HIGH AVAILABILITY IN IP ROUTING

SESSION RST-3212

Phil Harris

pharris@cisco.com

Recuerde siempre:

- Apagar su teléfono móvil/pager, o usar el modo "silencioso".
- Completar la evaluación de esta sesión y entregarla a los asistentes de sala.
- Ser puntual para asistir a todas las actividades de entrenamiento, almuerzos y eventos sociales para un desarrollo óptimo de la agenda.
- Completar la evaluación general incluida en su mochila y entregarla el miércoles 8 de Junio en los mostradores de registración. Al entregarla recibirá un regalo recordatorio del evento.

Agenda

- High Availability Overview
- Non-Stop Forwarding
- Fast Convergence
- HA Deployment Summary

HIGH AVAILABILITY OVERVIEW

High Availability Overview

Cisco.com

Availability Definition:
The Proportion of Time That a System/Network Can be Used for Productive Work

$$\text{Availability \%} = \frac{\text{MTBF}}{\text{MTBF+MTTR}}$$

- **Availability reduces with downtime and Mean Time To Repair (MTTR)**
- **Common causes of “downtime” include:**
  - Hardware failure
  - Network failure
  - Operating system error/failure
  - Application error
  - Human error
  - Security breach or attack
  - System overload
  - Power/Environment

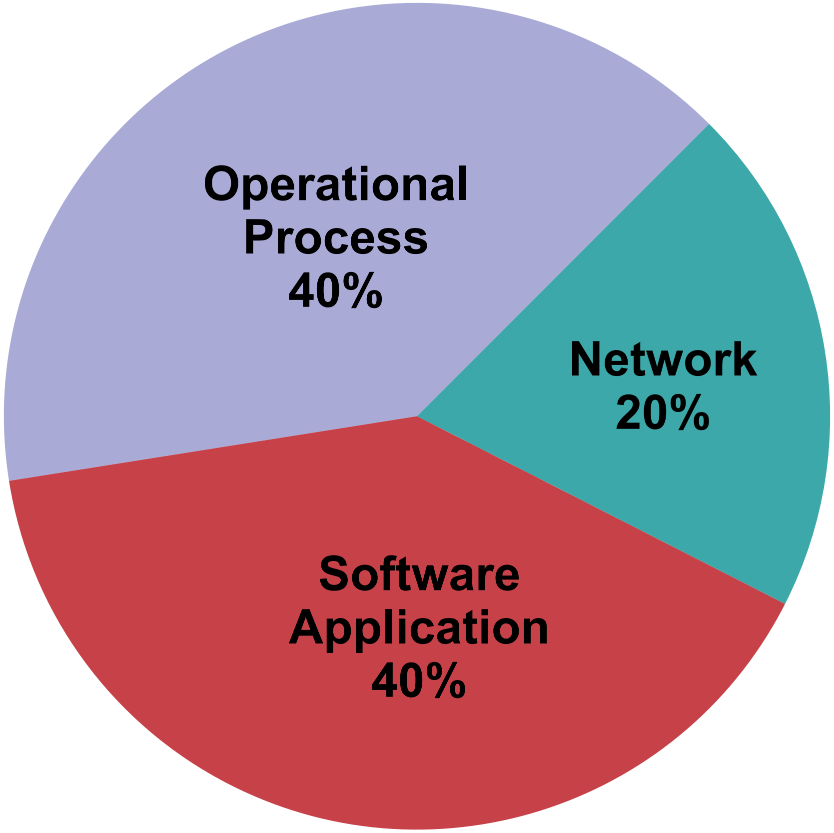
What Is High Availability?

Availability	DPM	Downtime per Year (24x365)			
99.000%	10000	3 Days	15 Hours	36 Minutes	Reactive
99.500%	5000	1 Day	19 Hours	48 Minutes	
99.900%	1000		8 Hours	46 Minutes	Proactive
99.950%	500		4 Hours	23 Minutes	
99.990%	100			53 Minutes	Predictive
99.999%	10			5 Minutes	"High Availability"
99.9999%	1			30 seconds	

DPM = Defects per Million (Hours of Running Time)

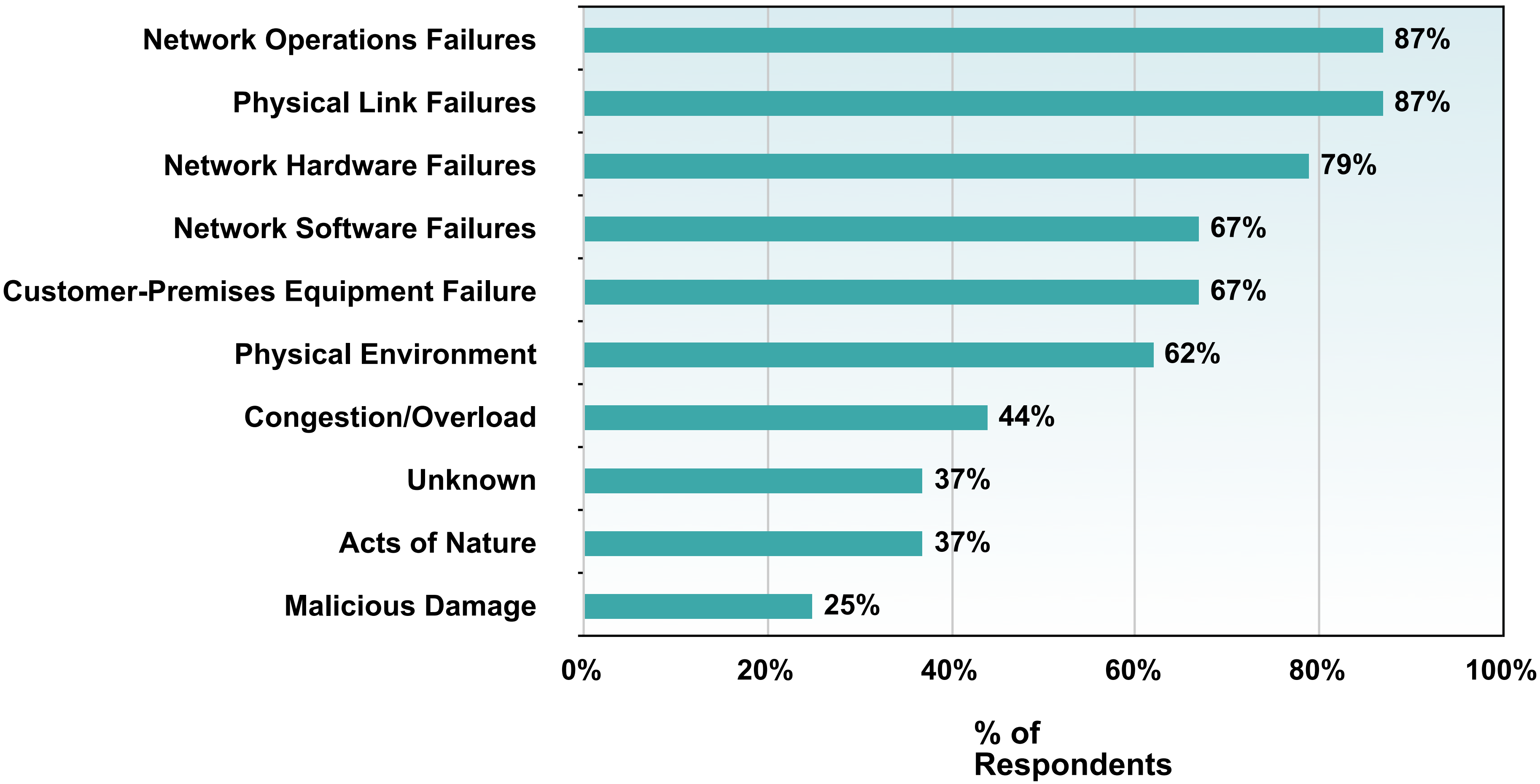
Causes of Unscheduled Network Downtime

- Change
- Communication
- Process
- Design
- Hardware
- Software
- Link
- Power/env
- Resource utilization



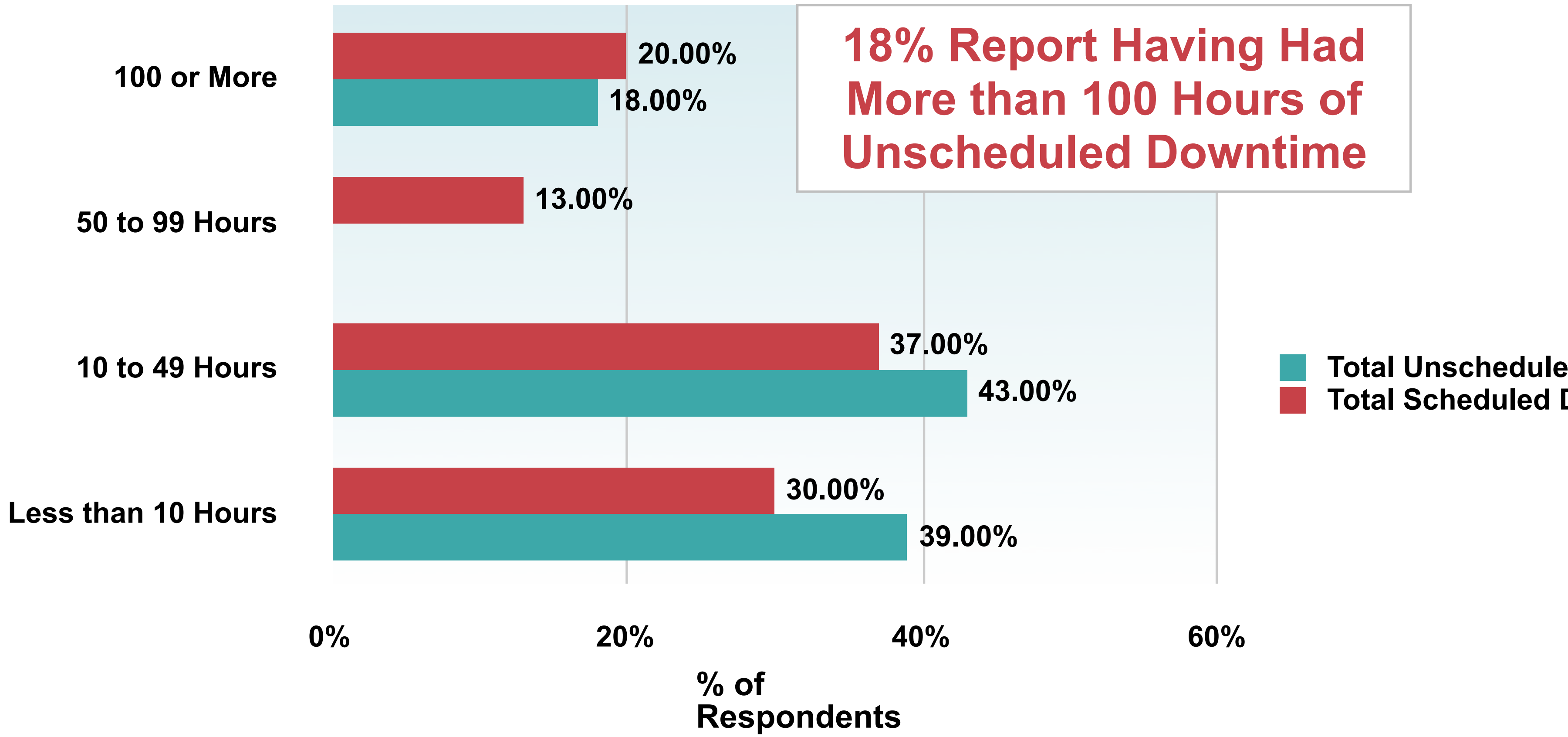
Source: Gartner

Causes of Unscheduled Downtime



Source: Sage Research, IP Service Provider Downtime Study: Analysis of Downtime Causes, Costs, and Containment Strategies, August 17, 2001, Prepared for Cisco SPLOB

Duration of Downtime

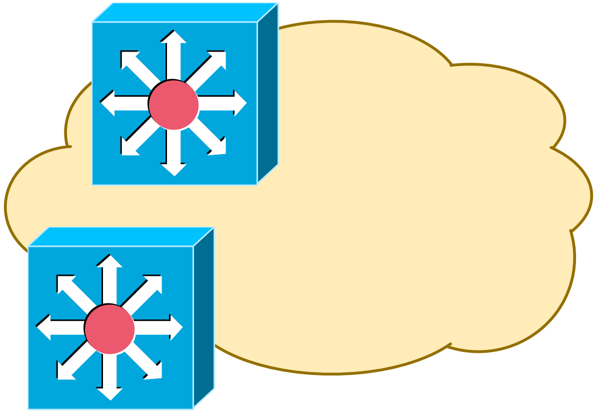


Source: Sage Research, IP Service Provider Downtime Study: Analysis of Downtime Causes, Costs, and Containment Strategies, August 17, 2001, Prepared for Cisco SPLOB

Hardware

Redundancy Options

Highly Available Networks Tend to Have Both



−
- Failover redundant modules only
- Operating system determines failover

+
- Typically cost effective

+
- Often only option for edge devices (point to point)

+
- All modules are redundant
- Protocols determine failover

−
- Increased cost and complexity

+
- Load balancing

The Culture of Availability

- Identify gaps
- Root-cause failure analysis
- Availability modeling
- Availability metrics
- Priority and ROI analysis
- Quality improvement

What Is Your Availability Level?

Analyze the Gaps: Reactive ~99%

- Few, if any, identified processes (except maybe to fix problems as reported by users)
- Low tool utilization
- Low level of consistency (HW, SW, config, design)
- No quality-improvement processes

What Is Your Availability Level?

Analyze the Gaps: Proactive ~99.9%

- Good change management processes including what-if analysis and change validation
- Fault and configuration management tools
- Improved consistency (HW, SW, config, design)
- Typically no quality improvement process

What Is Your Availability Level?

Analyze the Gaps: Predictive ~99.99+%

- Consistent processes for fault, configuration, performance, and security
- Fault, configuration, performance, and workflow process tools
- Excellent consistency (HW, SW, config, design)
- HA culture of quality improvement

PROTOCOL-INDEPENDENT FEATURES

IP Event Dampening

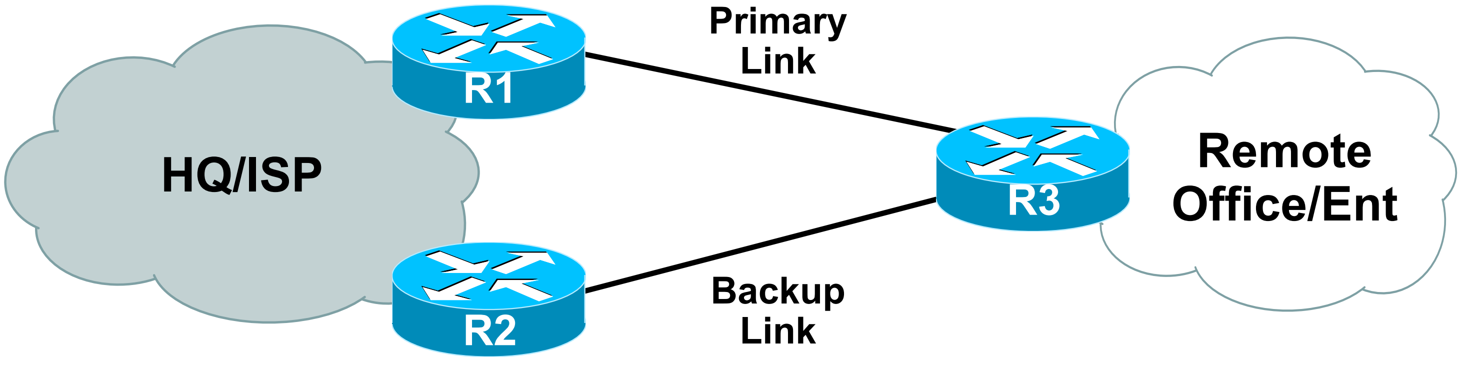
- Prevents routing protocol churn caused by constant interface state changes
- Supports all IP routing protocols
  - Static routing, RIP, EIGRP, OSPF, IS-IS, BGP
  - In addition, it supports HSRP and CLNS routing
  - Applies on physical interfaces and can't be applied on subinterfaces individually
- Available in 12.0(22)S, 12.2(13)T

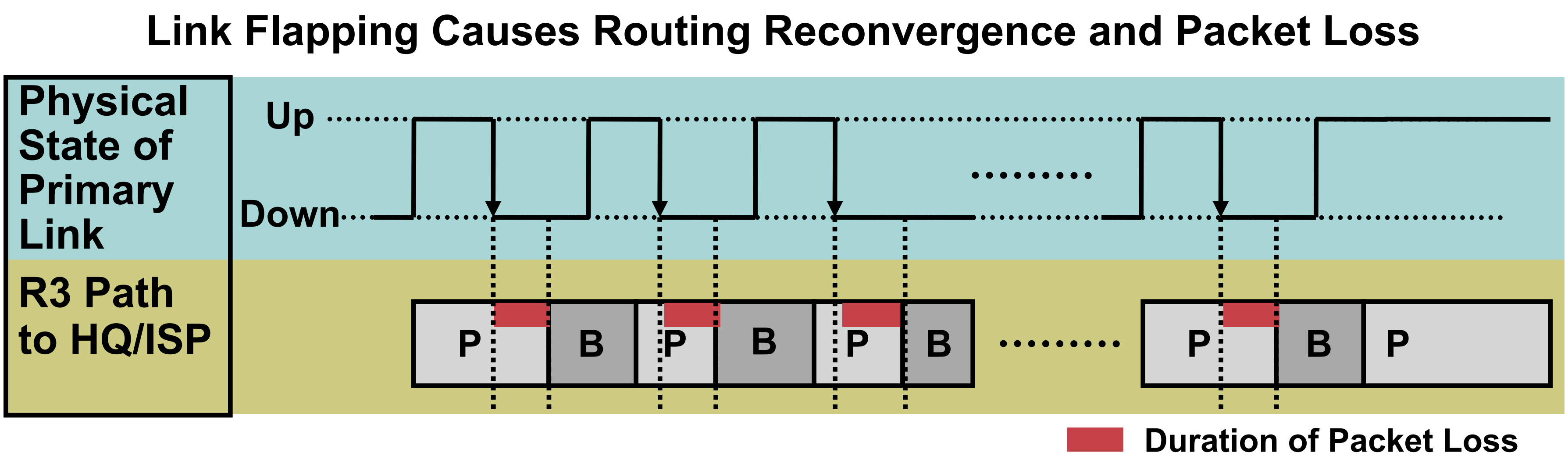
IP Event Dampening: Concept

- Takes the concept of BGP route-flap dampening and applies it at the interface level, so all IP-routing protocols can benefit
- Tracks interface flapping, applying a “penalty” to a flapping interface
- Puts the interface in “down” state from routing protocol perspective if the penalty is over a threshold tolerance
- Uses exponential decay algorithm to decrease the penalty over time and brings the interface back to “up” state

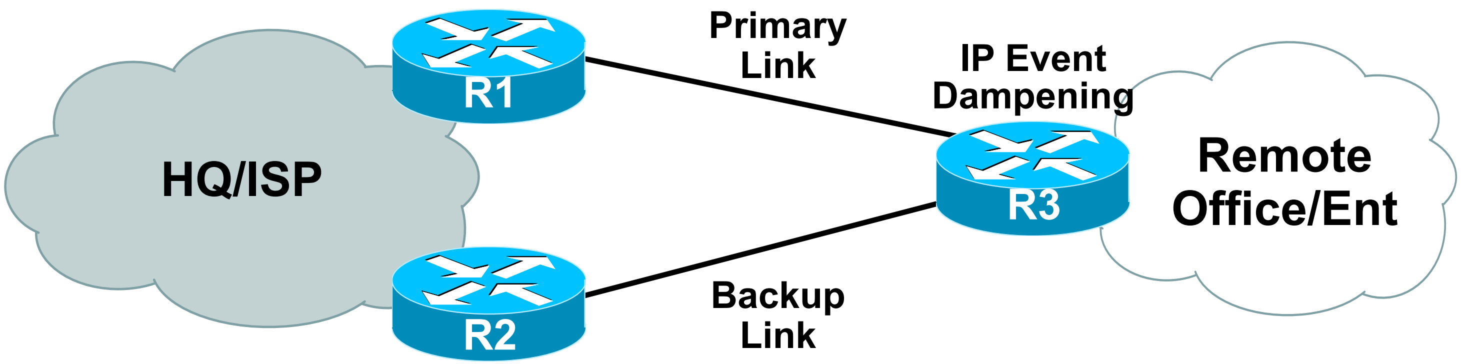
IP Event Dampening: Deployment

Cisco.com

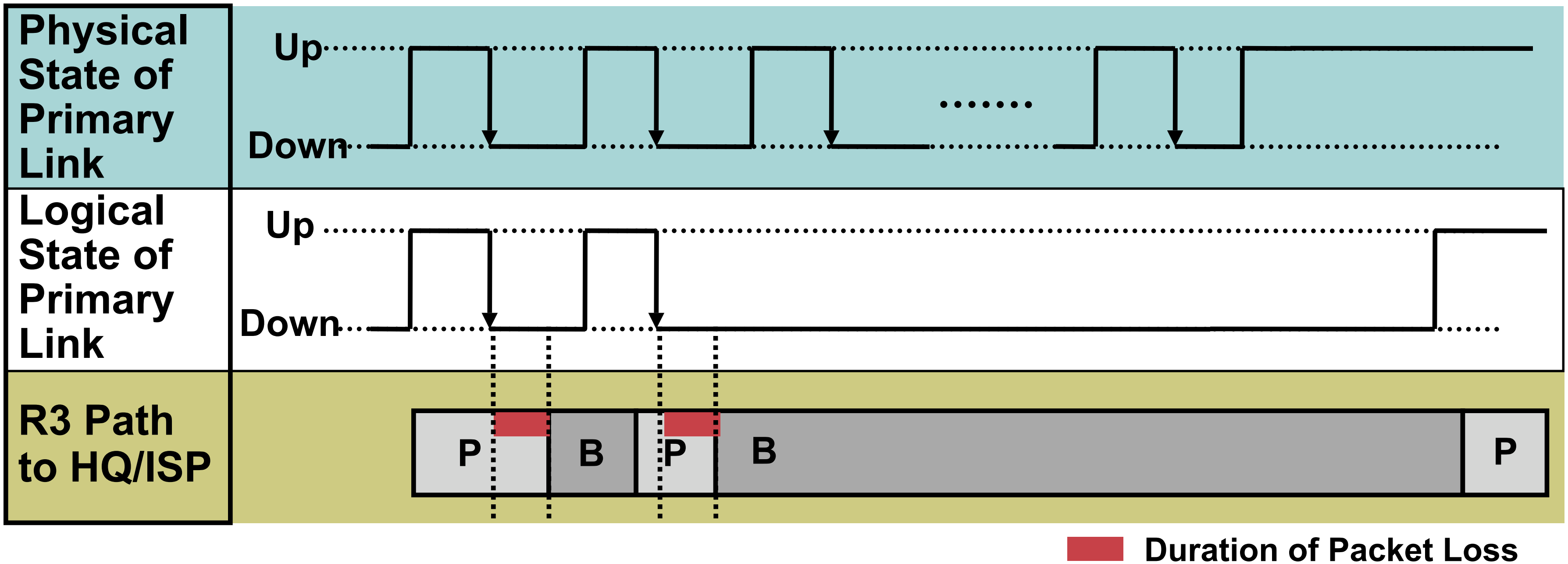




IP Event Dampening: Deployment



IP Event Dampening Absorbs Link-Flapping Effects on Routing Protocols

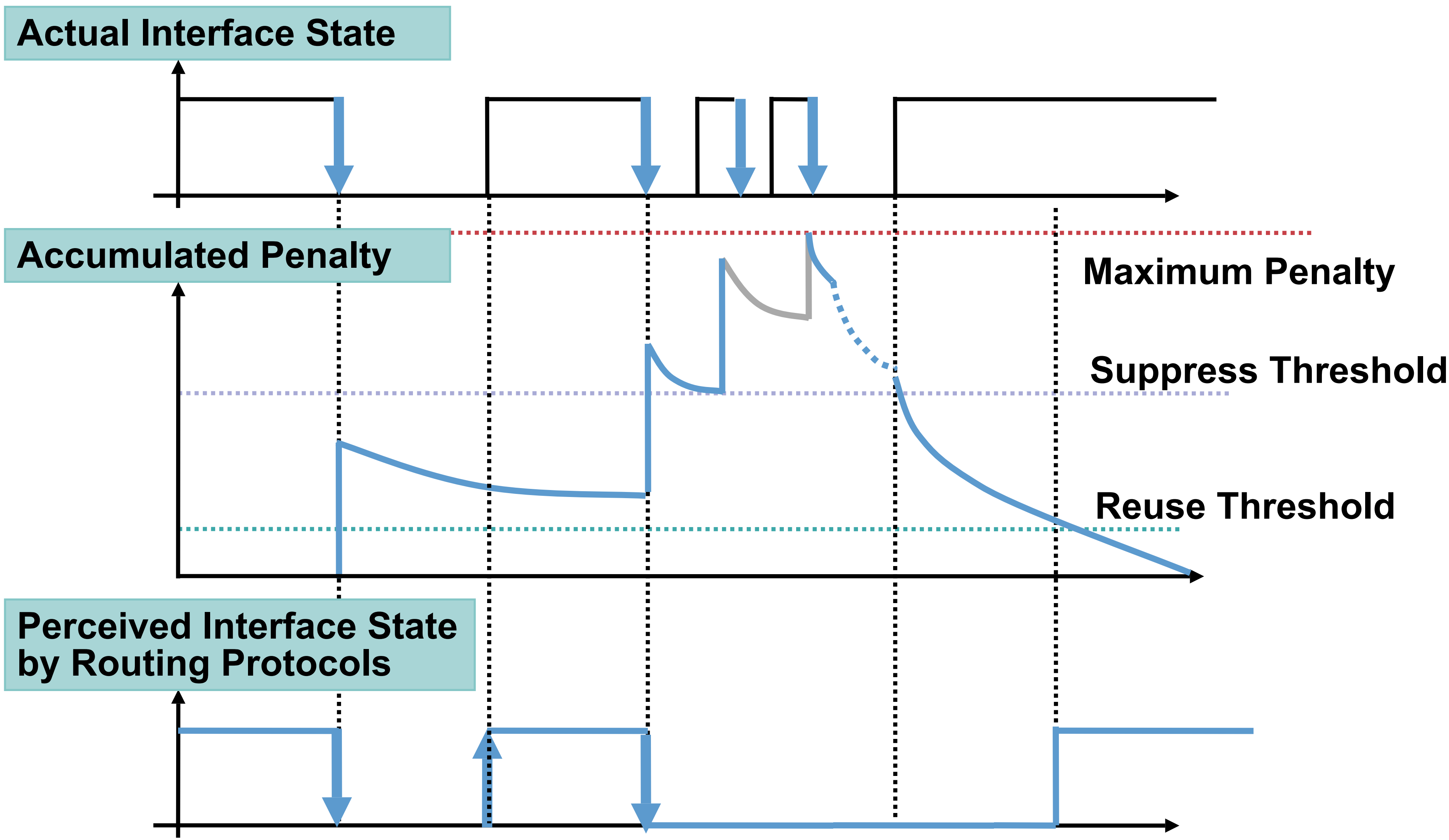


IP Event Dampening: Algorithm


```
interface Serial 0
        dampening [half-life reuse-threshold] [suppress-threshold max-
         suppress [restart-penalty]]
```

- **Penalty**: a numeric value applied to the interface each time it flaps
- **Half-life**: amount of time that must elapse without a flap to reduce penalty by half
- **Reuse-threshold**: if penalty goes below this limit, the interface is reintroduced to the routing protocols
- **Suppress-threshold**: if penalty exceeds this value, interface is suppressed from routing protocols' perspective
- **Max-suppress**: maximum amount of time an interface can be suppressed
- **Restart-penalty**: determines initial penalty (if any) to be applied to interface when system boots

IP Event Dampening: Algorithm Illustration

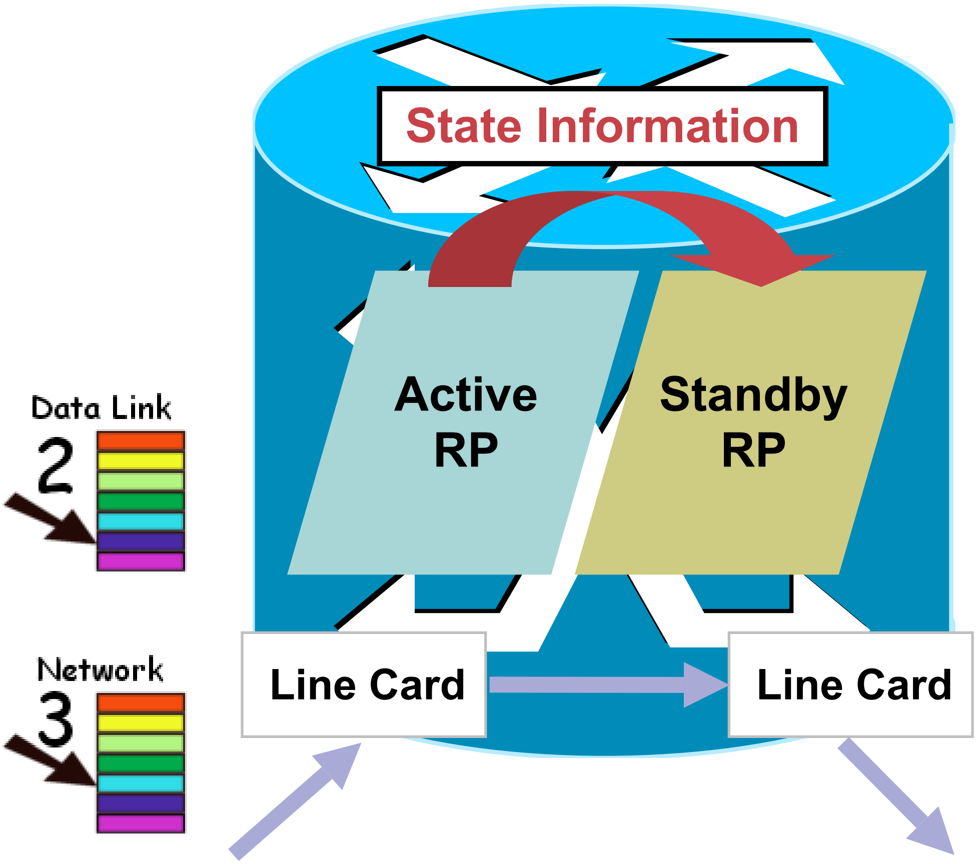


NON-STOP FORWARDING

GRACEFUL RESTART AND STATEFUL SWITCHOVER

Cisco Non-Stop Forwarding with Stateful Switchover (NSF/SSO)

- **Standby route processor (RP) takes control of router after a hardware or software fault on the active RP**
- **SSO allows standby RP to take immediate control and maintain connectivity protocols**
- **NSF continues to forward packets until route convergence is complete**
- **GR (graceful restart) reestablishes the routing information bases without churning the network**



NSF/SSO Software Design Goals

- **Provide a scalable solution**
  - Architecture must scale with workloads and features and meet network requirements
- **Minimize state that must be synchronized**
  - Minimize impact of HA on service
- **Detect and react to failures quickly**
  - Continuously monitor active components
  - Continuously verify operation of standby components

Enabling SSO

- **Perform this step on 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

- **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

NSF: Routing Protocol Requirements

- Adjacencies MUST NOT be reset when switchover is complete
  - Protocol state is not maintained
- Switchover MUST be completed before dead/hold timer expires
  - Else peers will reset the adjacency and reroute the traffic
- FIB MUST remain unchanged during switchover
  - Current routes marked as “dirty” during restart
  - “Cleaned” once convergence is complete
  - Transient routing loops or black holes MAY be introduced if the network topology changes before the FIB is updated

Enhancements to Routing Protocols

- **Neighbor routers must know that an NSF router can still forward packets**

  Call this “NSF aware” as opposed to “NSF capable”

- **Enhancements to ISIS, OSPF, EIGRP, and BGP designed to prevent route flapping**

NSF and SSO Support

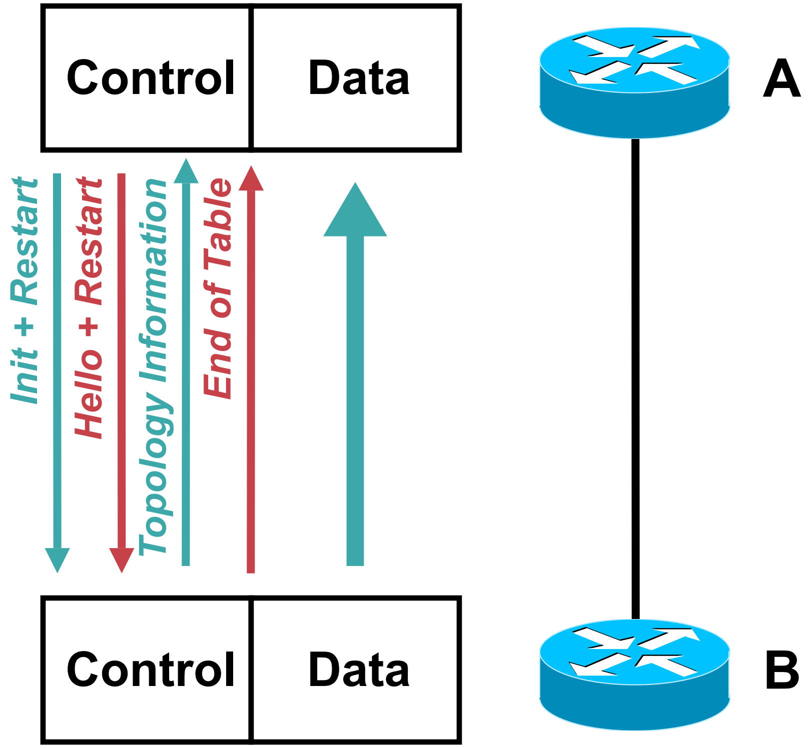
- **EIGRP**
  - NSF capable—12.2(18)S
  - NSF aware—12.2(15)T
- **BGP**
  - NSF capable—12.0(22)S, 12.2(18)S
  - NSF aware—12.2(15)T
- **OSPF**
  - NSF capable—12.0(22)S, 12.2(18)S
  - NSF aware—12.2(15)T
- **IS-IS**
  - NSF capable—12.0(22)S, 12.2(18)S
  - NSF aware—12.2(15)T

NON-STOP FORWARDING

EIGRP

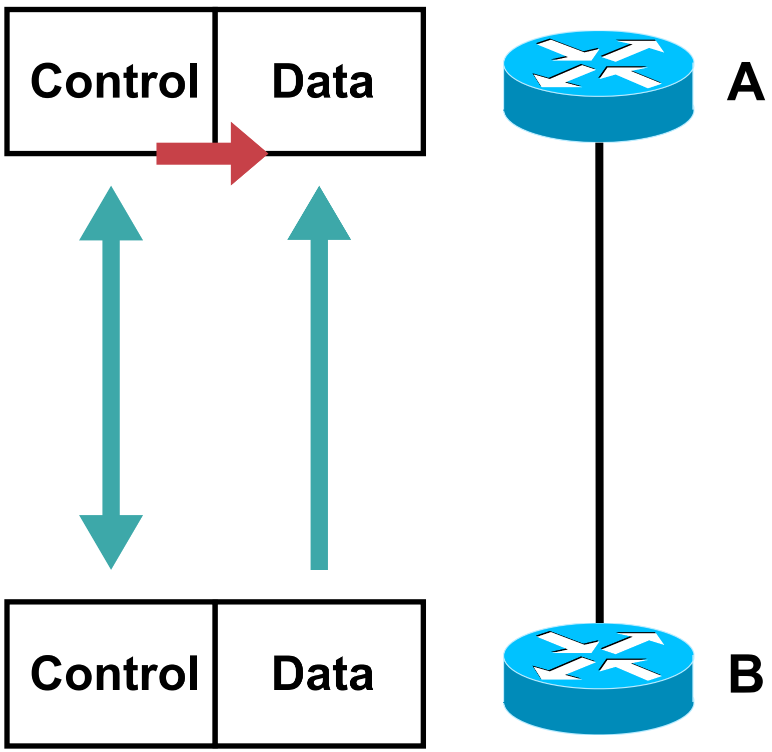
EIGRP GR/NSF Fundamentals

- The signal in EIGRP is an update with the *initialization* and *restart* (RS) bits set
- "A" sends its hello's with the restart bit set until Graceful Restart is complete
- "B" transmits the routing information it knows to A
- When "B" is finished sending information, it sends a special end of table signal so "A" knows the table is complete



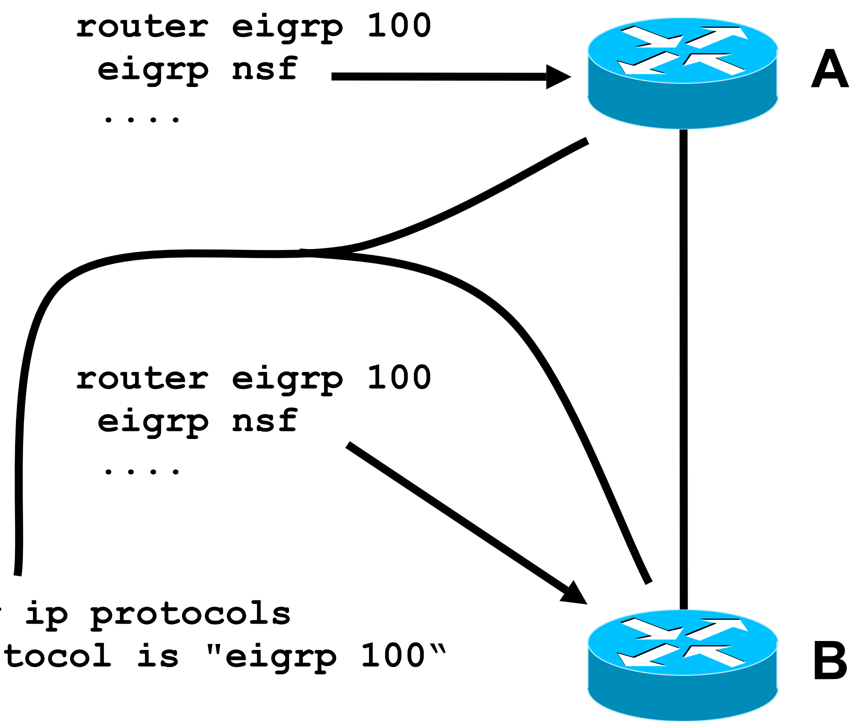
EIGRP GR/NSF Fundamentals

- When "A" receives this end of table marker, it recalculates its topology table, and updates the local routing table
- When the local routing table is completely updated, EIGRP notifies CEF
- CEF then updates the forwarding tables, and removes all information marked as stale



EIGRP GR/NSF Fundamentals

- Use the *eigrp nsf* command under the *router eigrp* configuration mode to enable graceful restart
- *Show ip protocols* can be used to verify graceful restart is operational
- http://www.cisco.com/en/US/products/sw/iosswrel/ps1839/products_feature_guide09186a0080160010.html



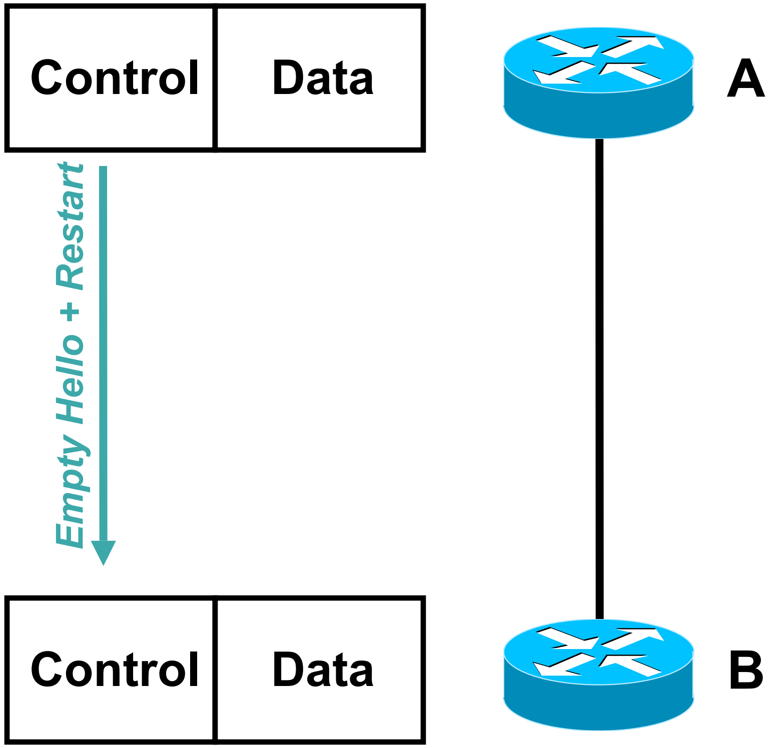

```
router#show ip protocols
Routing Protocol is "eigrp 100"
....
Redistributing: eigrp 100
EIGRP NSF-aware route hold timer is 240s
Automatic network summarization is in effect
Maximum path: 4
....
```

NON-STOP FORWARDING

OSPF

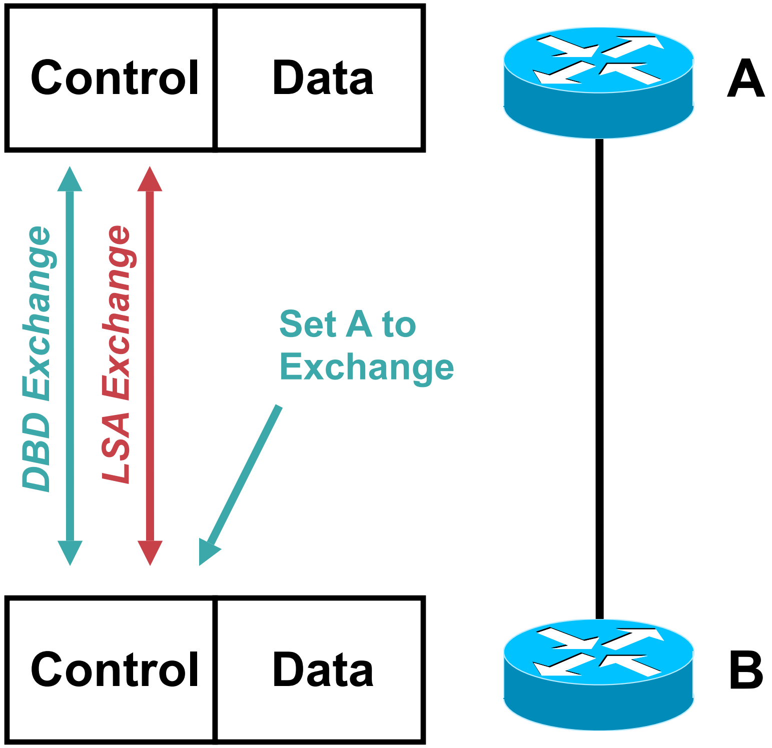
OSPF GR/NSF Fundamentals

- OSPF uses an extension to the hello packets called link-local signaling
- The first hello “A” sends to “B” has an empty neighbor list; this tells “B” that something is wrong with the neighbor relationship
- “A” sets the restart bit in its hello, which tells “B” that “A” is still forwarding traffic, and would like to resynchronize its database



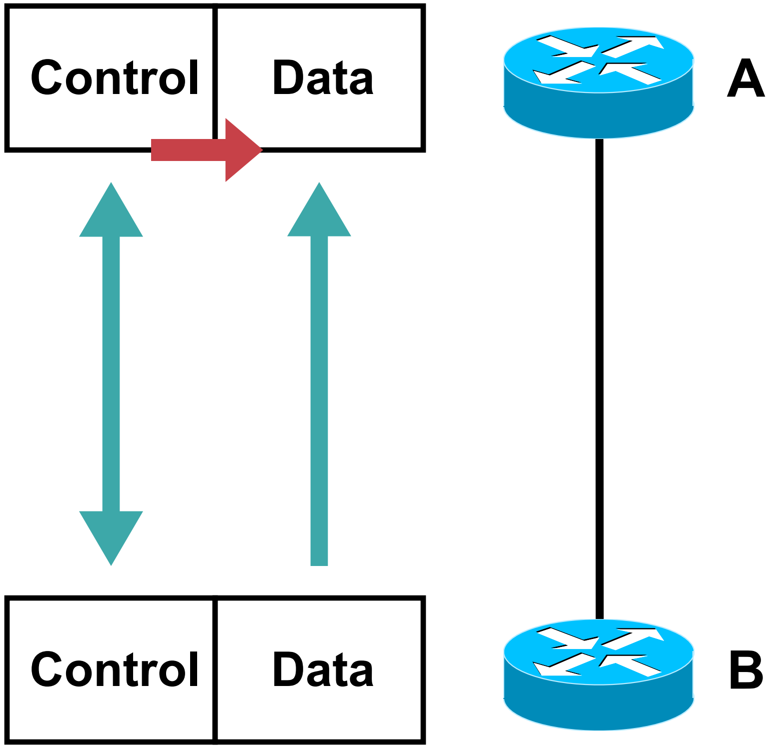
OSPF GR/NSF Fundamentals

- "B" moves "A" into the exchange state, and uses out of band signaling (OOB) to resynchronize their databases
- This process is the same as initial database synchronization, but it uses different packet types



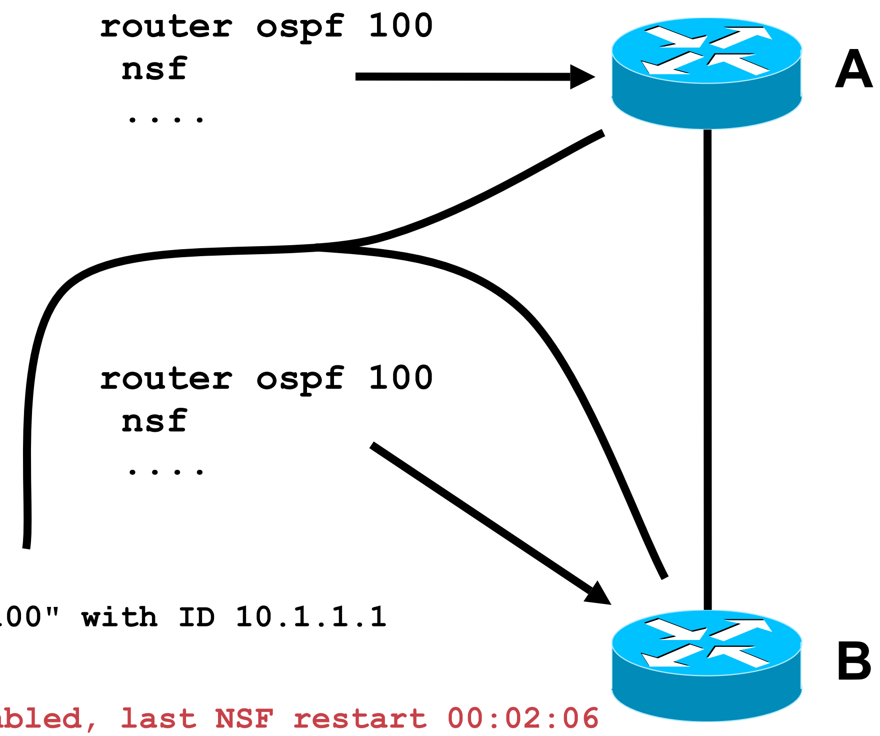
OSPF GR/NSF Fundamentals

- When "A" and "B" have resynchronized their databases, they place each other in full state, and run SPF
- After running SPF, the local routing table is updated, and OSPF notifies CEF
- CEF then updates the forwarding tables, and removes all information marked as stale



OSPF GR/NSF Fundamentals

- 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



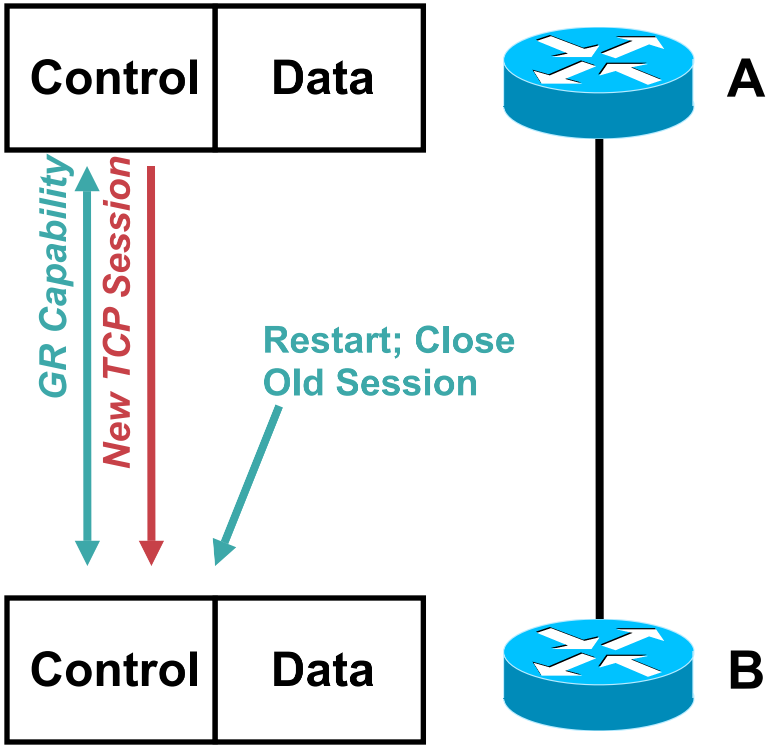

```
router#sh ip ospf
Routing Process "ospf 100" with ID 10.1.1.1
....
Non-Stop Forwarding enabled, last NSF restart 00:02:06
ago (took 44 secs)
router#show ip ospf neighbor detail
 Neighbor 3.3.3.3, interface address 170.10.10.3
 ....
 Options is 0x52
 LLS Options is 0x1 (LR), last OOB-Resync 00:02:22 ago
```

NON-STOP FORWARDING

BGP

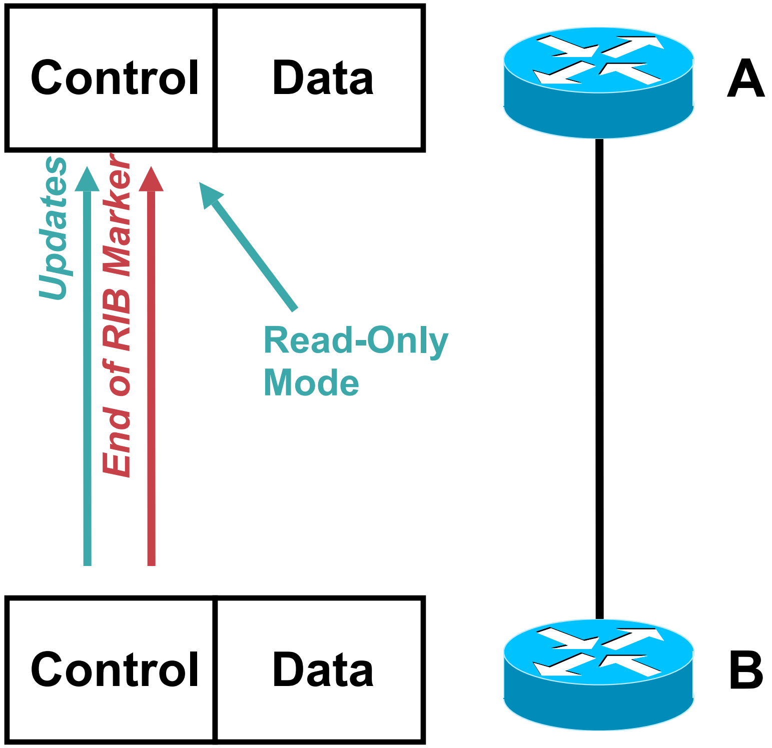
BGP GR/NSF Fundamentals

- When the BGP peering session is brought up, the graceful restart capability is negotiated; if both peers state they are capable of GR, it's enabled on the peering session
- When "A" restarts, it opens a new TCP session to "B", using the same router ID
- "B" interprets this as a restart, and closes the old TCP session



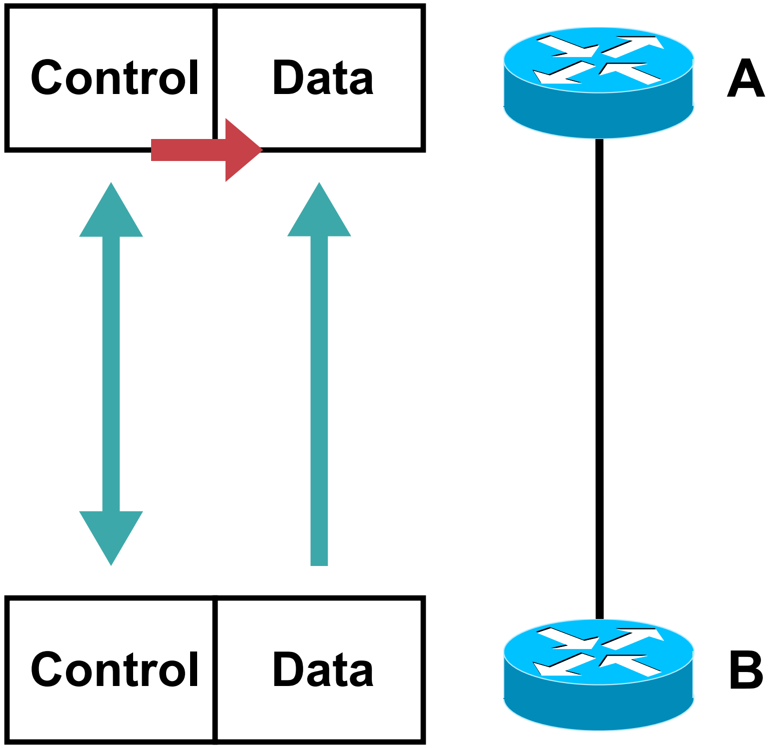
BGP GR/NSF Fundamentals

- "B" transmits updates containing its BGP table
- "A" goes into read-only mode, and does not run the bestpath calculations until its "B" has finished sending updates
- When "B" has finished sending updates, it sends an end of RIB marker, which is an update with an empty withdrawn NLRI TLV



BGP GR/NSF Fundamentals

- When "A" receives the end of RIB marker, it runs bestpath, and installs the best routes in the routing table
- After the local routing table is updated, BGP notifies CEF
- CEF then updates the forwarding tables, and removes all information marked as stale

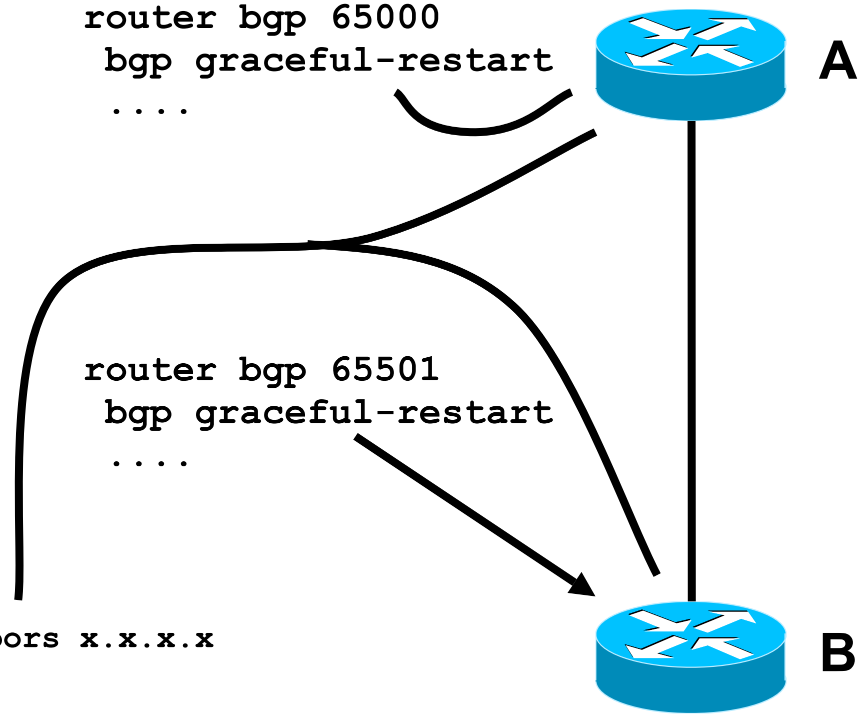


BGP GR/NSF Fundamentals

- 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 Capabilty:advertised and received
Remote Restart timer is 120 seconds
Address families preserved by peer:
IPv4 Unicast, IPv4 Multicast
```

FAST CONVERGENCE TUNING
LINK-STATE PROTOCOLS

Network Convergence

- Network convergence is the time needed for traffic to be rerouted to the alternative or more optimal path after the network event
- Network convergence requires all affected routers to process the event and update the appropriate data structures used for forwarding

Network Convergence

- **Network convergence is the time required to:**
  - **Detect event has occurred**
  - **Propagate the event**
  - **Process the event**
  - **Update related forwarding structures**

FAST CONVERGENCE TUNING

EVENT DETECTION (LINK STATE)

Event Detection: Subsecond Hellos

- **At what frequency should hellos be issued?**
  - How many interfaces involved?
  - What is the current resource utilization?
  - How fast does a change need to be detected?
- **Are subsecond hellos the most effective method?**
  - Will Layer 1/Layer 2 provide faster notification? (POS/serial)
  - Tune Layer 1 to detect as fast as possible without causing excessive flapping

OSPF Subsecond Hellos

- **Supported: 12.0(23)S, 12.2(18)S, 12.2(15)T**
- **Operation:**

  Dead interval—minimum one second

  Hello multiplier is used to specify how many hellos to send within one second

  Hello interval will be advertised as zero second

- **Configuration:**

  `ip ospf dead-interval minimal hello-multiplier` *`value`*

  Value—range 3–20

Fast Hellos: Scaling Issues

Scaling Is a Major Issue

30 Interfaces x 10 Neighbors/Interface = 300 Neighbors

---

20 Hello Packets per Second on Each Interface

Router has to Generate 200 Hello's per Second

---

300 Neighbors Each Send 20 Hello's per Second to this Router

Router has to Accept and Process 6000 Hello's per Second

Router has to Deal with 6200 Hello's per Second

FAST CONVERGENCE TUNING

EVENT PROCESSING (LINK STATE)

SPF Overview

- **Full SPF**
  - Triggered by the change in router or network LSA
  - SPT tree is recomputed
  - All LSA types (type 1/2/3/4/5/7) are processed
- **Partial SPF**
  - Triggered by the change in type-3/4/5/7 LSA
  - If triggered by type 3 (Summary LSA created by ABR):
    - all type-3 LSAs that contribute to the certain destination are processed
  - If triggered by type 5/7 (External Information created by ASBR):
    - all type-5/7 LSAs that contribute to the certain destination are processed
  - If triggered by type 4 (Information about ASBR's, created by ABR):
    - all type-4 LSAs that announce a certain ASBR and all type-5/7 LSAs are processed

SPF Execution Time

- **SPF calculation time**
  - Full spf:
    - Depends on:
      - Number of nodes/links in the area
      - Number of type-3/4/5/7 LSAs
    - Some experimental numbers (GSR/7500)
      - 50 nodes fully-connected topology ~ 10ms
      - 100 node fully-connected topology ~ 25ms
      - 500 nodes ~ 50 ms
      - 1000 nodes ~ 100 ms
  - Partial SPF:
    - Fast—less then 0.5 ms

SPF Triggers

- **Router/network LSA triggers full SPF**
  - Some changes does not represent a topology change:
    - Stub network UP/DOWN
    - IP address change on link
  - During the full SPF the whole SPT is rebuilt
    - Change in the topology may not require the whole SPT rebuild
    - Major part of the tree may stay the same in many cases

Incremental SPF: Overview

- **Incremental SPF**
  - Modified Dijkstra algorithm
  - Keep the unchanged part of the tree
  - Rebuild only the affected parts of the tree
  - Reattach the affected parts of the tree to the unchanged part of the tree

Incremental SPF: Overview

- **Analyze the changes in the newly-received LSA**

  All new or changed LSAs received during the spf-wait interval are put in a NEW_LSA_LIST

- **LSA can carry:**

  Good news—a better path to the node becomes available

  Bad news—current best path to the node becomes worse (or is lost)

  No news—no topological change

- **The iSPF algorithm determines what to do based on the type of information in the LSA**

Incremental SPF: Overview

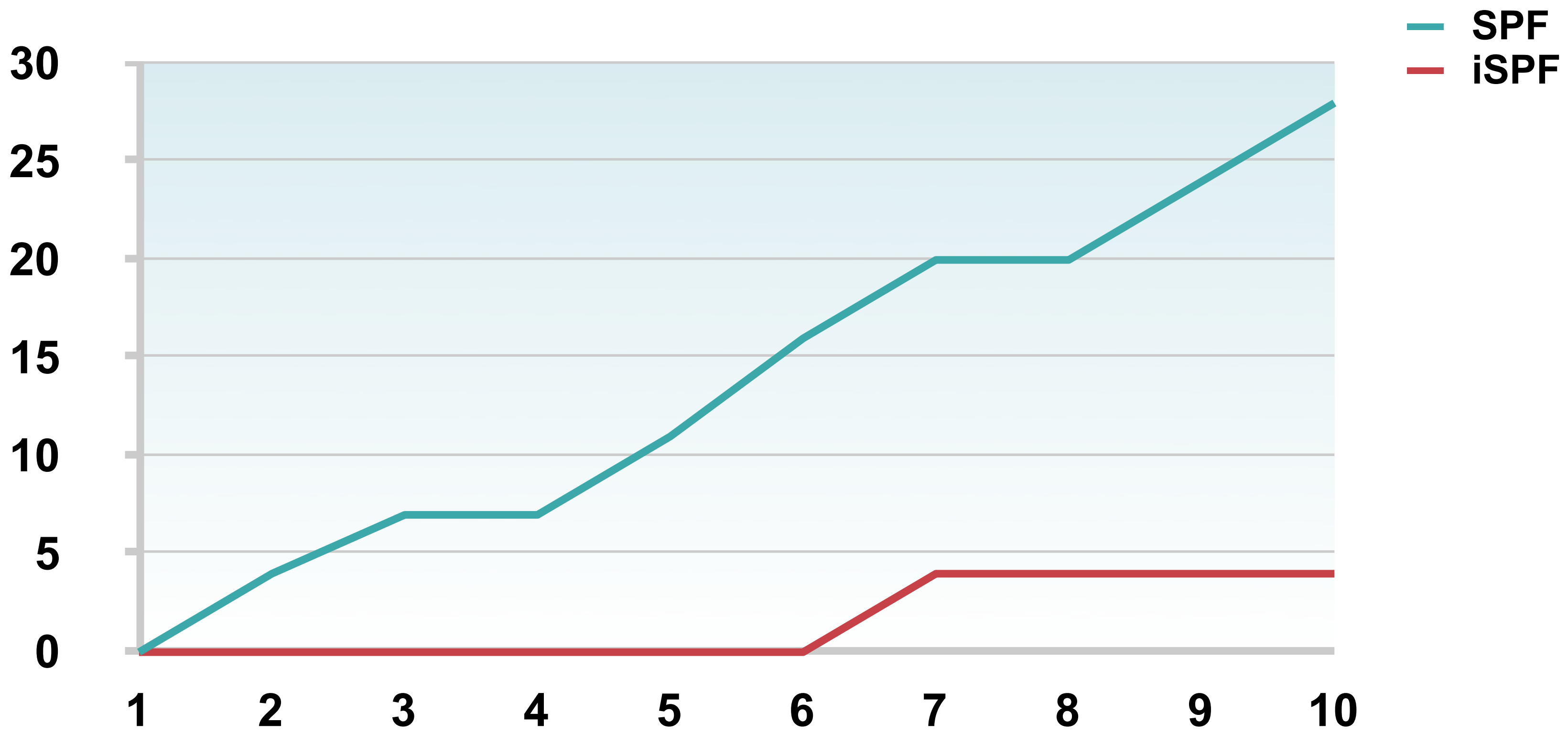
- The gain from iSPF depends on how far (topologically) the change happens from the calculating node
- If the change affects only a small part of Shortest Path Tree (SPT), the gain is significant
  - We were able to run SPF and update the Routing Table for the 1000 node network in less then 10 ms
- If the change is close to the calculating node it is likely a larger portion of the SPT will be affected, reducing the impact of iSPF

Incremental SPF: Overview

- There are always nodes closer to the topology change and nodes that are more remote
- Flooding takes some time—nodes that are most remote from the change are usually notified last
- If full SPF runs on all nodes regardless of the change, then routers notified last about it will converge last (giving that it takes same amount of time to run SPF on each node)
- With iSPF, the more remote the node is from the change, the less work it needs to do during iSPF, resulting in faster network-wide convergence

Incremental SPF: Convergence Times

Time It Takes to Run the SPF with a Link Flap



OSPF Incremental SPF

- **Supported: 12.0(24)S, 12.2(18)S, 12.3(2)T**
- **Configuration:**

```
router ospf <process number>
  ispf

sh ip ospf

Routing Process "ospf 1" with ID 170.99.99.99 and Domain ID
0.0.0.1

 Supports only single TOS(TOS0) routes

 Supports opaque LSA

 It is an area border and autonomous system boundary router

 Redistributing External Routes from,

 SPF schedule delay 5 secs, Hold time between two SPFs 10 secs

 Incremental-SPF enabled

 Minimum LSA interval 5 secs. Minimum LSA arrival 1 secs
```

FAST CONVERGENCE TUNING

BORDER GATEWAY PROTOCOL (BGP)

BGP Convergence Tuning

- BGP and IGP convergence tuning have a different focus
  - IGP convergence—rebuild the topology quickly following an event
  - BGP convergence—transfer large amounts of prefix information very quickly
- The magnitude of time involved is also different
  - IGP—subsecond
  - BGP—seconds to minutes
- Fast IGP convergence plays a role in maintaining availability for BGP prefixes
  - Often topological changes can result in no BGP changes, the IGP updates the next-hop information for BGP prefixes

BGP Convergence: Peer Groups

- Peer groups are more than a configuration simplification
- Update is formatted once for peer group leader, replicated for additional peers, provided they are in sync
- Update replication is much faster than update formatting

Convergence: Test Environment

- 7206 VXR w/ NPE-300 and 256MB DRAM
- Cisco IOS® 12.0(15)S1 and 12.0(23)S
- Single eBGP peering on which prefixes are received, then advertised over 50 iBGP sessions
- BGP is converged when table version for all peers is equal and the BGP InQ and BGP OutQ are 0
- Connectivity to all peers over same Fast Ethernet interface

Convergence: Path MTU Discovery

- The default TCP maximum segment size (MSS) is set to 536, based on the expected minimum MTU of 576
- Current typical minimum MTU size is 1500 when Fast Ethernet is used, and 4470 when ATM and POS are deployed
- If a TCP MSS of 1460 is used, 3x more prefixes will fit in a single UPDATE, for a TCP MSS of 4430, 8x more prefixes will fit in the UPDATE
- Configuration:

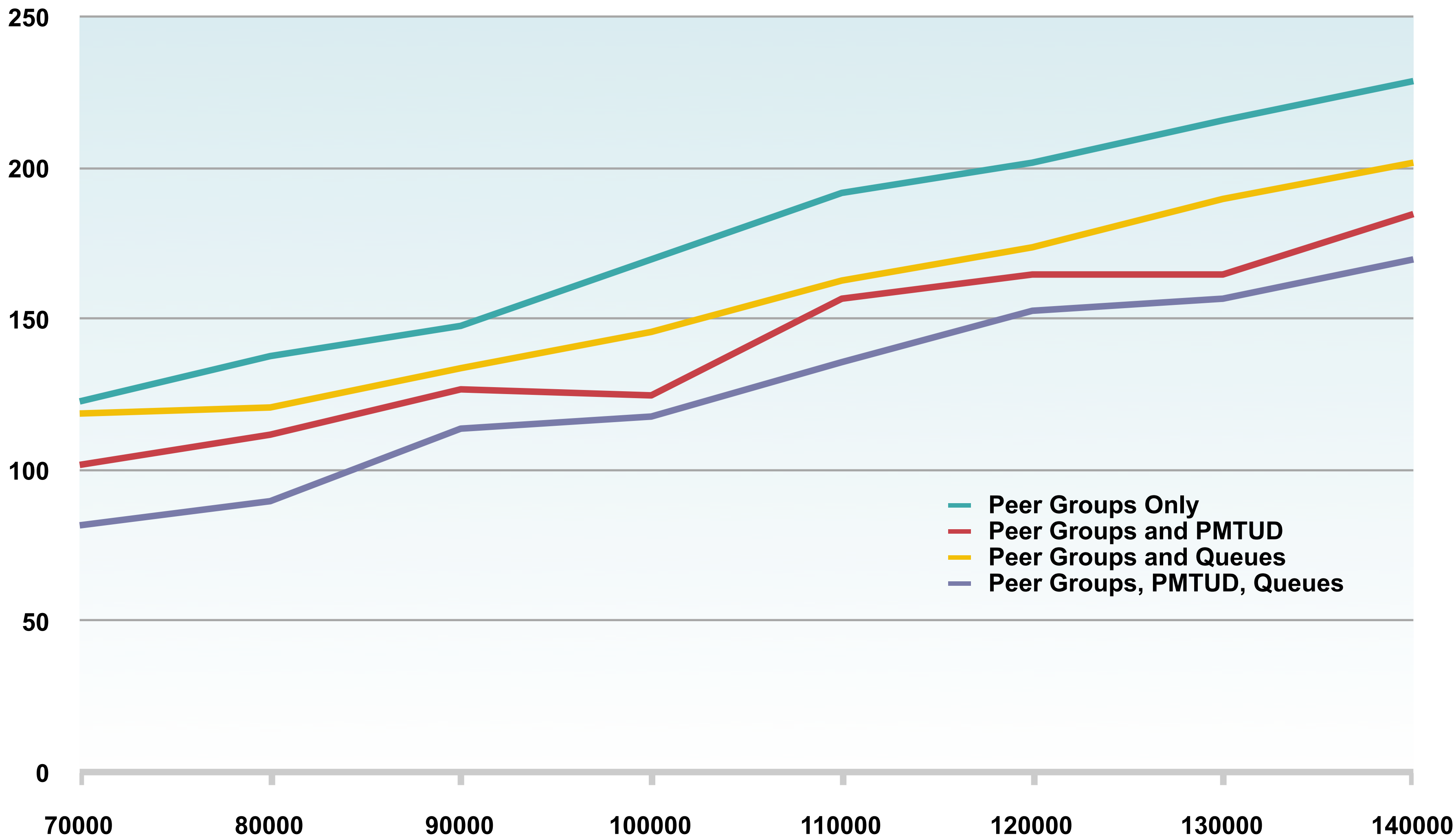
```
ip tcp path-mtu-discovery
```

BGP Convergence: Packet Drops

- The use of peer groups greatly increases the rate at which the router can send BGP UPDATE messages
- The returning TCP ACKs can overflow the input hold queue, resulting in lost ACKs and TCP backoff
- Will result in peers losing sync with peer-group leader!

$$\text{Hold Queue Size} = \frac{\text{Window Size}}{2 * \text{MSS}} * \text{Peer Count}$$

Convergence: Peer Groups/PMTUD/Queues



HIGH-AVAILABILITY DEPLOYMENT

SUMMARY

Getting to 4 Nines

Roadblocks to 4 Nines (99.99%)

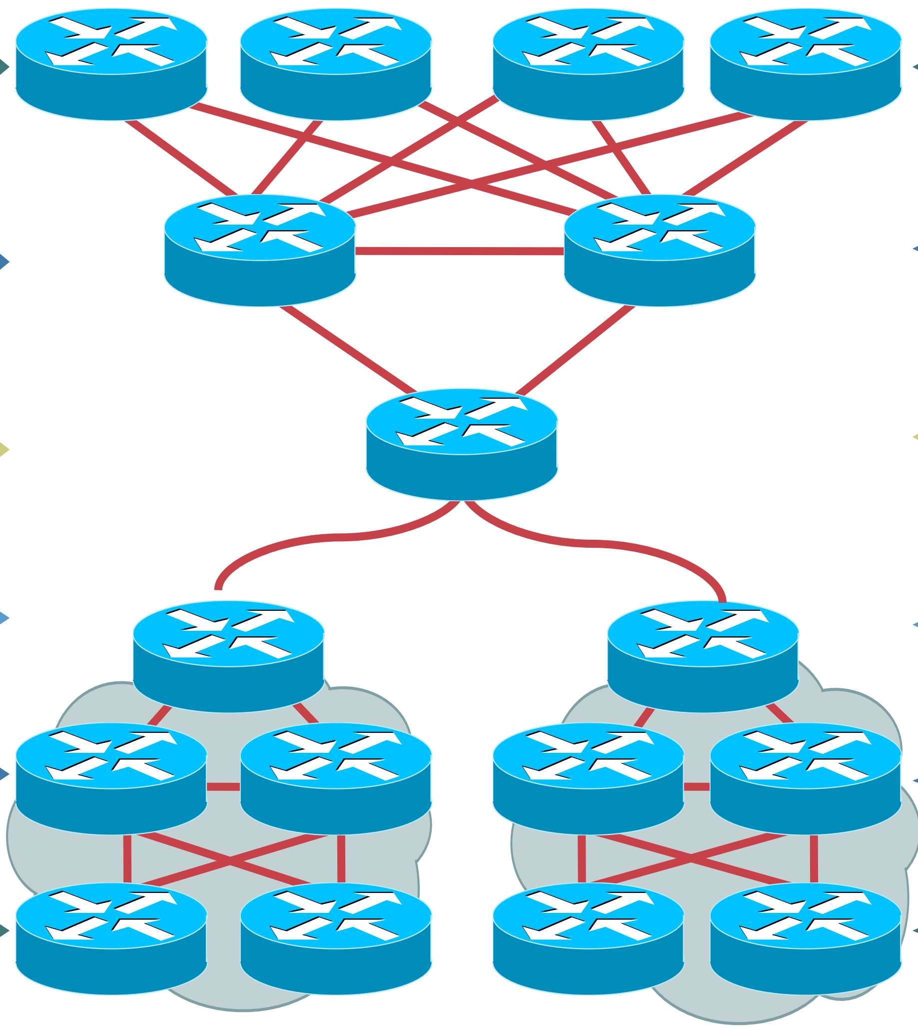
- Single point of failure
  (edge card, edge router, single trunk)
- Outage required for hardware and software upgrades
- Long recovery time for reboot or switchover
- No tested hardware spares available on site
- Long repair times due to a lack of troubleshooting guides and process
- Inappropriate environmental conditions

Getting to 5 Nines

Roadblocks to 5 Nines (99.999%)

- High probability of redundancy failure (failure not detected—redundancy not implemented)
- High probability of double failures
- Long convergence time for rerouting traffic around a failed trunk or router in the core
- Rely on manual operations

NSF/SSO: Deployment Strategies



Service Provider Core	HA Features May Provide Some Benefit
Service Provider Distribution	Good Position for NSF-Aware Routers
Service Provider Access	Primary Deployment for SSO/NSF Capable Routers
Enterprise Access	Deployment of SSO/NSF Capable or Aware Routers
Enterprise Distribution	NSF-Aware Routers
Enterprise Core	HA Features May Provide Some Benefit

Q AND A

Complete Your Online Session Evaluation!

Por favor, complete el formulario de evaluación.

Muchas gracias.

Session ID: RST – 3212

HIGH AVAILABILITY IN IP ROUTING