Best Practices to Deploy High-Availability in Service Provider Edge and Aggregation Architectures

BRKSPG-2402

www.ciscolivevirtual.com
Abstract

As Service Providers are deploying value-added triple-play or quadruple play services to maintain or generate a higher average revenue per user, overall Service Availability becomes increasingly important. High Availability techniques such as Fast Convergence or MPLS TE FRR have focused on raising the availability of the network core in the past. Recently, these techniques are being increasingly deployed in Ethernet Aggregation networks, for example by introducing MPLS TE FRR in the aggregation. Also, additional high-availability mechanism are being developed to enhance the resilience of the IP Edge against failures. Examples of new developments include IP Fast-Reroute, BGP Prefix Independent Convergence for both the Core and Edge, or even stateful application inter-chassis redundancy mechanisms to overcome single-system outages. This Session aims to provide the audience with best current practices to increase service availability by deploying Cisco High-Availability mechanisms in both the Aggregation and the IP Edge. Traditional HA techniques such as NSF/SSO, BFD, Fast convergence or NSR are reviewed. The details of new technologies such as IP FRR, BGP PIC are discussed in depth. Furthermore, advanced topics such as achieving HA for Layer 4-7 services or stateful inter-chassis redundancy solutions are introduced. The Session also provides the best current practices of deploying the tools offered by the Cisco High-availability toolset, in particular the deployment of MPLS TE FRR in the aggregation. Furthermore, possible stateful and stateless clustering approaches are introduced, which SPs may use to increase the availability of their IP Edge architecture.
## Glossary

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NHAT</td>
<td>Next hop address tracking</td>
</tr>
<tr>
<td>ACL</td>
<td>Access Control List</td>
</tr>
<tr>
<td>ACT</td>
<td>Active</td>
</tr>
<tr>
<td>APS</td>
<td>Automatic Protection Switching</td>
</tr>
<tr>
<td>ARP</td>
<td>Address Resolution Protocol</td>
</tr>
<tr>
<td>AS</td>
<td>autonomous System</td>
</tr>
<tr>
<td>ATM</td>
<td>Asynchronous Transfer Mode</td>
</tr>
<tr>
<td>BFD</td>
<td>Bi Directional Forwarding Detection</td>
</tr>
<tr>
<td>BNG</td>
<td>Broadband Network Gateway</td>
</tr>
<tr>
<td>BW</td>
<td>Bandwidth</td>
</tr>
<tr>
<td>CC</td>
<td>Continuity Check</td>
</tr>
<tr>
<td>CC</td>
<td>control connection</td>
</tr>
<tr>
<td>CDR</td>
<td>call detail record</td>
</tr>
<tr>
<td>CE</td>
<td>Customer Edge</td>
</tr>
<tr>
<td>CE</td>
<td>Customer Edge</td>
</tr>
<tr>
<td>CF</td>
<td>Checkpoint facility</td>
</tr>
<tr>
<td>CFM</td>
<td>Configuration and Fault Management</td>
</tr>
<tr>
<td>CLI</td>
<td>Command Line Interface</td>
</tr>
<tr>
<td>CM</td>
<td>Chassis Manager</td>
</tr>
<tr>
<td>CP</td>
<td>Control Plane</td>
</tr>
<tr>
<td>CPLD</td>
<td>Complex Programmable Logic Device?</td>
</tr>
<tr>
<td>CSC</td>
<td>Carrier’s Carrier</td>
</tr>
<tr>
<td>DHCP</td>
<td>Dynamic Host Configuration Protocol</td>
</tr>
<tr>
<td>DP</td>
<td>Data Plane</td>
</tr>
<tr>
<td>DPM</td>
<td>Defects per Million</td>
</tr>
<tr>
<td>DSLAM</td>
<td>DSL Access Multiplexer</td>
</tr>
<tr>
<td>E2E</td>
<td>End to end</td>
</tr>
<tr>
<td>ECMP</td>
<td>equal cost multipath</td>
</tr>
<tr>
<td>EEM</td>
<td>Embedded Event Manager</td>
</tr>
<tr>
<td>EOAM</td>
<td>Ethernet OAM</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EOBC</td>
<td>Ethernet out of band management</td>
</tr>
<tr>
<td>ESP</td>
<td>Embedded Services Processor</td>
</tr>
<tr>
<td>EVC</td>
<td>Ethernet Virtual Circuit</td>
</tr>
<tr>
<td>EVDO</td>
<td>Evolution Data Only</td>
</tr>
<tr>
<td>FECP</td>
<td>Forwarding Engine Control Processor</td>
</tr>
<tr>
<td>FIB</td>
<td>Forwarding Information Base</td>
</tr>
<tr>
<td>FM</td>
<td>Forwarding Manager</td>
</tr>
<tr>
<td>FR</td>
<td>Frame Relay</td>
</tr>
<tr>
<td>FRR</td>
<td>Fast Re Route</td>
</tr>
<tr>
<td>FSOL</td>
<td>First Sign of Life</td>
</tr>
<tr>
<td>FWLB</td>
<td>Firewall Loadbalancing</td>
</tr>
<tr>
<td>GEC</td>
<td>Gigabit Ether Channel</td>
</tr>
<tr>
<td>GLBP</td>
<td>Global Load Balancing Protocol</td>
</tr>
<tr>
<td>GRE</td>
<td>Generic Route Encapsulation</td>
</tr>
<tr>
<td>GW</td>
<td>Gateway</td>
</tr>
<tr>
<td>HA</td>
<td>High Availability</td>
</tr>
<tr>
<td>HSRP</td>
<td>Hot Standby Routing Protocol</td>
</tr>
<tr>
<td>HW</td>
<td>Hardware</td>
</tr>
<tr>
<td>IETF</td>
<td>Internet Engineering Task Force</td>
</tr>
<tr>
<td>IF</td>
<td>Interface</td>
</tr>
<tr>
<td>IG</td>
<td>Internal Gateway Protocol</td>
</tr>
<tr>
<td>IOCP</td>
<td>Input Output control Processor</td>
</tr>
<tr>
<td>IOS</td>
<td>Internet Operating System</td>
</tr>
<tr>
<td>IP</td>
<td>Internet Protocol</td>
</tr>
<tr>
<td>IPC</td>
<td>Inter process communication</td>
</tr>
<tr>
<td>ISG</td>
<td>Intelligent Services Gateway</td>
</tr>
<tr>
<td>ISPF</td>
<td>incremental Shortest Path First</td>
</tr>
<tr>
<td>ISSU</td>
<td>in service software upgrade</td>
</tr>
<tr>
<td>IWF</td>
<td>Interworking function</td>
</tr>
</tbody>
</table>
### Glossary (Cont.)

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>L2TP</td>
<td>Layer 2 transport protocol</td>
<td>NIC</td>
<td>network interface card</td>
</tr>
<tr>
<td>LAC</td>
<td>L2TP access concentrator</td>
<td>Nr</td>
<td>receive sequence number</td>
</tr>
<tr>
<td>LACP</td>
<td>Link aggregation control Protocol</td>
<td>Ns</td>
<td>send sequence number</td>
</tr>
<tr>
<td>LAN</td>
<td>Local Area Network</td>
<td>NSF</td>
<td>non stop forwarding</td>
</tr>
<tr>
<td>LC</td>
<td>Linecard</td>
<td>NSR</td>
<td>non stop routing</td>
</tr>
<tr>
<td>LDP</td>
<td>label Distribution Protocol</td>
<td>NVRAM</td>
<td>non volatile random access memory</td>
</tr>
<tr>
<td>LFA</td>
<td>loop free alternate</td>
<td>OAM</td>
<td>operations, administration and maintenance</td>
</tr>
<tr>
<td>LI</td>
<td>Lawful Intercept</td>
<td>OCE</td>
<td>Object Chain Element</td>
</tr>
<tr>
<td>LMI</td>
<td>Local management interface</td>
<td>OIR</td>
<td>online insertion and removal</td>
</tr>
<tr>
<td>LNS</td>
<td>L2TP network Server</td>
<td>OS</td>
<td>operating system</td>
</tr>
<tr>
<td>LOS</td>
<td>Loss of signal</td>
<td>PADR</td>
<td>PPP active discovery</td>
</tr>
<tr>
<td>LSDB</td>
<td>link state database</td>
<td>PE</td>
<td>provider edge</td>
</tr>
<tr>
<td>LSP</td>
<td>label switched path</td>
<td>PIC</td>
<td>prefix independent convergence</td>
</tr>
<tr>
<td>LTE</td>
<td>long term evolution</td>
<td>PIM</td>
<td>protocol independent multicast</td>
</tr>
<tr>
<td>MC LAG</td>
<td>multi chassis link aggregation</td>
<td>PPP</td>
<td>Point to point protocol</td>
</tr>
<tr>
<td>mcast</td>
<td>multicast</td>
<td>PS</td>
<td>power supply</td>
</tr>
<tr>
<td>MD5</td>
<td>message Digest algorithm 5</td>
<td>PSN</td>
<td>Packet Switched Network</td>
</tr>
<tr>
<td>MFIB</td>
<td>multicast forwarding information base</td>
<td>PTA</td>
<td>PPP termination and aggregation</td>
</tr>
<tr>
<td>MLD</td>
<td>multicast listener discovery</td>
<td>PVRSTP</td>
<td>Per VLAN rapid spanning tree</td>
</tr>
<tr>
<td>MME</td>
<td>mobile management entity</td>
<td>PW</td>
<td>pseudowire</td>
</tr>
<tr>
<td>MoFRR</td>
<td>Multicast Only fast reroute</td>
<td>QFP</td>
<td>Quantum flow Processor</td>
</tr>
<tr>
<td>MPLS</td>
<td>Multiprotocol label switching</td>
<td>RADIUS</td>
<td>remote authentication dial in user service</td>
</tr>
<tr>
<td>MRIB</td>
<td>multicast routing information base</td>
<td>RF</td>
<td>redundancy facility</td>
</tr>
<tr>
<td>MSC</td>
<td>mobile switching centre</td>
<td>RMA</td>
<td>Return material authorisation</td>
</tr>
<tr>
<td>MSPP</td>
<td>Multi-service provisioning platform</td>
<td>RNC</td>
<td>radio network controller</td>
</tr>
<tr>
<td>MST</td>
<td>Minimum spanning tree</td>
<td>RPR</td>
<td>route processor redundancy</td>
</tr>
<tr>
<td>MTBF</td>
<td>mean time between failures</td>
<td>RSP</td>
<td>route switch processor</td>
</tr>
<tr>
<td>MTSO</td>
<td>mobile telephone switching office</td>
<td>RSVP</td>
<td>resource reservation protocol</td>
</tr>
<tr>
<td>MTTR</td>
<td>mean time to repair</td>
<td>SAA</td>
<td>service assurance agent</td>
</tr>
<tr>
<td>NAT</td>
<td>network address translation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>--------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SBC</td>
<td>session border controller</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SBY</td>
<td>standby</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SGW</td>
<td>SAE gateway</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIP</td>
<td>Session initiation protocol</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SLA</td>
<td>service level assurance</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SLB</td>
<td>server loadbalancing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SP</td>
<td>service Provider</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPA</td>
<td>Shared port adapter</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPF</td>
<td>shortest path first</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SRLG</td>
<td>shared risk link group</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSH</td>
<td>secure shell</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSO</td>
<td>stateful switchover</td>
<td></td>
<td></td>
</tr>
<tr>
<td>STP</td>
<td>spanning tree protocol</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SW</td>
<td>software</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T&amp;C</td>
<td>terms &amp; conditions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TCAM</td>
<td>ternary content addressable memory</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TE</td>
<td>traffic engineering</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TR</td>
<td>Traceroute</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UC</td>
<td>unified communications</td>
<td></td>
<td></td>
</tr>
<tr>
<td>uRPF</td>
<td>unicast reverse path forwarding</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VAI</td>
<td>virtual access interface</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VC</td>
<td>Virtual Circuit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VCCV</td>
<td>VC connection verification</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VIP</td>
<td>virtual IP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VLAN</td>
<td>virtual LAN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VMAC</td>
<td>virtual MAC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VPN</td>
<td>virtual private network</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VRF</td>
<td>virtual routing and forwarding table</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VRRP</td>
<td>virtual router redundancy protocol</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WAN</td>
<td>wide area network</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Agenda

- Motivation for High Availability in SP Aggregation Networks
- Network Level High Availability
- System High Availability
- Service High Availability
- Stateful Inter-chassis Redundancy
- Case Studies
- Summary & Conclusions
Motivation for High-Availability in the IP Edge and Aggregation
High Availability & Service Level Agreements

- Many SPs specify their SLAs in the T&Cs
- Important characteristic of both business and residential services
- Historically given for Core network, but expanding to end-end SLAs
- Metrics
  - Service Availability (averaged over time)
  - Mean time to repair (MTTR)
  - Packet Loss / Delay / Jitter

- Examples
  - AT&T
  - Sprint
  - Verizon Business
  - BT
  - Level 3
### What Is High Availability?

Two ways to state availability of a network:
- **Percentage Method**
- **DPM Method** = Defects per Million (Hours of Running Time)

<table>
<thead>
<tr>
<th>Availability</th>
<th>DPM</th>
<th>Downtime per Year (24x365)</th>
</tr>
</thead>
<tbody>
<tr>
<td>99.000%</td>
<td>10000</td>
<td>3 Days 15 Hours 36 Minutes</td>
</tr>
<tr>
<td>99.500%</td>
<td>5000</td>
<td>1 Day 19 Hours 48 Minutes</td>
</tr>
<tr>
<td>99.900%</td>
<td>1000</td>
<td>8 Hours 46 Minutes</td>
</tr>
<tr>
<td>99.950%</td>
<td>500</td>
<td>4 Hours 23 Minutes</td>
</tr>
<tr>
<td>99.990%</td>
<td>100</td>
<td>53 Minutes</td>
</tr>
<tr>
<td>99.999%</td>
<td>10</td>
<td>5 Minutes</td>
</tr>
<tr>
<td>99.9999%</td>
<td>1</td>
<td>30 Seconds</td>
</tr>
</tbody>
</table>

**Reactive**

**Proactive**

**Predictive**

"High Availability"
Availability Definitions

\[
\text{Availability} = \frac{\text{MTBF}}{\text{MTBF} + \text{MTTR}}
\]

- The uptime divided by the total time percentage time your network is operational
- **MTBF** is Mean Time Between Failure
  - When does it fail?
- **MTTR** is Mean Time To Repair
  - How long does it take to fix?
Calculated vs. Measured Availability

- **Calculated Availability** based on:
  - Network design
  - Component MTBF and MTTR (different underlying models, simulations)
  - Cisco uses Industry standards to compute Hardware MTBF
  
  Basic Availability Calculation Formulae:

  \[
  A_{\text{Series}} = \prod_{k=1}^{N} A_k = A_1 \times A_2 \times \ldots \times A_N
  \]

  \[
  A_{\text{Parallel}} = 1 - \prod_{k=1}^{N} (1 - A_k) = 1 - (1 - A_1) \times \ldots \times (1 - A_N)
  \]

- **Measured Availability** based on:
  - ICMP Reachability (E2E, Device)
  - Cisco Service Assurance Agent (SAA)
  - Trouble Ticket / Outage Log Analysis
  
  Observed Method: Shipping/RMA Method
Reduction of MTTR

- Stateful inter-chassis redundancy allows for additional resilience against
  System Failures
  Interface Failures

- Issue is not really MTBF of hardware modules, but rather
  Line failures / optical path failures
  Interface failures
  Power outages

- Goal of stateful inter-chassis redundancy is sub-second failover with state preservation for applications

<table>
<thead>
<tr>
<th>Product ID</th>
<th>MTBF (hrs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASR1000-RP2</td>
<td>380532</td>
</tr>
<tr>
<td>ASR1000-ESP20</td>
<td>335317</td>
</tr>
<tr>
<td>ASR1000-SIP10</td>
<td>287549</td>
</tr>
<tr>
<td>ASR1006</td>
<td>1986649</td>
</tr>
<tr>
<td>ASR1006-PWR-AC</td>
<td>570776</td>
</tr>
<tr>
<td>ASR1006-PWR-DC</td>
<td>357781</td>
</tr>
<tr>
<td>ASR1000-SIP40</td>
<td>283225</td>
</tr>
<tr>
<td>ASR1000-ESP40</td>
<td>118790</td>
</tr>
<tr>
<td>SPA-8X1GE-V2</td>
<td>482023</td>
</tr>
<tr>
<td>SPA-1X10GE-L-V2</td>
<td>411892</td>
</tr>
</tbody>
</table>
Device Availability Calculation

**IOS**
- **IF1**
  - PA-E3
  - PA-POS-OC3
- **IF2**
- **CPU**
  - NPE-400
- **Chassis and Backplane**
  - CISCO7206-VXR

**Cisco 7206**
- **P/S**
  - PWR-7200-AC

**Calculated MTBF Values from Cisco Database**

**IOS**
\[
\text{IOS} = \frac{30.000}{30.000 + 0.1} = 0.999997
\]

**IF1**
\[
\text{IF1} = \frac{1.120.000}{1.120.000 + 4} = 0.999996
\]

**IF2**
\[
\text{IF2} = \frac{600.000}{600.000 + 4} = 0.999993
\]

**CPU**
\[
\text{CPU} = \frac{490.000}{490.000 + 4} = 0.999992
\]

**BB**
\[
\text{BB} = \frac{460.000}{460.000 + 8} = 0.999983
\]

**P/S**
\[
\text{P/S} = \frac{750.000}{750.000 + 4} = 0.999995
\]

**System Availability**
\[
\text{System Availability} = 0.999997 \times \ldots \times 0.999983 \times (1-(1-0.999995)^2)
\]
\[
= 0.999961 = 99.9961\%
\]
Availability of R1 and R2 in Series  
= (0.999961*0.999961) = 0.99992175

Availability of Parallel Network Path (R1-R4)  
= 1 - ((1-0.999921)(1 - 0.999921)) = 0.999999994

Network Availability = 99.9999%  
Only Based on Device Availability Values

but not considered:  
- Links (WAN, LAN)  
- Computer NICs  
- Computer OS  
- Computer Applications
Cisco High-Availability Focus

System Level Resilience
- increase MTBF using resilient HW/SW
- minimise MTTR for system failures (RP, LCs and SW)
- Mitigate planned outages by providing hitless software upgrades

Network Level Resilience
- in the core and where redundant paths exist
- Deliver features for fast network convergence, protection & restoration

Embedded Management and Automation
- Embed intelligent event management for proactive maintenance
- Automation and configuration management to reduce human errors
Cisco HA Feature Toolbox: System Level

- RPR, SSO
- NSF, NSR
- SSO Multirouter APS
- Stateful NAT/IPSec/Firewall/SLB stateful failover within single chassis
- MPLS HA (L3VPN, L2VPN, InterAS, CSC, TE, FRR)
- IOS ISSU, dual IOS
Network Level Resiliency

- Network Design Resiliency
  - Dual-homing
  - APS, GEC
- Event Dampening
- Fast Convergence
  - iSPF Optimisation (OSPF, IS-IS)
  - BGP Optimisation
  - Fast BGP Convergence
- Graceful Restart (MBGP, OSPF, RSVP, LDP)
- EMCP, Anycast, dual RR
- VRRP/HSRP/GLBP/SLB/FWL
- MPLS High Availability
- LDP Graceful Restart
- MPLS/VPN NSF
- BFD
- MPLS FRR Path Protection
- MoFRR
- IP FRR
- Pseudowire Redundancy
- Spanning Tree (MST, PVRSTP...)
- IEEE 802.3ad (LACP)
Cost of High-Availability

- Designing a network for higher Service Availability comes at a cost
  - Redundant Network Elements
  - Redundant Links
  - Redundant System Components (route processors, forwarding processors, power supplies, etc.)

- Operational costs
  - Lower steady-state Utilisation levels
  - Increased configuration and management
  - Tighter maintenance windows
Cost of High-Availability: Example

- Large SP Network
  Residential Services (3-Play)
  10M Subscribers
  1.25 Mbps / subscriber

- Up to 96% increased CAPEX for full redundancy!
  Opex increased due to higher number of network elements

<table>
<thead>
<tr>
<th>Redundancy Scheme</th>
<th>Total Cost $M</th>
<th>Chassis Costs $M</th>
<th>Interface Costs (SPA, SFPs), $M</th>
<th>Number of nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>No redundancy</td>
<td>$1,232</td>
<td>$529</td>
<td>$704</td>
<td>4658</td>
</tr>
<tr>
<td>Access NW uplink redundancy (Agg1, Agg2, Agg3)</td>
<td>$1,250</td>
<td>$529</td>
<td>$721</td>
<td>4658</td>
</tr>
<tr>
<td>AN uplink redundancy</td>
<td>$1,563</td>
<td>$531</td>
<td>$1,032</td>
<td>4680</td>
</tr>
<tr>
<td>Access Network node redundancy (Agg1, Agg2, Agg3)</td>
<td>$2,423</td>
<td>$1,044</td>
<td>$1,379</td>
<td>9222</td>
</tr>
<tr>
<td>Edge link redundancy</td>
<td>$2,425</td>
<td>$1,044</td>
<td>$1,381</td>
<td>9222</td>
</tr>
<tr>
<td>Edge Node redundancy</td>
<td>$2,437</td>
<td>$1,056</td>
<td>$1,381</td>
<td>9296</td>
</tr>
</tbody>
</table>

Values for AN, Agg1, Agg2, Agg3 and Edge nodes only (No Pp-routers). Cumulative redundancy Schemes, GPL.
Network High Availability
HA Network Map

Access

Access Domain

Aggregation

AN <-> AGG <-> Edge

Edge

Core

L0/1

Failure Detection

Interrupts

Loss of Signal

Interrupts

Loss of Signal

Interrupts

Recovery

Module Redundancy

Path diversity / dual homing

Module Redundancy

Path diversity / dual homing

Module Redundancy

BRKSPG-2402

© 2012 Cisco and/or its affiliates. All rights reserved.
Multi-chassis LAG

- mLACP uses ICCP to synchronise LACP configuration & operational state between PoAs, to provide DHD the perception of being connected to a single switch
- All PoAs use the same System MAC Address & System Priority when communicating with DHD
  - Configurable or automatically synchronised via ICCP
- Every PoA in the RG is configured with a unique Node ID (value 0 to 7). Node ID + 8 forms the most significant nibble of the Port Number
- For a given bundle, all links on the same PoA must have the same Port Priority
Inter-chassis Communication Protocol

- ICCP allows two or more devices to form a ‘Redundancy Group’
- ICCP provides a control channel for synchronising state between devices
- ICCP uses TCP/IP as the underlying transport
  - ICCP rides on targeted LDP session, but MPLS need not be enabled
- Various redundancy applications can use ICCP:
  - mLACP
  - Pseudowire redundancy
- Under standardisation in IETF
BFD Protocol Overview

- Accelerates convergence by running fast keepalives in a consistent, standardised mechanism across routing protocols
- Lightweight hello protocol
- Neighbours exchange hello packets at negotiated regular intervals
- Configurable transmit and receive time intervals
- Unicast packets, even on shared media
- No discovery mechanism
- BFD sessions are established by the clients e.g. OSPF, IS-IS, EIGRP, BGP, …
- Client hello packets transmitted independently
BFD Operation Modes

- Async mode (no echo), periodic control packets sent
- Neighbour declared dead if no pkt is received for <interval * multiplier> period
- Session established using async control session
- When echo is negotiated, echo packets sent at negotiated rate, used for failure detection
- Control packets sent at low rate

Async Mode
- green is alive

Echo Mode
- orange is alive
- green is alive
Ethernet OAM Overview

- **E-LMI** - Provides protocol and mechanisms used for:
  - Notification of EVC addition, deletion or status to CE
  - Communication of UNI and EVC attributes to CE
  - CE auto-configuration
  - Notification of Remote UNI name and status to CE

- **IEEE 802.3ah**
  - OAM Discovery
  - Link Monitoring
  - Fault Signalling
  - Remote MIB Variable Retrieval
  - Remote Loopback

- **IEEE 801.3ag (CFM)**
  - Family of protocols that provides capabilities to detect, verify, isolate and report end-to-end ethernet connectivity faults
  - Protocols (Continuity Check, Loopback and Linktrace) used for Fault Management activities
Ethernet OAM Overview

- Ethernet LMI: Automated configuration of CE based on EVCs and bandwidth profiles
  - L2 connectivity management
- IEEE 802.3ah: When applicable, physical connectivity management between devices.
- IEEE 802.1ag: Connectivity Fault Management (CFM)
  - Uses Domains to contain OAM flows and bound OAM responsibilities
  - Provides per EVC connectivity mgmt and fault isolation
  - Three types of packets: Continuity Check, L2 Ping, L2 Traceroute
ITU-T Y.1731 Overview

- **OAM Functions for Fault Management**
  - Ethernet Continuity Check (ETH-CC) *(Y.1731 adds unicast CCM)*
  - Ethernet Loopback (ETH-LB) *(Y.1731 adds multicast LBM)*
  - Ethernet Linktrace (ETH-LT)
  - Ethernet Remote Defect Indication (ETH-RDI)
  - Ethernet Alarm Indication Signal (ETH-AIS)
  - Ethernet Locked Signal (ETH-LCK)
  - In addition: ETH-TEST, ETH-APS, ETH-MCC, ETH-EXP, ETH-VSP

- **OAM Functions for Performance Management**
  - Frame Loss Measurement (ETH-LM)
  - Frame Delay Measurement (ETH-DM)
Virtual Circuit Connection Verification (VCCV) Overview

- checks connectivity between egress and ingress PEs
- VCCV allows sending control packets in band of pseudowires (PW)
  - Signalling component: communicate VCCV capabilities as part of VC label
  - Switching component: cause the PW payload to be treated as a control packet
- VCCV capability is negotiated when the AToM tunnel is brought up
  - depends on the LDP peer and the VC type
- marks the payload as control packet for switching purpose; packet follows the PW data path
- Control packets sent over the AToM tunnels are intercepted by the egress PE

<table>
<thead>
<tr>
<th>Type 1</th>
<th>Type 2</th>
<th>Type 3 (TTL expiry)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(in-band vccv)</td>
<td>To signal in-band VCCV [RFC4385] using PW ID from PW Control Word</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Signal out-of-band VCCV using MPLS router alert label</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Manipulate and Signal TTL exhaust (TTL == 1) for multiple switching point PEs</td>
<td></td>
</tr>
</tbody>
</table>
Network Convergence - Why It Takes So Long

- Detection of Link layer failure
- Report failure to Route Controller
- Generate and flood an LSP
- Trigger and Compute an SPF
- Communicate new FIB entries to linecards
- Install new FIB entries into linecard HW path

Bottleneck
Network Fast Convergence Solutions

Detection of Link layer failure

Report failure to Route Controller

Generate and flood an LSP

Trigger and Compute an SPF

Communicate new FIB entries to linecards

Install new FIB entries into linecard HW path
Hierarchical CEF

- Optimises the data plane for sub-second convergence

- CEF Data Structure Enhancements
  Solves the FIB Download Convergence Bottleneck
  LSP and Prefix Independent
  Optimises FIB

- Hierarchical CEF Technologies
  MPLS-FRR
  IP-FRR
  BGP PIC Core
  BGP PIC Edge

- Non-Hierarchical CEF Technologies
  MPLS Path Protection
MPLS FRR 50 ms Convergence

- **Key Features**
  - Fast Convergence for Link and Node Failures
  - Supported Across all Network Topologies
  - MPLS-TE Traffic Management
    - SRLG
    - BW Reservation
    - Per Tunnel Traffic Statistics

- **Caveats**
  - Requires MPLS and MPLS-TE
  - No Protection for Ingress or Egress Tunnel Failures
  - Requires Pre-Computed Backup Paths
  - Requires “(n-1)!” Tunnels for Full Protection

- **Applicability**
  - Protecting Expensive Backbone Links
MPLS-FRR – CEF

Typical FIB Programming Rate - ~5000 – 10,000 CEF Updates per second
IP FRR-LFA: 50ms Convergence

- **Key Features**
  - 50 msec Convergence for Link and Node Failures
  - Works for MPLS and IP Only Environments
  - Simple
    - Automatic configuration of “Loop Free Alternate Paths” via OSPF or ISIS
    - No Tunnels

- **Caveats**
  - Requires a “Loop Free Path” for Protection
  - No Bandwidth Reservation
  - No Support for SRLG
  - New Feature

- **Applicability**
  - Strong Solution for Deployments with Cost Effective Bandwidth

---

No Convergence Required on Routers R2, R3, R4 and R5 to Maintain Green Traffic Flow!
IP FRR-LFA – CEF Enhancement

Failure

Repair

Cleanup After Repair – Assuming No Available Loop Free Path

Typical FIB Programming Rate – ~5000 – 10,000 CEF Updates per second
MPLS Path Protection

- **Key Features**
  - Optimised for Ring Topologies
  - Utilises Pre-Signaled Backup Tunnel
  - MPLS-TE Traffic Management
    - SRLG
    - BW Reservation
    - Per Tunnel Traffic Statistics

- **Caveats**
  - Requires MPLS and MPLS-TE
  - No Protection for Ingress or Egress Tunnel Failures
  - Convergence Dependant on IGP Prefixes and L2VPN LSPs Under Protection

- **Applicability**
  - Protecting Ring Topologies
MPLS Path Protection – CEF

Typical FIB Programming Rate - ~5000 – 10,000 CEF Updates per second

MPLS-FRR – IP and MPLS

One CEF Update Message per IGP Prefix and L2VPN LSP!

For Your Reference
BGP PIC Edge

- Optimises BGP Convergence for BGP Next-Hop Change
  - PE to CE Link Failures
  - PE Node Failures
  - CE Node Failures

- Applicability
  - PE Routers
  - Requires “bgp advertise-best-external” to enable
BGP PIC Edge – CEF

**Typical FIB Programming Rate** - ~5000 – 10,000 CEF Updates per second

**Failure**
- BGP Prefix FIB Entry
- BGP Prefix FIB Entry
- BGP Prefix FIB Entry

**Pre-Computed Backup Path**
- Load Balance OCE
- MPLS Label OCE

**Repair**
- BGP Prefix FIB Entry
- BGP Prefix FIB Entry
- BGP Prefix FIB Entry

**One CEF Update Message for Multiple Prefixes**
- Load Balance OCE
- MPLS Label OCE

**Cleanup After Repair**
- BGP Prefix FIB Entry
- BGP Prefix FIB Entry
- BGP Prefix FIB Entry

**PIC Edge – IP to MPLS**
- MPLS Label OCE
- Adjacency OCE - Interface
- MPLS Label OCE
- Adjacency OCE - Interface
- MPLS Label OCE
- Adjacency OCE - Interface
BGP PIC core – when IGP path to BGP Next-Hop changes
1. Examples: PE-P or P-P link failure, P node failure
   Sub-second convergence (prefix independent)
   vs. multiple seconds convergence (prefix and hardware dependent)
   Enabled by default since IOS XE 2.5.0 (cef table output-chain build favor convergence-speed)

BGP PIC edge – When BGP Next-hop changes
2. when remote PE node fails or no longer reachable.
3. when PE-CE link fails.
   Immediate to sub-second convergence (prefix independent)
   vs. multiple seconds convergence (prefix and hardware dependent)
BGP PIC Edge PE-CE Link Protection - Summary

- PE1 and PE2 precomputes bgp backup paths using bgp best-external approach

- When primary link PE1 - CE1 fails:
  
  PE1 holds on to the bgp local labels and re-routes CE1’s traffic to PE2 using labels advertised by PE2
  
  PE1 uses fixed timer to clean up stale local labels
  
  PE3 is expected to converge to start using PE2 as the BGP nexthop and IGP label for PE2 to send traffic from CE2 to CE1
BGP PIC Edge PE-CE Link Protection

The best BGP path to CE1 is through PE1

Traffic flow due to BGP best path
BGP PIC Edge PE-CE Link Protection

The best BGP path to CE1 is through PE1

Traffic flow due to BGP best path

The BGP pre-calculated Backup Path
BGP PIC Edge PE-CE Link Protection

- Traffic flow due to BGP best path
- The best BGP path to CE1 is through PE1
- The best BGP path to CE1 is through PE1
- The BGP pre-calculated Backup Path
- Announce BGP Route change
- Detects that link is down and CEF layer will switch to pre-computed backup path
- PE-CE Link Failure
BGP PIC Edge PE-CE Link Protection

The best BGP path to CE1 is now through PE2

Traffic flow due to PE-CE Link Failure

Announce BGP Route change

Announce BGP Route Change
PE1, PE2 and PE3 precomputes bgp backup

When node PE1 fails:
IGP notification on PE3 invalidates active path
PE3 switches to backup path
PE3 is expected to converge to start using PE2 as the BGP nexthop and IGP label for PE2 to send traffic from CE2 to CE1
BGP PIC Edge PE-Node Protection

The best BGP path to CE1 is through PE1

Traffic flow due to BGP best path
BGP PIC Edge PE-Node Protection

The best BGP path to CE1 is through PE1

Traffic flow due to BGP best path

The BGP pre-calculated Backup Path
Detects that PE1 is down and CEF layer will switch to pre-computed backup path.

IGP Signals that the Router is dead

PE-CE Node Failure

The BGP pre-calculated Backup Path
At next BGP next-hop scan the path through PE2 will become the best Path.
# IP/MPLS High-Availability Options: Scorecard

## Network Infrastructure - all transit links and nodes

<table>
<thead>
<tr>
<th></th>
<th>Fault Coverage</th>
<th>Recovery Time</th>
<th>Operational Simplicity</th>
</tr>
</thead>
<tbody>
<tr>
<td>IGP Fast Convergence (IGP FC)</td>
<td></td>
<td>O(x00ms)</td>
<td></td>
</tr>
<tr>
<td>Broke the barrier of &lt;200msec restoration time</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Covers all faults, including multiple failures</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IP/MPLS Fast ReRoute Loop Free Alternate (IPFRR LFA)</td>
<td></td>
<td>&lt;50ms</td>
<td></td>
</tr>
<tr>
<td>Provides local protection (link, node) with &lt;50msec recovery</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tool to improve on IGP FC for most topologies (triangle, square, mesh)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MPLS TE Fast ReRoute (TE FRR)</td>
<td></td>
<td>&lt;50ms</td>
<td></td>
</tr>
<tr>
<td>Provides local protection (link, node, path) with &lt;50msec recovery</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## Service Edge - edge node and access links

<table>
<thead>
<tr>
<th></th>
<th>Fault Coverage</th>
<th>Recovery Time</th>
<th>Operational Simplicity</th>
</tr>
</thead>
<tbody>
<tr>
<td>BGP Prefix Independent Convergence (BGP PIC)</td>
<td></td>
<td>O(x00ms)</td>
<td></td>
</tr>
<tr>
<td>IP/IPVPN scale independent recovery in line with IGP FC and FRR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Applicable to all BGP based services (IPv4, IPv6, VPNv4, VPNv6)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Feasible to deliver very tight E2E Service Availability SLAs without increasing operational complexity
System High Availability
System High-Availability with Hardware Redundancy

- Redundant hardware components
  - Power Supplies
  - Route Processors
  - Forwarding Processors
  - Switching Matrix
  - SPA Interface Cards

- Interface Redundancy typically achieved using IEEE 802.3ad / LACP or APS
- Hardware Redundancy needs to be complemented by Software redundancy Features
- Cisco Platforms supporting hardware redundancy

CRS-1  ASR 9000  ASR 5000  ASR 1000  Cisco 12000  Cisco 7600
Distributed Forwarding Plane for Performance

- Up to Eight Linecards (Autonomous Forwarding)

Distributed IOS® XR based Control Plane for Scale

- Dual Route Switch Processors (RSPs)
- Dual-Core CPU on Each Linecard

Active/Active Switch Fabric for HA

- Non-blocking Memory-less Fabric
- Service Intelligence with Hi / Lo Priorities, Unicast & Multicast Recognition, and VoQ’s
- Redundant EOBC, Fan Trays, Power supplies
Example: ASR 1000 System Redundancy

**Data Plane**

- **Embedded Services Processor (active)**
  - FECP
  - QFP subsystem
  - Crypto assist

- **Route Processor (active)**
  - RP
  - Interconn.

- **Route Processor (standby)**
  - FECP
  - QFP subsystem
  - Crypto assist

- **IoP**
  - SPA Agg.

- **Passive Midplane**

**Control Plane**

- **Embedded Services Processor (active)**
  - FECP

- **Route Processor (active)**
  - RP

- **Route Processor (Standby)**
  - FECP

- **IoP**
  - SPA Agg.

- **Passive Midplane**

**Network Connectivity**

- ESI, (Enhanced Serdes) 10/40Gbps
- SPA-SPI, 11.2Gbps
- Hypertransport, 10Gbps
- GE, 1Gbps
- SPA Control
- SPA Bus
ASR 1000 Software Architecture

- Runs Control Plane
- Generates configurations
- Populates and maintains routing tables (RIB, FIB…)

- Provides abstraction layer between hardware and IOS (manages ESP redundancy)
- Maintains copy of FIB and interface list
- Communicates FIB status to active & standby ESP (or bulk-download state info in case of restart)

- Communicates with Forwarding manager on RP
- Provides interface to QFP Client / Driver

- Maintains copy of FIBs
- Programs QFP forwarding plane and QFP DRAM
- Statistics collection and communication to RP

- Implements forwarding plane
- Programs PPEs with forwarding information
ASR 1006 High Availability Infrastructure

- HA operates in a similar manner to other protocols on the ASR 1000
- Reliable IPC transport used for synchronisation
Which Events Trigger Failovers?

- The following events may trigger failovers on the RP/ESP
  1. Hardware component failures
  2. Software component failures
  3. Online Insertion and Removal (OIR)
  4. CLI-initiated failover (e.g. reload command, force-switchover command)
1. Failover Triggers: Hardware Failures

What hardware failures?

a. CPUs: RP-CPU, QFP, FECP, IOCP, interconnect CPU, I²C Mux, ESP Crypto Chip,

b. Memory: NVRAM, TCAM, Bootflash, RP SDRAM, FECP SDRAM, resource DRAM, Packet buffer DRAM, particle length DRAM, IOCP SDRAM, …

c. Interconnects: ESI Links, I²C links, EOBC Links, SPA-SPI bus, local RP bus, local FP bus …

d. Other: heat-sinks, …

Detected using

- CPLD interrupts / register bits within O(ms) controlled by CM_RP
- Watchdog timers: low level watchdogs running in O(min) that can initiate a reset (e.g. RP)
- JTAG: RP can program CPLD on other modules. Test interconnects and other boards (primarily for RMAd hardware)

Interrupts generated by hardware failures initiate fail-over events

Hardware failures are typically fatal such that modules need to be replaced!
2. Failover Triggers: Software Failures

What Software Failures?

a. Kernel: Linux on RP / ESP / SIP
b. Middleware: Chassis Manager (CM), Forwarding Manager (FM)

c. IOS

d. SPA drivers

Detected using

Kernel: the kernel supervises middleware or SPA driver processes (kmonitor()). It always knows if a process is healthy

IPC: between 2 IOS (and only for IOS)

Kernel will take the module down in a controlled manner

IOS, CM_{ESP}, CM_{SIP}, FM_{ESP}, QFP Driver/Client, IM_{SIP} are not re-startable!

Also setting register bits to initiate fail-over for ESP or RP

Note: some other processes are re-startable (CM_{RP}, FM_{RP}, SSH, Telnet…)

Kernel will try to re-start the processes in this case
RP Failover Procedure

Act

Failure

Detect RP

Restart

Establish ESI links

State information

New RP act information

Close ESI links (ESP)

RP (slot 1)

RP (slot 0)

FM (Esp)

CM (Esp)

Kernel

IOS

If not received in time, send restart message.

Update HW component file system

.state information

Failure detect RP

Restart

Close ESI links (ESP)

Update HW component

Restore ESI link

For Your Reference
RP\text{act} Failover Procedure (cont.)

- **Failover Procedure (cont.)**

  - **Check updated state and discard old state**
  - **Check updated state and discard old state**
  - **Forwarding State information**
  - **Service recovered**

  - **Start kernel**
  - **Start CM**
  - **Start IOS**
  - **H/W initialisation**
  - **Initialise EOBC**
  - **Start FM**
  - **Other RP information**
  - **Run Mastership**
  - **Detect RP\text{sby}**
  - **Forwarding State information**
  - **ESI link status**
  - **SBY**

- **Take-over control using checkpointed state**

- **RP (slot 0)**
  - CM\_RP
  - FM\_RP
  - Kernel
  - IOS

- **RP (slot 1)**
  - CM\_RP
  - FM\_RP
  - Kernel
  - IOS

- **Failover**
**ESP_{act} Failover Procedure**

Service Recovered with momentary packet loss

- **Detect ESP_{act} failure**
  - State information of failed ESP
  - Reconfigure ESI link w/ RPs
  - Change state of ESI link w/ new ESP_{act}
  - Disable ESI link w/ failed ESP

- **Interrupt**

- **Resend state information**

- **ESI link status**

- **For Your Reference**
ESP_{act} Failover Procedure (cont.)

- Restart
- H/W initialisation
- Initialise EOBC
- Wait for RP_{act}
- Detect RP_{act}
- Activate ESI-link
- Start kernel
- Start FM
- Start CM
- SBY

Download software packages

Register with CM_{RP}

Other-ESP information (e.g., mastership)

RP_{act} information

For Your Reference
## Cisco Software High-Availability Support

- **Stateful Switchover (SSO)** support for features provides the synchronisation of dynamic feature state between hardware modules.
- Configuration synchronisation ensures that the running config is synchronised on the route processors.

<table>
<thead>
<tr>
<th>Dynamic State Preservation</th>
<th>ASR 1000</th>
<th>ASR 9000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Connectivity Protocols</td>
<td>FR, PPP, MLPPP, HDLC, 802.1Q, BFD (BGP, IS-IS, OSPF)</td>
<td>BFD (OSPF, BGP, IS-IS, Static)</td>
</tr>
<tr>
<td>Routing &amp; IP Services</td>
<td>RP, HSRP, IPv6 NDP, uRPF, SNMP, GLBP, VRRP, NSR (MP-iBGP, eBGP), ISSU, GRE</td>
<td>NSF (ISIS, OSPF, BGP), NSR (ISIS, OSPFv2, OSPFv3, BGP)</td>
</tr>
<tr>
<td>Multicast</td>
<td>IPv4 Multicast (IGMP), IPv6 Multicast (MLD, PIM-SSM, MLD Access group)</td>
<td>NSF Multicast, BFD for PIM</td>
</tr>
<tr>
<td>MPLS Protocols</td>
<td>MPLS L3VPN, MPLS LDP, VRF-aware BFD,</td>
<td>NSF (LDP, T-LDP, RSVP-TE)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NSR (LDP), BFD for MPLS FRR, VRF-aware BFD</td>
</tr>
<tr>
<td>Broadband</td>
<td>PPPoE, L2TP (LAC, LNS), DHCPv4/v6, AAA, session state (virtual templates), ISG, ANCP, LI</td>
<td>Roadmap (including nV)</td>
</tr>
<tr>
<td>Security</td>
<td>SSO, Stateful Inter-chassis redundancy for FW / NAT</td>
<td>Roadmap</td>
</tr>
<tr>
<td>SBC</td>
<td>SSO</td>
<td>Roadmap</td>
</tr>
</tbody>
</table>
Non Stop Forwarding (NSF)

- Routers to maintain forwarding state when communication between them is lost.
- Routing sessions are established with NSF aware peers. Upon HA event, neighbouring peers maintain forwarding until routing sessions are reestablished.
- Copy of FIB maintained on secondary and used on failure for continuously traffic flow.
- Requires neighbouring routers to be NSF aware.

Traffic is forwarded continuously.
Non Stop Routing (NSR)

- Routers to maintain routing state and forwarding state when communication between them is lost.
- Routing sessions are maintained between processors on a failure, allowing routing sessions to stay up with Peer.
- Copy of FIB maintained on secondary and used on failure for continuously traffic flow.
- No need for neighbouring routers to be NSF aware or capable. Can give high reliability without upgrading CE.
BGP NSR

- Implemented by hardening code for
  - BGP RIB checkpointing
  - BGP TCP interaction

- Only supported for “ipv4 vrf <VRF name>” address family for PE-CE eBGP sessions at this time
  - `router bgp <asn>`
  - `address-family ipv4 vrf RED`
  - `neighbor x.x.x.x ha-mode sso`

  Roadmap: support for “ipv4, ipv6, ipv6 vrf, vpnv4, vpnv6” address families for RR, internet peering, MP-BGP between PEs and RR, 6vPE applications

- No peer session flaps on VPNv4 CE when RP switches over
- Route updates during RP switchover are announced to VPNv4 CE peers
  - NO delays which
  - Prevents data black-holes during RP switchover as in the case of graceful restart peers
OSPFv2 NSR

- Provides the ability to perform hitless RP switchovers when OSPF is used as the routing protocol (Expect zero-traffic loss across such HA events)

- To enable OSPFv2 Non-Stop Routing (NSR)
  ```
  router ospf <process id> [vrf <vrf name>]
  nsr
  ```

- Activated on a per-process basis (for both ipv4 or ipv4 VRF for PE-CE sessions)

- Depends on the forwarding plane’s ability to retain state across control plane restarts and RP switchovers

- Alleviates dependency on OSPFv2 protocol extensions (NSF)
  - Neighbouring routers are unaware that a router is NSR-capable
  - Neighbouring routers are unaware that a router has gone through an RP switchover

- Provides near-transparent RP switchover capability
  - OSPF adjacencies remain up
  - Minimal state refreshed by the restarting router post switchover
  - Scalable to larger link state database sizes and number of neighbours

- NSR deployment does not need other routers to be NSR-capable
Service High-Availability
High Availability for Advanced Service Models

- Many SP Services already go beyond standard L3VPN / L2VPN / transport services
- Increasing subscriber management capabilities and L4-L7 services
- Examples:
  - Subscriber Management
  - Multicast
  - Session Border Controller
  - Firewall
  - IPSec
  - LI
- Some Services can be made highly-available using Intra-chassis redundancy (e.g. IPSec, Firewall, NAT, PPPoX, L2TP)
- Stateless inter-chassis redundancy available for BNG
- Stateful Inter-chassis redundancy available for NAT, Firewall and SBC on the Cisco ASR 1000
L3VPN Key HA Technologies

- Physical
  Circuit Diversity
  Multihoming

- Link Detection
  IP Event Dampening
  BFD

- Chassis Redundancy
  NSF/NSR

- Routing Protocols and Convergence
  BGP PIC Core
  BGP PIC Edge
  IP-FRR
  MPLS Path Protection
L3VPN Key HA Technologies

- **CPE**
  - BFD for PE-CE Link Detection
  - NSF/NSR for Chassis HA
  - PE Multihoming
    - Intra-Site PE for PE Diversity
    - Inter-Site for SP Facility Diversity

- **Access**
  - Circuit Diversity - Physical Diversity for Multihomed CPE
    - Physical Circuit Diversity is Not the Default
    - Must be Requested from the SP

- **Edge**
  - BFD for PE-CPE / PE-P Link Detection
  - NSF/NSR for Chassis HA
  - IP Event Dampening for PE-CPE
  - IP-FRR for PE-P
    - For Cost Effective PE-P Bandwidth
  - BGP PIC Core
  - BGP PIC Edge for Multi-Homed CPE
L2VPN – Pseudowire Redundancy

- **Active-Standby PW Access Circuit Redundancy**
  L2TPv3 and MPLS Support

- **Detection Mechanisms**
  - IGP Convergence for Remote PE Failure
  - LDP Signalling for PE-CE Failure
  - LDP Timeout for Remote PE Software Failure

![Diagram of L2VPN - Pseudowire Redundancy](image)
MPLS Pseudowire Status Signalling Procedure

- PW Status Signalling method selected if supported by both peers.
- PEs exchange label mapping messages upon PW configuration.
- Simple Label Withdraw status method will be used if one of the peers doesn’t support PW Status Signalling.
- PW label won’t be withdrawn unless AC is administratively down or the PW configuration is deleted.
- PW state set to “down” if the Label mapping is not available.
- Capability is on by default.
Multicast High-Availability Behaviour

- Before failure
  Multicast state is synchronised from $RP_{\text{act}}$ to $RP_{\text{sby}}$
    - Configuration
    - MLDv1/v2 state information
    - PIM or MRIB state are NOT synchronised
  MFIB also synched to $ESP_{\text{act}}$ and $ESP_{\text{sby}}$

- After failure
  - $RP_{\text{sby}}$ sends out PIM hellos to all neighbours
  - PIM neighbours re-send PIM state
  - Newly active RP re-builds the PIM state
  - IGP reconverges to assure uRPF check
  - MFIB and ESP updates proceed to incorporate refreshed PIM state

- $ESP_{\text{act}}$ continues to forward multicast traffic based on its version of the MFIB
  - Forwarding of multicast packets is NOT disrupted
Multicast only Fast Re-Route (MoFRR)

- **Receiver**
  - Multicast join on primary path
  - Multicast join on backup path
- **Data packets** are received from the primary and secondary paths.
- The redundant packets are discarded at topology merge points due to RPF checks.
- **Failure:**
  - Interface chance on where packets are accepted
  - Backup path interfaces become ‘active’

**Configuration and Restrictions**
- Dependency on ECMP and will not work without it
- Disabled by default and enabled through a cli
- Applicable to IPv4 multicast only and not IPv6 multicast
- Works only for SM S,G and SSM routes
- Works where the rpf lookups are done in a single vrf
- Extranet routes are not supported
- Both primary and secondary paths should exist in the same multicast topology.
Stateful Application Switchover: PPP

- Copies state information for PPP, PPPoE, and PPPoEoVLAN Sessions
- Switch-over is transparent to peers
  - Sessions are not torn-down / re-established.
- PPP, PPPoE, and PPPoEoVLAN Session States:
  - Configuration (through config synch), including QoS configuration, ACLs
  - Session identifiers
  - PADDR frame (cached)
  - RADIUS session attributes
  - Physical interface
  - VAI identifier
  - MD5 signature
- Statistics can be synchronised on ASR 1000!
Stateful Application Switchover: L2TP

- \( \text{RP}_{\text{act}} \) synchronises state with \( \text{Rp}_{\text{sby}} \)
  
  State includes configuration, PPP session IDs, L2TP CC sequence numbers etc.

  Sequence numbers (Ns, Nr) for L2TP Control Connections (CC) are only synched once for a packet window of X (i.e. once every X L2TP control packets)
BNG Service Edge High Availability

- PPP Smart Server Selection allows user to configure specific PADO delay for a received PADI packet
  - Can be configured per bba-group or based on circuit-id/remote-id

- In case of an outage of a BNG in the cluster, other BNG stand ready to accept subscriber sessions
  - Detection of failure possible at both ends of PPPoE session because of missing keepalives
  - Subscriber sessions have to be re-established

- Allows BNG redundancy with predictable behaviour
Stateful Inter-chassis Redundancy
Motivation for Stateful Application Inter-chassis Redundancy

- Current Intra-chassis HA typically protects against
  - Control Plane (RP) Failures
  - Forwarding Plane (ESP) failures
  - Interface failures can be mitigated using link bundling (e.g. GEC)

- Any other failures may result in recovery times O(hours)

- Inter-chassis redundancy provides additional resilience against
  - Interface Failures
  - System failures
  - Site failures (allowing for geographic redundancy)
Stateful System Redundancy Models

- Different deployment models
  - 1+1 – one system is actively processing and passing traffic, the other in standby mode.
  - 1:1 – two systems are actively processing and passing traffic, and backing each other up.
  - N+1 – N systems are actively processing and passing traffic, and share a single standby.

- System vs. Application
  - Is the inter-chassis resilience applicable to ALL of the features / functions configured on the system, or only for a particular application?
    - System-level: provide resilience for ALL applications and traffic configured on a system.
    - Application-Level: provide resilience for a particular application and its traffic.

- Hot-standby vs. Cold-standby
  - Cold-standby: FIB / adjacency updates are NOT synchronised between active and standby system.
  - Hot-standby: forwarding/state information is synchronised between active and standby system.

- Different Approaches can also be categorised into
  - Control plane active-standby / active-active
  - Forwarding plane active-standby / active-active
System Level Redundancy

- Example: VSS / nV
- Failover Granularity at the System Level
- Control-plane active-standby
  Active RP considers ‘remote’ linecards under its control
- Forwarding-plane active-active
- No application granularity for failover
  Need to ensure all features are SSO capable

Application Level Redundancy

- Example: RG Infra
- Failover Granularity at the Application Level (NAT, Firewall, SBC etc)
- Control plane active-active
  Each RP only considers its own linecards, but synchronises application state
- Forwarding-plane active-active
- E.g. can have one set of firewall services resilient, and other set of firewall services non-resilient

RP
Fabric LC
RP
Fabric LC
Failover
RP
act
RP
sby
RP
act
RP
sby
Failover
ESP SIP
ESP SIP
Failover
FW
ESP
FW
SIP SIP
nV Edge Overview

- 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.

For redundancy purpose, minimal two control plane and two data plane links are required.
Control Plane HA Model

- Only one Active RSP, Only one standby RSP at a given time, which are located on two different chassis
  - SSO/NSF/NSR works exactly the same way as two RSPs on the same chassis
  - Reliable out of band control channel between two chassis
  - IOS-XR control plan can tolerate hundreds of msec latency*, although the latency can impact overall service convergence time
- Virtual Chassis is always on as long as there is one chassis and one RSP alive

* Practically, recommend maximum 10msec latency between two chassis
Data Plane Forwarding Model

- Inter-chassis data links simulate the switch fabric, which provide the data connection between two chassis. It has similar features as switch fabric, for example, fabric qos. Packet load balancing over inter-chassis links is same as regular link bundle: per-flow based.

- Keep the existing IOS-XR two-stage forwarding model → no forwarding architecture change for single chassis vs. nV Edge system.

- In case of ECMP or link bundle paths cross two chassis, it prefers the local port instead of load balancing packet to the other chassis. This is to reduce the inter-chassis link usage as much as possible. However, this feature (local rack preference) could be turned off by user CLI.

- Only single Multicast copy is sent over inter-chassis link. Multicast replication is done on egress line cards and fabric on the local chassis.
Inter-Chassis Control Plane and Data Plane Packet Format

- Inter-chassis control plane link
  Ethernet snap with special ethertype and internal mac addresses
  Work over L2 circuit as well assuming it’s port mode: transparently forward every packet
  Recommend L1 link, with upto 10msec latency

- Inter-chassis data plane link
  Regular Ethernet frame, with 802.1q tag (VLAN=1)
  In theory, it can work over L2 circuit, but it’s never tested and won’t be supported officially
Data Forwarding

Ingress LC
Data Plane
Lookup
Ingress Forwarding Lookup ➔ L2/L3/Mcast regular lookup

Inter-Chassis Load Balance ➔ Load balance across multiple inter-chassis links

Inter-Chassis LC
Data Plane
Encapsulation

Chassis 0

Inter-Chassis Link bundle

Chassis 1

Egress LC
Data Plane
LOOKUP
Egress Forwarding Lookup ➔ L2/L3/Mcast regular lookup

Inter-Chassis LC
Data Plane
Decapsulation

Inter-Chassis Encapsulation

Inter-Chassis Decapsulation

1 2 3 4 5

1 2 3 4 5
Introduction to RG-Infra

- RG Infra is the IOS Redundancy Group Infrastructure to enable the synchronisation of application state data between different physical systems
  - Does the job of RF/CF between chassis

- Infrastructure provides the functions to
  - Pair two instances of RG configured on different chassis for application redundancy purposes
  - Determine active/standby state of each RG instance
  - Exchange application state data (e.g. for NAT/Firewall)
  - Detect failures in the local system
  - Initiate & manage failover (based on RG priorities, allows for pre-emption)

- Assumptions
  - Application state has to be supported by RG infra (ASR 1000 currently supports NAT, Firewall, SBC)
  - Connectivity redundancy solved at the architectural level (need to ‘externalise’ the redundant ESI links of the intra-chassis redundancy solution)
Redundancy Groups Functions

- Registers applications as clients
- Registers (sub)interfaces / {SA/DA}-tuplets in case of firewall
- Determines if traffic needs to be processed or not
  E.g. for Firewall: if a subset of sessions are associated with a RG in active state, then the Firewall application will perform normal processing for those sessions and actively sync the session state to another device that has the same RG in STANDBY state.
  For Firewall sessions that are associated with a RG in STANDBY state, the session information will be synchronised from a device that has the RG in ACTIVE state.
- Communicates control information between RGs using a redundancy group protocol
  Advertisement of RGs and RG state
  Determination of peer IP address
  Determination of presence of active RG
- Synchronises application state data using a transport protocol
- Manages Failovers!
Redundancy Groups Functions - Details

- Configuration of stateful system redundancy
  - Priority (similar to HSRP priority for RG state determination)
  - Preemption, Name

- RG State control
  - Init, Active, Standby, disabled
  - Communicating state changes to other software entities in the system (e.g. QFP software)

- Synchronisation management
  - Synchronisation state tracking (standby has to request bulk-updates from active)
  - Determines when synchronisation is started (e.g. ensures transport is available)

- Peer Management
  - Maintain information about peers

- Fault Handling
  - Changing priorities of RG (may affect RG state)
  - Fault event dampening
  - Logging
  - Integration with Enhanced Object tracking / BFD

- Transport Connectivity
  - Knows via which interface application state is synchronised
  - Can be different for application state data and RG control messages
# Intra-chassis vs. Inter-Chassis Redundancy

<table>
<thead>
<tr>
<th>Function / Method</th>
<th>Stateful Intra-chassis</th>
<th>Stateful Inter-chassis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hardware redundancy</td>
<td>ESP, RP</td>
<td>ESP, RP, Interfaces</td>
</tr>
<tr>
<td>Redundant connectivity</td>
<td>Internal ESI links</td>
<td>Redundant links to neighbour nodes</td>
</tr>
<tr>
<td>Redundancy control</td>
<td>RF/CF</td>
<td>RG</td>
</tr>
<tr>
<td>State synchronisation</td>
<td>IPC over EOBC</td>
<td>External GEC</td>
</tr>
<tr>
<td>Failure detection mechanism</td>
<td>Interrupts</td>
<td>BFD, Hellos</td>
</tr>
<tr>
<td>Failover mechanism</td>
<td>Chassis Manager</td>
<td>RG Protocol (HSRP like)</td>
</tr>
</tbody>
</table>
Possible RG-Infra Redundancy Models

Active-Standby
- All application traffic associated with a SINGLE RG instance
- Failures would switch all traffic over to the standby chassis

Active-Active
- Multiple RG instances configured per system
- Subset of traffic associated with a particular RG instance
- Single failure only affects subset of overall application traffic

2+1 Active-Standby
- 2 or more chassis loadshare application traffic, backed up by a single standby system
- Subset of traffic associated with a particular RG instance on different chassis
- Single failure only affects subset of overall application traffic
Case Studies
Case Study: Highly Available IP Architecture for Mobile

One Second Convergence Requirement

EvDO/LTE VRF
1xRTT VRF
EvDO/LTE VRF
1xRTT VRF

CSR
Agg1
Agg2

MPLS VPN
S-VLAN 1
S-VLAN 2

MTSO

LTE Core
MME
SGW
MSC
RNC
CDMA Core

Internet Core

EoMPLS Backhaul

Local VLANs or T1s
L2 Domain
L3 Domain

Transport VLANs / Static Routes / BGP PIC Edge / BFD protection

OSPF/RIP/VRRP

Service Termination

FE
T1
QFP
GE
10 GE

MPLS VPN

Local VLANs or T1s

L2 Domain

L3 Domain

Transport VLANs / Static Routes / BGP PIC Edge / BFD protection

OSPF/RIP/VRRP

Service Termination

FE
T1
QFP
GE
10 GE
Case Study: Highly Available IP Architecture for Mobile - Transport

- Static routes for cellsite reachability
- BGP PIC Edge for Layer-3 convergence
- VRRP for MTSO
Steady state:
- CSR distributes flows across both Agg’s using ECMP.
- Traffic could flow across Agg inter switch links.
- Each Agg handles traffic related to all services from the cell-site.

Case Study: Highly Available IP Architecture for Mobile – Steady-state Traffic Flows
Case Study: Highly Available IP Architecture for Mobile – Link Failure

- **Steady state:** Traffic flows distributed across both Agg.
- **Failure:** GE link from MSPP to Agg1 fails.
- **Action:**
  - BFD session to Agg1 times out at CSR.
  - Agg1 next hop removed from forwarding table.
  - Traffic flows resume across existing path to Agg2.
- **Results:** Traffic flows to Agg1 via Agg2.

No changes to LAN side connectivity
Case Study: Highly Available IP Architecture for Mobile – Aggregation Switch Failure

- **Steady state:** Traffic flows distributed across Agg.
- **Failure:** Agg1 power outage.
- **Action:**
  - BFD and VRRP sessions time out
  - BGP and OSPF neighbours drop due to BFD
  - BGP PIC Edge ensures sub-second convergence
  - Traffic flows resume across existing path thru Agg2.
- **Results:** Traffic flows via Agg2 to end hosts.
Case Study: Highly Available IP Architecture for Mobile – CSR Failure

- **Steady state**: Traffic flows distributed across CSR.
- **Failure**: CSR power outage.
- **Action**: BFD sessions time out, BGP neighbours drop due to BFD, Mobile handsets resync to neighbouring cell site
- **Results**: Mobile handset voice connectivity is maintained.
Summary
Summary

- Motivation for High Availability in SP Aggregation Networks
- Network Level High Availability
- System High Availability
- Service High Availability
- Stateful Inter-chassis Redundancy
- Case Studies
- Summary & Conclusions
Key Takeaways

- High-Availability becoming increasingly deployed in Aggregation Networks
  Motivated by experiences with MPLS Core Networks

- Many high-availability techniques deployed in the core are now applied in MPLS aggregation networks
  MPLS TE FRR, BFD, EOAM, Pseudowire Redundancy …

- Service High Availability requires comprehensive approach including the deployment of
  Network level resiliency
  System Level resiliency
  L4-7 service resiliency

- Stateful Inter-chassis redundancy increasingly being considered to provide geographic redundancy for applications.

- High Availability comes at a cost (CAPEX & OPEX)!
Q & A
Complete Your Online Session Evaluation

Complete your session evaluation:

- Directly from your mobile device by visiting www.ciscoliveaustralia.com/mobile and login by entering your username and password
- Visit one of the Cisco Live internet stations located throughout the venue
- Open a browser on your own computer to access the Cisco Live onsite portal

Don’t forget to activate your Cisco Live Virtual account for access to all session materials, communities, and on-demand and live activities throughout the year. Activate your account at any internet station or visit www.ciscolivevirtual.com.
Recommended Reading

- N. Stringfield et. Al, “Cisco Express Forwarding”,
- D. C. Lee, “Enhanced IP Services for Cisco Networks”,
- A. Sayeed, M. Morrow, “MPLS and Next-Generation Networks”,
- V. Bollapragada et. Al, “Inside Cisco IOS Software Architecture “,
- R. Wood, “Next-generation Network Services”,
- K. Lee, F. Lim, B. Ong, “Building Resilient IP Networks”,
- T. Szigeti, C. Hattingh, “End-to-End QoS Network Design: Quality of Service in LANs, WANs, and VPNs;,
Whitepapers on CCO

- Cisco IOS High Availability

- Campus Network for High Availability Design Guide

- Cisco Validated Designs

- ASR 1000
  Cisco ASR 1000 Series Aggregation Services Routers
  Cisco ASR 1000 Series: ISSU Deployment Guide and Case Study
  Cisco Unified Border Element (SP Edition) on Cisco ASR 1000 Series
  Cisco Unified WAN Services: Services, Security, Resiliency, and Intelligence