HIGH AVAILABILITY IN IP ROUTING

SESSION RST-4312

Agenda

Cisco.com

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

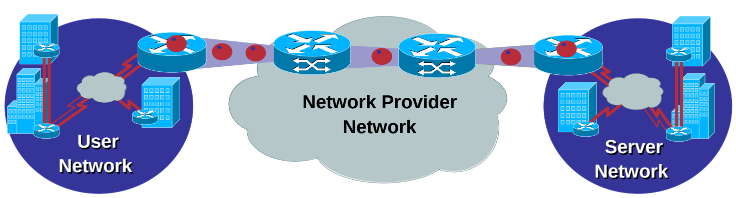
HIGH AVAILABILITY OVERVIEW

Availability Definitions

Cisco.com

Availability

- **The probability that an item (or network, etc.) is operational, and functional as needed, at any point in time**
- **Or, the expected or measured fraction of time the defined service, device, or area is operational; annual uptime is the amount (in days, hrs., min., etc.) the item is operational in a year**



Availability Definitions

Cisco.com

Availability

- **Availability = (MTBF—MTTR)/MTBF**
  - Useful definition for theoretical and practical
- **MTBF is mean time between failure**
  - What, when, why, and how does it fail?
- **MTTR is mean time to repair**
  - How long does it take to fix?

What Is High Availability?

Cisco.com

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	

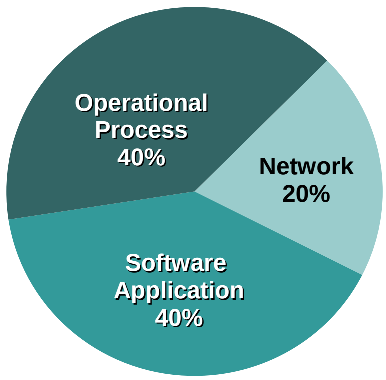
DPM = Defects per Million (Hours of Running Time)

Causes of Unscheduled Network Downtime

Cisco.com

- Change
- Communication
- Process
- Design

- Hardware
- Software
- Link
- Power/env
- Resource utilization

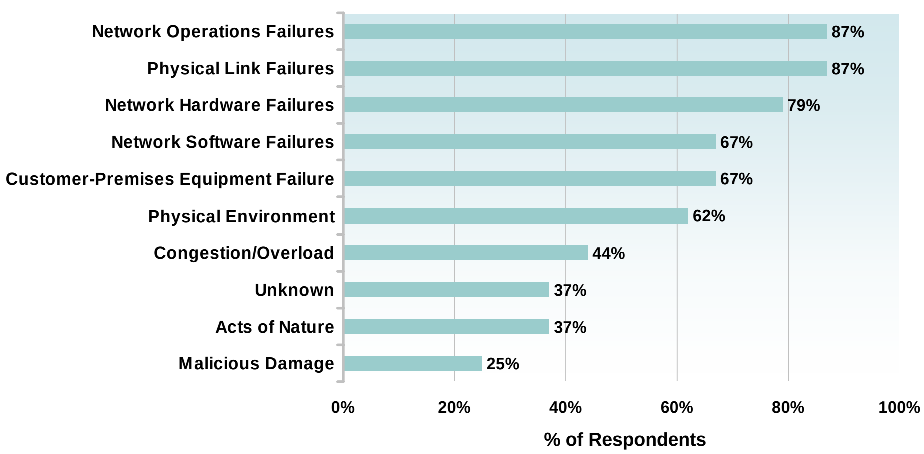


Source: Gartner

RST-4312
9723_05_2004_c2

Causes of Unscheduled Downtime

Cisco.com

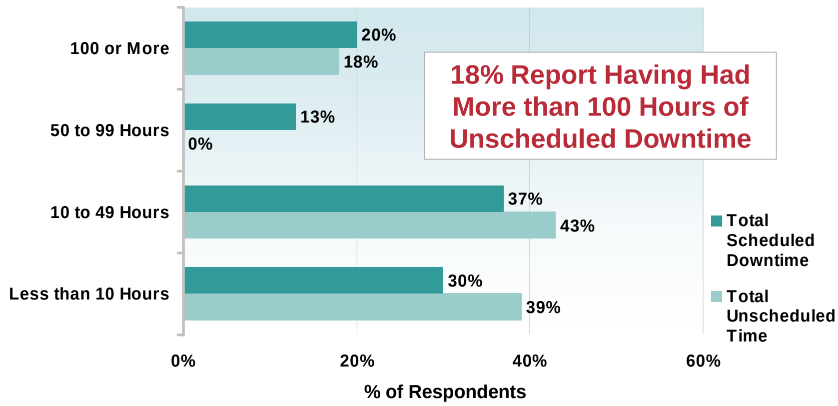


Source: Sage Research, IP Service Provider Downtime Study: Analysis of Downtime Causes, Costs, and Containment Strategies, August 17, 2001, Prepared for Cisco SPLOB

RST-4312
9723_05_2004_c2

Duration of Downtime

Cisco.com



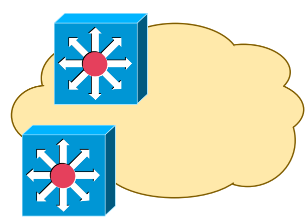
Source: Sage Research, IP Service Provider Downtime Study: Analysis of Downtime Causes, Costs, and Containment Strategies, August 17, 2001, Prepared for Cisco SPLOB

Hardware

Cisco.com

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

Cisco.com

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

What Is Your Availability Level?

Cisco.com

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?

Cisco.com

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

RST-4312
9723_05_2004_c2

What Is Your Availability Level?

Cisco.com

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

RST-4312
9723_05_2004_c2

PROTOCOL-INDEPENDENT FFEATURES

IP Event Dampening

Cisco.com

- **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**

 © 2004 Cisco Systems, Inc. All rights reserved.

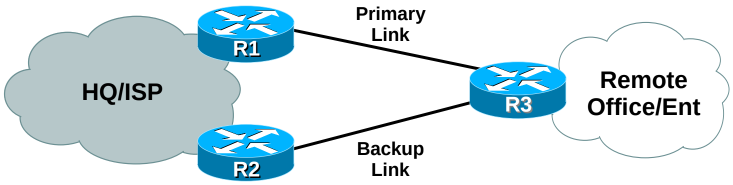
IP Event Dampening: Concept

Cisco.com

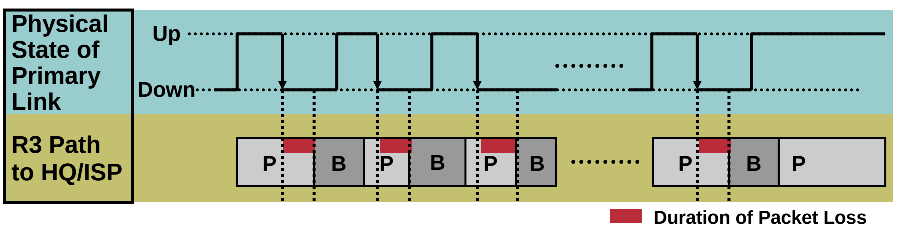
- Takes the concept of BGP route-flap dampening and applies it at the interface level, so all IP-routing protocols can benefit
- Tracks interface flapping, applying a "penalty" to a flapping interface
- Puts the interface in "down" state from routing protocol perspective if the penalty is over a threshold 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

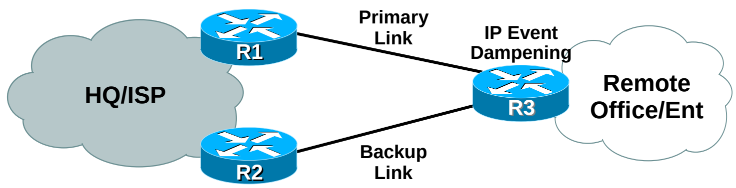


Link Flapping Causes Routing Reconvergence and Packet Loss

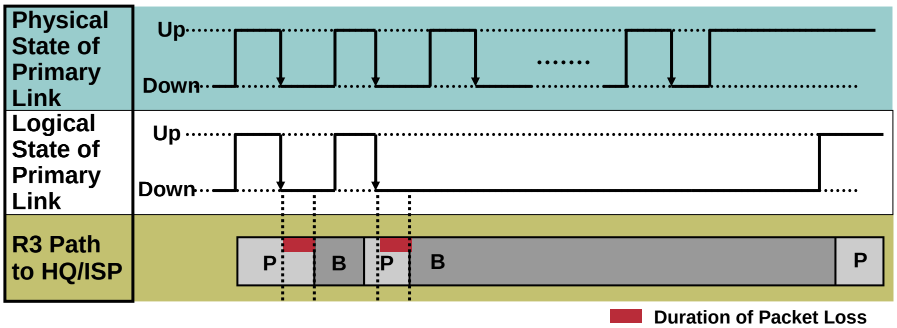


IP Event Dampening: Deployment

Cisco.com



IP Event Dampening Absorbs Link-Flapping Effects on Routing Protocols



IP Event Dampening: Algorithm

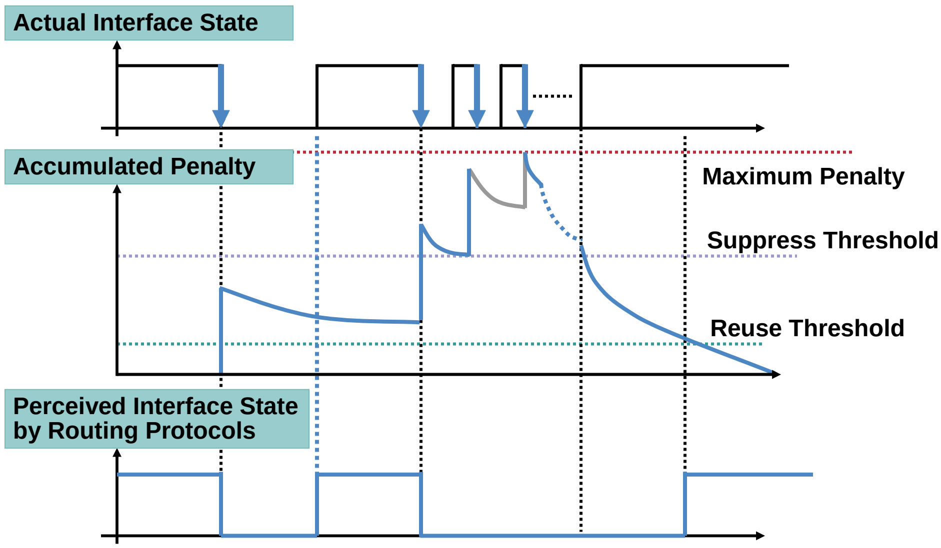
Cisco.com

```
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

Cisco.com



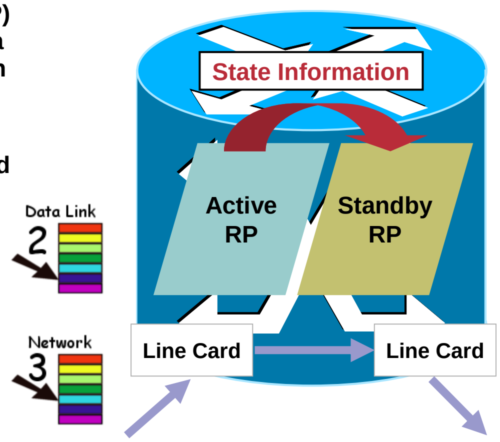
NON-STOP FORWARDING

GRACEFUL RESTART AND STATEFUL SWITCHOVER

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

Cisco.com

- 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

Cisco.com

- **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

Cisco.com

- **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

Cisco.com

- **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

Cisco.com

- 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

Cisco.com

- 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

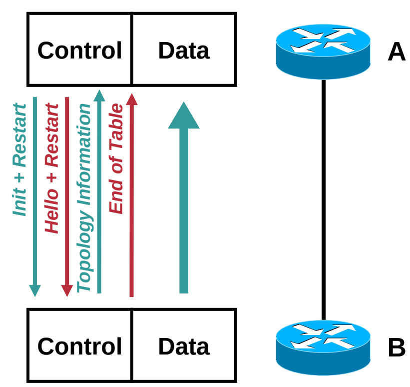
  NSF aware—12.2(15)T

NON-STOP FORWARDING

EIGRP

EIGRP GR/NSF Fundamentals

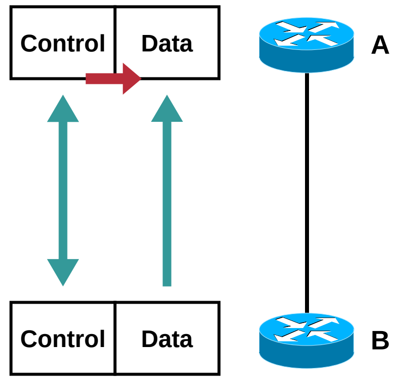
- The signal in EIGRP is an update with the ***initialization*** and ***restart*** (RS) bits set
- A sends its hellos with the restart bit set until GR is complete
- B transmits the routing information it knows to A
- When B is finished sending information, it sends a special end of table signal so A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



EIGRP GR/NSF Fundamentals

Cisco.com

- 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

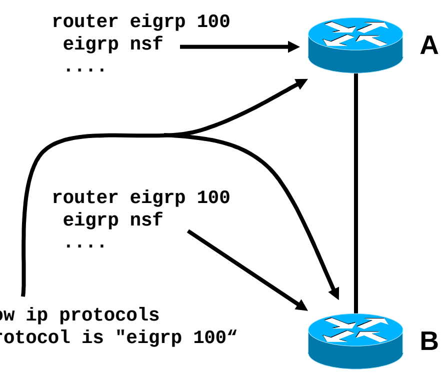


EIGRP GR/NSF Fundamentals

Cisco.com

- 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 eigrp 100
 eigrp nsf
 ....
```




```
router eigrp 100
 eigrp nsf
 ....
```

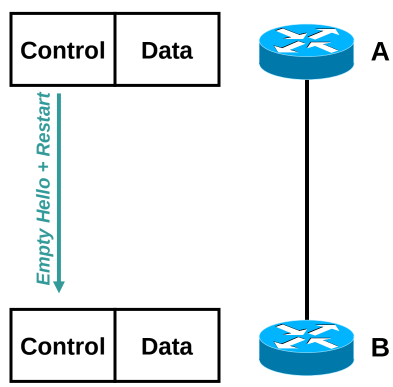
```
router#show ip protocols
Routing Protocol is "eigrp 100“
....
Redistributing: eigrp 100
EIGRP NSF-aware route hold timer is 240s
Automatic network summarization is in effect
Maximum path: 4
....
```

NON-STOP FORWARDING

OSPF

OSPF GR/NSF Fundamentals

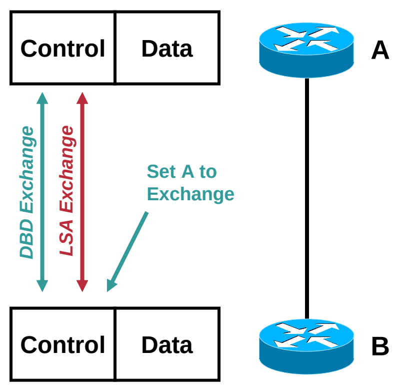
- OSPF uses an extension to the hello packets called link-local signaling
- The first hello A sends to B has an empty neighbor list; this tells B that something is wrong with the neighbor relationship
- A sets the restart bit in its hello, which tells B that A is still forwarding traffic, and would like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



OSPF GR/NSF Fundamentals

Cisco.com

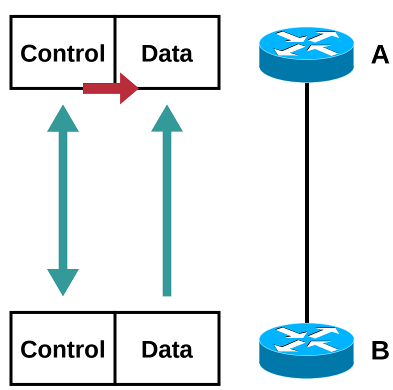
- B moves A into the exchange state, and uses out of band signaling (OOB) to resynchronize their databases
- This process is the same as initial database synchronization, but it uses different packet types



OSPF GR/NSF Fundamentals

Cisco.com

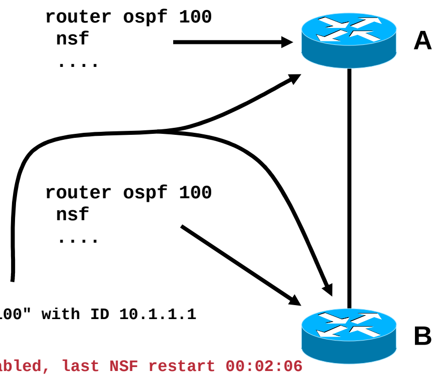
- When A and B have resynchronized their databases, they place each other in full state, and run SPF
- After running SPF, the local routing table is updated, and OSPF notifies CEF
- CEF then updates the forwarding tables, and removes all information marked as stale



OSPF GR/NSF Fundamentals

Cisco.com

- 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
```


OSPF GR/NSF Fundamentals

Cisco.com

- Out-of-band resynchronization is described in draft-nguyen-ospf-oob-resync-00.txt

  http://www.ietf.org/internet-drafts/draft-nguyen-ospf-oob-resync-04.txt

- The link-local signaling extensions to OSPF's hello packets are described in draft-nguyen-ospf-lls-00.txt

  http://www.ietf.org/internet-drafts/draft-nguyen-ospf-lls-04.txt

- The process of restarting using the above drafts is described in draft-nguyen-ospf-restart-00.txt

  http://www.ietf.org/internet-drafts/draft-nguyen-ospf-restart-04.txt

- OSPF graceful-restart documentation:

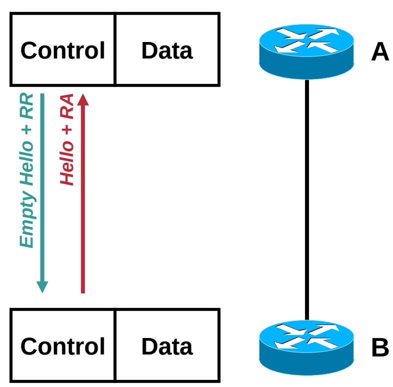
  http://www.cisco.com/en/US/partner/products/sw/iosswrel/ps1839/products_feature_guide09186a0080153edd.html

NON-STOP FORWARDING

IS-IS

IS-IS GR/NSF Fundamentals

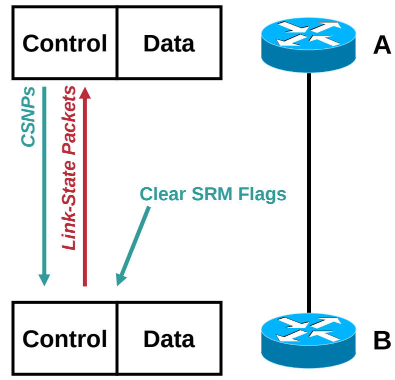
- IS-IS adds a new TLV to the hello packet, the restart option; the restart-option TLV contains a Restart Request (RR) bit and a Restart Acknowledgement (RA) bit
- A transmits its hellos with an empty neighbor list, and the RR-bit set
- B transmits hellos to A with the RA-bit set



IS-IS GR/NSF Fundamentals

Cisco.com

- B then clears the flags which indicate routing data that needs to be transmitted to A (the SRM flags)
- A and B then use IS-IS' normal synchronization process using Complete Sequence Number Packets (CSNPs) to describe their databases, and exchanging Link-State Packets (LSPs)

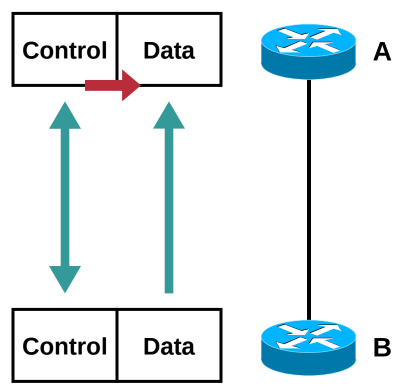


RST-4312
9723_05_2004_c2

IS-IS GR/NSF Fundamentals

Cisco.com

- 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 IS-IS notifies CEF
- CEF then updates the forwarding tables, and removes all information marked as stale

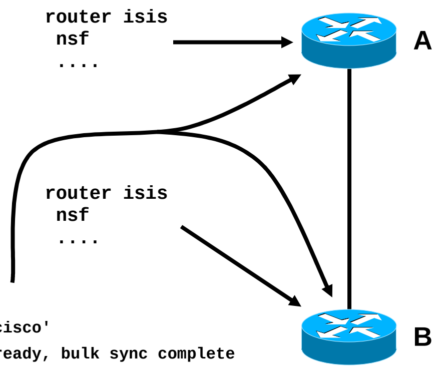


RST-4312
9723_05_2004_c2

IS-IS GR/NSF Fundamentals

Cisco.com

- Use the *nsf* command under the *router isis* configuration mode to enable graceful restart
- *Show isis nfs* 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
```


IS-IS GR/NSF Fundamentals

Cisco.com

- IS-IS Graceful Restart is described in draft-shand-isis-restart-01.txt

  http://www.ietf.org/internet-drafts/draft-shand-isis-restart-05.txt

- IS-IS Graceful Restart documentation:

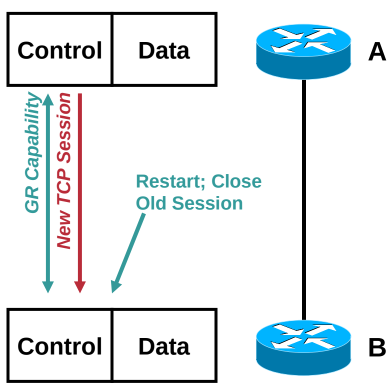
  http://www.cisco.com/en/US/partner/products/sw/iosswrel/ps1839/products_feature_guide09186a00801541c7.html

NON-STOP FORWARDING

BGP

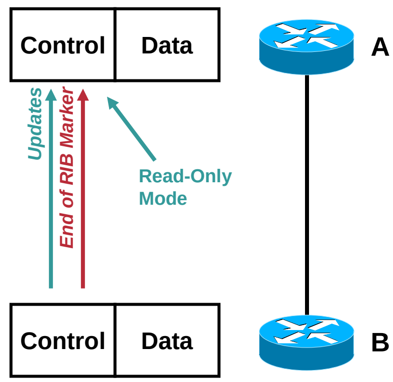
BGP GR/NSF Fundamentals

- 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

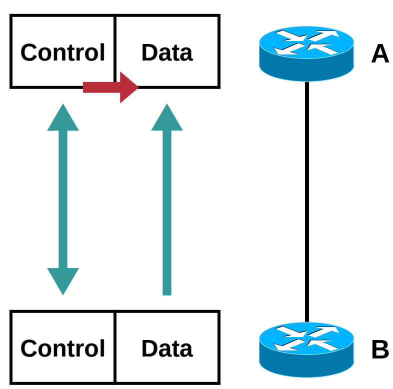
- B transmits updates containing its BGP table (it's local RIB out)
- A goes into read-only mode, and does not run the bestpath calculations until its B has finished sending updates
- When B has finished sending updates, it sends an end of RIB marker, which is an update with an empty withdrawn NLRI TLV



BGP GR/NSF Fundamentals

Cisco.com

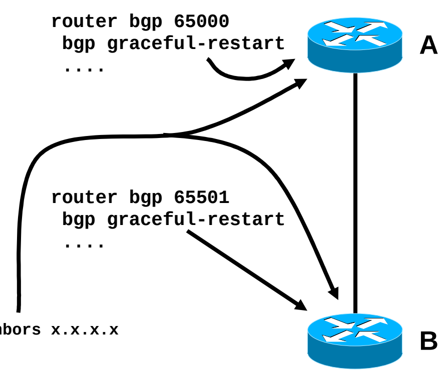
- When A receives the end of RIB marker, it runs bestpath, and installs the best routes in the routing table
- After the local routing table is updated, BGP notifies CEF
- CEF then updates the forwarding tables, and removes all information marked as stale



BGP GR/NSF Fundamentals

Cisco.com

- 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#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
```

RST-4312
9723_05_2004_c2

BGP GR/NSF Fundamentals

Cisco.com

- BGP Graceful Restart is described in draft-ietf-idr-restart-09.txt

  http://www.ietf.org/internet-drafts/draft-ietf-idr-restart-09.txt

- Cisco's implementation of BGP Graceful Restart:

  http://www.cisco.com/en/US/partner/products/sw/iosswrel/ps1839/products_feature_guide09186a008015fede.html

  http://www.cisco.com/en/US/partner/tech/tk826/tk364/technologies_white_paper09186a008016317c.shtml

RST-4312
9723_05_2004_c2

FAST CONVERGENCE TUNING

LINK-STATE PROTOCOLS

RST-4312
9723_05_2004_c2

Network Convergence

Cisco.com

- **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**

RST-4312
9723_05_2004_c2

Network Convergence

Cisco.com

- 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

IS-IS Subsecond Hellos

Cisco.com

- **Supported: 12.0(5)T**
- **IS-IS hello interval**

  `isis hello-interval { seconds | minimal }`

  Time between two consecutive IIH PDUs

  Decrease to detect rapid adjacency changes
- **IS-IS hello multiplier**

  `isis hello-multiplier value`

  Value—range 3–20

  Calculates hold time before neighbor is lost

  Increase when frequent packet loss on interval

RST-4312
9723_05_2004_c2

Fast Hellos: Scaling Issues

Cisco.com

Scaling Is a Major Issue

300 Interfaces x 10 Neighbors/Interface = 3000 Neighbors

Three Hello Packets per Second on Each Interface

Router has to Generate 900 Hellos per Second

3000 Neighbors Each Send Three Hellos per Second to this Router

Router has to Accept and Process 9000 Hellos per Second

Router has to Deal with 9900 Hellos per Second

One Hello Every 10,000th of a Second

RST-4312
9723_05_2004_c2

FAST CONVERGENCE TUNING

EVENT PROPAGATION (LINK STATE)

Network Convergence: OSPF Event Propagation

- **Initial LSA generation delay**

  OSPF_LSA_DELAY_INTERVAL—500 ms delay

  Only router and network LSA generation delayed

- **Recurring LSA origination delay**

  MinLSInterval

  The minimum time between distinct originations of any particular LSA; the value of MinLSInterval is set to five seconds

- **LSA arrival throttling**

  MinLSArrival

  "For any particular LSA, the minimum time that must elapse between reception of new LSA instances during flooding. LSA instances received at higher frequencies are discarded. The value of MinLSArrival is set to one second."

OSPF LSA Generation Exponential Backoff

Cisco.com

- LSAs generated quickly after the initial event
- Repeated events increase the regeneration delay
- Supported: 12.0(25)S, 12.2(18)S, 12.3(2)T
- Configuration:

  timers throttle lsa all *lsa-start lsa-hold lsa-max*

  timers lsa arrival *timer*

  all values are in ms

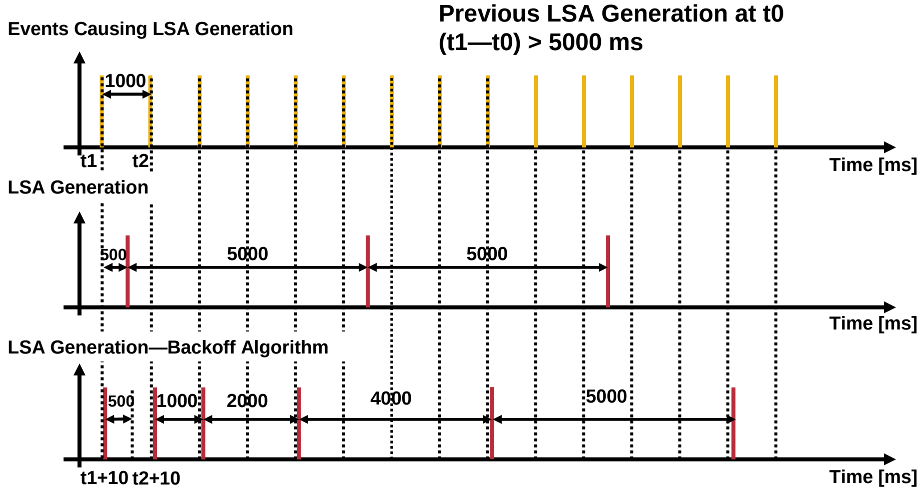
Note: MinLSArrival Must Be <= lsa-hold

LSA Generation Exponential Backoff

Cisco.com

timers throttle lsa all 10 500 5000

Previous LSA Generation at t0
(t1—t0) > 5000 ms



IS-IS LSP-Generation Exponential Backoff

Cisco.com

- IS-IS has not initial delay or arrival throttling
- The parameter ordering different from OSPF!
- Supported: 12.0(11)ST, 12.0(19)S, 12.0(5)T
- Configuration:

  lsp-gen-interval *lsp-max lsp-start lsp-hold*

  - lsp-max—(sec)
  - lsp-hold—(msec)
  - lsp-start—(msec)

Network Convergence: Event Propagation

Cisco.com

- LSA/LSP has to be processed on each node
  - Detect if the LSA/LSP is newer/older
  - If the LSA/LSP is newer, detect if it carries any change
    - Number of links in the LSA/LSP (link comparison)
    - Size of the database (search)
  - If change detected schedule SPF (full/partial)
  - Install in the database
- Above processing time is rather small
  - Router LSA with 10 links: 0.5 ms
  - Router LSA with 100 links: 1 ms

Network Convergence: Event Propagation

Cisco.com

- If the LSA/LSP is declared newer, it's flooded over a certain set of interfaces
    - Excluding the neighbor from which the LSA/LSP has been received
- The LSA/LSP is not flooded immediately
    - Link-state update packets are paced
    - Pacing timer is 33 ms by default (jittered by 10%)

Network Convergence: OSPF Event Propagation

Cisco.com

- With default values and no retransmission each node can add 33-ms delay to the event propagation
- Supported: 12.2(4)T, 12.2(18)S, 12.0(25)S
- Configuration:
    - default values are 33 msec/66 msec

```
timers pacing flood value
timers pacing retransmission value
```

Network Convergence: IS-IS Event Propagation

Cisco.com

- IS-IS flooding interval is configurable per interface
- The default rate is one LSP every 33 ms

  interface x/y

  lsp-interval *value*

- Aggressive fast-convergence tuning may result in SPF being performed prior to flooding the LSP that triggered SPF
- Supported: 12.0(27)S, 12.3(7)T
- Configuration:

  router isis <tag>

  fast-flood *lsp-number*

FAST CONVERGENCE TUNING

EVENT PROCESSING (LINK STATE)

OSPF SPF Exponential Backoff

Cisco.com

- **The SPF-DELAY and SPF-HOLDTIME protect the router as the cost of convergence time**
- **Supported: 12.0(25)S, 12.2(18)S, 12.3(2)T**
- **Configuration:**

```
Timers throttle spf spf-start spf-hold spf-max
```

all values are in ms

RST-4312
9723_05_2004_c2

IS-IS SPF Exponential Backoff

Cisco.com

- **Supported: 12.0(5)T**
- **Configuration:**

```
spf-interval spf-max spf-start spf-hold
```

<spf-max> Maximum time between SPF calculations (sec)

<spf-start> Time between first trigger and first SPF (msec)

<spf-hold> Time between first and second SPF (msec)

```
prc-interval prc-max prc-start prc-hold
```

<prc-max> Maximum time between PRC calculations (sec)

<prc-start> Time between first trigger and first PRC (msec)

<prc-hold> Time between first and second PRC (msec)

RST-4312
9723_05_2004_c2

SPF Overview

Cisco.com

- **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/ all type-3 LSAs that contribute to the certain destination are processed
  - If triggered by type 5/7 all type-5/7 LSAs that contribute to the certain destination are processed
  - If triggered by type 4 all type-4 LSAs that announce a certain ASBR and all type-5/7 LSAs are processed

SPF Execution Time

Cisco.com

- **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

Cisco.com

- **Router/network LSA triggers full SPF**
  - Some changes does not represent the 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

Cisco.com

- **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

Cisco.com

- **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 news in the LSA**

RST-4312
9723_05_2004_c2

Incremental SPF: Overview

Cisco.com

- **Once the incremental SPF analyzes all new LSAs received during the wait interval a standard Dijkstra will be executed**

  iSPF populates the CANDIDATE_LIST (TENTATIVE_LIST)

  If the CANDIDATE_LIST is empty—no work to do
- **If no change in topology (stub link changed), CANDIDATE_LIST will be empty**

  Stub(s) would be recalculated and some route(s) may need to be updated

RST-4312
9723_05_2004_c2

Incremental SPF: Overview

Cisco.com

- 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 RT 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

Cisco.com

- 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

- **Stub link down event (IP prefix lost):**

FULL SPF:

```
Sep 25 14:07:37.795:   OSPF: Begin SPF at 187751.852ms, process time 149100ms
Sep 25 14:07:37.795:   spf_time 2d04h, wait_interval 10s
Sep 25 14:07:37.839:   OSPF: End SPF at 187751.896ms,Total elapsed time 44ms
Sep 25 14:07:37.839:   Intra: 44ms, Inter: 0ms, External: 0ms
Sep 25 14:07:37.839:   R: 506, N: 786, Stubs: 620
Sep 25 14:07:37.839:   SN: 0, SA: 0, X5: 0, X7: 0
Sep 25 14:07:37.839:   SPF suspends: 0 intra, 0 total
```

INCREMENTAL SPF:

```
Sep 25 14:06:27.715:   OSPF: Begin SPF at 187681.772ms, process time 149016ms
Sep 25 14:06:27.715:   spf_time 2d04h, wait_interval 10s
Sep 25 14:06:27.719:   OSPF: End SPF at 187681.776ms,Total elapsed time 4ms
Sep 25 14:06:27.719:   Incremental-SPF: 0ms
Sep 25 14:06:27.719:   Intra: 0ms, Inter: 0ms, External: 0ms
Sep 25 14:06:27.719:   R: 0, N: 0, Stubs: 1
Sep 25 14:06:27.719:   SN: 0, SA: 0, X5: 0, X7: 0
Sep 25 14:06:27.723:   SPF suspends: 0 intra, 0 total
```


Incremental SPF: Convergence Times

- **Linkup event—part of the network becomes reachable:**

FULL SPF:

```
Sep 25 14:27:13.463:   OSPF: Begin SPF at 188927.520ms, process time 149760ms
Sep 25 14:27:13.463:   spf_time 2d04h, wait_interval 5s
Sep 25 14:27:13.515:   OSPF: End SPF at 188927.572ms,Total elapsed time 52ms
Sep 25 14:27:13.515:   Intra: 48ms, Inter: 0ms, External: 0ms
Sep 25 14:27:13.515:   R: 488, N: 758, Stubs: 598
Sep 25 14:27:13.515:   SN: 0, SA: 0, X5: 0, X7: 0
Sep 25 14:27:13.515:   SPF suspends: 0 intra, 0 total
```

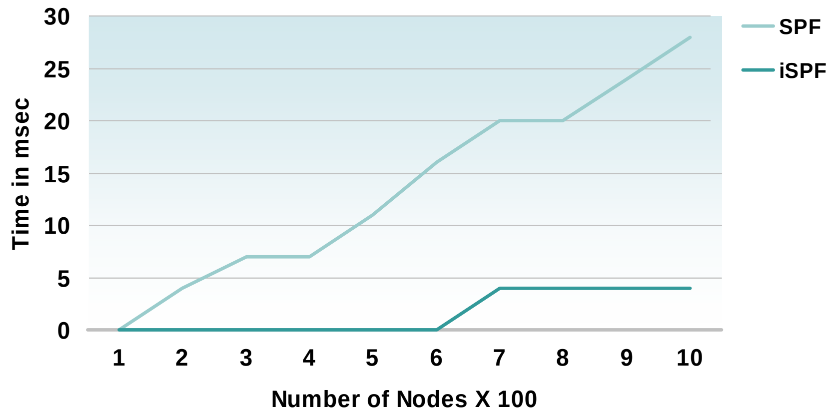
INCREMENTAL SPF:

```
Sep 25 14:23:13.467:   OSPF: Begin SPF at 188687.524ms, process time 149612ms
Sep 25 14:23:13.467:   spf_time 2d04h, wait_interval 5s
Sep 25 14:23:13.479:   OSPF: End SPF at 188687.536ms,Total elapsed time 12ms
Sep 25 14:23:13.479:   Incremental-SPF: 0ms
Sep 25 14:23:13.479:   Intra: 8ms, Inter: 0msExternal: 0ms
Sep 25 14:23:13.479:   R: 18, N: 29, Stubs: 22
Sep 25 14:23:13.479:   SN: 0, SA: 0, X5: 0, X7: 0
Sep 25 14:23:13.479:   SPF suspends: 0 intra, 0 total
```

Incremental SPF: Convergence Times

Time It Takes to Run the SPF with the Transit-Link Flap

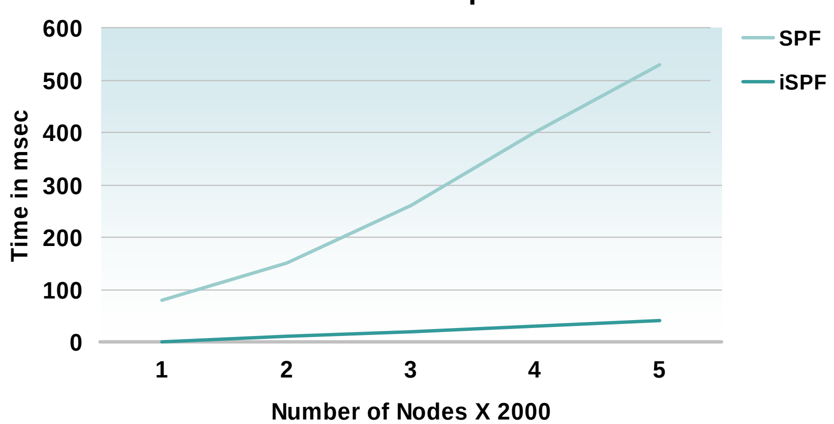
Link Flap



Incremental SPF: Convergence Times

Time It Takes to Run the SPF with the Transit-Link Flap

Link Flap



OSPF Incremental SPF

Cisco.com

- **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
```


IS-IS Incremental SPF

Cisco.com

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

```
router isis <tag>
     ispf { level-1 | level-2 | level-1-2 } [seconds]
```

- **'show isis spf-log detail'**

```
level 1 SPF log
When     SPF Duration  Nodes  Count     Last trigger LSP   Triggers
00:03:25   F   0         1     13       A.00-00           PERIODIC NEWAREA
00:03:15   F   0         5      6       C.00-00           NEWADJ TLVCONTENT
00:01:34   F   0         6      2       D.00-00           NEWLSP TLVCONTENT
00:00:16   I   0         0      1       D.00-00           TLVCONTENT
```

FAST CONVERGENCE TUNING

LINK STATE SUMMARY

Link State Exponential Backoff

Cisco.com

- **Set the SPF and PRC INITIAL WAIT TIME to 1 ms**
- **SPF and PRC INCREMENT:**

  Build a baseline of the time normally required to run SPF in the network; this will generally be around 50 ms

  Set the increment to this plus some padding, 5 to 10 ms
- **SPF and PRC MAXIMUM WAIT TIME:**

  If the normal SPF time is under 100 ms, set the maximum wait to one second

  If it's higher than 100 ms, set it to:
  (S x P)/1000
  S = normal SPF time
  P = maximum percentage of processor utilization for SPF

Link State Exponential Backoff

Cisco.com

- **Set the link state generation initial wait time to 5 ms**
  - This dampens some of the faster link flaps in the network
  - Consider using IP event dampening to quell link flaps, as well
- **Set the increment and the maximum wait times to the same values as you've set the SPF and PRC timers**
  - No point in generating LSPs faster than the routers will actually process them!
- **Tune carrier delay down to 0, IP event dampening will handled any instability from a flapping link**
- **Remember: exponential backoff is NOT dampening!**

FAST CONVERGENCE TUNING

BORDER GATEWAY PROTOCOL (BGP)

BGP Convergence Tuning

Cisco.com

- 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

Cisco.com

- 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
- 12.0(24)S provides support for dynamic update groups, which groups peers dynamically to provide the update replication

Convergence: Test Environment

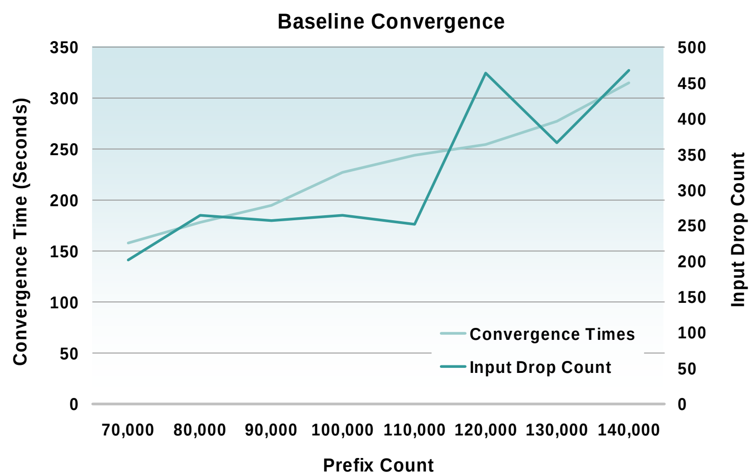
Cisco.com

- 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

RST-4312
9723_05_2004_c2

Convergence: Baseline Testing

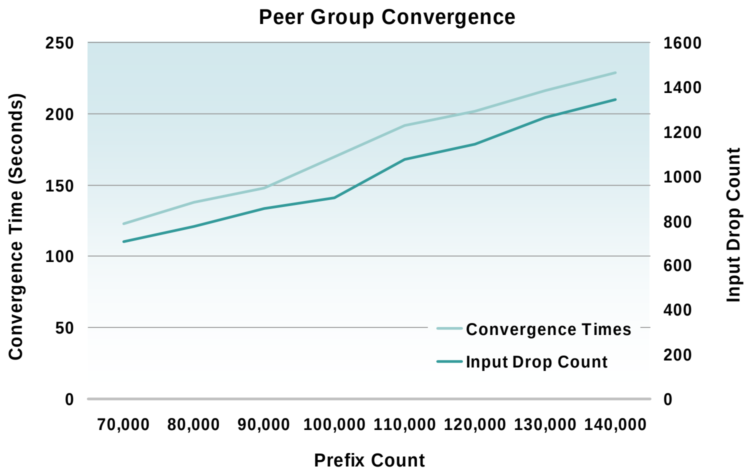
Cisco.com



RST-4312
9723_05_2004_c2

Convergence: Peer Groups

Cisco.com



RST-4312
9723_05_2004_c2

Convergence: Path MTU Discovery

Cisco.com

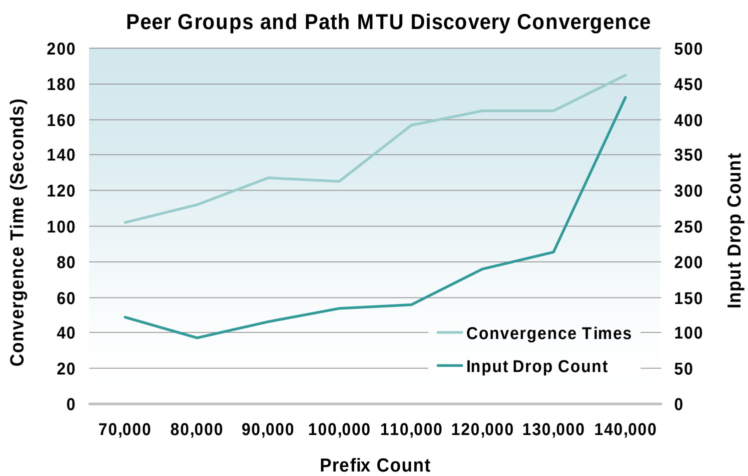
- **The default TCP maximum segment size (MSS) is set to 536, based on the expected minimum MTU of 576**
- **Current typical minimum MTU size is 1500 when Fast Ethernet is used, and 4470 when ATM and POS are deployed**
- **If a TCP MSS of 1460 is used, 3x more prefixes will fit in a single UPDATE, for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

  ip tcp path-mtu-discovery

RST-4312
9723_05_2004_c2

Convergence: Peer Group and PMTUD

Cisco.com



RST-4312
9723_05_2004_c2

BGP Convergence: Packet Drops

Cisco.com

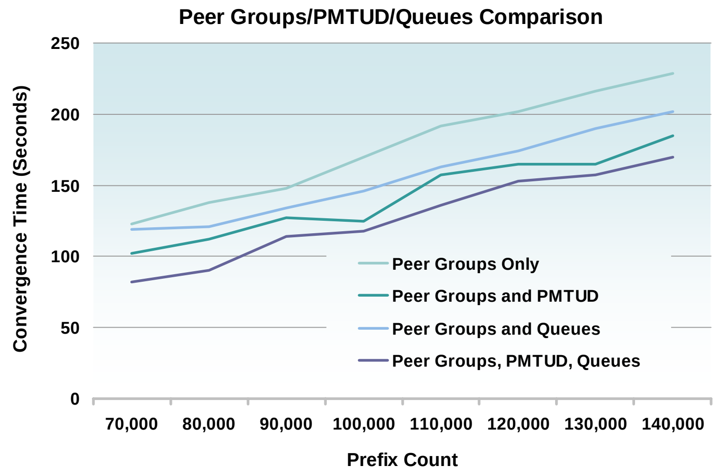
- The use of peer groups greatly increases the rate at which the router can send BGP UPDATE messages
- The returning TCP ACKs can overflow the input hold queue, resulting in lost ACKs and TCP backoff
- Will result in peers losing sync with peer-group leader!

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

RST-4312
9723_05_2004_c2

Convergence: Peer Groups/PMTUD/Queues

Cisco.com



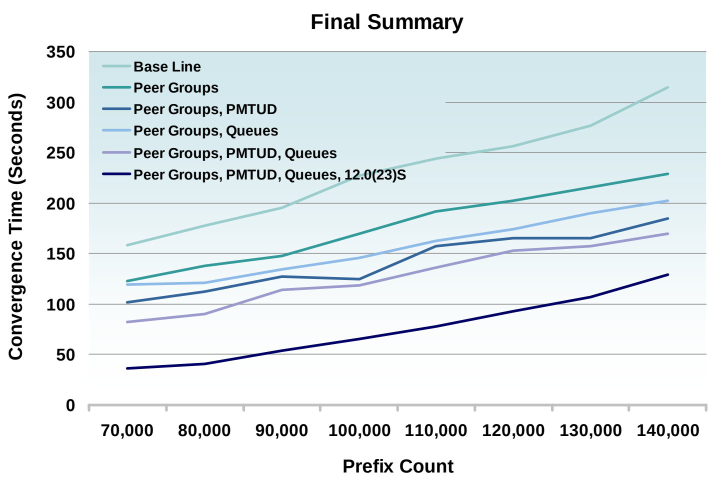
BGP Convergence: Update Packing

Cisco.com

- BGP UPDATES are based on a set of attributes and a list of prefixes that share that particular set of attributes
- Prior to 12.0(19)S, BGP UPDATE messages were not packed optimally
- Waiting until all prefixes are received from a peer prior to processing them and sending updates can further increase the ability to pack efficiently
- Fully supported in 12.0(23)S

Convergence: Full Optimization

Cisco.com



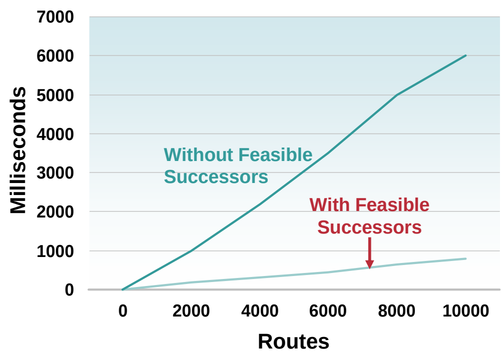
FAST CONVERGENCE TUNING

ENHANCED INTERIOR GATEWAY PROTOCOL (EIGRP)

EIGRP Feasible Successors

Cisco.com

- Whether an alternate path is a feasible successor or not makes a large difference in convergence
- In this test, switching from the best path to a feasible successor takes less than one second; switching to some other neighbor takes about six seconds
- IT'S IMPORTANT TO CONSIDER NOT ONLY THE BEST PATHS THROUGH AN EIGRP NETWORK, BUT ALSO THE FEASIBLE SUCCESSORS

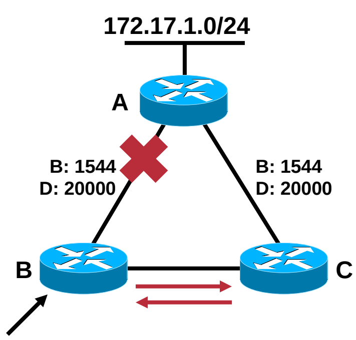


RST-4312
9723_05_2004_c2

EIGRP Feasible Successors

Cisco.com

- In this network, B and C have equal cost paths to A; neither one will see the other as a feasible successor because the feasible distance is equal to the reported distance
- If either link fails, at least one query/reply will be required to converge




```
router-b#sho ip eigrp topo 172.17.1.0
IP-EIGRP (AS 100): Topology entry for 172.17.1.0/24
  State is Passive, Query origin flag is 1, 1 Successor(s), FD is 2172416
  Routing Descriptor Blocks:
  172.17.2.1 (Serial0/0), from 208.0.8.4, Send flag is 0x0
      Composite metric is (2172416/18944), Route is Internal
      ....
  172.17.1.0 (Serial0/3), from 172.17.3.1, Send flag is 0x0
      Composite metric is (2684416/2172416), Route is Internal
      ....
```

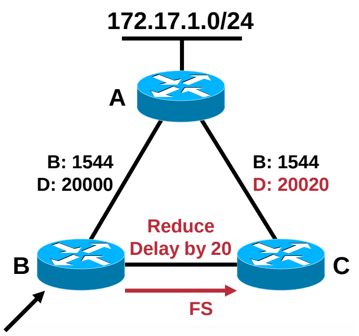
RST-4312
9723_05_2004_c2

Presentation_ID.scr

EIGRP Feasible Successors

Cisco.com

- Increasing the C to A metric, and decreasing the C to B metric by the same amount, will allow B to become C's feasible successor
- But this only works one way; there's no way to make B and C point at each other as feasible successors of each other




```
router-b#sho ip eigrp topo 172.17.1.0
IP-EIGRP (AS 100): Topology entry for 172.17.1.0/24
  State is Passive, Query origin flag is 1, 1 Successor(s), FD is 2172416
  Routing Descriptor Blocks:
  172.17.2.1 (Serial0/0), from 208.0.8.4, Send flag is 0x0
      Composite metric is (2172416/18944), Route is Internal
      ....
  172.17.1.0 (Serial0/3), from 172.17.3.1, Send flag is 0x0
      Composite metric is (2684416/2167296), Route is Internal
      ....
```


EIGRP Feasible Successors

Cisco.com

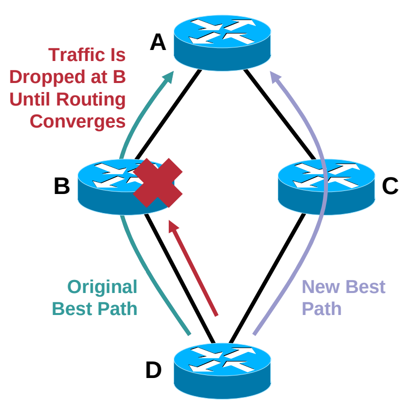
- Whether the next best path is considered loop free by EIGRP (a feasible successor) or not has a large impact on convergence times
- Don't just consider the best path from every point in your network, but also the next best path
- Determine how best to set up your path metrics to improve convergence performance
- No special configuration required to enable fast convergence, it's based on the network design
- ALWAYS USE THE DELAY METRIC TO ENGINEER YOUR ROUTING, NEVER THE BANDWIDTH METRIC

OPERATIONAL PROCESS IMPROVEMENT

OPERATIONAL FEATURES

Graceful Shutdown

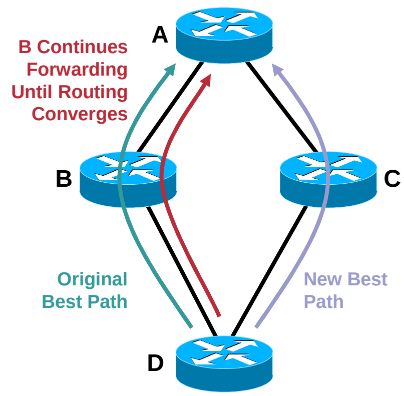
- You want to bring B down for maintenance; the routing protocol will reroute around B
- The packets "in flight" will be lost when B is taken off line, though—and this could be a lot of packets, if these are high-speed links
- It's better to get A and D to route around B while B can still forward traffic, so B can be taken off line gracefully



Graceful Shutdown

Cisco.com

- Graceful shutdown will allow the routing protocols to tear their adjacencies down without impacting the forwarding tables for some short period of time
- Once the protocol has torn its adjacencies down, it will then clean up the forwarding tables



Graceful Shutdown

Cisco.com

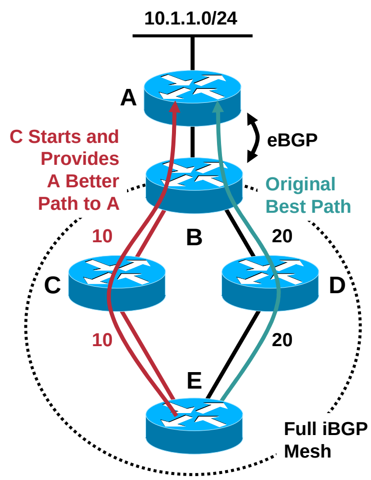
	Protocol Specific Signaling
BGP	CEASE NOTIFICATION
EIGRP	UPDATE with INIT-Bit Set
ISIS	LAN Interfaces: IIH with an Empty "IS Neighbors" TLV Point-to-Point Links: "p2p Adjacency State" TLV Set to "Init"
OSPF	Hello with an Empty Neighbor List

GRACEFUL SHUTDOWN IS CURRENTLY UNDER DEVELOPMENT!
IS-IS SUPPORTED: 12.0(27)S

Wait for BGP

Cisco.com

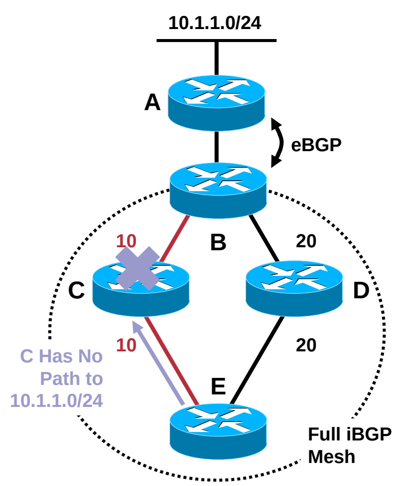
- E is learning 10.1.1.0/24 through iBGP from D with a next hop of A
- E examines the path to A, and finds an IGP route through D to A; it installs this route in the routing table
- C is now inserted into the circuit; after a few seconds, the IGP has converged, and E now chooses C as the best path to A



Wait for BGP

Cisco.com

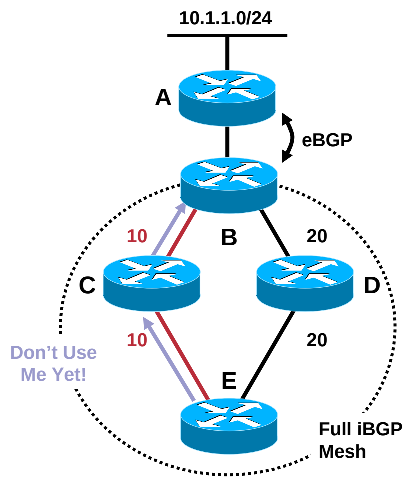
- However, BGP takes much longer to converge if C is accepting full routes (about 150,000 routes) from A; at least five minutes
- When E forwards packets to C for 10.1.1.1, C hasn't finished building its BGP tables, so it doesn't know how to reach this destination
- C drops the packets



Wait for BGP

Cisco.com

- Instead, once the IGP has converged, C signals its IGP neighbors that they should not route this direction
- The IGP remains in this state until BGP notifies the IGP it has converged
- E will continue using D as its best path to A, even though a better one is available, until BGP converges on C



RST-4312
9723_05_2004_c2

Wait for BGP

Cisco.com

- OSPF uses `max-metric router-lsa [on-startup [wait-for-bgp]]` to configure this feature

  Available in 12.2T

  http://www.cisco.com/en/US/partner/products/sw/iosswrel/ps1839/products_feature_guide09186a0080087c09.html

- IS-IS uses `set-overload-bit [on-startup [wait-for-bgp]]` to configure this feature

  Available in 11.3

  http://www.cisco.com/en/US/partner/tech/tk472/tk474/technologies_tech_note09186a00800a4bb1.shtml

RST-4312
9723_05_2004_c2

HIGH-AVAILABILITY DEPLOYMENT

SUMMARY

RST-4312
9723_05_2004_c2

Getting to 4 Nines

Cisco.com

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

RST-4312
9723_05_2004_c2

Getting to 5 Nines

Cisco.com

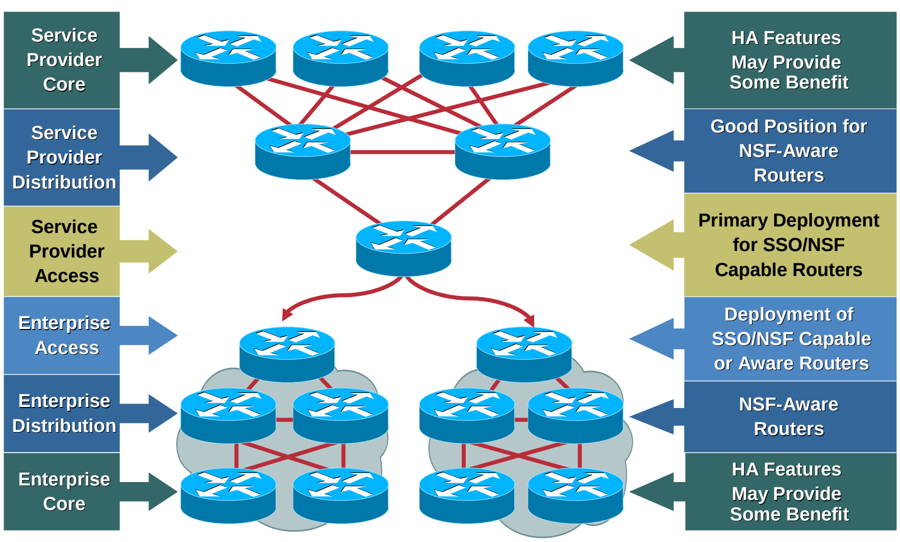
Roadblocks to 5 Nines (99.999%)

- 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

RST-4312
9723_05_2004_c2

NSF/SSO: Deployment Strategies

Cisco.com



RST-4312
9723_05_2004_c2

Q AND A

RST-4312
9723_05_2004_c2

Recommended Reading

Cisco.com

- Continue your Networkers learning experience with further reading for this session from Cisco Press.
- Check the Recommended Reading flyer for suggested books.

Available on-site at the Cisco Company Store

RST-4312
9723_05_2004_c2

Complete Your Online Session Evaluation!

WHAT: Complete an online session evaluation and your name will be entered into a daily drawing

WHY: Win fabulous prizes! Give us your feedback!

WHERE: Go to the Internet stations located throughout the Convention Center

HOW: Winners will be posted on the onsite Networkers Website; four winners per day