Routing/Signalling Non-Stop Forwarding And Increased Network and Node Availability

Cisco.com

David Ward

2002.02

Cisco Systems

Topics

Cisco.com

- Why? What is the problem?
- Solution emerging
- BGP
- OSPF & ISIS
- Multicast
- MPLS
- RIB
- Summary: What happens to my network?

Problem

1. **Must Increase Network and Node availability during planned or unplanned software restart or hardware change**

2. **Upgrade/Downgrade of Software is currently high impact**

3. **Packets stop flowing (FULL Stop Forwarding)**

4. **Topology changes seen in network**

   - **Flapping routes**
   - **Traffic oscillations**
   - **Increased load**

5. **Decreasing maintenance window durations**

Potential solution

1. **Recover State from network peers/adjacencies:**
   - **Compare w/ checkpoint of peer state to optimize restart timing**
   - **Do not try and keep TCP (transport) state - unnecessary**

2. **Negotiate as part of protocol state to optimize for min outage duration**

3. **Non Stop Forwarding**
   - **SLAs can be preserved**
   - **No sudden oscillation of traffic in the network**

4. **No Topology Change: NO FLAPPING ROUTES**

5. **Unicast, Multicast, MPLS specific approaches**

Potential solution

Cisco.com

6. **Detect outage, restart and reconverge within time of specified neighbor or peer ‘downtime’**

7. **Have control knobs to abort NSF**

8. **Find solutions that do and do not require protocol extensions**
   - **Generally, those that do converge faster**

9. **Don’t carry or announce more information with each prefix**

10. **Don’t require new Update/LSA/LSP stuffing properties**

Topics

Cisco.com

- Why? What is the problem?
- BGP
- OSPF & ISIS
- Multicast
- MPLS
- RIB
- Summary: What happens to my network?

BGP Protocol Behavior and Extensions

Cisco.com

1. **Routing table recoverable from peers without keeping TCP state**
   - **Socket is closed**

2. **Works for Board failover, software restart, peer reset of any flavor**
   - **Can limit number of graceful restarts if peer flapping**

3. **Dynamically negotiated during peer initialization and/or once peering has been established**

4. **Extensions Required:**
   1. **End of RIB marker**
   2. **Graceful Restart Capability**
      - **Per AF/SAF**
      - **Also MPLS labels**

BGP behavior

Cisco.com

For the Graceful Restart capable *receiving* speaker

1. Mark all routes from down peer as *stale* and start purge timer.

2. Peer comes back
   - stop purge timer
   - start timer to wait for End Of RIB marker

3. EOR timer fires or when an EOR is received from the restarting speaker
   - remove all routes that are still marked stale

4. If purge timer fires (i.e the neighbor did not come back up)
   - remove all stale routes.

5. If peer comes back but does not have Graceful Restart capability,
   - purge all stale routes from this peer.

BGP NSF protocol behavior

Cisco.com

For the GR capable *restarting* speaker

1. Configure the RIB to hold onto routes
   - 'specified' period before deleting them.

2. Wait for End of RIB marker from all peers that are GR capable.

3. Once all EORs have been rcvd or the update-delay timer fires
   - calculate the best paths,
   - update the RIB and also generate updates to all peers
   - Additionally, if IGP also restarted, wait for IGP to converge

4. Send End of RIB to all GR capable peers after sending updates.

Topics

- Why? What is the problem?
- Solution emerging
- BGP
- OSPF & ISIS
- Multicast
- MPLS
- RIB
- Summary: What happens to my network?

IGP Goals

Cisco.com

- IGPs must be able to recover/rebuild state information from neighbors before adjacency timers expire
- All IGP neighbors must be NSF-aware for fastest recovery.
  - There are other protocol enhancements for NSF unaware neighbors that are slower
- IGP must re-install the routes within purge timer.
  - Configurable
  - The above mechanism applies to MPLS

High level flow of ospf restart

Cisco.com

1. Read config to find if NSF is to be done.
2. On all of the up interfaces send Hello with ReSync bit set.
3. Wait for N seconds to learn about all the neighbors.
4. Start LSDB re-synchronisation with all of the neighbors.
5. Once mark/sweep completed generate router LSA.
6. Run SPF algorithm.
7. Flush all stale self originated LSA's.

OSPF NSF will be aborted

Cisco.com

- **On restarting router**
  - **When any of the neighbors are NSF incapable.**
  - **Resync doesn't get completed in Dead Interval.**
  - **When there are more than one neighbors attempting NSF**
- **On adjacent router**
  - **When there are more than one neighbors attempting NSF.**
  - **Change in LSDB when NSF is in progress (configurable).**

ISIS NSF Version 1

Cisco.com

- **Required: Recovery before hello timers expire.**
  - **Must be able to send Hello or LSA**
- **No extensions to protocol**
  - **Interoperates with NSF unaware adjacencies**
- **Relies on heuristic to determine "convergence"**
  - **Interval between NSF Restarts**
  - **Max route lifetime following restart**

ISIS: NSF State Machine.1

Cisco.com

- **Phase 1**
  - **- Read ADJ chkpt table: adjacencies restored**
- **Phase 2**
  - **- Enable IIH I/O, send hellos to restored adj**
- **Phase 3**
  - **- Read LSP chkpt table :LSPs restored, 2 local LSPs**
- **Phase 4**
  - **- Enable PDU I/O**
- **Phase 5**
  - **- Request checkpointed LSPs on non-DR circuits**
- **Phase 6**
  - **- Restore LSPs**
- **Phase 7**
  - **- Check DR LAN circuits**

ISIS: NSF State Machine.2

Cisco.com

- Phase 8
  - – - Restore LAN circuits
- Phase 9
  - • - Sync p2p circuits: P2P CSNP/PSNP handshake complete

Phase 9 continued

- – P2P LSPs restored

- Phase 10
  - - Listen for new LSPs: CONFIGURABLE Duration

Phase 11

- – L1 spf ... L2 spf

Phase 13

- – Sync local checkpoint tables

ISIS NSF restart complete

ISIS: NSF Version 2

Cisco.com

- **Requires neighbors to implement protocol changes**
  - **Similar to BGP's End of RIB**
- **Doesn't require any state preservation.**
  - **Hello contains two new options:**
    1. **RR - Restart Request**
    2. **RA - Restart ACK**
- **How it Works**
  1. **Receipt of an RR causes adjacency refresh.**
  2. **RA is used to ack the database synchronization.**

Topics

Cisco.com

- Why? What is the problem?
- Solution emerging
- BGP
- OSPF & ISIS
- Multicast
- MPLS
- RIB
- Summary: What happens to my network?

Multicast NSF Approach

Cisco.com

1. **If control plane restarts, line cards will forward on the old MFIB**
   - **Only long enough for the control plane to failover or restart.**

2. **During NSF, rate limit will be applied to all multicast traffic**
   - **Avoid Internet packet gun, just in case**

Overview of design changes for NSF:

Cisco.com

1. Checkpointing of critical state locally

2. Protocol DONE messages
   - Similar to BGP End of RIB

3. NSF-aware process restart behavior

What's "Checkpointed"?

Cisco.com

- **PIM *G/M -> RP* mapping**
- **RPF-RIB**

What ISN'T Checkpointed

- **PIM Topology: Recoverable from network**
- **IGMP State: Recoverable from network**
- **MSDP SA Cache: Recoverable from network**

Detecting Loops during Multicast NSF

Cisco.com

- **We can detect and shut down loops during NSF:**

  MFIB entries that have updated will use PIM-Assert to designate a forwarder (even if unicast has not converged).

  MFIB entries marked OLD remove interfaces from the O-List if they receive matching data on it.

Multicast Control plane Restart

Cisco.com

1. RP restarts both unicast & multicast
2. *Forwarding* detects fault, initiates NSF action
    a. marks MFIB entries OLD
    b. rate-limits total multicast
    c. starts timer for MFIB OLD sweep
3. All control processes restart
    MRIB in NSF mode, holds back MFIB updates.
4. MURIB restores routes from Unicast
5. IGMP gets updates and propagates to MRIB
6. IGMP sends DONE
7. PIM gets updates from net & IGMP via MRIB

Mcast Control Plane Restart (continued)

Cisco.com

8. PIM updates MRIB and sends DONE

9. MRIB gets DONE from PIM, then:

    a. Sends MFIB updates

10. MFIB receives updates and reinstall at route granularity

11. MFIB timer expires and sweeps OLD entries

12. Unicast reconverges receive DONE

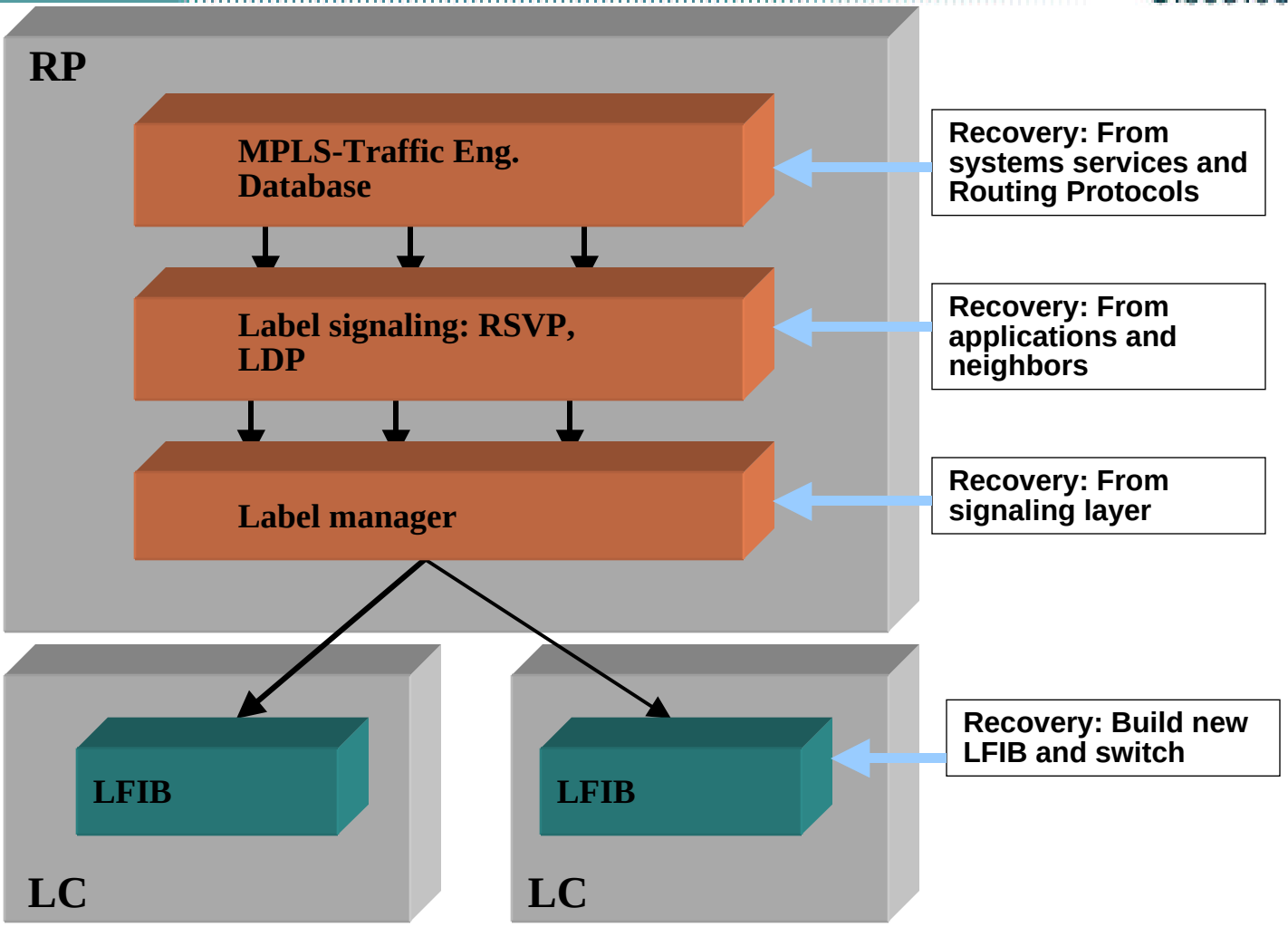
13. Rate limit in forwarding removed

Topics

Cisco.com

- Why? What is the problem?
- Solution Paradigm emerging
- BGP
- OSPF
- ISIS
- EIGRP
- Multicast
- MPLS
- RIB
- Summary: What happens to my network?

MPLS High Availability Design



MPLS Software NSF: If desired

- Can use FRR for node protection
  - May not want to due to no autoplacement
- Traffic Engineering Database recovers state from IGPs, BGP
- RSVP, LDP extensions
  - Refresh messaging and softstate machinery
  - Graceful restart signalling
  - Local checkpointing of data: Ingress/Egress interface mapping
- Generate and switch to new LFIB (label FIB)

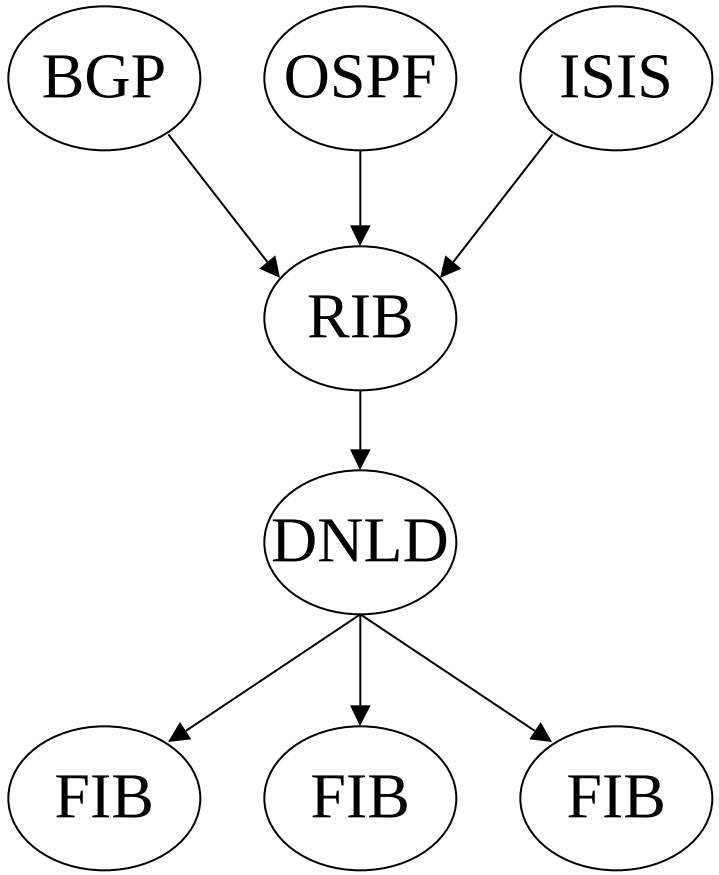
Topics

Cisco.com

- Why? What is the problem?
- Solution Paradigm emerging
- BGP
- OSPF & ISIS
- Multicast
- MPLS
- RIB
- Summary: What happens to my network?

RIB recovery issues

1. **Downloader failure/restart**
2. **Protocol failure/restart**
3. **RIB failure/restart**



DWNLD restart

Cisco.com

1. RIB detects failure, purges associated state
2. DWNLD restarts, connects to RIB
3. FIB clients informed of restart, mark routes for deletion
4. DWNLD retrieves all routes from RIB and downloads to FIB clients
5. DWNLD informs FIB clients when download complete
6. FIB clients purge marked routes

Protocol restart

Cisco.com

1. RIB detects failure of protocol, marks protocol's routes for deletion.
2. Purge timer set for protocol's routes (duration set by protocol)
3. Protocol restarts, reconverges, inserts routes in RIB,
4. Protocol informs RIB when finished inserting: DONE
5. RIB purges remaining marked routes:
   - After all protocols have finished inserting
   - Or purge timer expires

RIB restart.1

Cisco.com

1. RIB restarts
   - Starts convergence timer
   - Withhold update to DWNLD
2. Protocols detect restart of RIB
3. DWNLD detects RIB failure; reconnects
   - DWNLD informs FIB clients of restart
   - FIB clients create new table
   - FIB clients mark routes for deletion
4. Protocols insert routes in RIB, inform RIB when finished

RIB restart continued

Cisco.com

5. DWNLD retrieves all routes from RIB and downloads to FIB clients
6. RIB informs DWNLD of convergence after:
   - All protocols have finished inserting
   - Or convergence timer expires
7. RIB resumes normal operation
8. DWNLD informs FIB clients download complete
9. FIB switches to new table
10. FIB clients purge marked routes

Topics

Cisco.com

- **Why? What is the problem?**
- **Solution Paradigm emerging**
- **BGP**
- **OSPF & ISIS**
- **Multicast**
- **MPLS**
- **RIB**
- **Summary: What happens to my network?**

Summary - marketing slide

- **Routing and Signalling protocol NonStopForwarding enables:**
  - **Nondisruptive planned and unplanned Software restarts**
  - **Upgrade/Downgrade of Software**
  - **Upgrade/Replacement of Hardware**
  - **Eliminate traffic oscillations**
- **Configurable knobs for local tolerance of graceful restart timing**
- **Non Stop Forwarding is key for SW and HW Maintenance**
- **No Route Flap internally or externally**

NOTE: EIGRP NSF similar to ISIS NSF v2

Summary: What happens to the network

- See new STATE of BGP peers: STALE
  - TCP connection dropped
- No extra Withdraw/Announce packets due to HA extensions
- No longer Withdraw/Announce for planned or unplanned outage
  - Don't thrash network if unnecessary
  - Let network know only if something goes wrong
- Not Required that every router runs NSF code
  - Network can handle transition router by router
- No new blackholes or routing loops are necessarily created
- Not all Address families or subaddress families must be NSF enabled in network