High Availability for IPTV

Søren Andreasen
CCIE #3252
Cisco DK
Agenda

• IPTV Architecture overview
• High Availability (HA) in the core and edge
  – Router Level HA with SSO
  – Protocol HA with NSF/GR (IGP, BGP, Multicast)
  – MPLS HA (LDP NSF/GR, BGP + label NSF, FRR, PW Redundancy)
    – Fast Convergence (Multicast FC, Fast channel change, LDP-IGP sync, LDP session protection)
• Video Source Redundancy
• Conclusion
Tripleplay world

Policy and Service Management Layer

Portals
Broadband Policy Manager
Billing
Identity
Subscriber Database
Policy Definition

Policy and Service Management Layer

IP / MPLS / Ethernet Aggregation

Distributed: L2 PW, VPLS L3 VPN
Centralized: H-VPLS L3 VPN

Service Control Engine

IP / MPLS Core

Access

Residential

Ethernet

ETTx
Cable
DSL
PON

Local CO PE-R / PE-S

Metro CO PE-R

Service/Application Layer

iFrame Cache
Hosted Business Apps (Storage, Centrex, Security, Gaming)
VoD
VoIP
Video Broadcast
Agenda

• IPTV Architecture overview

• High Availability (HA) in the core and edge
  – Router Level HA with SSO
  – Protocol HA with NSF/GR (IGP, BGP, Multicast)
  – MPLS HA (LDP NSF/GR, BGP + label NSF, FRR, PW Redundancy)
  – Fast Convergence (Multicast FC, Fast channel change, LDP-IGP sync, LDP session protection)

• Video Source Redundancy

• Conclusion
High Availability

- End-to-end protection against network failures
- Preserve end-to-end network service connectivity
- Isolate control plane failures from forwarding plane
- Separation of control and forwarding plane should allow forwarding to continue while control plane recovers (NSF)
HA Mechanisms

• Component and Device Level Resiliency
  – Hardware and software component resiliency
    • Distributed line cards, route processors, Modular operating software
  – Stateful Switch-Over between RPs (SSO)
  – Control/forwarding plane decoupling; Non-Stop Forwarding (NSF)
  – Graceful Restart of protocols (GR)

• Network Level Resiliency
  – Optimized convergence algorithms, speeding up network recovery
Device Level HA: NSF With SSO

- **Stateful Switch-Over (SSO):** zero interruption to protocol sessions
  - Active RP synchronizes information with standby RP
  - Session state maintained for high availability-aware protocols on standby RP
  - Standby RP takes control when active RP is compromised

- **Non-Stop Forwarding (NSF):** minimal or no packet loss
  - Packet forwarding continues during reestablishment of peering relationships
  - No route flaps between participating neighbor routers
SSO in action

```
HA-Router(config)# redundancy

HA-Router(config-red)# mode ?
  rpr       Route Processor Redundancy
  rpr-plus  Route Processor Redundancy Plus
  sso       Stateful Switchover (Hot Standby)

HA-Router(config-red)# mode sso

HA-Router# show redundancy
Active GRP in slot 0:
Standby GRP in slot 9:
Preferred GRP: none
Operating Redundancy Mode: SSO
Auto synch: startup-config running-config
switchover timer 3 seconds [default]
```
Agenda

• IPTV Architecture overview
• High Availability (HA) in the core and edge
  – Router Level HA with SSO
  – Protocol HA with NSF/GR (IGP, BGP, Multicast)
  – MPLS HA (LDP NSF/GR, BGP + label NSF, FRR, PW Redundancy)
  – Fast Convergence (Multicast FC, Fast channel change, LDP-IGP sync, LDP session protection)
• Video Source Redundancy
• Conclusion
GR/NSF Fundamentals

- NonStop Forwarding (NSF) is a way to continue forwarding packets while the control plane is recovering from a failure.
- Graceful Restart (GR) is a way to rebuild forwarding information in routing protocols when the control plane has recovered from a failure.
• Router A loses its control plane for some period of time.

• It will take some time for Router B to recognize this failure, and react to it.
GR/NSF Fundamentals (uden GR/NSF)

- During the time that A has failed, and B has not detected the failure, B will continue forwarding traffic through A.
- Once the control plane resets, the data plane will reset as well, and this traffic will be dropped.
- NSF reduces or eliminates the traffic dropped while A’s control plane is down.
• If A is NSF capable, the control plane will not reset the data plane when it restarts.

• Instead, the forwarding information in the data plane is marked as stale.

• Any traffic B sends to A will still be switched based on the last known forwarding information.
GR/NSF Fundamentals

• While A’s control plane is down, the routing protocol hold timer on B counts down.

• A has to come back up and signal B before B’s hold timer expires, or B will route around it.

• When A comes back up, it signals B that it is still forwarding traffic, and would like to resync.

Hold Timer: 6
GR/NSF Design Goals

- No Link Flap
- CEF table sync to the standby RP
- LC CEF Table not cleared on switchover
- Packet forwarding during switchover while routing is converging on Standby RP
- Maintain peer relationship, no adjacency flapping with peers
- Limit restart to be local event, not network wide

Ultimate Goal: Achieve 0% Packet Loss
IS-IS GR/NSF

- Use the `nsf` command under the `router isis` configuration mode to enable graceful restart.
- *Show isis nsf* can be used to verify graceful restart is operational.

```
router#show isis nsf
NSF is ENABLED, mode 'cisco'
RP is ACTIVE, standby ready, bulk sync complete
NSF interval timer expired (NSF restart enabled)
Checkpointing enabled, no errors
Local state:ACTIVE, Peer state:STANDBY HOT, Mode:SSO
```
**OSPF GR/NSF**

- Use the `nsf` command under the `router ospf` configuration mode to enable graceful restart.

- **Show ip ospf** can be used to verify graceful restart is operational.

```
router#sh ip ospf
Routing Process "ospf 100" with ID 10.1.1.1

Non-Stop Forwarding enabled, last NSF restart 00:02:06 ago (took 44 secs)

router#show ip ospf neighbor detail
Neighbor 3.3.3.3, interface address 170.10.10.3

Options is 0x52
LLS Options is 0x1 (LR), last OOB-Resync 00:02:22 ago
```
BGP GR/NSF

- Use the `bgp graceful-restart` command under the `router bgp` configuration mode to enable graceful restart.
- `show ip bgp neighbors` can be used to verify graceful restart is operational.

```
router bgp 65000
  bgp graceful-restart
  ....
```

```
router bgp 65501
  bgp graceful-restart
  ....
```

```
router#show ip bgp neighbors x.x.x.x
  ....
  Neighbor capabilities:
  ....
  Graceful Restart Capability: advertised and received
  Remote Restart timer is 120 seconds
  Address families preserved by peer:
  IPv4 Unicast, IPv4 Multicast
```
Multicast HA: PIM/SSM in the Core

How Triggered PIM Join(s) work when Active Route Processor Fails:

1. Active Route Processor receives periodic PIM Joins in steady-state
2. Active Route Processor fails
3. Standby Route Processor takes over
4. “Special” PIM Hello is sent out
5. Triggers adjacent PIM neighbors to resend PIM Joins refreshing state of distribution tree(s) preventing them from timing out
Multicast HA: IGMP at the edge

Provides Seamless Forwarding of Multicast Traffic During Supervisor Switchovers

How it works:

1. Global sync of IGMP Data Structures occurs when standby Supervisor comes online
2. Periodic syncs are performed as changes occur
3. Active Supervisor fails
4. Standby Supervisor takes over
5. Synced IGMP Data Structures are utilized to provide Layer-2 Non-Stop Forwarding (NSF) of multicast traffic
Agenda

• IPTV Architecture overview
• High Availability (HA) in the core and edge
  – Router Level HA with SSO
  – Protocol HA with NSF/GR (IGP, BGP, Multicast)
  – MPLS HA (LDP NSF/GR, BGP + label NSF, FRR, PW Redundancy)
  – Fast Convergence (Multicast FC, Fast channel change, LDP-IGP sync, LDP session protection)
• Video Source Redundancy
• Conclusion
MPLS HA – Component Resiliency

• MPLS HA features extend NSF with SSO for:
  – LDP (for L3VPNs and L2VPNs)
  – BGP (L3VPNs)
  – RSVP (for TE)

• Minimal disruption to MPLS forwarding plane due to route processor control plane failures
  – Includes MPLS control plane failures (LDP, BGP, RSVP)
MPLS HA – Network Resiliency

- **MPLS control plane protocol enhancements** to improve failure detection time and network convergence
- **Graceful Restart (GR)**
  - LDP, BGP, RSVP
- **Fast Convergence**
  - LDP (IGP sync), LDP session protection
- **TE FRR**
  - Link protection
  - Node protection
- **Pseudowire Redundancy**
MPLS Graceful Restart + NSF/SSO

No MPLS HA support

MPLS HA Support: NSF/SSO + Graceful Restart
LDP NSF/GR

- LDP NSF/SSO/GR is also known as LDP NSF
- LDP NSF allows a route processor to recover from disruption in LDP control plane without losing its MPLS forwarding state
- LDP NSF works with LDP sessions between directly connected peers as well as with targeted sessions
- LDP HA Key Elements
  - Checkpointing local label bindings (synkronisering mellem RP’s)
    - On devices with route processor redundancy
  - LDP graceful restart capability
    - On participating PEs, RRs, and P routers
LDP NSF/GR Key Elements: Checkpointing Local Label Bindings

- The checkpointing function is enabled by default.
- The checkpointing function copies active RP’s LDP local label bindings to the backup RP.
- For the first round, all the labels are copied from active to backup RP.
- Periodic incremental updates are done to reflect new routes that have been learned or routes that have been removed and/or when labels are allocated or freed.
- Label bindings on backup RP are marked checkpointed.
- This marking is removed when it becomes active RP.
LDP NSF/GR Operation

- LDP paths established, LDP GR negotiated
- When RP fails on LSRb, communication between peers is lost; LSRb encounters a LDP restart, while LSRa and LSRc encounter an LDP session reset
- LSRa and LSRc mark all the label bindings from LSRb as stale, but continue to use the same bindings for MPLS forwarding
- LSRa and LSRc attempt to reestablish an LDP session with Rb
- LSRb restarts and marks all of its forwarding entries as stale; at this point, LDP state is in restarting mode
- LSRa and LSRc reestablish LDP sessions with Rb, but keep their stale label bindings; at this point, the LDP state is in recovery mode
- All routers re-advertise their label binding info; stale flags are removed if a label has been relearned; new LFIB table is ready
Enable MPLS Nonstop Forwarding on an Interface that Uses BGP as the Label Distribution Protocol (Inter-as and CsC Environment)

```
mpls bgp forwarding
```
MPLS TE Fast Re-Route (FRR)

- IP routing protocols (e.g., OSPF, BGP) may be tuned to convergence within a few seconds
- Some traffic (e.g. voice) will require more aggressive convergence time
  - Typically 50 ms or less
- MPLS TE FRR offers protection against network failures
  - Link protection
  - Node protection
FRR Link Protection

• Creation of Next-hop Backup Tunnel
  – Parallel path around protected link to next hop
• On link failure detection
  Point of Local Repair (PLR) swaps label and pushes backup label
  – Traffic sent over backup path
• PLR notifies TE Head End (HE), which triggers global TE path re-optimization
FRR Node Protection

- Creation of Next-next-hop Backup Tunnel
  - Parallel path around protected node to next-next-hop
- Node failure triggers Point of Local Repair (PLR) to swap label and push backup label
  - Traffic sent over backup path around failed node
- PLR notifies TE Head End (HE), which triggers global TE path re-optimization
FRR in triple play with P2MP-TE

**Basic RSVP-TE P2MP Tunnel Setup**

### Source
- Multicast Packet
- Multicast Packet + MPLS Label

### Service Edge
- Layer 2 Switch

### Core
- Primary RSVP Tunnel
- MPLS Core
- Backup RSVP Tunnel

### Distribution/Access
- Layer 2 Switch

### Receiver
- Multicast Packet
- Source
- Receiver

© 2006 Cisco Systems, Inc. All rights reserved.
Multiple Points of Failure Still Exist

Source | Service Edge | Core | Distribution/Access | Receiver
---|---|---|---|---
Source | Layer 2 Switch | PE | | Receiver
PE | CE | | Layer 2 Switch |
Primary RSVP Tunnel
Backup RSVP Tunnel

MPLS Core
Pseudo Wire HA

- Failure in the provider core mitigated with link redundancy and FRR
- PE router failure—Another PE or NSF/SSO
- PE-CE Attachment Circuit failure—need pair of attachment Circuits end-to-end
- CE Router failure—redundant CEs
pe1(config)#int e 0/0.1
pe1(config-subif)#encapsulation dot1q 10
pe1(config-subif)#xconnect <PE3 router ID> <VCID> encapsulation mpls
pe1(config-subif-xconn)#backup peer <PE4 router ID> <VCID>
MPLS HA

- Link failures—solved using redundancy in network design and MPLS TE FRR and/or Fast IGP
- Node failures—solved using redundant nodes and meshing of connections
  - Also using MPLS TE FRR node protection
- Line card failures
  - Hardware redundancy—line card redundancy (1+1, 1:N, 1:1)
- PE router (RP) failures
  - Dual RP systems NSF/SSO
  - HA software provides seamless transition...
LDP and BGP use TCP as a reliable transport mechanism for its protocol messages.
The TCP session between two LDP/BGP peers may go down due to HW/SW failure (RP switchover).
Use IGP, BGP, LDP NSF/GR
• The TCP session between two LDP peers may go down due to HW/SW failure (RP switchover)
• If PE3 fails, traffic will be dropped
• Use PW-redundancy, LDP NSF/GR and IGP NSF/GR
Agenda

• IPTV Architecture overview

• High Availability (HA) in the core and edge
  – Router Level HA with SSO
  – Protocol HA with NSF/GR (IGP, BGP, Multicast)
  – MPLS HA (LDP NSF/GR, BGP + label NSF, FRR, PW Redundancy)
    – Fast Convergence (Multicast FC, Fast channel change, LDP-IGP sync, LDP session protection)

• Video Source Redundancy

• Conclusion
Fast Convergence:

- Too much to cover, not possible in 45 mins
- Layer 2 tuning
  - SONET alarms
  - Interface timers
- CEF optimizations
- IGP FC
  - Fast IGP timers, iSPF, fast flooding, fast hellos, etc
- BGP FC
  - Next Hop Tracking, BGP timers
- BFD (bidirectional forwarding)
- Multicast Fast Convergence
  - Multicast timers, fast channel change
- LDP
  - LDP (IGP sync), LDP session protection
Multicast Subsecond Convergence

First, what it’s NOT:

A single feature or set of CLI’s

So What is it?

Various components seamlessly working together to provide an end-to-end fast convergence solution

Objective:

Achieve and meet requirements for sub-second channel change, pause, resume, etc.
### Multicast Subsecond Convergence:
The Video Challenge - Channel Change Time & Source Redundancy

<table>
<thead>
<tr>
<th>Network Components:</th>
<th>Channel Change Delay and Slow Convergence Contributors:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set Top Box (STB)</td>
<td>1. Set Top Box (STB) performance in processing channel change request</td>
</tr>
<tr>
<td>DSLAM</td>
<td>2. Network Signaling latency to join or resume new IP video feed</td>
</tr>
<tr>
<td>Multicast Distribution Tree Building</td>
<td>3. Delay incurred if using encryption</td>
</tr>
<tr>
<td>primary, backup</td>
<td>4. DSLAM Signaling latency to terminate previous IP video feed</td>
</tr>
<tr>
<td>Video Sources</td>
<td>5. Network Signaling latency to rebuild tree(s) during link/node failures in the core</td>
</tr>
<tr>
<td></td>
<td>6. Time for video feed to recover during primary source failover</td>
</tr>
</tbody>
</table>
**Multicast Subsecond Convergence**

### Areas Addressed Today:

1. **Set Top Box (STB) performance in processing channel change request**
   - Vendor dependent

2. **Network Signaling latency to join or resume new IP video feed**
   - Multicast Fast Join/Leave for Fast Channel Change

3. **Delay incurred if using encryption**
   - Vendor dependent

4. **DSLAM Signaling latency to terminate previous IP video feed**
   - Vendor dependent

5. **Network Signaling latency to rebuild tree(s) in the core**
   - IGP timers
   - **RPF Interval timer** - `ip multicast rpf interval <seconds> [list <acl> | route-map route-map>]`
   - **PIM RPF Failover timer** - `ip multicast rpf backoff <min> <max> [disable]`
   - **PIM Router Query interval timer** - `ip pim query-interval <period> [msec]`
   - PIM Triggered Join(s)
   - IGMP High Availability

6. **Time for video feed to recover during primary source failover**
   - Anycast Sources (and other related methodologies)
Fast Join/Leave for Faster Channel Change

Problem Description:
In networks where bandwidth is constrained between multicast routers and hosts (like in xDSL deployments), fast channel changes can easily lead to bandwidth oversubscription, resulting in a temporary degradation of traffic flow for all users.

Solution:
Reduce the leave latency during a channel change by extending the IGMPv3 protocol.

Benefits:
• Faster channel changing without BW oversubscription
• Improved diagnostics capabilities
Multicast Fast Join/Leave for Faster Channel Change

How it works:
- Relies on IGMPv3
- Router tracks both User and Channel(s) being watched
- When user leaves channel no one else is watching, router immediately prunes the channel off the interface compared to IGMPv2 (up to 3 seconds) and IGMPv1 (up to 180 seconds)!

Configuration:
```
interface Ethernet 0
ip pim sparse-mode
ip igmp version 3
ip igmp explicit-tracking
```

First introduced in 12.0(29)S
Fast Convergence: LDP-IGP sync => LDP Down but IGP Still Up

1. Data stream between CE1 and CE2 traversing SP cloud
2. LDP Adjacency goes down between P4 and P3 but IGP still points to P3 as the best path
3. LDP Adjacency goes down between P4 and P3 but IGP still points to P3 as the best path
LDP down but IGP Still Up with IGP-LDP Synch Feature

1. Data stream between CE1 and CE2 traversing SP cloud
2. LDP Adjacency goes down between P4 and P3 but IGP still points to P3 as the best path
3. P1 chooses P2 as the preferred IGP path because of the lower metric
IGP-LDP Sync

- IGP-LDP Synch features synchronizes the state of IGP and LDP between two routers
- If an LDP adjacency between two routers goes down then these routers would advertise a max-metric for the link connecting them via IGP
- This way upstream router can choose an alternate path if available
- Following configuration needs to be enabled on the routers
  - `P4(config)# router ospf 1`
  - `P4(config-router)# mpls ldp sync`
  - `P4(config)# mpls ldp igp sync holddown <msecs>`
  - If no holddown is configured then no timeout is applied i.e. IGPs will wait forever.
Fast Convergence: LDP Session Protection

- There are two discovery mechanisms for LDP peers
  1. **LDP Basic Discovery** — Discovery of directly connected neighbors via link hellos
  2. **LDP Extended Discovery** — Discovery of non-directly connected neighbors via Targeted Hellos
- If the session protection is enabled, the establishment of a directly connected LDP session triggers Extended Discovery with the neighbor
- With targeted hellos, session stays up even when the link goes down
- No need to re-establish an LDP session with the link neighbor and relearn prefix label bindings when the link recovers

```
p1(config)#mpls ldp session protection
  duration Period to sustain session protection after loss of link discovery
  for Access-list to specify LDP peers
  vrf VRF Routing/Forwarding instance information
```
Agenda

• IPTV Architecture overview

• High Availability (HA) in the core and edge
  – Router Level HA with SSO
  – Protocol HA with NSF/GR (IGP, BGP, Multicast)
  – MPLS HA (LDP NSF/GR, BGP + label NSF, FRR, PW Redundancy)
  – Fast Convergence (Multicast FC, Fast channel change, LDP-IGP sync, LDP session protection)

• Video Source Redundancy

• Conclusion
### Video Source Redundancy: Two Approaches

<table>
<thead>
<tr>
<th>Primary-Backup “Heartbeat”</th>
<th>Hot-Hot</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Two sources, One active (src’ing content), Second in standby mode (not src’ing content)</strong></td>
<td><strong>Two sources, both are active and src’ing multicast into the network</strong></td>
</tr>
<tr>
<td>Heartbeat mechanism used to communicate with each other</td>
<td>No Protocol between the two sources</td>
</tr>
<tr>
<td><strong>Only one copy is on the network at any instant</strong></td>
<td><strong>Two copies of the multicast packets will be in the network at any instant</strong></td>
</tr>
<tr>
<td>Single Multicast tree is built per the unicast routing table</td>
<td>Two Multicast trees on almost redundant Infrastructure</td>
</tr>
<tr>
<td>Uses required bandwidth</td>
<td>Uses 2x network bandwidth</td>
</tr>
<tr>
<td><strong>Receiver’s functionality simpler:</strong></td>
<td><strong>Receiver is smarter:</strong></td>
</tr>
<tr>
<td>Aware of only one src, fail-over logic handled between sources.</td>
<td>Is aware/configured with two feeds (s1,g1), (s2,g2) / (<em>,g1), (</em>,g2)</td>
</tr>
<tr>
<td><strong>This approach requires the network to have fast IGP and PIM convergence</strong></td>
<td><strong>This approach does not require fast IGP and PIM convergence</strong></td>
</tr>
</tbody>
</table>
Video Source Redundancy:
Heartbeat Model
Video Source Redundancy: Hot-Hot Video Delivery Model
How is source redundancy achieved in the network?

- Enable SSM on all routers
- Have R1 and R2 advertise same prefix for each source segment.
- R3 and R4 follow best path towards source based on IGP metrics.
- Let’s say R3’s best path to SF is through R1. The source in SF now suddenly fails.
- R3’s IGP will reconverge and trigger SSM joins towards R2 in NY.
Conclusion

• High Availability involves many more components in Metro Ethernet, DSL/PON deployments, Authentication/Application Servers
  – Distributed DSLAMs

• Slowest component is a bottleneck for the entire system

• Careful when using High Availability versus Fast Convergence. Race Conditions possible!!