoon

Egress Typhoon

Ingress Typhoon

Egress Typhoon

FIA

FIA

FIA

Switch Fabric

ASIC

MUX FPGA

CPU

8x55 G

Switch Fabric

RSP0

Switch Fabric

RSP1
LC Architecture – Modular Ethernet MOD160

- Supported MPA
  - 1x40GE
  - 2x40GE
  - 2x10GE
  - 4x10GE
  - 20xGE

- Typhoon

- FIA

- Switch Fabric

- ASIC

- CPU

- Modular line card

- 8x55 G

- Switch Fabric

- RSP0

- RSP1
LC Architecture – Modular Ethernet MOD80

Supported MPA
1x40GE
2x10GE
4x10GE
20xGE

Supported MPA
1x40GE
2x10GE
4x10GE
20xGE

Modular line card
Packet Flow Overview

Same as existing system: Two-stage IOS-XR packet forwarding
Uniform packet flow: All packet go through central fabric on the RP
ASR 9001 “Iron Man” Overview

4.2.1 release

Two Modular bays
Supported MPA: 20xGE(4.2.1), 2/4x10GE (4.2.1), 1x40GE (4.3.0), 2x40GE (not supported on Iron man)

Redundant (AC or DC) Power Supplies Field Replaceable

GPS, 1588 BITS

Console, Aux, Management

Fixed 4x10G SFP+ ports

EOBC ports for nV Cluster (2xSFP+)

Fan Tray Field Replaceable
System Architecture Overview

It has both central RP and LC CPU like big chassis. But it only have central switch fabric, no LC fabric.
What’s ASR 9000 nV Edge System?
Super, Simple Resiliency and more Capacity

Leverage existing IOS-XR CRS multi-chassis SW infrastructure
Simplified/Enhanced for ASR 9000 nV Edge

CRS Multi-Chassis

ASR 9000 nV Edge

Single control plane, single management plane, fully distributed data plane across multiple* physical chassis → one virtual nV system

Super, Simple network resiliency, and extensible node capacity

*Maximum two physical chassis per nV Edge System in phase 1
nV Edge Overview

Control Plane EOBC Extension (L1 or L2 connection)
One or two 10G/1G from each RSP

- Control plane EOBC extension is through special 1G or 10G EOBC ports on the RSP. External EOBC could be over dedicated L1 link, or over port-mode L2 connection.
- Data plane extension is through regular LC ports (it can even mix regular data ports and inter-chassis data plane ports on the same LC).
- Doesn’t require dedicated fabric chassis → flexible co-located or different location deployment, lower cost.

External EOBC link fail won’t cause RP failover as long as it has alternative EOBC link.

Inter-chassis data link (L1 connection)
10G or 100 G bundle (up to 32 ports)

Regular 10G or 100G data ports

Special external EOBC 1G/10G ports on RSP

Internal EOBC
Satellite – Host Control Plane
Satellite discovery and control protocol

Discovery Phase
• A CDP-like link-level protocol that discovers satellites and maintains a periodic heartbeat

Control Phase
• Used for Inter-Process Communication between Host and Satellite
• Cisco proprietary protocol over TCP socket for the time being.
Satellite Operation (1) – End User View

Virtual satellite interface/sub-int sample CLIs

interface GigabitEthernet 100/0/0/1
ipv4 address 1.1.1.1 255.255.255.0
interface GigabitEthernet 100/0/0/2.100 l2transport
encapsulation dot1q 100
rewrite ingress tag push dot1q 2

Satellite access port

Satellite uplink port is treated as internal “fabric” port

Satellite access port is represented by virtual “nv” interface on the Host. User configure this virtual interface just as regular local L2/L3 interface or sub-interface on the Host

All satellite configuration is done on the Host
Satellite Operation (2) – Packet Flow

- Satellite ONLY does local connect between access and fabric port
- Satellite access port to normal port communication
- No local switching/routing on satellite, all forwarding is via Host
- Satellite ONLY does local connect between access port and fabric, NOT between access ports. No MAC learning involved
- Advanced features are processed on the Host chassis satellite virtual port
- Minimal mandatory features could be applied to satellite directly, including basic QoS, multicast replication, OAM performance measurement, SyncE. However, the configuration is still done on the Host
First Satellite Hardware – ASR 9000v

- **Power Feeds**
  - Redundant -48vDC Power Feeds
  - Single AC power feed

- **44x10/100/1000 Mbps Pluggables**
  - Full Line Rate Packet Processing and Traffic Management

- **Field Replaceable Fan Tray**
  - Redundant Fans
  - ToD/PSS Output
  - Bits Out

- **4x10G SFP+**
  - Initially used as Fabric Ports ONLY (could be used as access port in the future)
  - Plug-n-Play In-Band Management
  - Automatic Discovery and Provisioning
  - Co-Located or Remote Distribution

- **1 RU ANSI & ETSI Compliant**
- **LEDs**
Summary

- New RSP440/Typhoon LC
- New ASR9001
- Nv Satellite
- Nv Cluster
Otázky a odpovědi

- Twitter  www.twitter.com/CiscoCZ
- Talk2Cisco  www.talk2cisco.cz/dotazy
- SMS  721 994 600

- Zveřejněme Vás na Ptali jste se… v sále LEO
  1.den 17:45 – 18:30
  2.den 16:30 – 17:00
Prosíme, ohodnoťte tuto přednášku.