Architektura
Carrier Routing System

T-SP4
David Jakl – Cisco
Program

• Overview
• Architecture
• Line Card
• Switch Fabric
• Multi-chassis
• Summary
CRS Overview
CRS-1
Carrier Routing System 1

Core, Peering, L3 (L2) Edge Router
- 40Gbps/slot Full Duplex
- Redundant 3-stage non-blocking Beneš Switch Fabric
- IOS XR - Highly modular, Microkernel-based
- SDR - Secure Domain Routers
- IPoDWDM 10GE/40G, Tunable, G.709, (E)FEC, 40G over 10G DWDM System (single lambda)

Control Plane:
- Redundant Route Processors (RP → PRP)
- More RPs for scale (DPRs) for "Process Placement“ or SDR

PLIMs (Front):
- Modular: SIP-800/6xSPA: 1/10GE, POS 155M→10G, E3, ATM
- Fixed: 1/10GE, POS 2.5G→40G, 10GE/40G IPoDWDM
- Services: CGSE = Carrier Grade Services Engine

Forwarding Line Cards (Back):
- MSC40 – Core/Peering/Edge (4/8/16/MC)
  - 2M routes, H-QoS 8000 queues/port, 2000 intfs
- FP40 – Thin Core/Peering (4/8) +licenses
  - 2M routes, 8 queues/port, 100 intfs

4-slot 320Gbps
8-slot 640Gbps
16-slot 1280Gbps

May 2004
>8000 CRS Chassis
>400 Customers

Today Multi-chassis
→ 10.2 (8xLCC) → 92 Tbps (72xLCC)

© 2012 Cisco and/or its affiliates. All rights reserved.
Cisco Expo
CRS-3
Carrier Routing System 3
Powered by QuantumFlow Array

140Gbps/slot Full Duplex
- 140G Switch Fabric (4/8/16/MC)

PLIMs (Front):
1x100GE Line Rate, CFP (L4)
14x10GE Line Rate, LAN/WAN PHY, XFP
20x10GE Oversubscribed 140G, LAN/WAN PHY, XFP

2012: 100GE OTN IPoDWDM PM-QPSK over 10G DWDM System (single lambda)

Forwarding Line Cards (Back):
MSC140 – Edge (4/8/16/MC)
  4M routes, H-QoS 64000 queues/port, 12000 intfs
FP140 – Core/Peering (4/8/16/MC) +licenses
  1/4M routes, 8 queues/port, 250 intfs
LSP140 – MPLS Core (limited IP) (4/8/16/MC) +licenses

HW Ready: E-OAM, Video monitoring, Time stamping, SyncE

CRS-1 → CRS-3
- NO Chassis/Power/Cooling/RP upgrade
- Switch Fabric upgrade: Hitless = Zero packet loss
- 40G & 140G Cards in the same chassis

TODAY Multi-chassis MAX
→ 35.8 (8xLCC) → 322 Tbps (72xLCC)
CRS-3 PLIMs

<table>
<thead>
<tr>
<th>CRS-3</th>
<th>#100GE LR</th>
<th>#10GE LR</th>
<th>#10GE OS</th>
</tr>
</thead>
<tbody>
<tr>
<td>4-slot</td>
<td>4</td>
<td>56</td>
<td>80</td>
</tr>
<tr>
<td>8-slot</td>
<td>8</td>
<td>112</td>
<td>160</td>
</tr>
<tr>
<td>16-slot</td>
<td>16</td>
<td>224</td>
<td>320</td>
</tr>
<tr>
<td>MC 8 LCC</td>
<td>128</td>
<td>1792</td>
<td>2560</td>
</tr>
<tr>
<td>MC 72 LCC</td>
<td>1152</td>
<td>16128</td>
<td>23040</td>
</tr>
</tbody>
</table>

1x 100GE

14x 10GE

20x 10GE

CFP
CRS in 2012 and 2013

Subject to change

2012

- CRS-3 16-slot B2B (2+0) Multichassis
- 100GE SR10 CFP
- 100GE IPoDWDM PLIM
- 4x40GE OTN PLIM – CFP: SR4, LR4, FR
- Tunable DWDM XFP
- FlexPLIM – 6x10GE LAN/WAN/OTN + 4xSPA

2013

- CRS-3 8-slot B2B (2+0) Multichassis
- CGSE+ 80Gbps
- 400G/slot – Switch Fabrics, Line Cards, PLIMs
CRS Architecture
**CRS Architecture**

**PLIM** – Physical Layer Interface Module  
**MSC** – Modular Service Card  
**RP** – Route Processor

**Switch Fabric**  
- 4 or 8 redundant planes  
- Multi-Chassis option

![Diagram of CRS Architecture](image-url)

- **Packets In**
- **Packets Out**

**Definitions**

- **PLIM** – Physical Layer Interface Module
- **MSC** – Modular Service Card
- **RP** – Route Processor
- **Switch Fabric**
  - 4 or 8 redundant planes
  - Multi-Chassis option
- **FABRIC S1**
- **FABRIC S2**
- **FABRIC S3**
CRS High Level Physical Architecture
Line Card Chassis (LCC)

- **RP** – System Master
- **PLIM & MSC** – Packet Forwarding
- **Switch Fabric** – LC/LC & RP/LC Path
- **Fans and Controllers** - Cooling
- **Air Intake and Exhaust** - Cooling
- **Power Supplies** – AC or DC

**Front View**
- 16 Slot
- 8 Slot
- 4 Slot

**Rear View**
- 4 Slot
- 8 Slot
- 16 Slot
CRS Line Card
CRS-1 PLIM

- 1 → 4 PLIM ASICs (PLAs) depending on card type
- Nominally 40 Gbps for CRS-1
- Oversubscription allowed when all Ethernet
- 96 Gbps aggregate bandwidth into PSE = no bandwidth bottleneck even when oversubscribed

<table>
<thead>
<tr>
<th>PLIM</th>
<th>PLAs</th>
<th>BW to PSE (per PLA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4xOC192 POS</td>
<td>2</td>
<td>48 Gbps</td>
</tr>
<tr>
<td>16xOC48 POS</td>
<td>4</td>
<td>24 Gbps</td>
</tr>
<tr>
<td>1xOC768 POS</td>
<td>1</td>
<td>96 Gbps</td>
</tr>
<tr>
<td>4/8xTenGE</td>
<td>2</td>
<td>48 Gbps</td>
</tr>
<tr>
<td>SIP-800</td>
<td>2</td>
<td>48 Gbps</td>
</tr>
</tbody>
</table>
CRS Line Card Architecture

- **Ingress PSE**
  - RX Forwarding & Features
  - Input Queuing
  - Fabric Queuing
  - Segmentation to Cells
  - Shaping

- **EgressQ**
  - Output Queuing

- **Egress PSE**
  - TX Forwarding & Features
  - Multicast Replication

- **2 FabricQs**
  - Cell Reassembly & Queuing into PSE
CRS PSE = Packet Switching Engine ASIC

- 188 Parallel Processing Engines (PPE)
  Independent operation, not pipelined
  All PFEs can access forwarding resources
- Micro-coded for Service Flexibility
- Performs per packet operations
  IPv4, IPv6, MPLS, and Multicast lookups
  Statistics
  ACLs and Netflow accounting
  Policing, Marking and WRED
- Access to memories and TCAMs
- Adds 8-byte buffer header
CRS-1 PSE Forwarding ASIC Architecture

- Cluster – 12 PPEs
- Packet Buffer
- Instruction Memory

Input Stream Interface

Packets from PLAs

Packet Distributor

Packet Processing Engine

Packets

MUX to access Resource Fabric

Resources

Cluster – 12 PPEs

Packet Buffer

Instruction Memory

Input Stream Interface

Packets from PLAs

Packet Distributor

Packet Processing Engine

Packets

MUX to access Resource Fabric

Resources
Packet Path within PSE
Assign Packet (Header) to a PPE
Packet Path within PSE
Perform lookup and features using Resources
Packet Path within PSE
Recombine header and tail and send to IngressQ
**IngressQ**

- **Input Shaping Queues**
  - Per Interface
  - HP & LP per class if configured (max 8k queues)

- **Fabric Destination Queues**
  - HP & LP for every FabricQ in system
  - 4 queues for every MSC in entire system
  - Queue determined by Ingress QoS or Fabric QoS

- **Segmentation of packets into cells**

- **45 Gbps limit between Shape Queues and Fabric Destination Queues (140 Gbps in CRS-3)**

- **Discard bitmap**
CRS Fabric Cells

- 136 byte cells with
  - 12 byte header, 120 byte payload, 4 byte CRC
- 1 or 2 packets per cell
  - Packets must start on a 30 byte boundary
  - Packets sharing a cell must be same priority and cast
  - Entire cell travels over 1 fabric plane
- Round Robin among 8 fabric planes (4 for 4 slot CRS)
FabricQ Queues

- Cells reassembled into packets
- Packets queued prior to PSE
- Unicast queues
  - Per type of service (HP/AF/BE)
  - Per output interface
- Multicast queues
  - Per type of service
- Raw (to CPU) queues
  - 8 queues
Discard Bitmap Concept
Drop on Ingress when a FabricQ queue is congested
Egress PSE

- Identical hardware to Ingress PSE
- Performs lookup for output adjacency
- Performs output features: Policing, Marking, WRED, Netflow, ACLs, Statistics, …
- HW multicast replication for multiple ports on same line card
Ingress PSE selects
- Output Line Card
- Output Interface
- Fabric Dest Q (HP/LP)
- FabricQ (FQ0 or FQ1)
- IngressQ Queue
  - HP, AF, or BE

Egress PSE selects
- Output Interface
- Output Queue
- Adjacency
- Dest MAC address
Hardware Multicast Replication

- HW Replication within fabric planes
  For cells going to multiple line cards
  No performance impact for additional replications

- HW Replication on Egress PSE
  For multiple ports on a line card

- Efficient scale for high fan-out
  No increase in load on MSC

1 plane of 8 (4) shown
EgressQ

• Per interface/sub-interface queuing
• 8k queues
• 1GB packet buffer
• P2PMDRR
• MQC configuration
  Strict HP queue
  Bandwidth guarantees
  Shaping
  Bandwidth remaining
• 3 Level Hierarchy
  Port – Highest Level Queuing Engine
  Group – Middle Level Queuing Engine
  Queue – Lowest Level Queuing Engine
CGSE Architecture

Service Engine PLIM

Octeon CPUs

Same PLA as SIP-800

Modular Services Card

iPSE

IngressQ

EgressQ

ePSE

FabQs

Accel FPGA

Accel FPGA

Modular Services Card

Service Engine PLIM

Octeon CPUs

Same PLA as SIP-800

Accel FPGA
CGSE Packet Flow
CRS-3 Line Card

- Output Queuing
- Output Features Multicast Replication
- Cell Reassembly

Forwarding Lookup
Input Features

Queuing for Fabric
Cell Segmentation
Input Shaping

from PLIM

160G
2x100G

PSE

IngressQ

160G
141G

Intel CPU
Sub-system

EgressQ

FabricQ

FabricQ

120G
2x80G

160G

160G

100G

100G

113G

113G

from Fabric

to Fabric
CRS Switch Fabric
CRS-1 Switch Fabric Overview

3 stage switching fabric

40 Gbps

100 Gbps

8 independent fabric planes (4 on CRS-1/4)
2.5x speedup through fabric
Support for 72 chassis & 1296 RP/MSC clients

High and low priority cells
Set on control by default
Set on transit pkts via CLI
Vital Bit for IPC
Decisions at Each Stage in 16 slot
For each of the 8 planes

Send to any S2. No need to look at header

Look at cell header. Send to S3 based on chassis/LC

Look at cell header. Send to specific LC and FabricQ
CRS-1/16 Switch Fabric
Full View of 1 Plane

Fabric Plane (1 of 8)

Ingress LC & DRP x 8

RP X 2

Ingress LC & DPR x 8

Egress LC & DRP x 8

(16 LCs x 4 links) + (2 RPs x 2 links)

= 1 x 2.5G links

= 18 x 2.5G links

Fabric speedup from S2 to S3

(16 LCs x 4 links) + (2 RPs x 2 links)
Queuing Inside the Fabric
Each destination and HP/LP queued separately
Switch Fabric Multicast Replication

- CRS-1 provides efficient multicast replication via 3 operations
  S2 can replicate cells to registered (via FGID) S3 SEAs
  S3 can replicate cells to registered (via FGID) FabricQs
  Egress PSE can replicate packets for each output port

- 1 million Fabric Group IDs
  Program fabric for mcast

- S2 and S3 lookup FGID
CRS Multi-chassis
S2 stage moves into Fabric Card Chassis (FCC)
Logical operation of fabric remains the same
Decisions at Each Stage in 2+1 Multi-chassis
For 1 the 8 planes

S1 and S3 Stages in Line Card Chassis

S2 Stage in Fabric Card Chassis

Logical decisions same as standalone system
CRS Multi-chassis Components
CRS Multi-chassis

- Single to Multi-chassis upgrade w/o packet loss
  - Moves fabric stage 2 to separate chassis
  - Upgrade LCC fabric cards
  - Add fabric card chassis (FCC)
  - Connect control GE
  - Connect LCC and FCC with fiber bundles

Optical Array Cable (100m)

72 Fiber Bundle

FCC Optical Connections
CRS Multi-chassis 2+1

LCC 0

8 S13 boards

FCC

8 S2 boards & OIMs

LCC 1
CRS 2+1 \(\rightarrow\) 2+0 (B2B) Multi-chassis
CRS Multi-chassis Fabric Plane Distribution

1 FCC
Max 9 LCC

2 FCC
Max 18 LCC

4 FCC
Max 36 LCC

8 FCC
Max 72 LCC
CRS Summary
CRS Packet Path Review
More PPEs
Faster PPEs

More memory
Higher scale

PLIM QoS
140G PLIMs

More Queues
8K -> 64K

Faster CPU
CRS-3 Queuing

- 2 queues/port; 75MB total
- 64k Input rate shaping queues; 1GB total
- 3072 High priority fabric destination queues
- S2 Queues per priority per fabric group
- S3 Queues per priority per fabric destination

**PLA**
- WRED
  - PSE
  - 8k shaped Queues
  - 64k Input rate shaping queues; 1GB total
  - 3072 High priority fabric destination queues
  - S2 Queues per priority per fabric group
  - S3 Queues per priority per fabric destination

**IngressQ**
- Discard Filter
  - Fabric Destination BP
  - 3072 Low priority fabric destination queues

**EgressQ**
- WRED
  - PSE
  - 8k shaped Queues
  - 64K queues, 16K groups; 1GB total

**PLA**
- 512 raw queues in fabricQ; .5GB total per fabricQ
- ~110 MTUs
CRS Summary

#1 The Fastest Industry Leading Core Router

Carrier Class Architecture

- Fully Modular
- Scalable
- Reliable
- Predictable
- Backward compatible
Odkazy

www.cisco.com/go/crs

- Data Sheets and Literature:

- Support Documentation: