



# Cisco IOS High Availability (HA) - Non-Stop Forwarding with Stateful Switchover (NSF/SSO)

## Technical Overview

# Agenda

- **Cisco IOS® HA Introduction**
- **Cisco IOS NSF/SSO**
- **Summary**
- **Questions**

# Introduction to Cisco IOS High Availability



# Cisco IOS High Availability Strategy: Based on Customer Needs

- **Overarching requirement is to provide continuous access to applications, data, and content from anywhere and anytime**
- **Nonstop application delivery**
  - End-to-end**
  - Systems approach**
  - Target every potential cause of downtime with functionality, design, or best practice to mitigate the impact**

# Cisco Approach

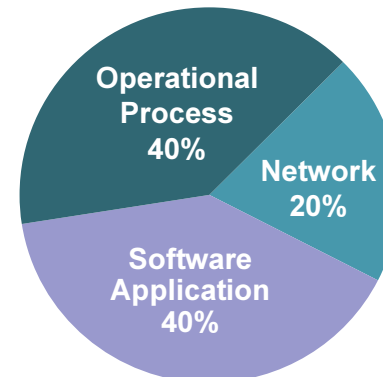
Understand	Identify	Design and Engineer	Best Practice
<ul style="list-style-type: none"> <li>• Research and development</li> <li>• IP network expertise, experience</li> <li>• Listening and responding to customer requests</li> <li>• Standards development/participation</li> <li>• Field interaction</li> <li>• Customer partnership</li> </ul>	<ul style="list-style-type: none"> <li>• Identify Service Affecting Conditions                             <ul style="list-style-type: none"> <li>Hardware</li> <li>Software</li> <li>Traffic and Protocol</li> <li>Performance</li> <li>Management</li> <li>Security</li> </ul> </li> <li>• Leverage broad experience, deployment history, and analysis</li> <li>• Industry experiences shared</li> </ul>	<ul style="list-style-type: none"> <li>• Focused solutions                             <ul style="list-style-type: none"> <li>Targeted to specific, known causes of downtime</li> </ul> </li> <li>• Engineered solutions                             <ul style="list-style-type: none"> <li>Architected for long term, infrastructure reuse</li> </ul> </li> <li>• Service related, rather than platform related</li> <li>• Considers network topology and design—eliminate single points of failure</li> <li>• Automation and embedded intelligence</li> </ul>	<ul style="list-style-type: none"> <li>• Test solution for compliance with objectives</li> <li>• Institute best practices and designs using high availability solutions</li> <li>• Continual process</li> <li>• Evolutionary</li> </ul>

# Where is the Exposure?

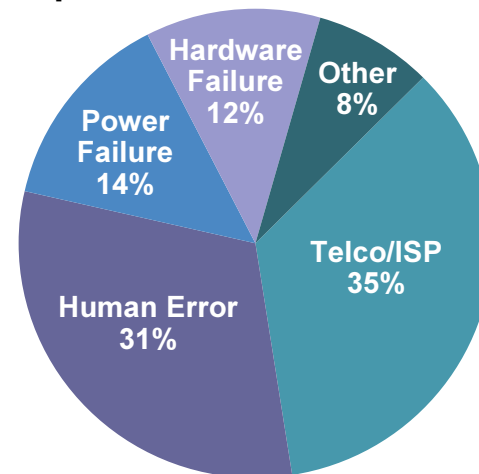
## Most Common Causes of Network Outages

- **Network and software applications**
  - Hardware failures
  - Software failures
  - Link failures
  - Power/environment failures
  - Resource utilization issues
- **Operational processes**
  - Network design issues
  - Lack of standards (hardware, software, configurations)
  - No fault management or capacity planning
  - Inadequate change of management, documentation, and staff training
  - Lack of ongoing event correlation
  - Lack of timely access to experts/knowledge

Sources of Network Downtime\*



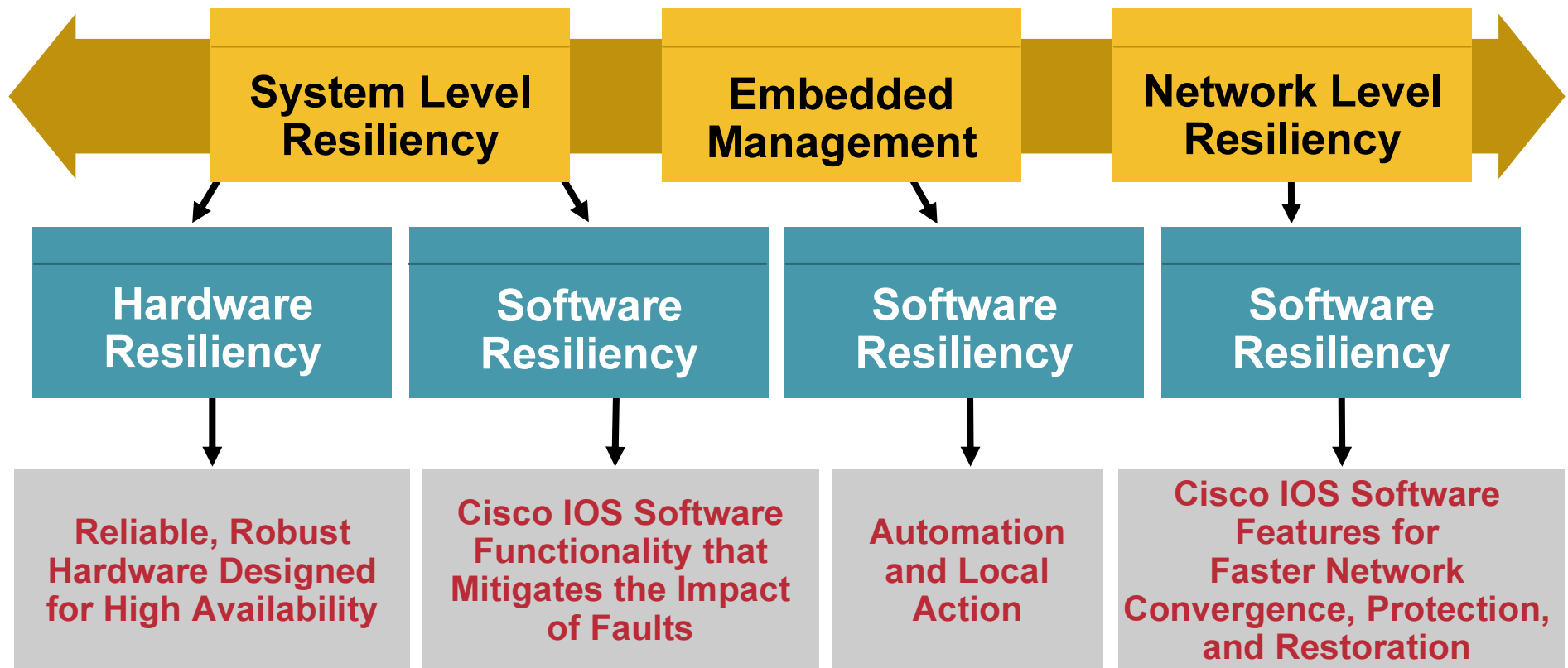
Common Causes of Enterprise Network Downtime\*\*



Pie Graph Source: \*Gartner Group

\*\*Yankee Group The Road to a Five-Nines Network, 2/2004

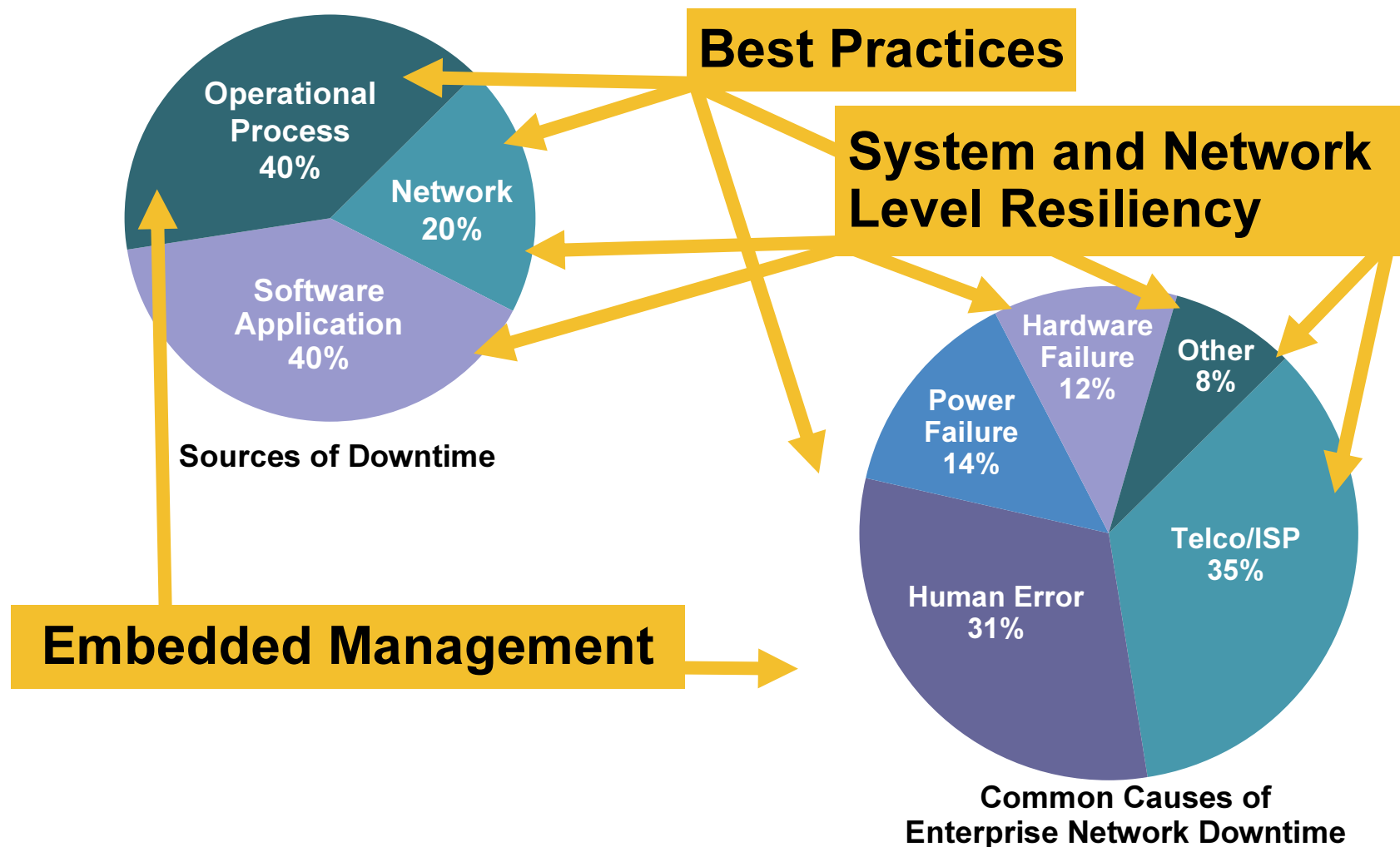
# Systematic, End-to-End Approach: Targeting Downtime



**Investment Protection Is a Key Component**

# Mitigating the Exposure: Targeting Downtime

## Most Common Causes of Downtime





# Cisco IOS High Availability: Focused Approach

- Start with resilient hardware foundation

↑ MTBF

Eliminate single points of failure

- System-level resiliency** at critical network edges

↓ MTTR for system failures including route processors, Line Cards and software component failures

Reduce outage for both dual and single route processor systems

Mitigate planned outages by providing in-service software upgrade

Protect remote devices while creating more efficient and cost-effective resiliency

- Network-level resiliency** in the core and where redundant paths exist

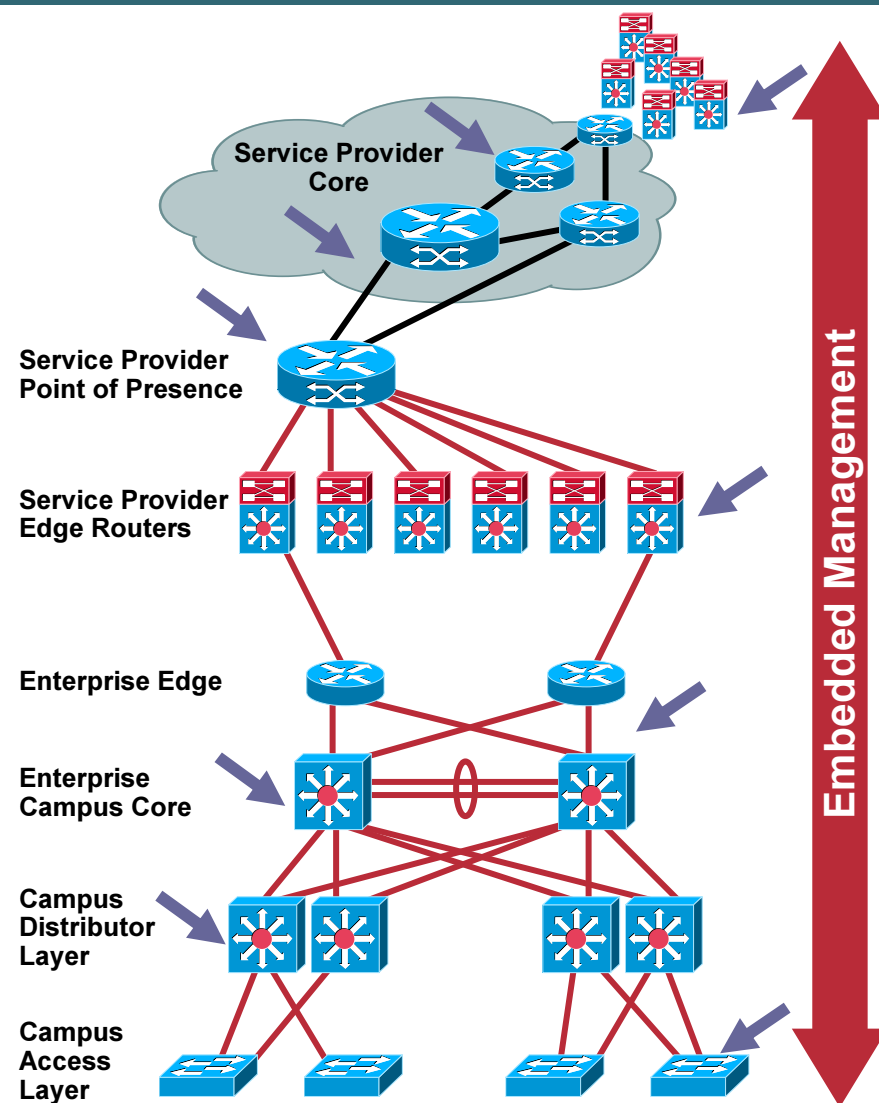
Deliver industry-leading features for fast network convergence, protection and restoration

- Account for services and protocols

- Embedded management** and automation

Embed intelligent event management for proactive maintenance

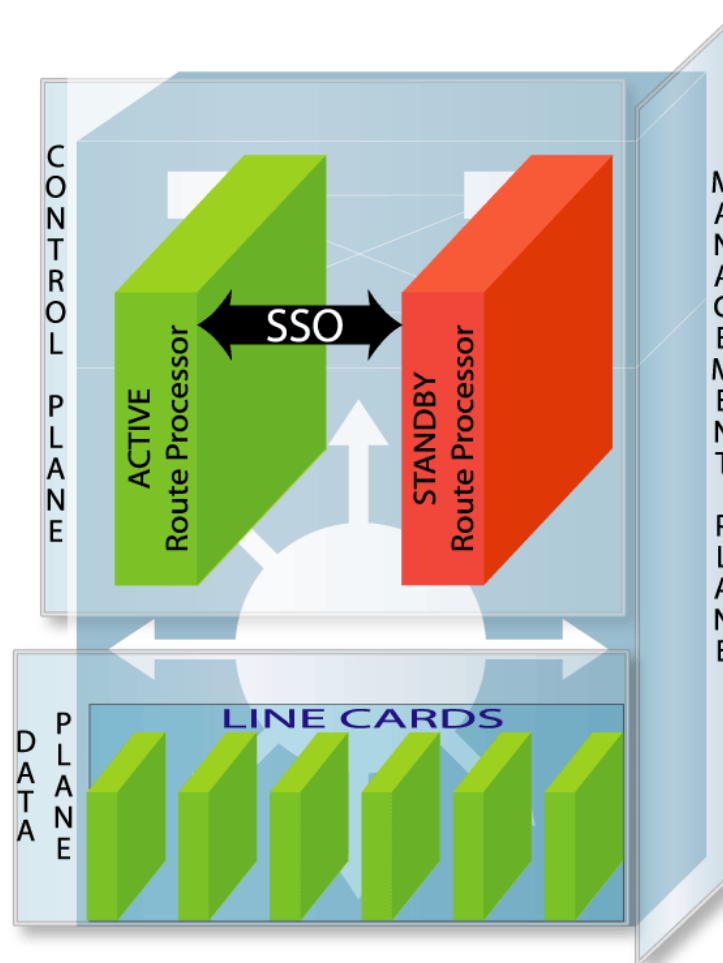
Automation and configuration management to reduce human errors



# System Level Resiliency Overview: Reduce MTTR

## Eliminate Single Points of Failure for Hardware and Software Components

- **Control/data plane resiliency**
  - Separation of control and forwarding plane
  - Control plane handles signaling and network awareness
  - Forwarding/data plane rapidly switches packets
  - Seamless restoration of route processor control and data plane failures
- **Link resiliency**
  - Reduced impact of line card hardware and software failures
- **Planned outages**
  - Seamless software and hardware upgrades
- **Fault isolation and containment**
  - Process independence
  - Target single processor systems



# Cisco IOS NSF/SSO



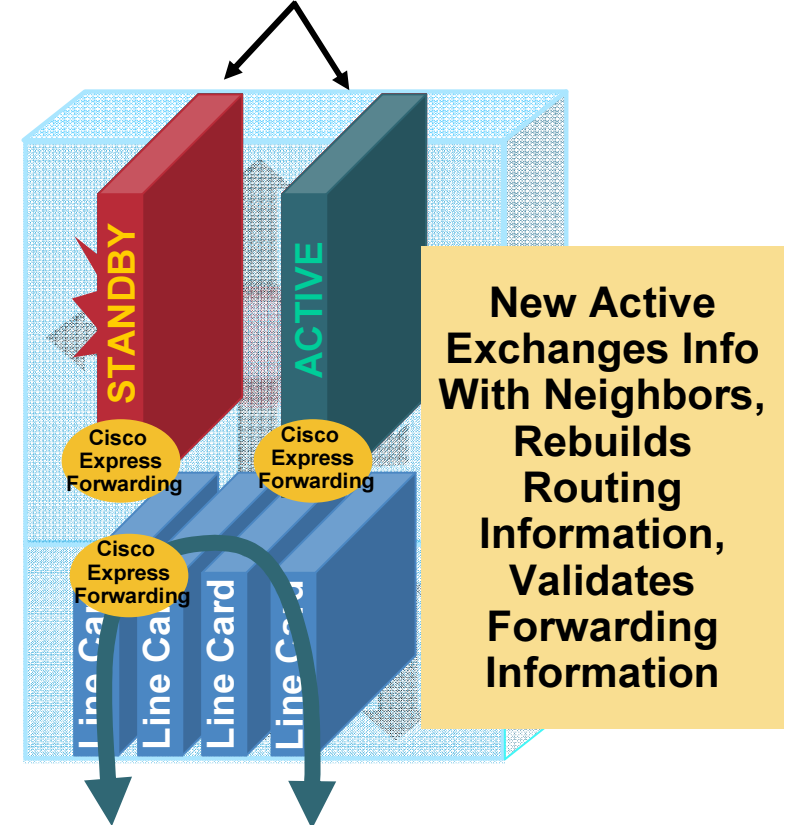
# System-Level Resiliency: Cisco Non-Stop Forwarding with Stateful Switchover

- **Network edge is critical**
  - Service Provider and Enterprise edge
  - Often a single point of failure
- **Cisco NSF with SSO increases availability at key edge points**
- **Employs dual route processors**
- **Cisco SSO maintains connectivity for Layer 2 protocols**
- **Non-stop forwarding of packets while control plane is reestablished and routing information is validated**

Packet forwarding continues using current forwarding information base (FIB)

Layer 3 (BGP, OSPF, IS-IS) recovers routing information from neighbors, rebuilds routing information base (RIB) and updates FIB

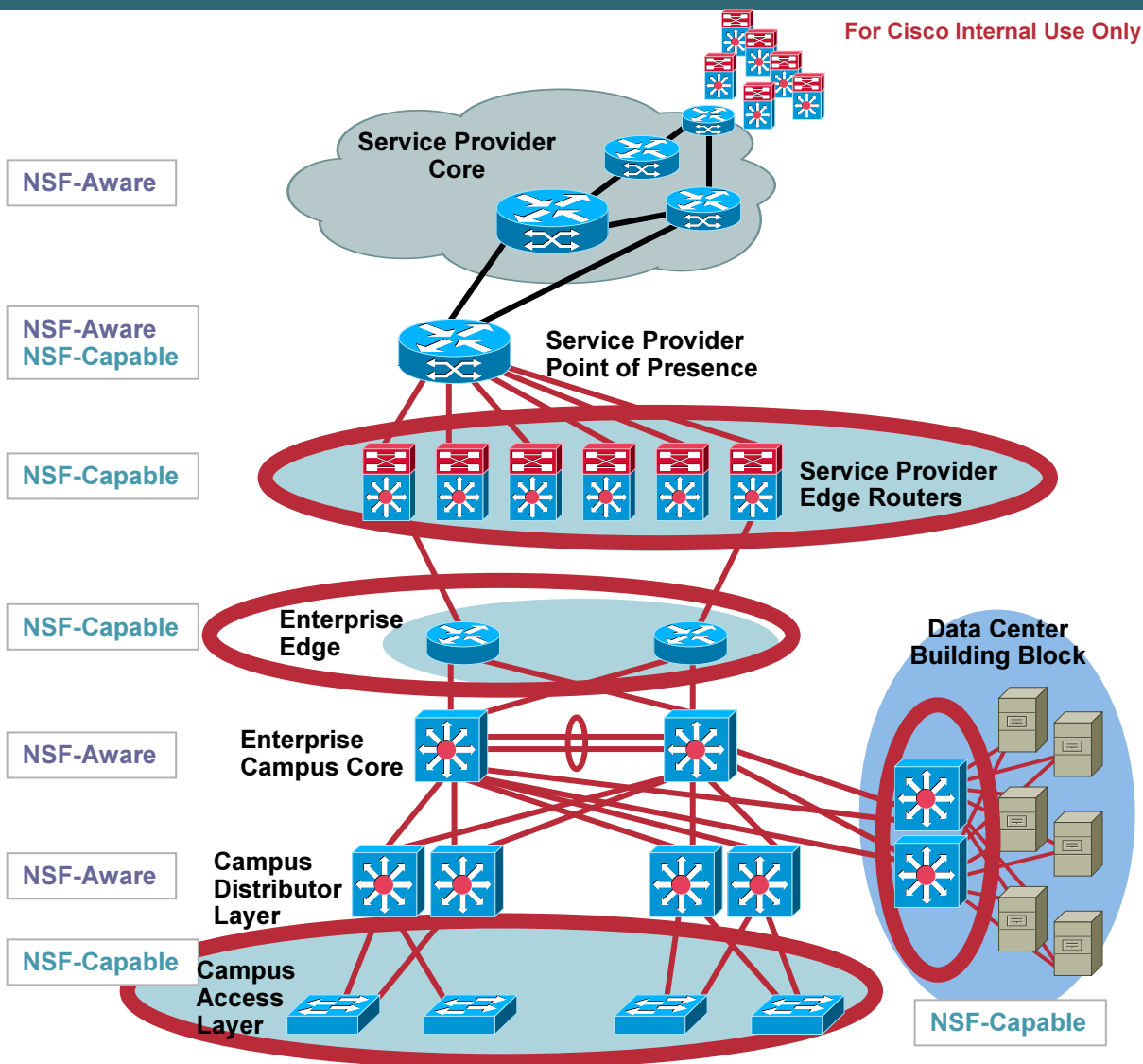
## Redundant Route Processors



Asynchronous Transfer Mode (ATM),  
Ethernet, Point-to-Point Protocol (PPP),  
MLPPP, Frame Relay, Layer 2 LAN  
Protocols cHDLc

# Cisco NSF Awareness

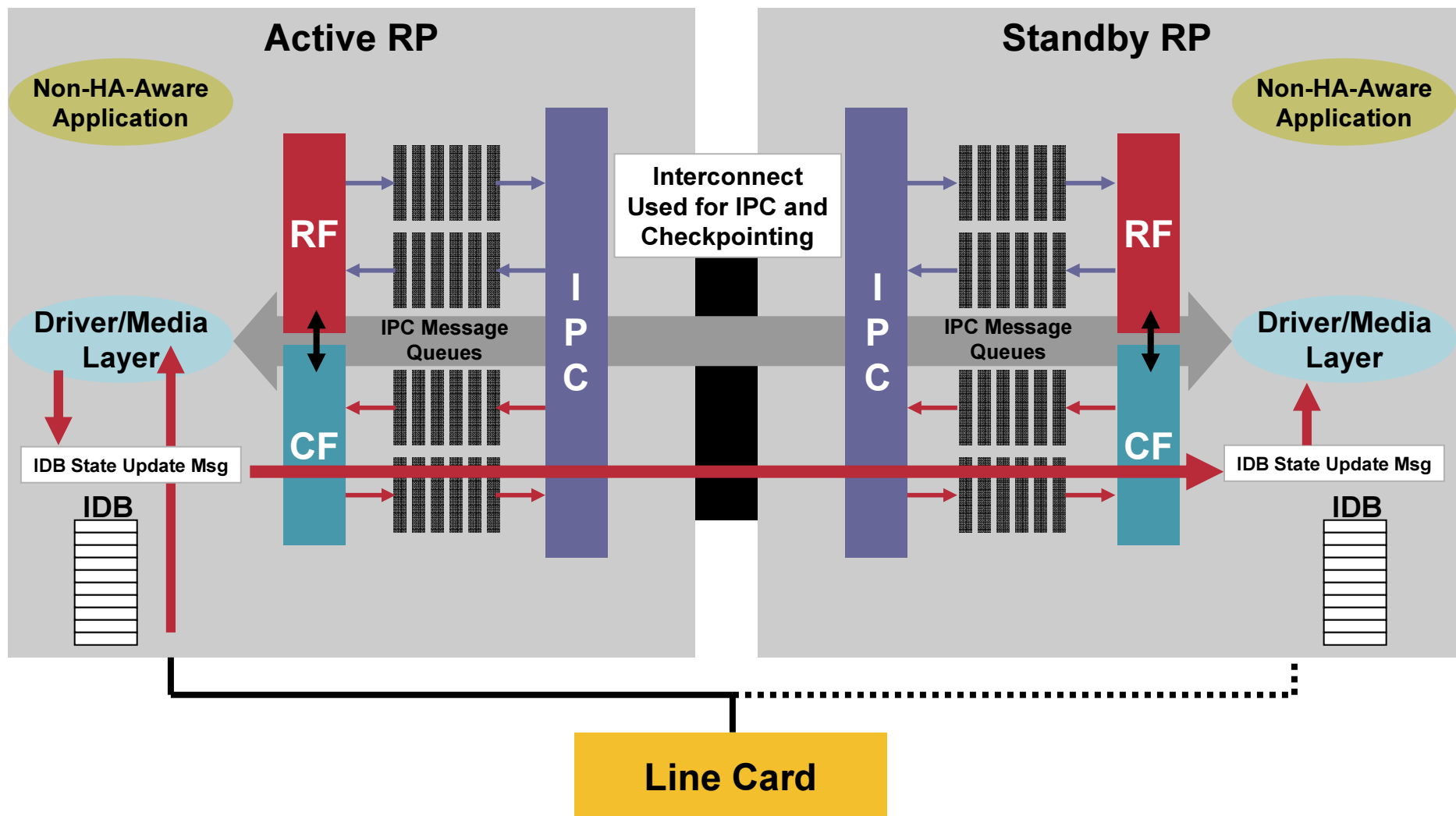
- An NSF-capable router continuously forwards packets during a switchover
- An NSF-aware router assists NSF-capable routers by:
  - Not resetting adjacency
  - Supplying routing information for verification after switchover
- NSF-capable and NSF-aware peers cooperate using Graceful Restart extensions to BGP, OSPF, IS-IS, and EIGRP protocols
- NSF capability at key edge nodes that are single points of failure
- NSF awareness for deployments with redundant topologies having alternate paths
- Cisco advantage:
  - End-to-end NSF awareness across product portfolio (Cisco 1700 Series Router to CRS-1) maximizes the benefits of NSF



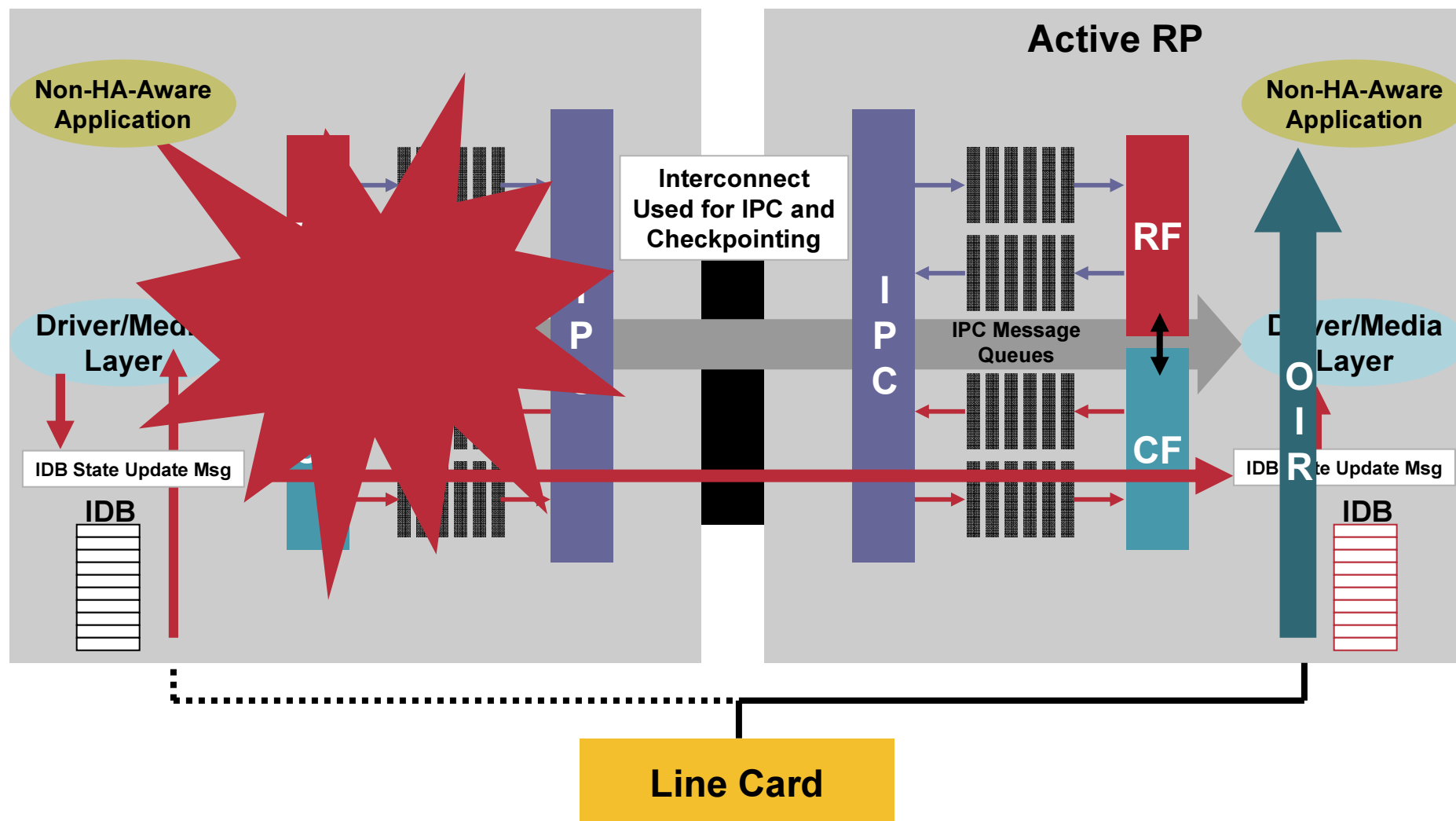
# SSO Infrastructure: Major Components

- **RF (Redundancy Facility)**  
Monitors and reports RP transitions on the Active and Standby RPs
- **CF (Checkpointing Facility)**  
Allows clients to send state updates from the Active to the Standby peer
- **IPC (Inter-Process Communication)**  
Used as the transport for CF, RF and Config Sync
- **Media layer/platform drivers**  
Platform independent/dependent code to maintain IDB state on the Standby
- **Config Sync**  
Maintains the same configuration on the Standby as on the Active
- **SNMP**  
New MIBs and Traps (Cisco-RF.mib), maintains switchover history, value of sysUpTime, syncs other stats and MIBs, provides sync engine
- **Fault Management**  
Embedded Event Manager: Programmable policy management for faults

# SSO Architecture: Non-HA-Aware Protocols



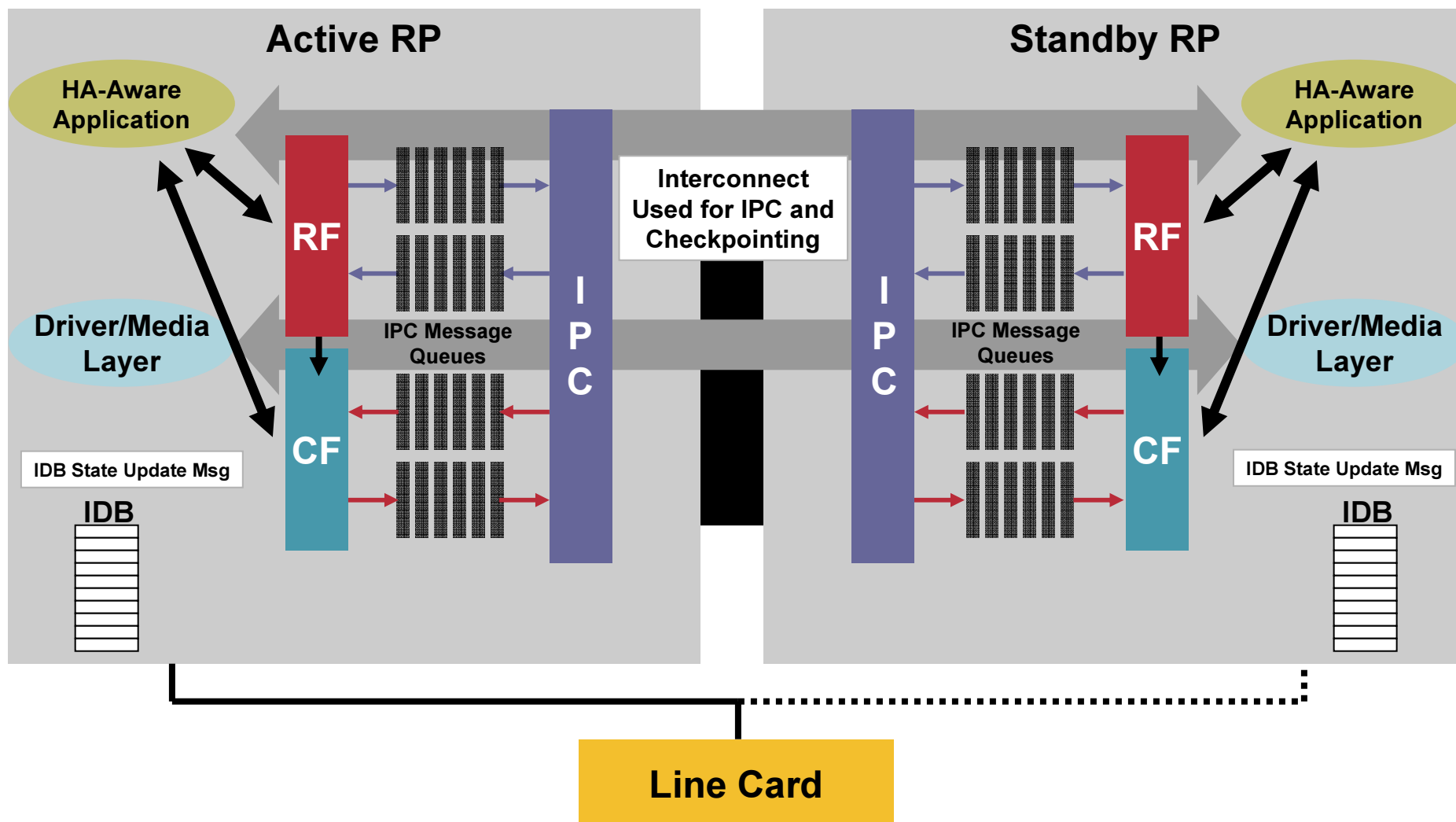
# SSO Architecture: Non-HA-Aware Protocols





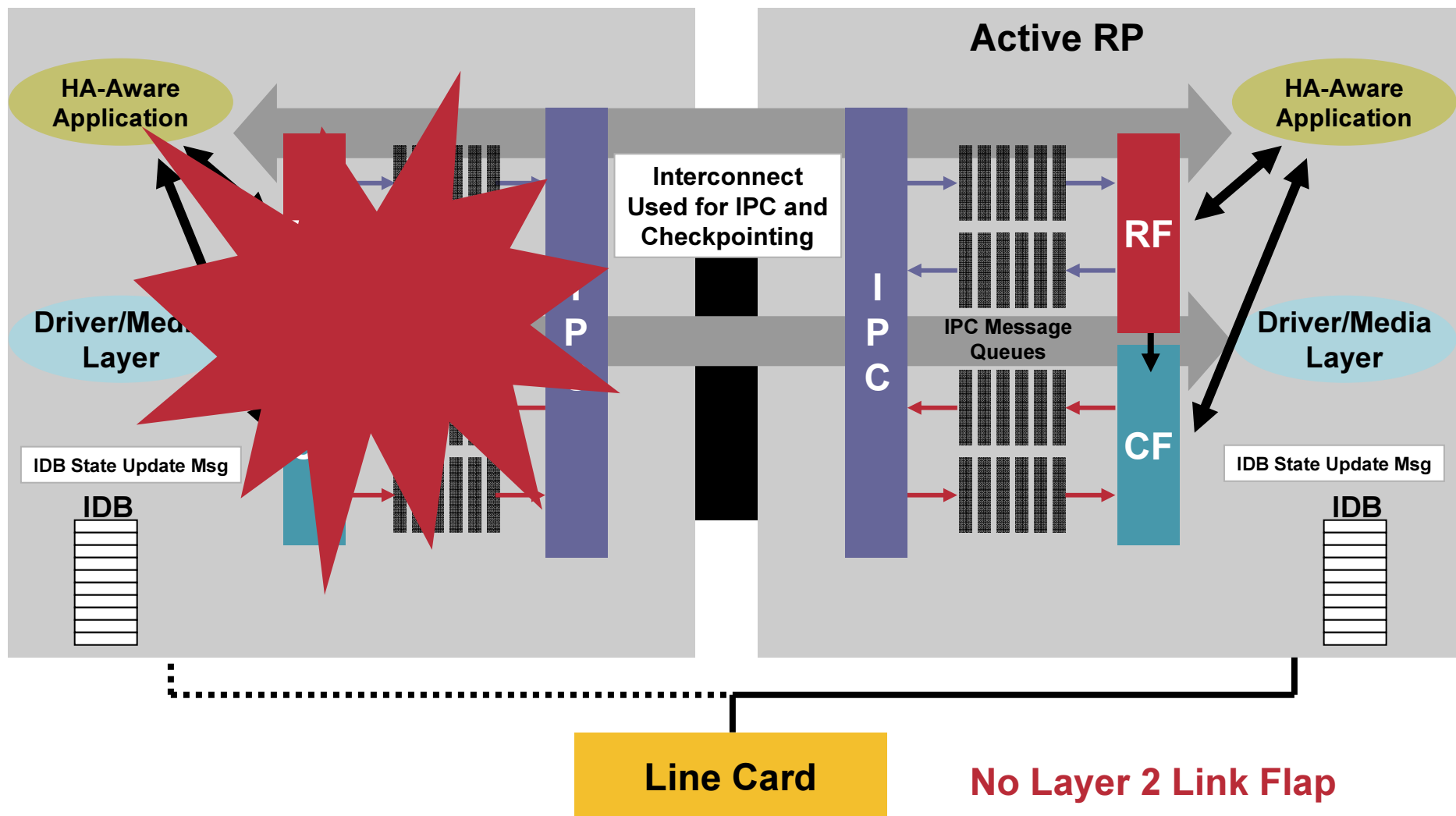
# SSO Architecture:

## HA-Aware, Stateful L2 Protocols (PPP, FR, HDLC ... )



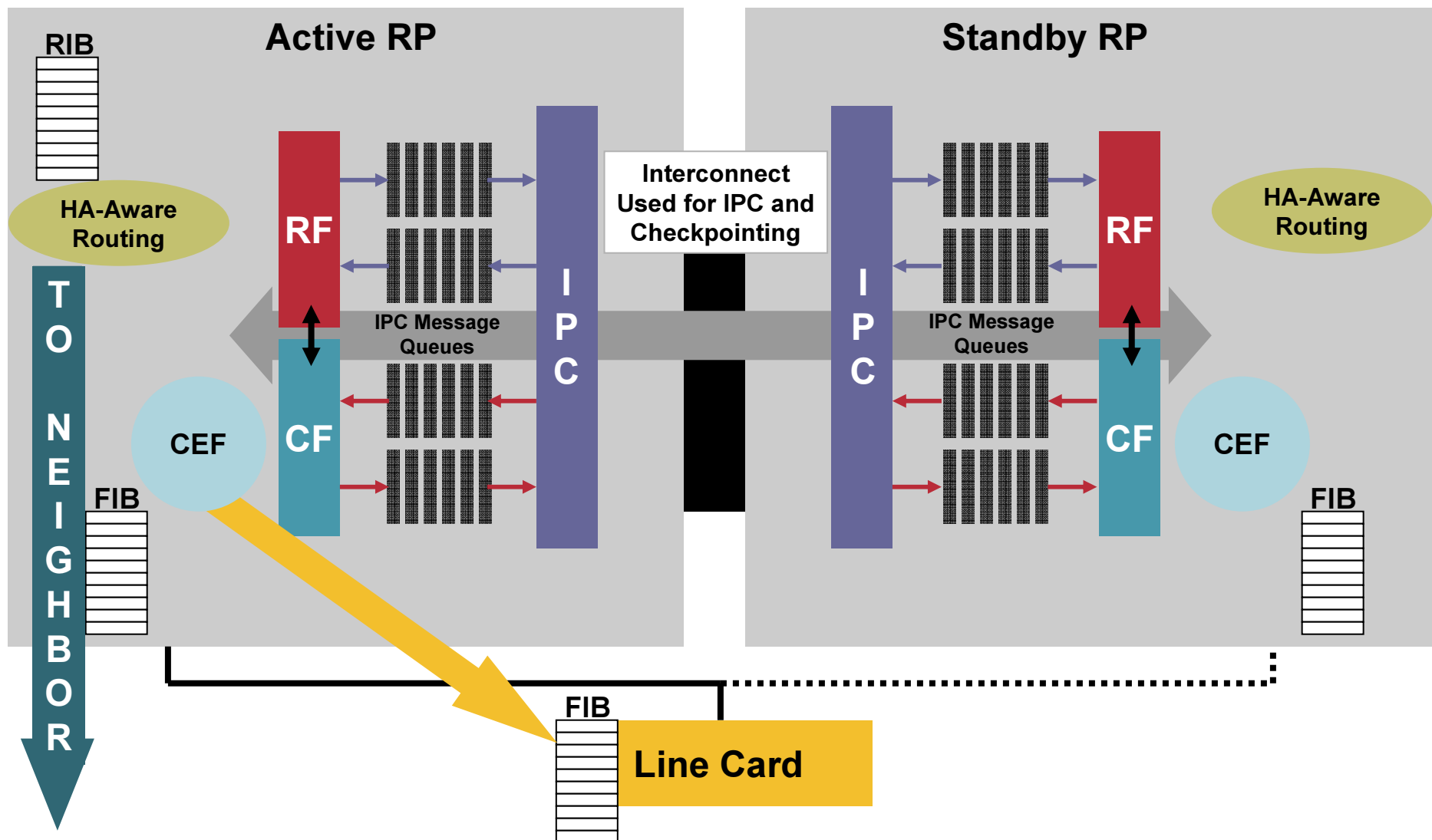
# SSO Architecture:

## HA-Aware, Stateful L2 Protocols (PPP, FR, HDLC ... )



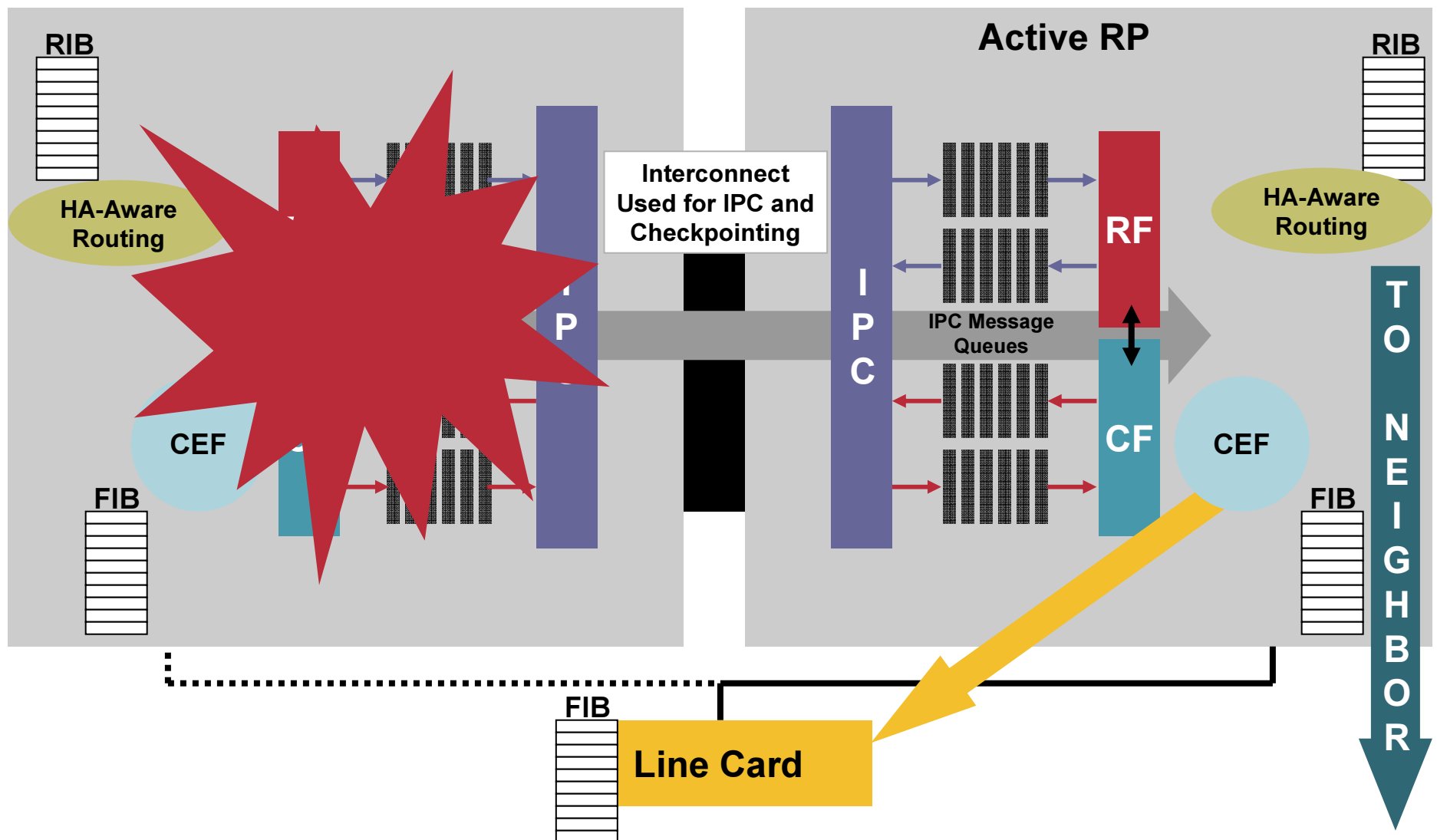
# SSO Architecture:

## HA-Aware, Non-Stateful Layer 3 Protocols (NSF)



# SSO Architecture:

## HA-Aware, Non-Stateful Layer 3 Protocols (NSF)



# Enabling SSO

- Enter redundancy configuration mode and set the redundancy configuration mode to SSO on both the active and standby RP

```
Router(config)# redundancy
Router(config-red)# mode sso
```

Note: standby will reset after this command

- This step is specific to the Cisco 7500 series devices only

```
Router(config)# hw-module slot slot-number image file-spec
```

slot-number—Specifies the active RSP slot where the Flash memory card is located

file-spec—Indicates the Flash device and the name of the image on the active RSP

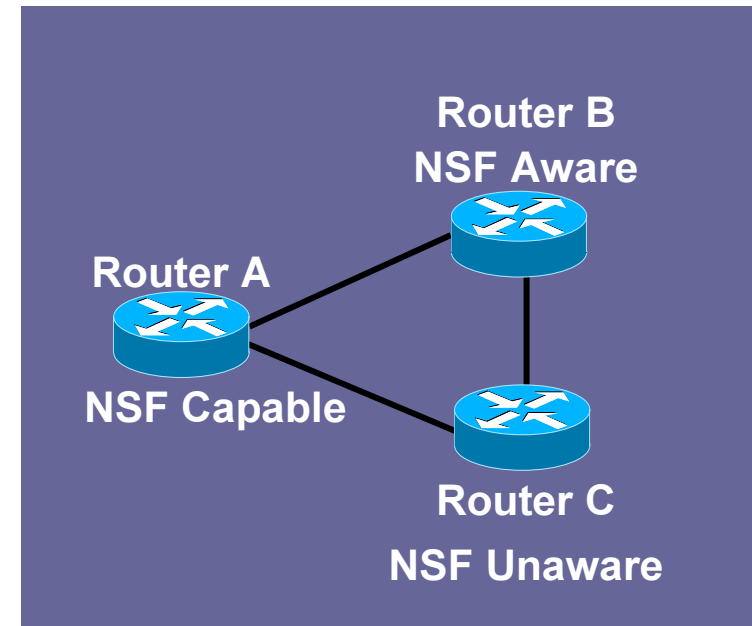
- Repeat command for standby RSP

```
Router(config)# hw-module slot slot-number image file-spec
```

# Make Sure We're on the Same Page

## Frequently Used Terms ...

- **NSF-capable router (restarting router)**  
A router that preserves its forwarding table and rebuilds its routing topology after an RP switchover; currently a dual RP router  
Ex: Cisco 7500,10000,12000
- **NSF-aware router (peer)**  
A router that assists an NSF capable during restart and can preserve routes reachable via the restarting router  
ex: Cisco 7200, 3600,2600,1700
- **NSF-unaware router**  
A router that is not capable of assisting an NSF-capable router during an RP switchover
- **NSF-capable router is NSF-aware, too!!!!**
- **SSO-aware or HA-aware**  
Cisco IOS subsystem—an HA client



### NSF—Non-Stop Forwarding

Cisco terminology and marketing name for feature set

### Graceful Restart (GR)

Term used in some protocol standards and drafts

# NSF Protocol Extensions

## Relevant Standards and Drafts

- The mechanisms used to provide continuous forwarding in the event of a route processor switchover are not completely standardized
- Cisco's implementation for BGP follows the specification described in draft-ietf-idr-restart-nn.txt

The latest version at the time of this writing was draft-ietf-idr-restart-10 and is a Proposed Standard

- Cisco's implementation for OSPF follows the specification described in the following IETF drafts:

<http://www.ietf.org/internet-drafts/draft-nguyen-ospf-lls-05.txt> (experimental)

<http://www.ietf.org/internet-drafts/draft-nguyen-ospf-oob-resync-05.txt> (informational)

<http://www.ietf.org/internet-drafts/draft-nguyen-ospf-restart-05.txt> (informational)

The current standard for OSPF Hitless Restart is RFC 3623 Graceful OSPF Restart. Cisco's implementation is currently (at the time of this writing) not interoperable with RFC 3623

- Cisco's NSF implementation for ISIS (ietf option) follows the specification described in RFC 3847 Restart Signaling for Intermediate System to Intermediate System (IS-IS)

A Cisco-specific stateful implementation is also supported and configurable

- Cisco's implementation for LDP follows the specification described in RFC 3478 Graceful Restart Mechanism for Label Distribution Protocol

# Relationship Building Exercise—1

**GR (NSF/SSO)  
Capable Router**

**GR (NSF) Aware Peer**

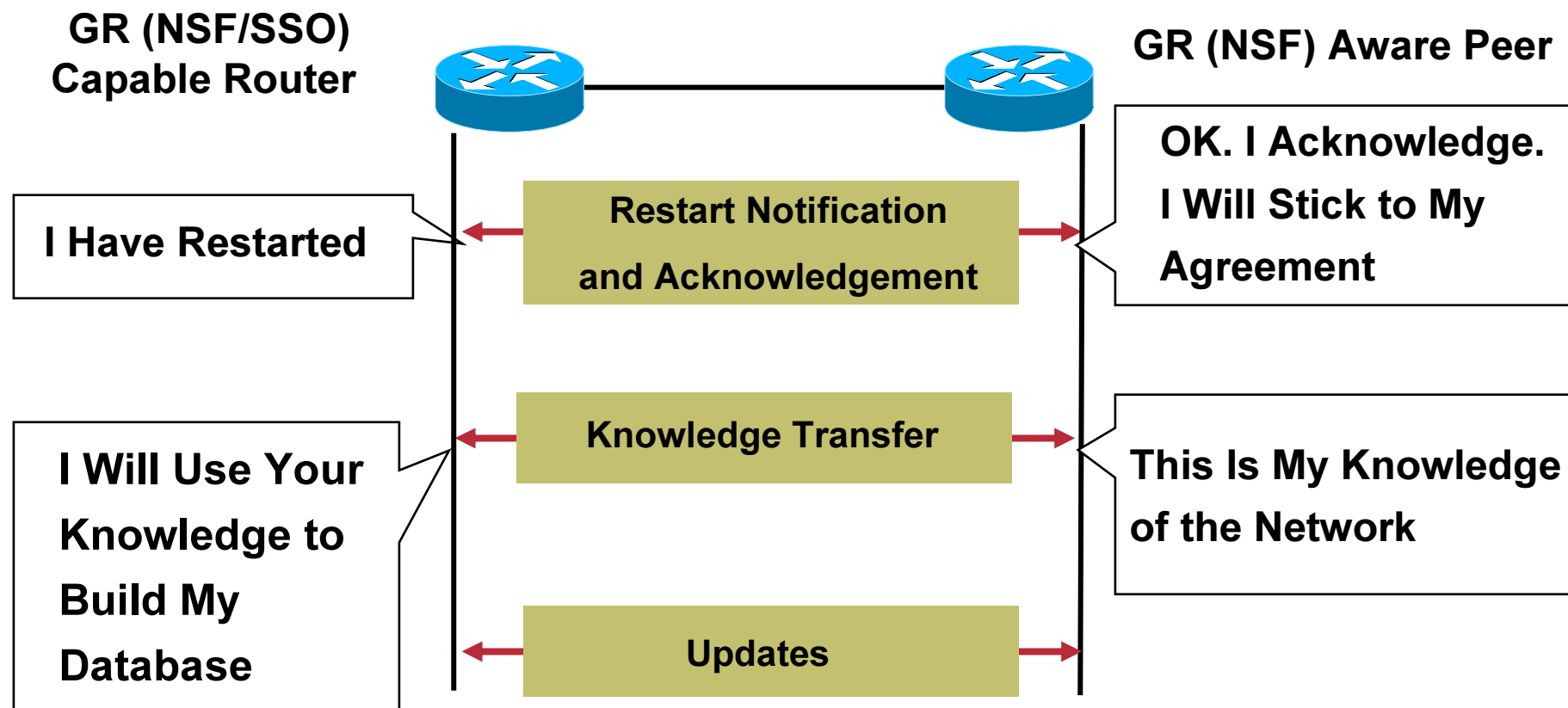
**“I Can Preserve My  
Forwarding Table  
During Restart”**

**Agreement**

**During Restart**  
– I Will Preserve My  
Forwarding Table  
– I Will Not Declare  
You Dead  
– I Will Not Inform  
My Neighbors



# Relationship Building Exercise—2



# Networks Without NSF/SSO and Graceful Restart

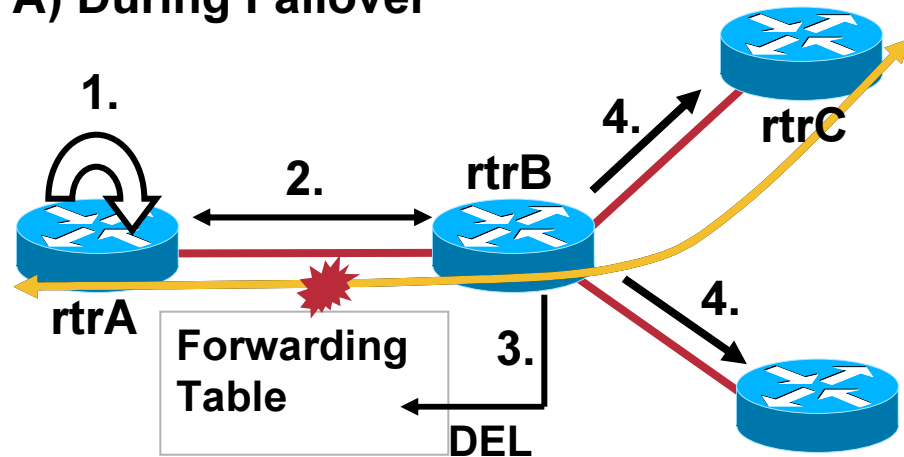
1

1. Router Restarts
2. Adjacency/Peer Relation Fails
3. Peer Removes All Associated Routes from the Routing Table
4. **Peer Informs the Neighbors about the Change**
5. Restarting Router Re-Establishes Adjacency
6. Peer Adds Associated Routes
7. **Peer Informs Neighbors about the Change**

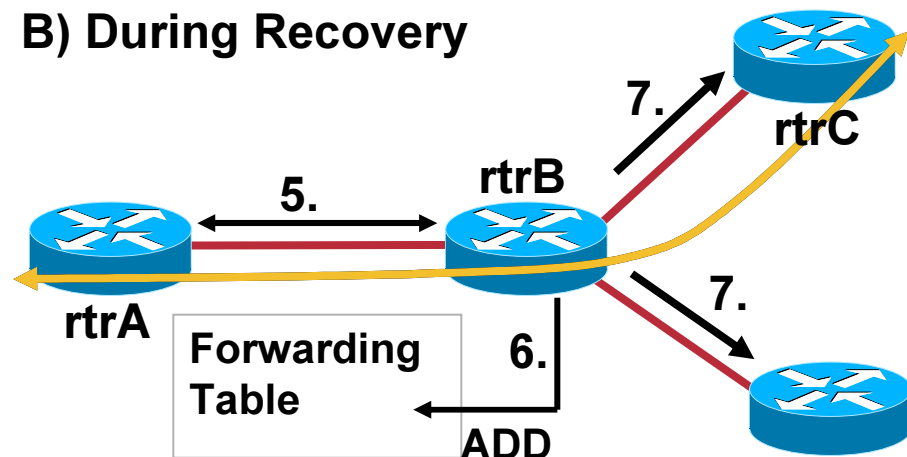
4, 7 Cause Route Flaps and Network Instability

Traffic Is Interrupted

## A) During Failover



## B) During Recovery

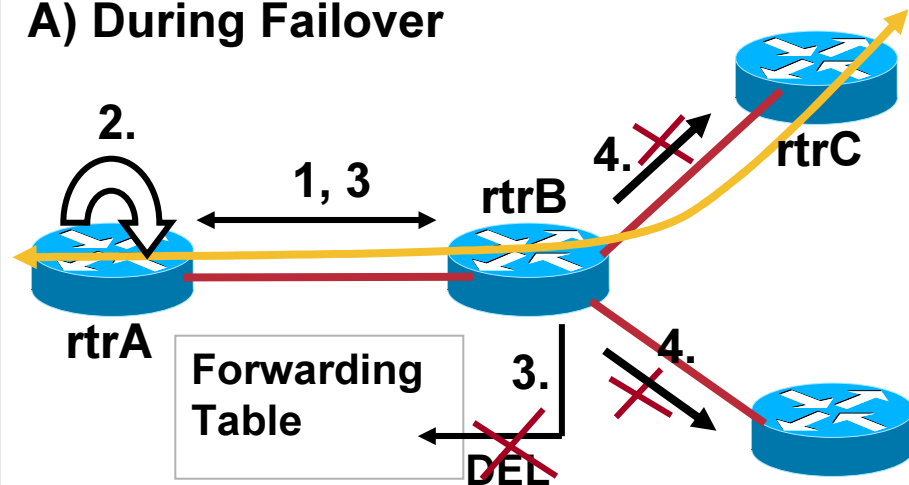


# Routing Protocol Operation with NSF/SSO and Graceful Restart

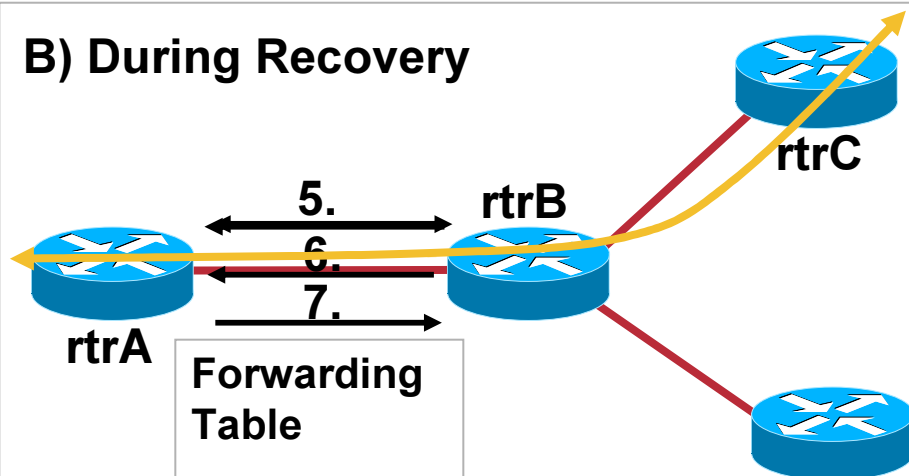
1. Routers Establish Peer Relation and Exchange Graceful Restart Capability
2. Router Restarts
3. Peer Relation Is Lost. Peer DOES NOT Remove Routes from Table
4. **Peer Does Not Inform Neighbors**
5. Restarting Router Re-Establishes Adjacency
6. Peer Updates Restarting Router with It's Routing Information
7. Restarting Router Sends Routing Updates to the Peer

**No Route Flaps During Recovery  
Traffic Flow Is Not Interrupted**

## A) During Failover



## B) During Recovery



# BGP NSF

## ① R1 established peering relationship with R2

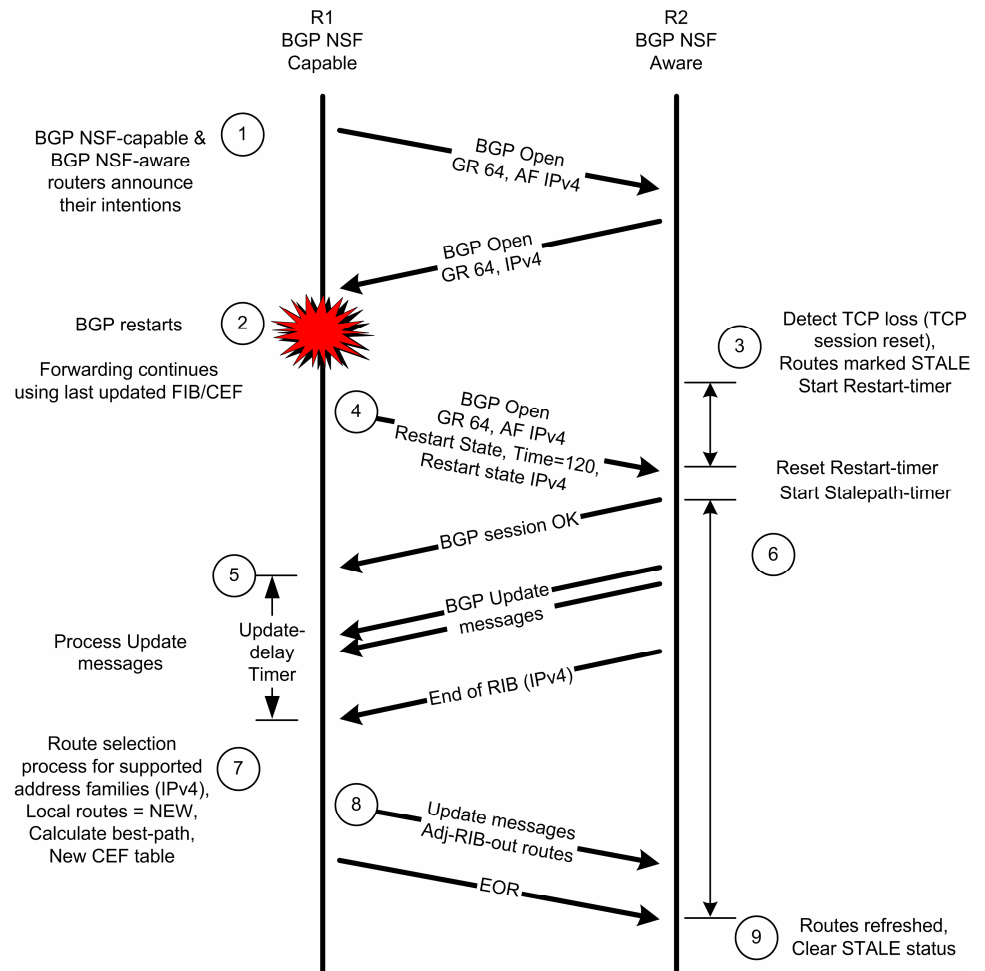
Open message (over TCP) contains GR capability code 64, AF IPv4, SAFI Unicast

R2 is BGP NSF-aware, so also includes GR 64 and AF IPv4

## ② Switchover happens

Newly active RP must acquire routing information base from peers;

Uses existing FIB and checkpointed CEF



# BGP NSF (Cont.)

- ③ R2 may detect that the TCP session has cleared; it immediately marks existing routes from R1 as stale

Starts restart-timer to wait for the OPEN from R1 (Default 120 seconds)

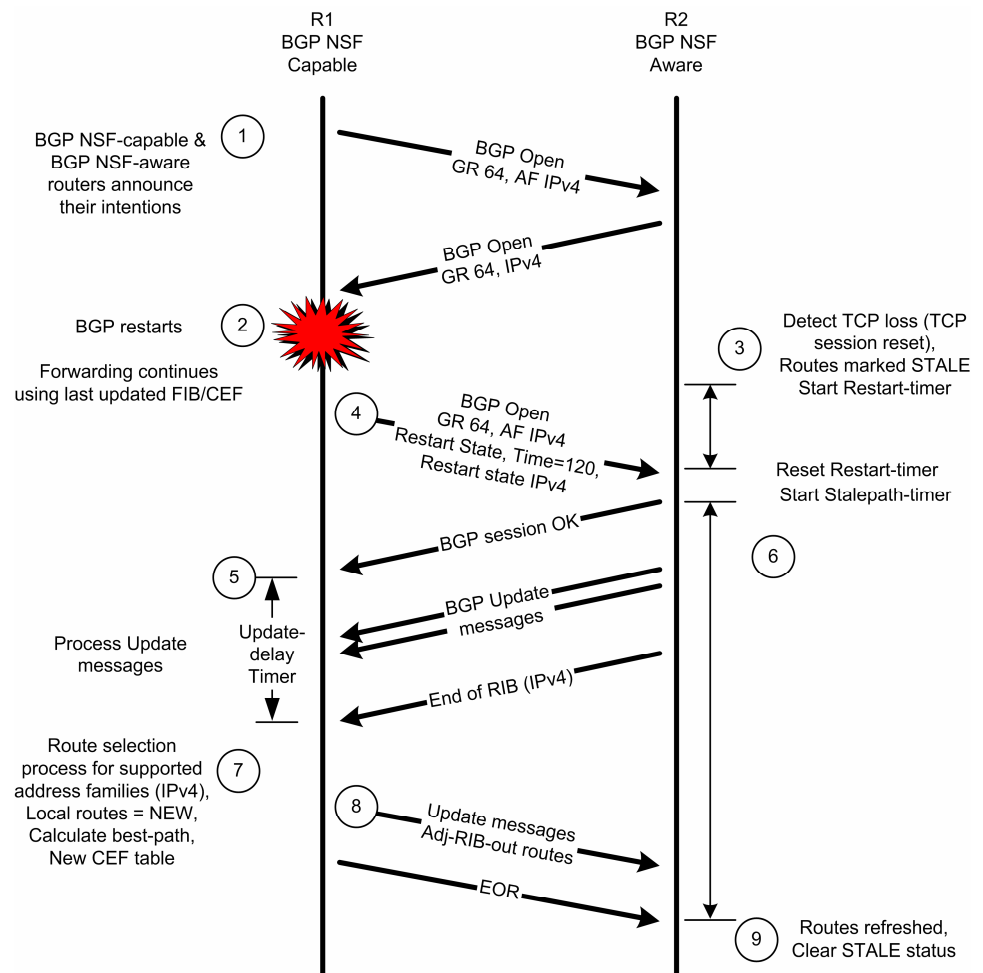
- ④ New TCP session established and OPEN sent with restart-state set

Also includes AF preserved (IPv4) and value of restart-timer

Open receipt, R2 starts Stale-path timer (Default 360 sec.)

- ⑤ New BGP session established

R2 knows forwarding state for AF IPv4 has been preserved



# BGP NSF (Cont.)

- ⑥ R1 starts an update delay timer and waits up to 120 seconds to receive end-of-RB (EOR) from peers**

R2 begins sending routing updates;  
R1 process accordingly

- ⑦ R1 will receive EOR when updates are complete**

Once EOR is received from all peers,  
it begins BGP route selection process

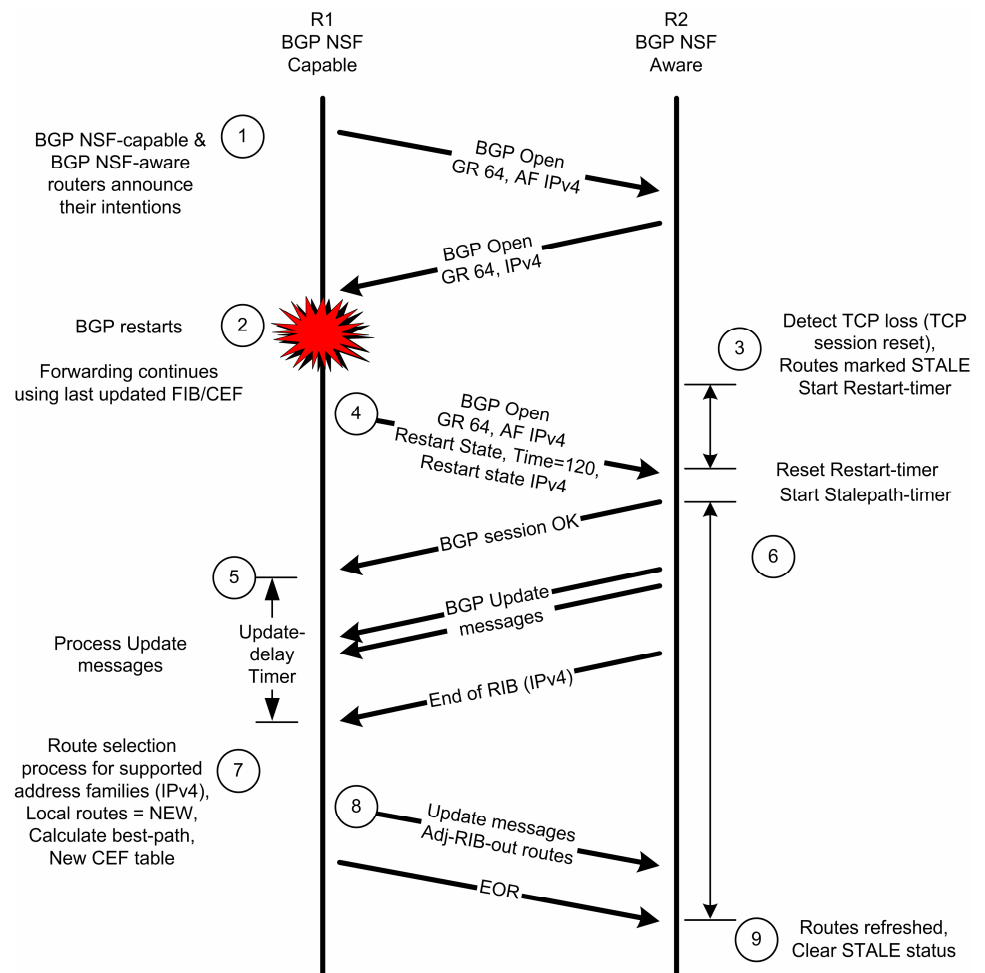
Following route selection process,  
BGP entries are refreshed

Any stale entries are removed

- ⑧ R1 then sends any updates it has to peers and ends with EOR**

- ⑨ R2: Following route selection process, BGP entries are refreshed**

Any stale entries are removed



# BGP NSF Configuration

- **BGP NSF (Graceful Restart) is configured under the global “router bgp” configuration command**

```
Router(config-route)# bgp graceful-restart
```

```
Router(config-route)# bgp graceful-restart restart-time n
```

```
Router(config-route)# bgp update-delay n
```

```
Router(config-route)# bgp graceful-restart stalepath-time n
```

**\*Note: bgp graceful-restart Must Be Configured for NSF Awareness Also**

# BGP NSF Timers

- **“bgp graceful-restart restart-time n”**

Maximum amount of time that a peer will wait for a reconnection of the TCP session and a new BGP OPEN message following the detection of a failure on the Restarting Router

If the TCP and BGP sessions are not re-established before this timer expires, the BGP session is deemed a failure, and normal BGP recovery procedures take effect

The default value for restart-time is 120 seconds

- **“bgp update-delay n”**

Applies to Cisco NSF-capable router

The update-delay specifies the time interval, after the first peer has reconnected, during which the restarting router expects to receive all BGP updates and the EOR marker from all of its configured peers

The default value of n is 120 seconds

If the restarting router has a large number of peers, each with a large number of updates to be sent, this value may need to be increased from its default value

- **“bgp graceful-restart stalepath-time n”**

Applies to NSF-aware peer(s) of the restarting router

Sets an upper limit on how long the peer will continue to use stale routes for forwarding after it has re-established the BGP session with the restarting router

The default value is 360 seconds

While this should allow an adequate amount of time to allow for complete convergence, on very large networks it may be necessary to increase this value



# OSPF NSF

## When Switchover Happens:

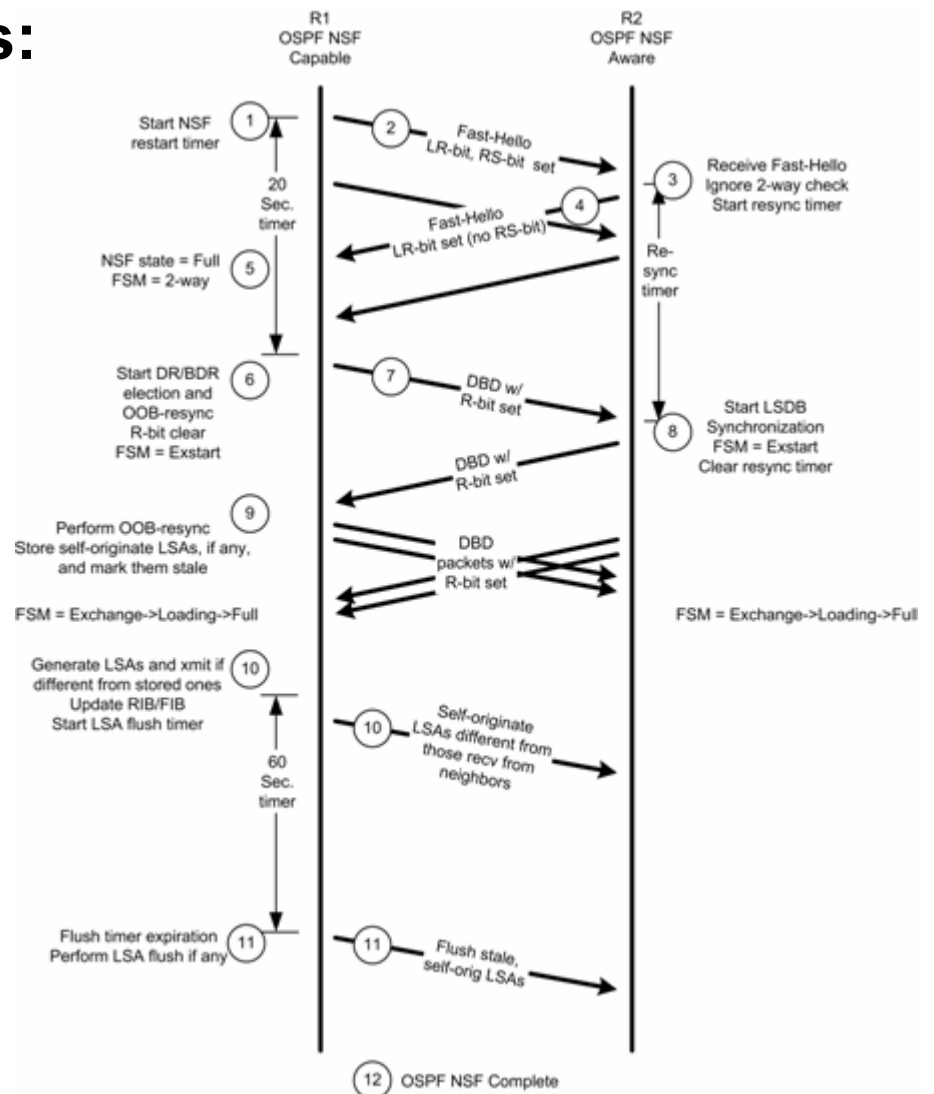
- 1 Existing routes marked stale; start NSF restart timer**

Timer used to wait until all Hellos are received from all neighbors before starting OOB-Resynchronization
- 2 R1 multicasts Hellos with RS bit set; LR-bit set**

LR bit indicates OOB Resynch capability; RS-bit means I'm restarting
- 3 R3 receives Hellos with RS-bit; knows the restart has occurred; ignores the two-way check**

OOB-Resync timer (Resync-Timeout) is started to limit a delay waiting for OOB resynchronization after receipt of Hello

Set to max of dead-interval or 40 seconds (configurable)



# OSPF NSF (Cont.)

## 4 Fast Hellos sent back from R2 to R1 without RS-bit

OSPF NSF-aware software always has LR-bit set

## 5 When R1 receives Hellos from R2, state goes to 2-way (but from NSF perspective state is considered Full)

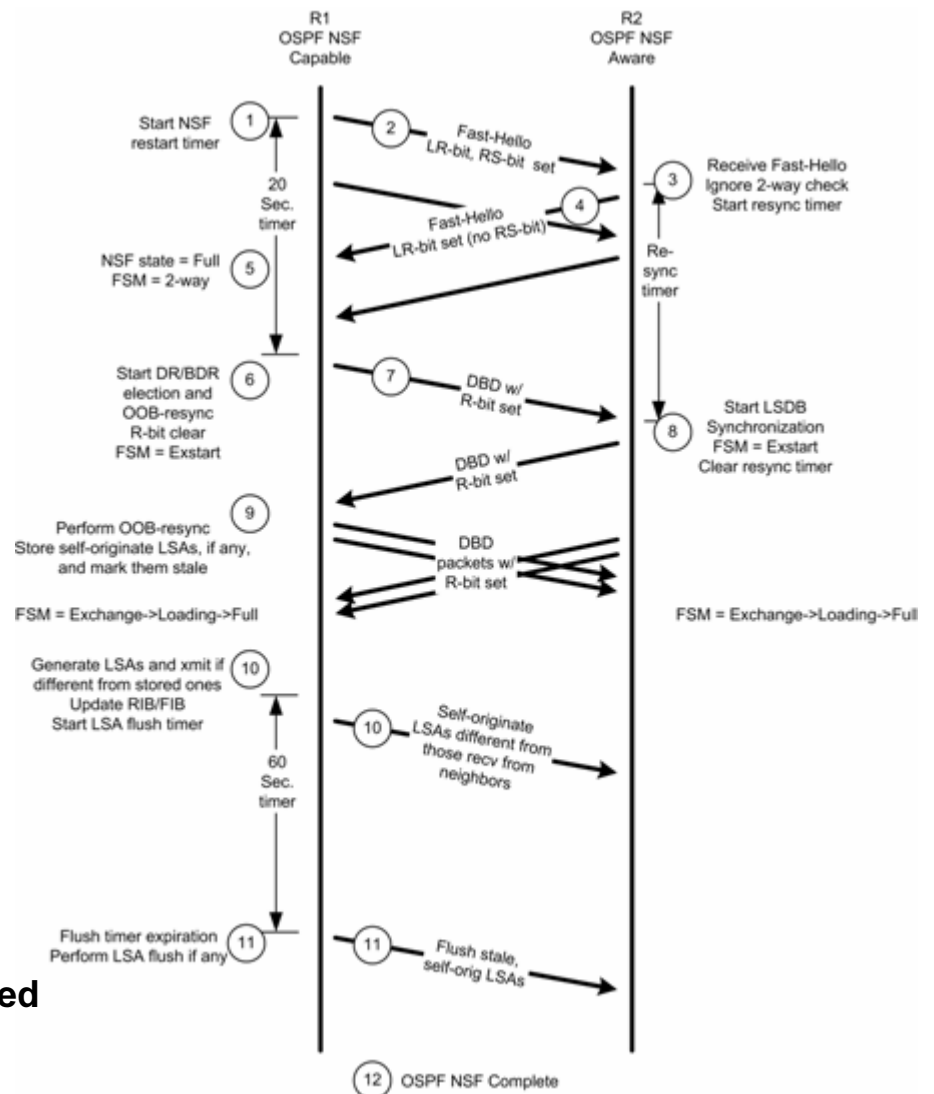
## 6 R1 waits for expiration of NSF Restart-Timer (20 Seconds)\*

“Wait time” ensures R1 learns all of its neighbors

Then starts DR/BDR election; state EXSTART and OOB-Resync begins

**Note:** Any Hello from NSF-Unaware Router Cancel NSF Is [enforce-global] Option Specified

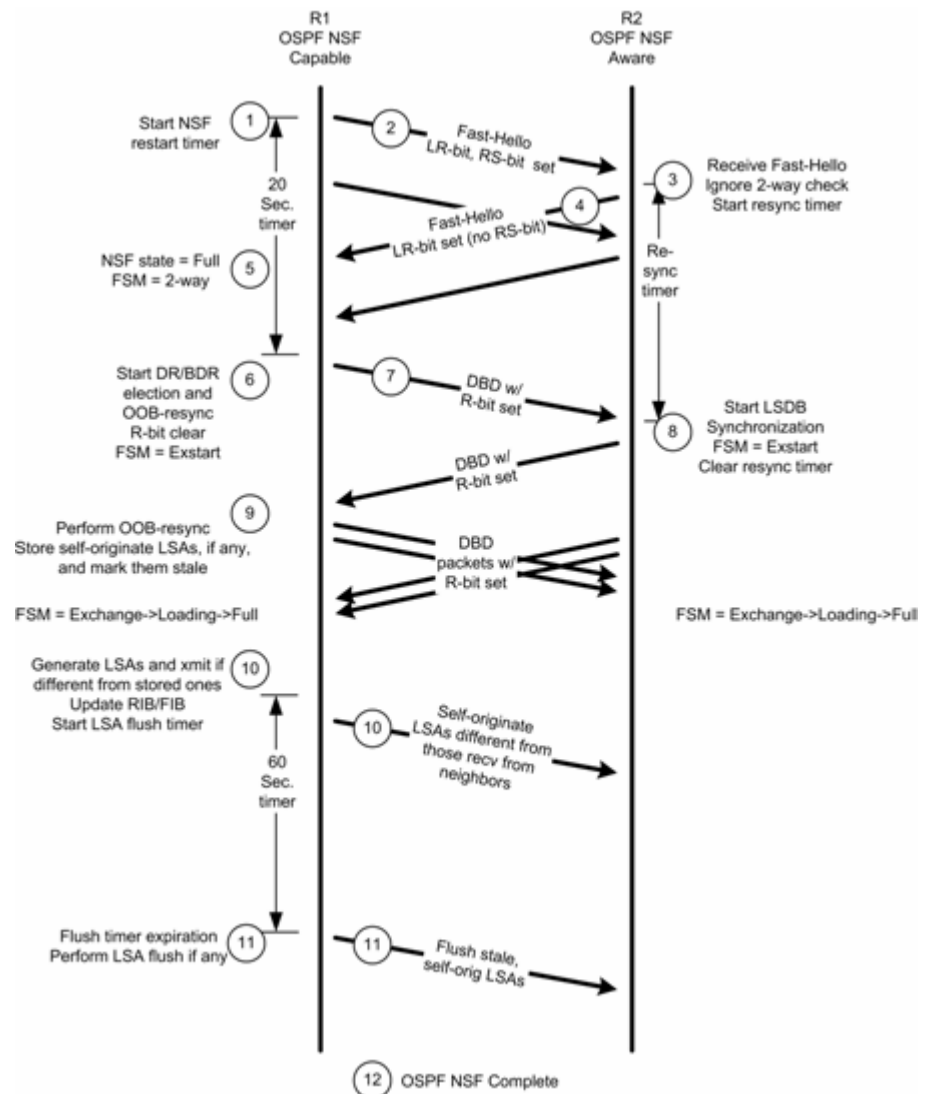
**\*Not Generally Changeable**



# OSPF NSF (Cont.)

- 7 R1 sends DBD packets with R-bit (meaning OOB-Resync procedure active)
- 8 R2 receives DBD packets from R1; goes to EXSTART; send DBD packets
- 9 Link State Database (LSDB) synchronization continues
- 10 OOB-Resync is complete; clear stale flags; generate any new and different LSAs
- 11 Flush timer expiration triggers flush of any remaining stale LSAs

Default 60 seconds



# Configure OSPF NSF

- **Enabling NSF Capability**

```
router(config)# router ospf 100
```

```
router(config-router)# nsf
```

- **You do not have to enable NSF awareness**

- **To optionally terminate OSPF NSF for entire router if an NSF-unaware peer is detected use:**

```
router(config)# router ospf 1
```

```
router(config-router)# nsf enforce global
```

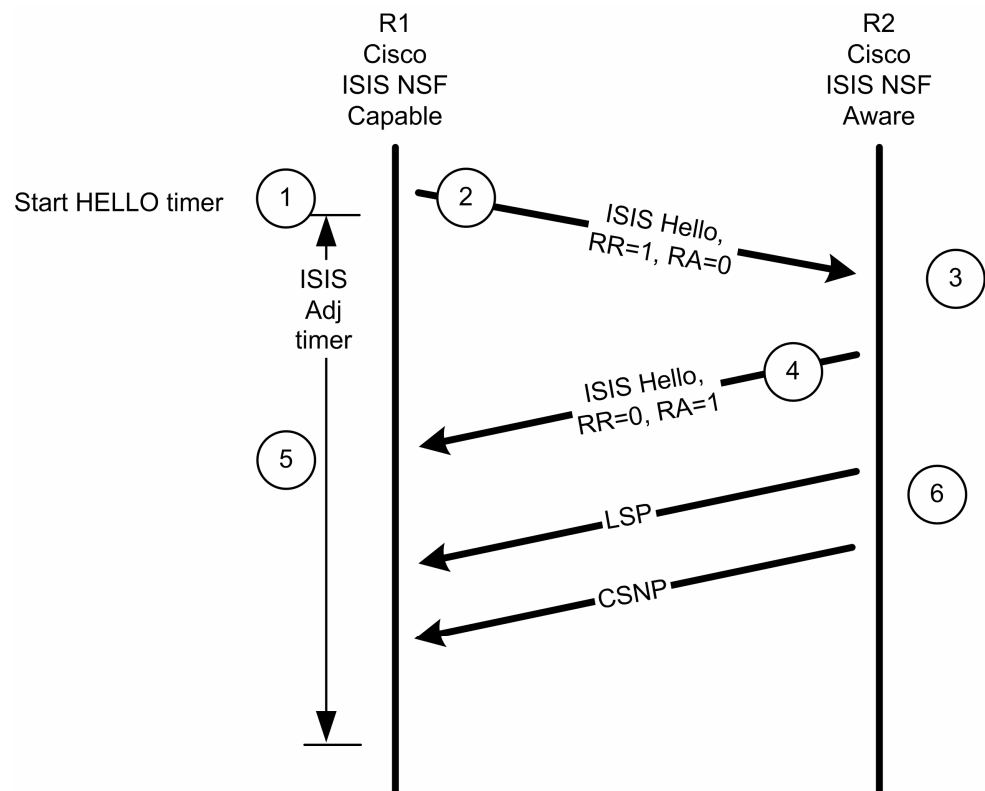
# IS-IS NSF

- 1 R1 Restarts
- 2 R1 sends Hello to R2 w/ new “restart option” TLV211, RR-bit set, RA-bit 0

Adjacency timer is started to limit the time for completion of database synchronization

- 3 R2 receives Hello and knows that R1 has restarted—Leaves adjacency “up”
- 4 R2 (is NSF-aware) sends Hello w/ TLV211, RR=0, RA=1
- 5 R1 receives Hello

Acknowledges the restart



\*RR= Restart Request; RA= Restart Acknowledgement

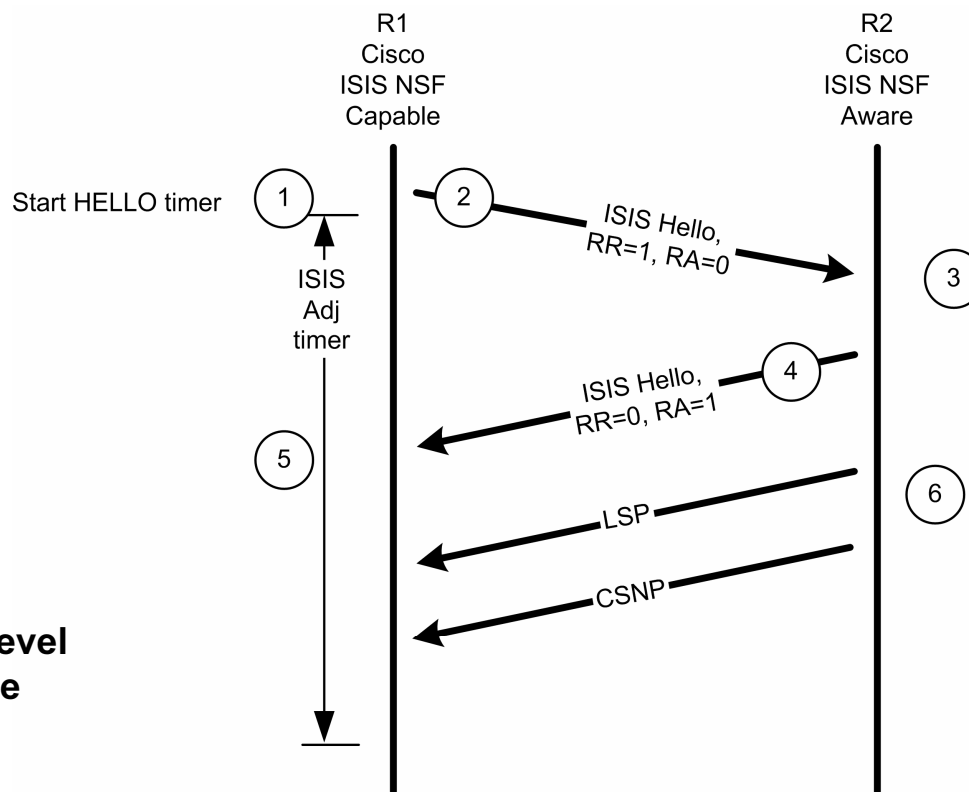
# IS-IS NSF (Cont.)

**6** R2 then clears the SRM flags that indicate routing data that needs to be sent to R1 and begins synchronization using complete sequence number PDUs and Link State PDUs

- When the synchronization is complete, the Adj timer is stopped
- SPF calculation is run; RIB updated, CEF notified
- CEF updates FIB, any routes not refreshed (stale) are removed

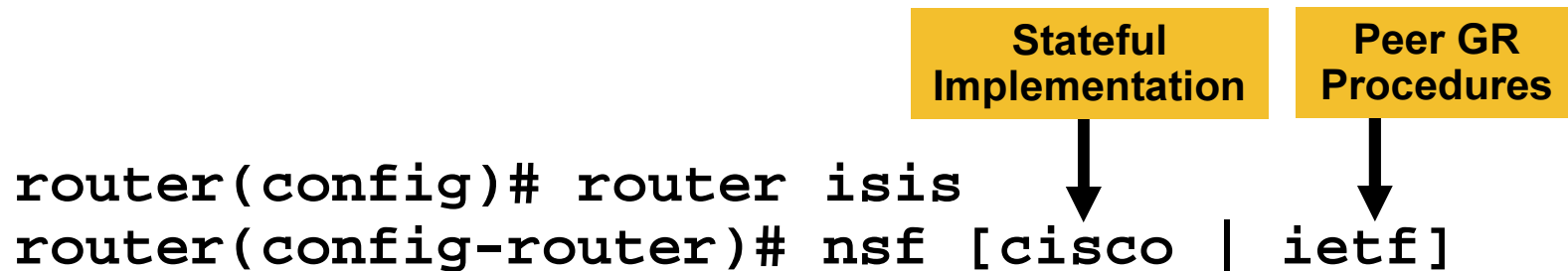
**\*Note—For LAN Interfaces, Level 1 and Level 2 Hellos Are Sent and Complete Sequence Number Protocol Data Units (CSNP) Synchronization Is Done Independently**

**For Point-to-Point Interfaces, Only a Single Hello Is Needed, but Synchronization Is Still Performed at Each Level; L1 and L2**



# IS-IS Stateful Implementation

- Cisco also support a stateful implementation of IS-IS
- Full adjacency and LSP info is checkpointed
- Following a switchover, newly active RP uses checkpointed data and quickly restores the routing information
- Using the Cisco stateful mode of operation ...
  - Hellos are sent from the newly Active RP prior to peer hold timer expiration
  - And with appropriate information so the neighbor is unaware of the restart
- Database resynchronization and verification still occurs
  - Done in an innovative way and performed without triggering a topology change



# IS-IS NSF Configuration and Timers

- **Enable IS-IS NSF**

```
router(config)# router isis
router(config-router)# nsf [cisco | ietf]
```

- **The following command limits the interval (in a 0-1440 minutes range) between two restarts; the default value is 5 minutes**

```
router(config)# router isis
router(config-router)# nsf interval 600
```

- **The following command sets the time (in a 1-60 seconds range) an NSF restart will wait for all interfaces with ISIS adjacencies to come up before completing the restart; the default value is 10 seconds**

```
router(config)# router isis
router(config-router)# nsf interface wait 20
```

- **In IETF mode and only in this mode, the following command sets the time (in seconds) NSF will wait for the LSP database to synchronize before generating and flooding its own LSP with the overload-bit set**

```
router(config)# router isis
router(config-router)# nsf t3 manual 60
```

- **If the “adjacency” keyword is used, this above mentioned t3 time would be determined from the adjacency holdtime advertised to neighbors prior to switchover**

```
router(config)# router isis
router(config-router)# nsf t3 adjacency
```



# EIGRP NSF

## 1 Switchover happens

R1 initiates Hellos from newly Active RP with RS-bit set

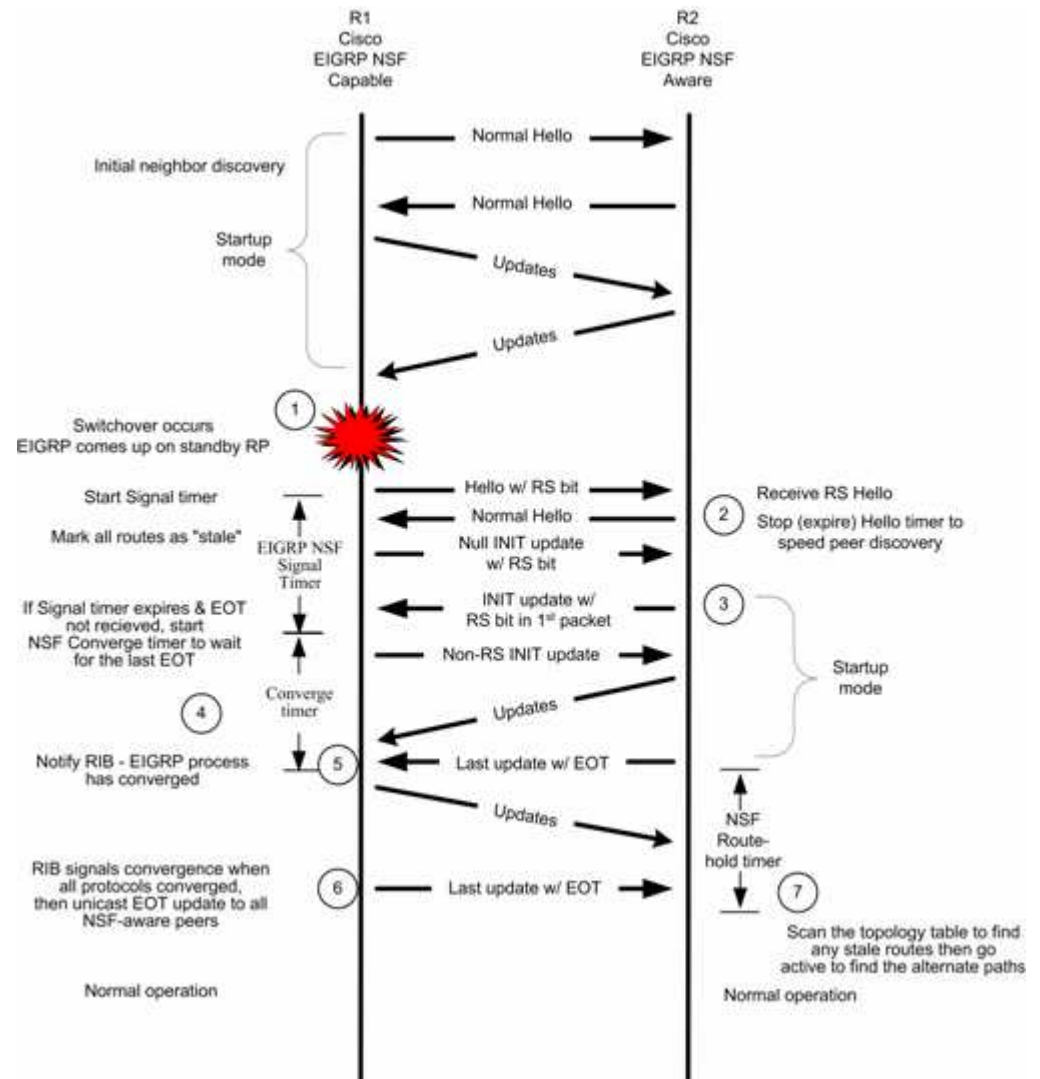
Restart (RS) bit indicates to peer that it should follow NSF extensions and NOT use normal adjacency discovery and startup method

## 2 R2 recognizes restart and retains forwarding state for the restarting router

Hello packets are returned immediately

## 3 R2 sends topology table w/ RS-bit in INIT indicating that it is aware of the restart and is helping to provide routing info

## 4 R1 uses signal-timer and Converge-timer to wait for EOT



# EIGRP NSF (Cont.)

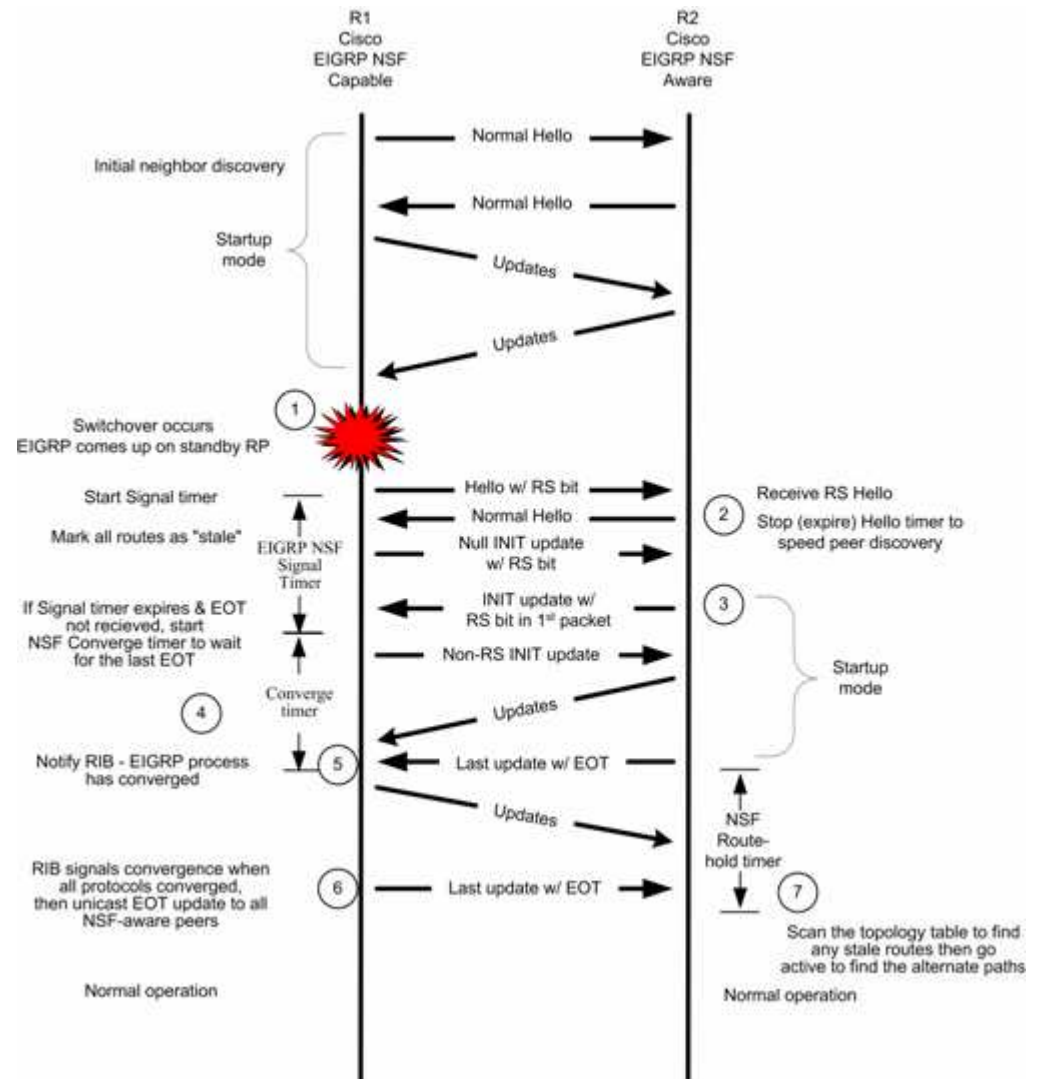
## 5 R1 receive End of Table marker

R2 starts route-hold timer to set upper bound to wait for EOT from R1

When R1 has received all EOT from peers it performs Diffusing Update Algorithm (DUAL) and notifies RIB of convergence

## 6 When RIB has received convergence notification from all protocols in use, EIGRP is notified and EIGRP sends EOT to peers

## 7 R2 then knows R1 has converged; R2 scans topology table to verify routes; goes Active to find alternate paths for any routes that are no longer available



# EIGRP Configuration

- **EIGRP NSF is disabled by default**

**router eigrp <AS-number>**

**[no] nsf**

- **Timers may be specified using the commands:**

**router eigrp <AS-number>**

**[no] timers nsf signal <seconds>**

**[no] timers nsf converge <seconds>**

**[no] timers nsf route-hold <seconds>**

# EIGRP NSF Timers

- **Signal timer**

Each EIGRP process starts a signal timer when it is notified of a switchover event  
Hellos with the RS bit set will be sent during this period

- **Converge timer**

The Converge timer may be used to wait for the last EOT update if all startup updates have not been received within the signal timer period

If an EIGRP process discovers no neighbor, or if it has received all startup updates from its neighbor within the signal timer period, the Converge timer will not be started

- **Route-Hold timer**

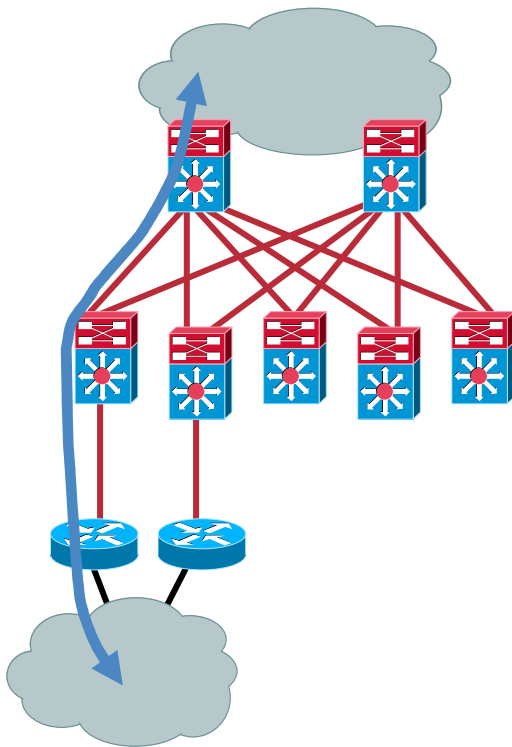
A NSF-aware peer will start the Route-Hold timer to wait for the EOT from the restarting router

At the end of this timer period, the peer will stop waiting, start scanning the topology table, and go active on those routes that were not updated by the restarting router

The Route-Hold timer may be tuned (shortened) so the peer can find alternate paths faster and avoid black holing traffic if the restart period is too long

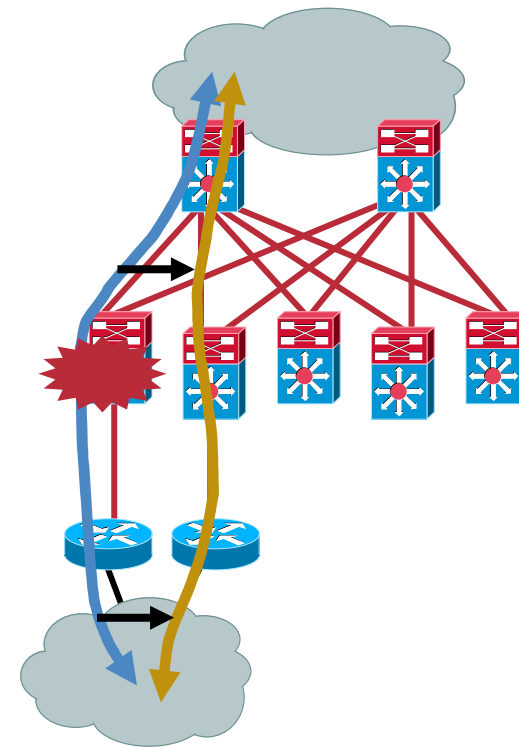
# IGP Timer Manipulation and NSF/SSO

- OSPF, ISIS, EIGRP maintains and verifies neighbor adjacency through periodic transmission of Hellos
- Different but similar terminology used—dead-interval, hold-timer, holdtime



**NSF/SSO Seeks to  
Preserve Traffic  
Forwarding**

**Fast Convergence  
Seeks to Shift  
Traffic Quickly to  
an Alternate Path**



# Dead-Interval or Hold Times



- Timers too short negate NSF
- Time begins after peer sends Hello—Hold-timer started  
Peer perspective determines if NSF will continue
- Some examples below:  
Recommendations is to go no less that 4 seconds if you want to retain forwarding through a router undergoing a switchover  
EIGRP similar—recommendation is 2 seconds hello, 6 second holdtime

First Hello Sent	ISIS (IETF)	ISIS (Cisco)	OSPF
Cisco 10000	2.020	2.016	2.016
Cisco 12000	2.200	2.292	2.276

First Hello Received	ISIS (IETF)	ISIS (Cisco)	OSPF
Cisco 10000	2.088	2.420	2.604
Cisco 12000	2.948	2.560	2.376

## 7500 Test

Dead Interval \ Hello Interval	1	2	3	4	5	6
4	✗					
8		✓				
12			✓			
16				✓		
20					✓	
24						✓

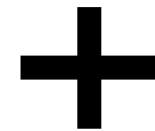
 Met Design Criteria
  Did Not Meet Design Criteria

# MPLS Nonstop Forwarding and Stateful Switchover

- **MPLS High Availability** targeted mainly toward Service Provider PE devices
- **MPLS HA features extend Cisco NSF with SSO capabilities for:**
  - Label Distribution Protocol (LDP)
  - MPLS Forwarding
  - Virtual Private Networks (VPNs)
- **Offers minimal disruption to MPLS forwarding plane due to route processor control plane failures**
  - Includes MPLS control plane failures (LDP,BGP)

## MPLS HA

**LDP  
MP-BGP**



## IP HA

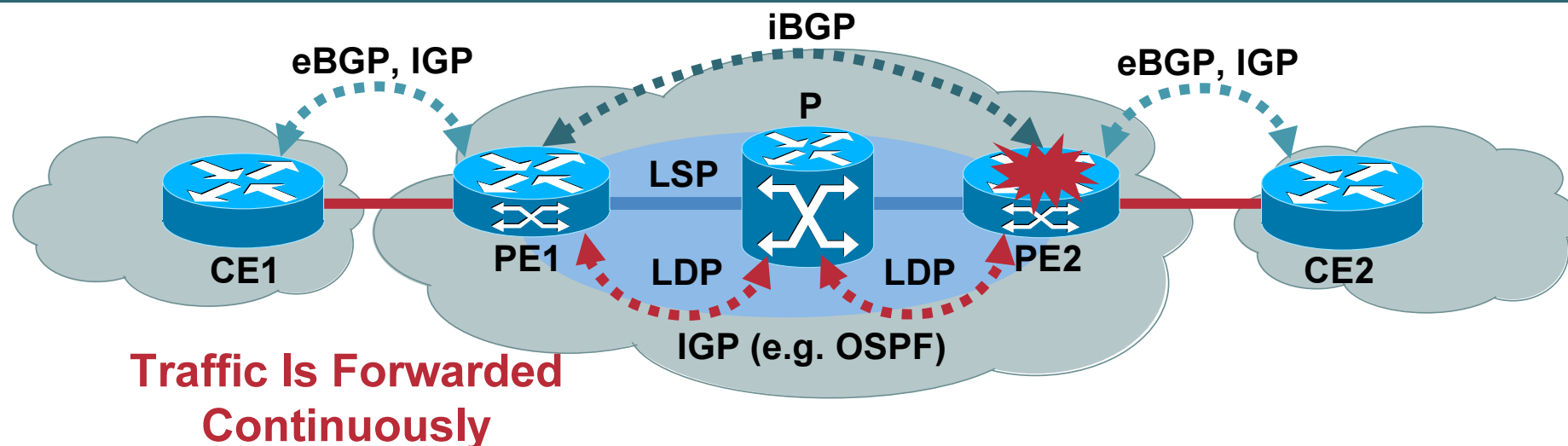
**BGP  
IS-IS  
OSPF  
EIGRP**

# MPLS NSF/SSO Operation

- **When a router that is capable of BGP Graceful Restart loses connectivity, the following happens to the restarting router:**
  - The router **establishes BGP sessions** with other routers and **relearns the BGP routes** from other routers that are also capable of Graceful Restart
  - The restarting router **accesses the checkpoint database to find the label that was assigned for each prefix**
    - If it finds the label, it advertises it to the neighboring router; if it does not find the label, it allocates a new label and advertises it
  - The restarting router removes any stale prefixes after a timer for stale entries expires
- **When a peer router that is capable of BGP Graceful Restart encounters a restarting router, it does the following:**
  - The peer router sends all of the routing updates to the restarting router
  - When it has finished sending updates, the peer router sends an end-of RIB marker to the restarting router
  - The peer router **does not immediately remove the BGP routes learned from the restarting router** from its BGP routing table; as it learns the prefixes from the restarting router, the peer **refreshes the stale routes if the new prefix and label information matches the old information**

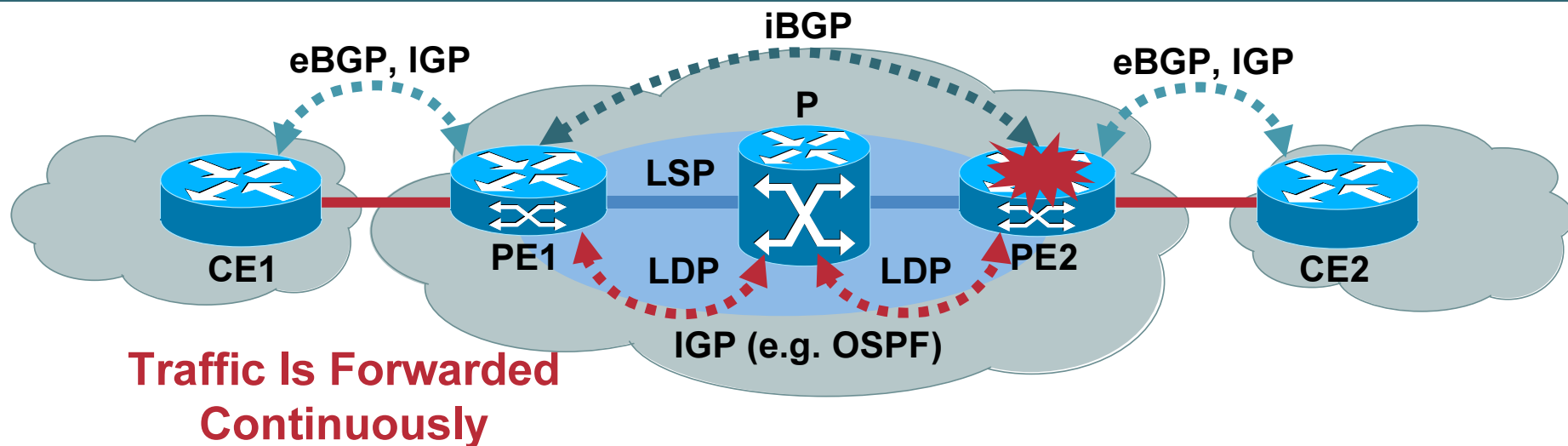


# MPLS VPN— BGP Graceful Restart Procedure



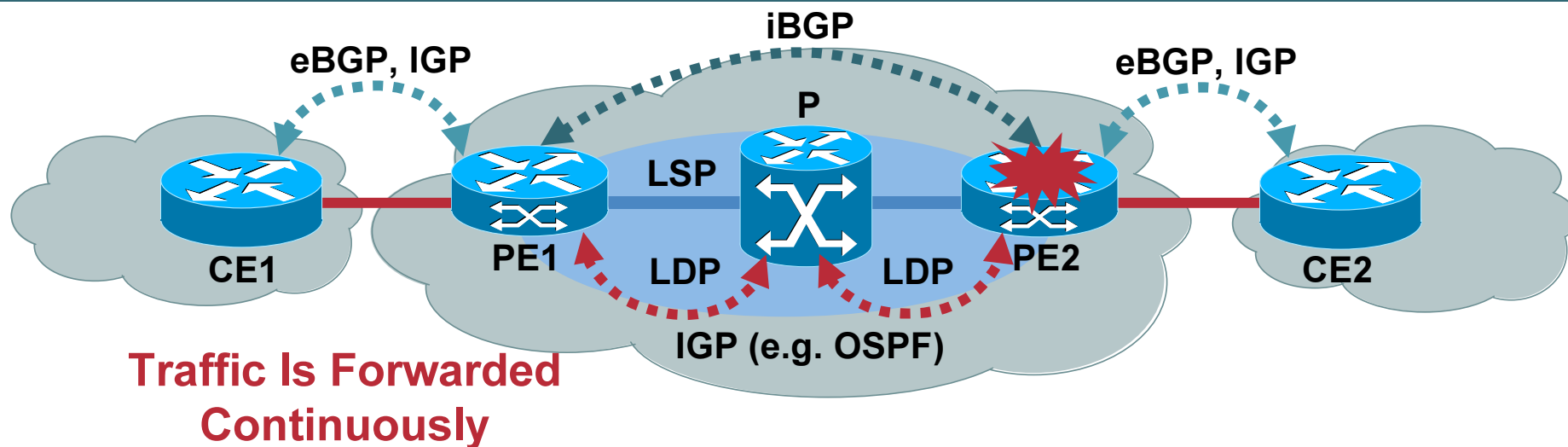
1. PE1 and PE2 are IBGP neighbors and exchange VPNv4 routes
2. Since PE1 and PE2 are configured for BGP GR, they also exchange the GR capability in the OPEN messages they send to each other during BGP session initialization
3. PE2 is the router which will be restarted (active RP fails, switches over to backup RP)
4. PE2 syncs the **local label to prefix mapping** in its BGP VPN table to the standby RP

# MPLS VPN— BGP Graceful Restart Procedure (Cont.)



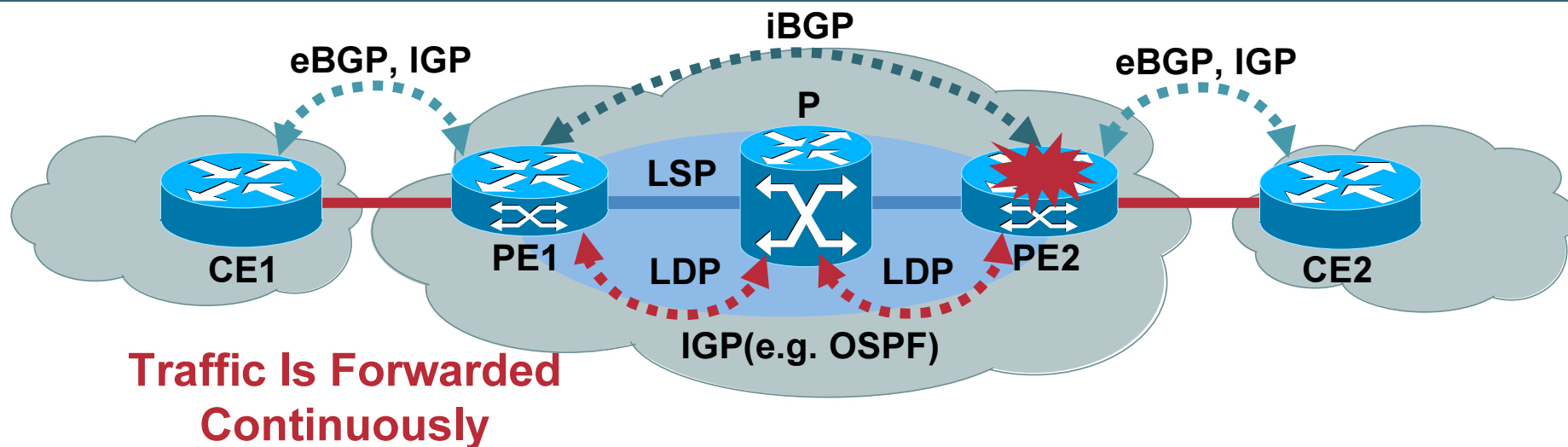
5. CEF table and the Label Forwarding Database (LFD) are also synced to the standby RP
6. Label switching database (LSD) which is responsible for label allocation to the LDM (Label Distribution modules) also syncs over blocks of allocated labels to the standby RP; it does this so that after switchover, the new active RP does not allocate the already allocated labels to another LDM or another prefix
7. Now a switchover happens on PE2
8. The BGP session between PE1 and PE2 goes down

# MPLS VPN— BGP Graceful Restart Procedure (Cont.)



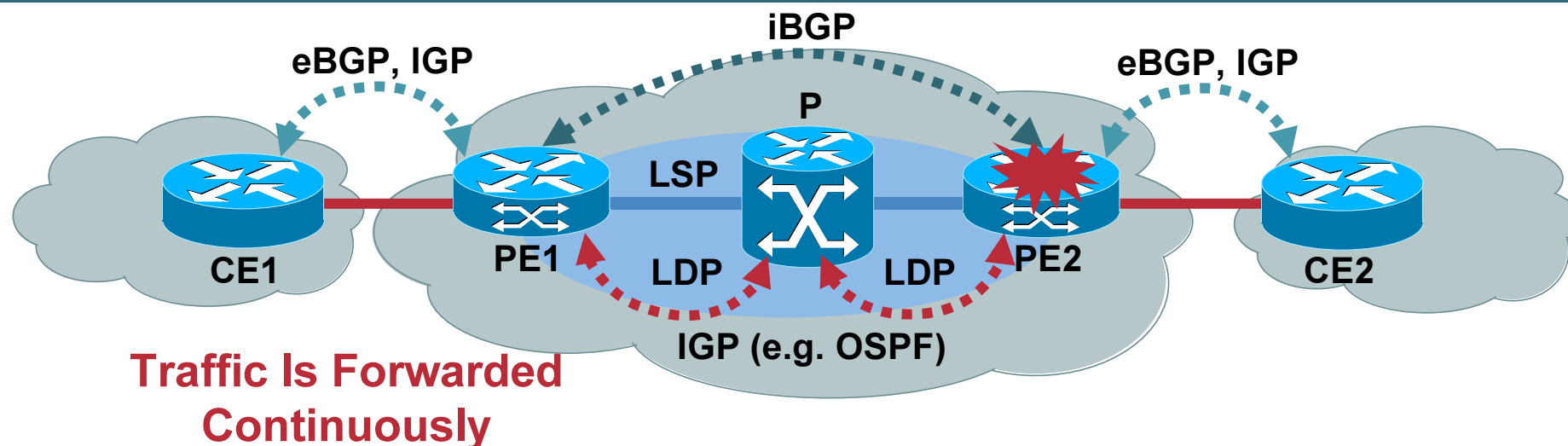
9. PE1 marks the entries in its BGP table which it learned from PE2 as stale but does not delete the routes from its VRF tables; hence it continues to forward traffic to PE2
10. PE2 also maintains its forwarding capability by maintaining its CEF and LFD on the line cards; hence it is capable of forwarding traffic arriving from CE and going towards PE1 as well as traffic coming from PE1 and going towards CE
11. The BGP session between PE1 and PE2 comes back up. PE1 needs to see the session come back up within the restart time (default 120s); if not, it is going to delete all the stale routes from the BGP table and hence the routing table

# MPLS VPN— BGP Graceful Restart Procedure (Cont.)



12. Once the BGP session comes back up, PE1 advertises all the routes in its Adj-RIB-out to PE2 along with the label mapping
13. PE2 receives these updates which contain the prefix to outgoing label mapping. BGP has synced over the prefix to incoming label mapping to the RP prior to switchover
14. BGP on PE2 will wait for all of its restarting peers to complete resending their updates to PE2; when all the updates are received, BGP starts its route selection process
15. BGP on PE2 now lets IPRM (IP Resource Manager) know that it has learned a new outgoing label for the prefix; IPRM is the module which sits in the middle of the LDM (LDP, TE, VPN, etc.) and MFI and handles the interaction between them

# MPLS VPN— BGP Graceful Restart Procedure (Cont.)



16. IPRM now installs the rewrite (incoming and outgoing labels for the prefix) into the LSD which then distributes it to the LFD on the LC
17. LFD on the LC installs the rewrite into the CEF on the LC
18. On PE2, after BGP has run its route selection process, it populates its Adj-RIB-out which it advertises to PE1
19. Once PE1 receives the updates from PE2, it removes the stale marking from the BGP prefixes; if PE1 does not receive these updates within the stalepath time (360s by default), it deletes all its stale entries from its BGP table and hence the routing table

# LDP NSF/SSO or Graceful Restart

- **Allows Label Switching Routers (LSRs) to maintain LDP and forwarding state when communication between them is lost**

**Due to LDP restart**

**Or LDP session reset**

- **Delivers higher availability for traffic that is switched using labels**
- **LDP NSF/SSO operation is similar to other NSF protocols**

**At least one LDP peer is LDP GR capable**

**Other is (at least) LDP GR-aware**

# LDP NSF (Graceful Restart)

- LDP uses TCP like BGP

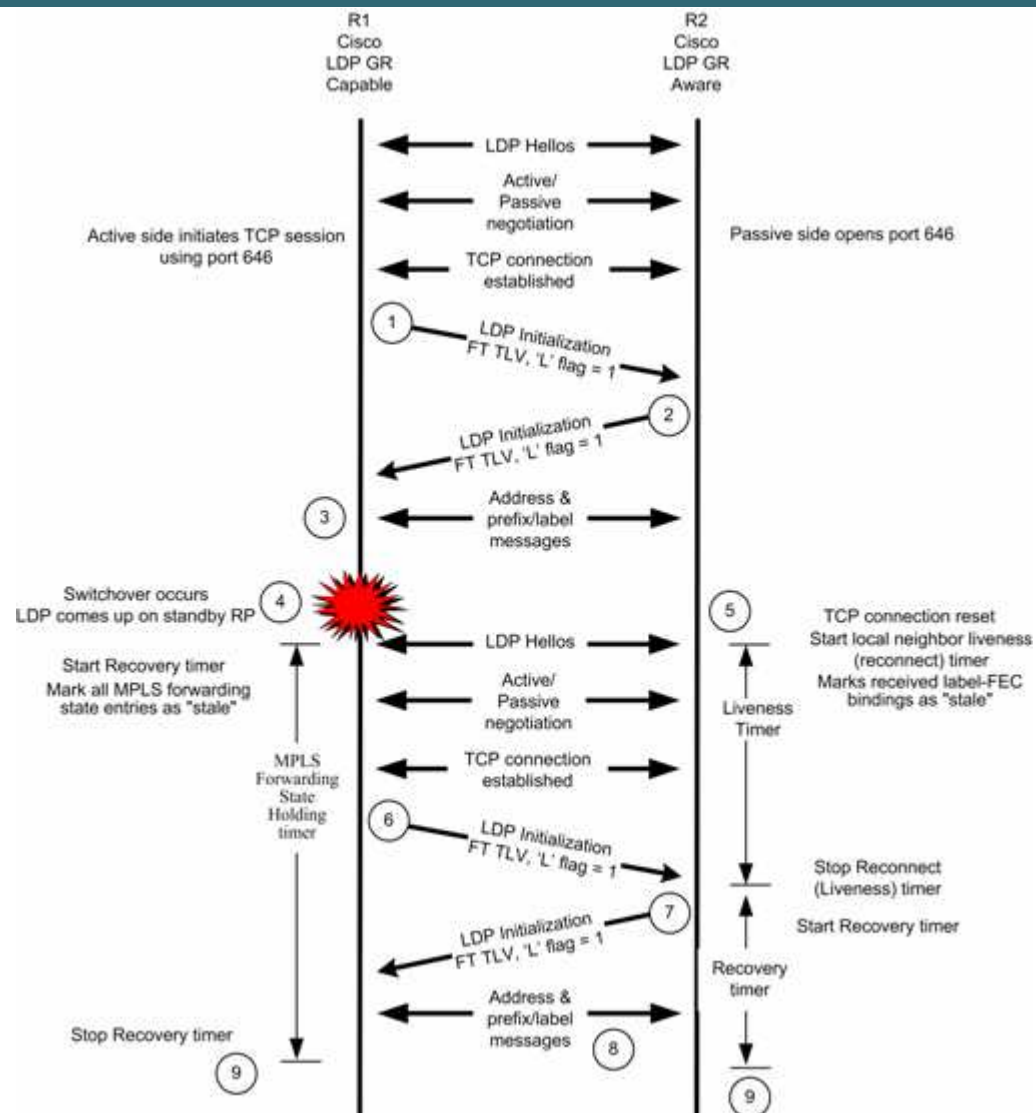
- LSR indicates it is capable of GR by including Fault Tolerant TLV in the LDP Init msg

The "L"-flag, learn from network indicates that LDP GR procedures are used

- R2 indicates its support for LDP GR procedures

- Label information exchanged

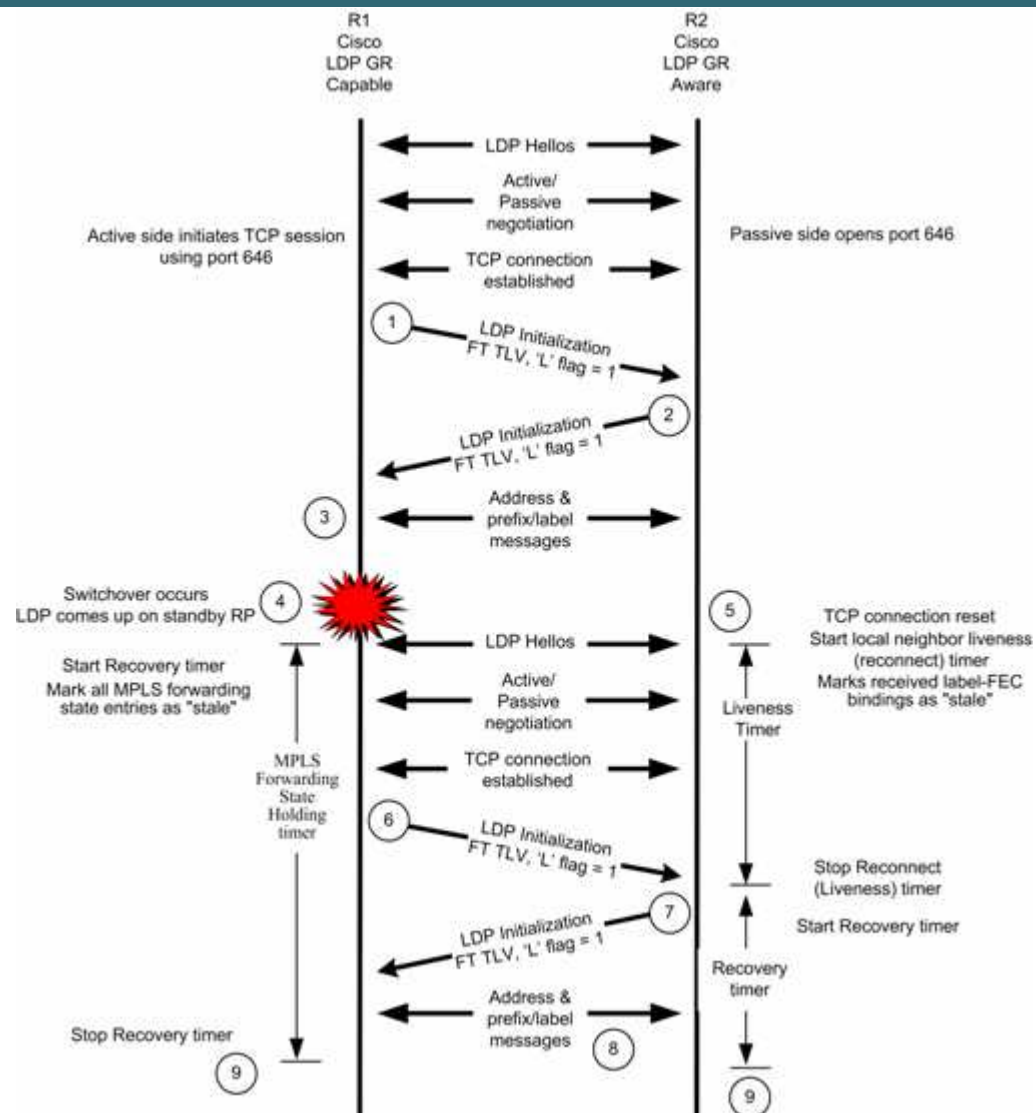
- Switchover happens



# LDP NSF (Graceful Restart) (Cont.)

- ④ R1 starts internal recovery timer—forwarding state holding timer; re-establishes TCP session
- ⑤ R2 recognizes failure of LDP session (TCP session reset or timeout)
  - Marks label bindings as stale; starts Liveness timer
- ⑥ Current value of forwarding-holding timer is used as the Recovery-time sent in the FT TLV; reconnect timer is always sent as 120 (release dependent)\*

\*More About Timers Later





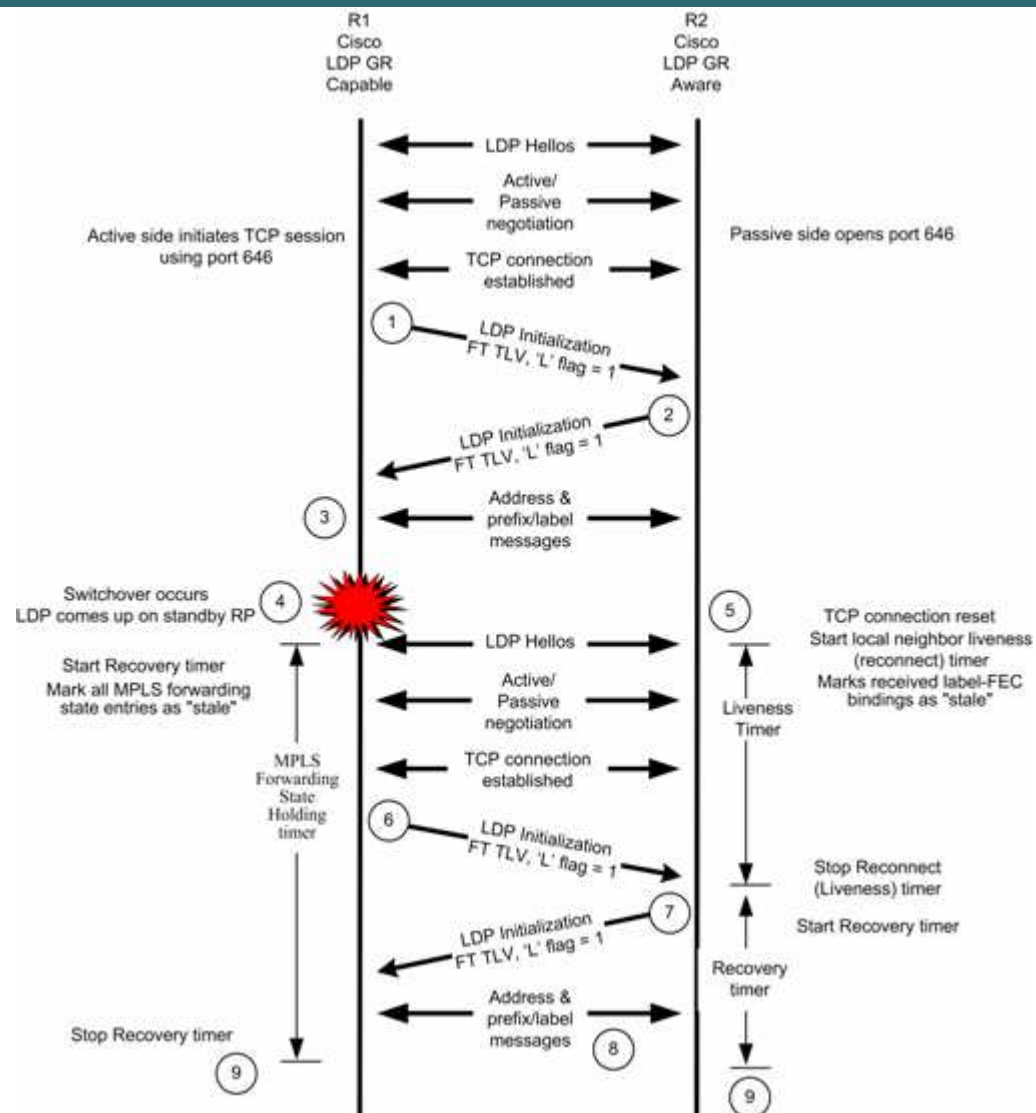
# LDP NSF (Graceful Restart) (Cont.)

## 7 R2: LDP session is established

If an LDP session is reestablished with the neighbor before the reconnect timer expires, the reconnect timer should be stopped, and the recovery timer started

## 8 LDP label binding are exchanged

## 9 If any Recovery timer expires, stale entries are removed



# LDP GR Timers

- **Reconnect Timeout**

Time (sent in milliseconds) that the sender of the TLV would like the receiver of that TLV to wait after the receiver detects the failure of LDP communication with the sender

The default value for this timer is 120 seconds—it cannot be configured

- **Recovery Time**

For a restarting LSR, the Recovery Time carries the time (sent in milliseconds) the LSR is willing to retain its MPLS forwarding state that it preserved across the restart; the time is from the moment the LSR sends the Initialization message that carries the FT Session TLV after restart

Setting this time to 0 indicates that the MPLS forwarding state was not preserved across the restart

This timer is controlled by setting “**forwarding-holding**” time\*; default is 600 seconds

- **Liveness Timer\***

Amount of time to wait for a session to re-establish—uses the lesser of the FT Reconnect Timeout, as was advertised by the restarting router, and a local timer, called the Neighbor Liveness Timer

If within that time the LSR still does not establish an LDP session with the neighbor, all the stale bindings **should** be deleted

Liveness Timer is started when the LSR detects that its LDP session with the neighbor went down

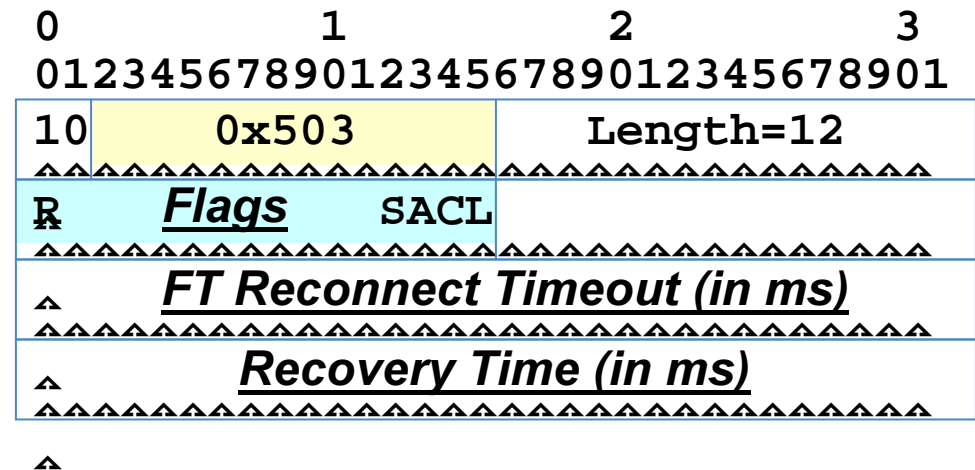
- **Max-Recovery time\***

Places an upper bound on the amount of time that an LSR is willing to hold stale label-FEC bindings after the LDP session has been reestablished; default is 120 seconds

\* These Timers Can Be Configured

# Another Way to Think of LDP Timers

- When waiting for a session to re-establish  
Use  $\text{Min}(\text{NbrLiveness}, \text{Rconn})$
- When waiting for state to be refreshed after session re-establishment  
Use  $\text{Min}(\text{MaxRecovery}, \text{Rcov})$



```
R1#show mpls ldp gr
LDP Graceful Restart is enabled
Neighbor Liveness Timer: 60 seconds
Max Recovery Time: 120 seconds
Forwarding State Holding Time: 300 seconds
Down Neighbor Database (0 records):
Graceful Restart-enabled Sessions:
  VRF default:
    Peer LDP Ident: 88.1.11.1:0, State: estab
```

```
mpls ldp graceful-restart timers neighbor-liveness 60
mpls ldp graceful-restart timers forwarding-holding 300
mpls ldp graceful-restart
```

8503000C

00010000

0001D4C0

00042DB4

Rconn  
120,000

Rcov  
273,844

Current  
Value

# LDP NSF (GR) Configuration

## 1. Enable GR globally (must do before enabling LDP)

```
mpls ldp graceful-restart
```

## 2. Enable MPLS (global)

```
mpls ip
```

## 3. Enable LDP on appropriate interfaces

```
mpls label protocol ldp
```

```
iguana(config)#mpls ldp graceful-restart
```

```
% Previously established LDP sessions may not have  
graceful restart protection.
```

```
iguana(config)#mpls ldp graceful-restart timers ?
```

```
forwarding-holding Forwarding State Holding time
```

```
max-recovery Max-Recovery time
```

```
neighbor-liveness Neighbor-Liveness time
```

# MPLS NSF/SSO Prerequisites

## MPLS HA Relies on the Underlying NSF/SSO Features We Have Been Discussing so Far for Operation

### 1. BGP NSF mechanisms must be enabled

BGP Graceful Restart allows a router to create MPLS forwarding entries for VPNv4 prefixes in NSF mode

Forwarding entries are preserved during a restart

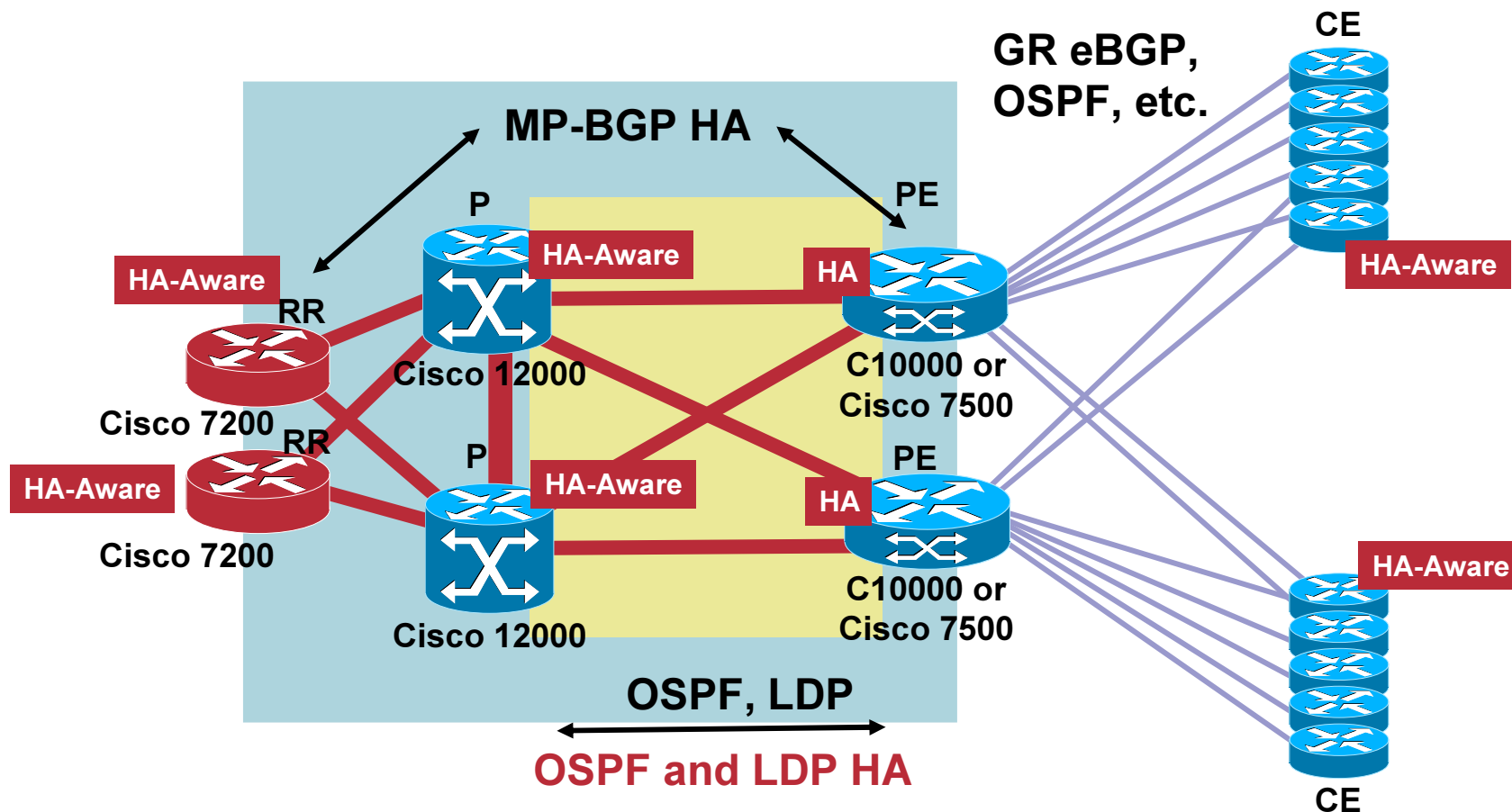
BGP also saves prefix and corresponding label information and recovers the information after a restart

### 2. NSF support for the label distribution protocol in the core network (LDP NSF or GR)

### 3. NSF support for the Internal Gateway Protocol (IGP) used in the core; i.e. OSPF or IS-IS

### 4. NSF support for the routing protocols between the PE and customer edge (CE) routers

# Deploying MPLS NSF/SSO

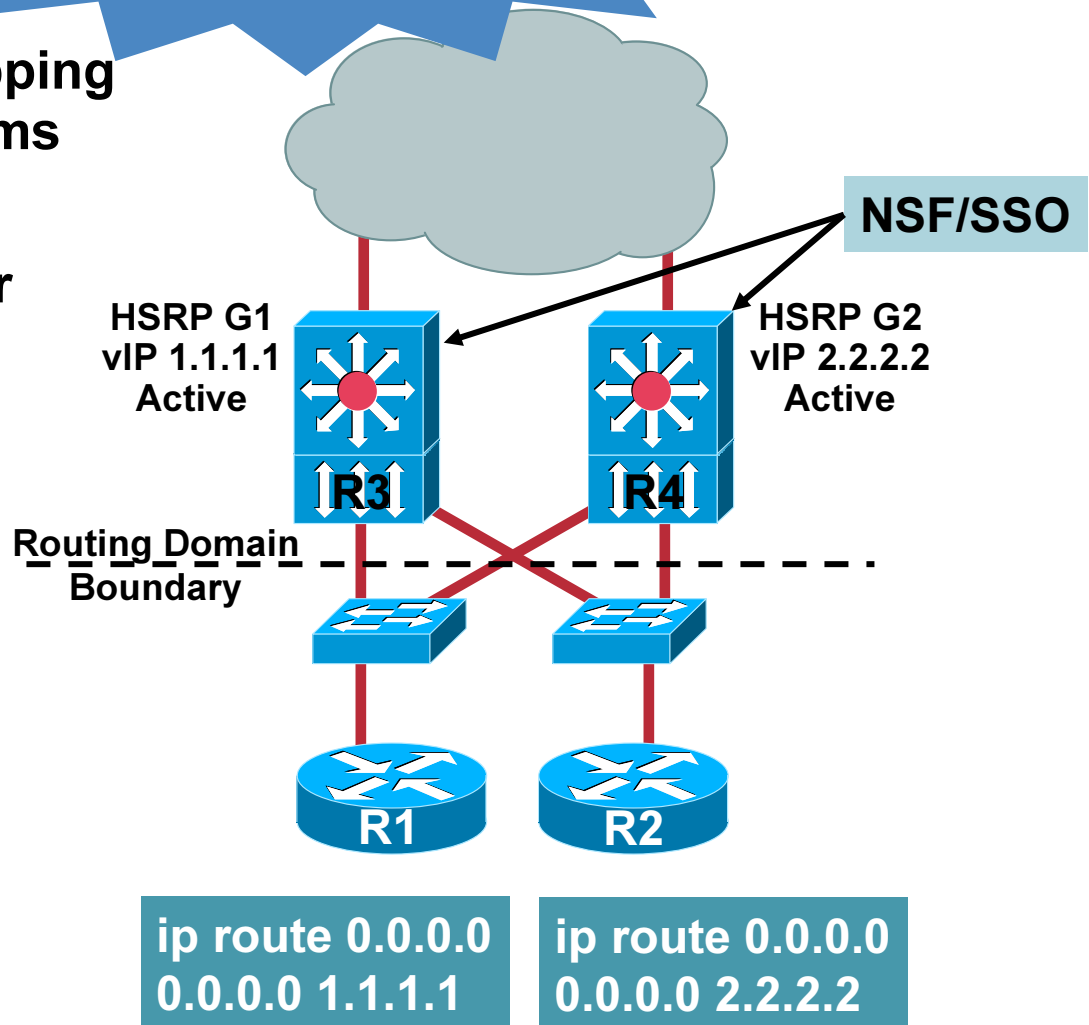


**HA Aware Devices Must Support Graceful Restart for Deployed Routing Protocol**

# HSRP/SSO

HSRP SSO (HA-Aware)  
Beginning with 12.2(25)S

- Be careful with overlapping redundancy mechanisms
- Traffic switches (quickly) on RP failover within R3 or R4



**\*Beware: This Design Is  
Not Multicast Friendly**

# Cisco IOS High Availability Summary

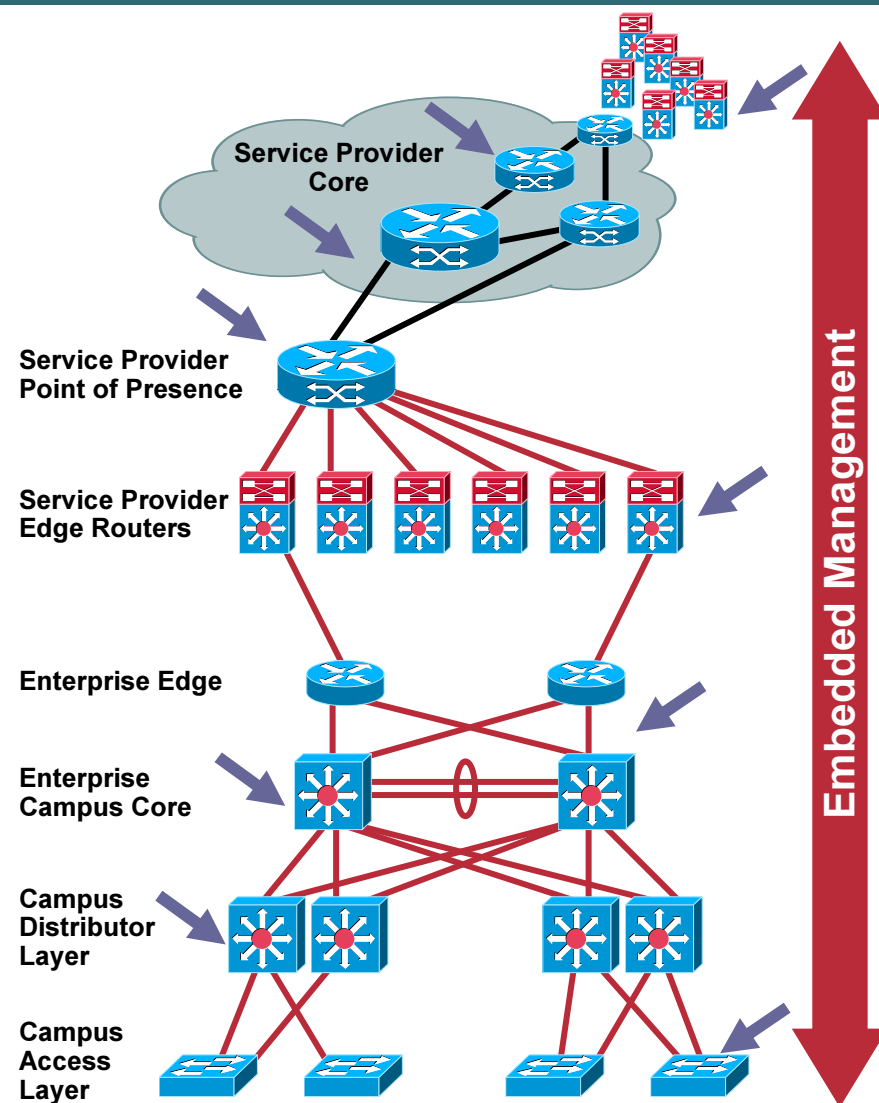




# Cisco IOS Software High Availability

## Non-Stop Application Delivery

- **End-to-end, systematic approach**
  - System level resiliency at critical network edges
  - Network resiliency in the core and where redundant paths exist
  - Accounts for services and protocols
  - Embedded management
- **Trusted partner**
  - System, application view, product breadth
  - Track record for investment protection
  - Best practice, design, support



# What Questions Do You Have?



