

Cisco Nonstop Forwarding for BGP: Deployment and Troubleshooting

Glossary

Nonstop Forwarding	Cisco Nonstop Forwarding (NSF) refers to ability of a router to maintain forwarding state and re-converge routing protocols in the background during a route processor switchover. Because Cisco made extensions to BGP, OSPF and IS-IS, Cisco NSF is a handy term to collectively refer to those extensions.
Stateful Switchover	Cisco Stateful Switchover (SSO) is the process by which information about platform, infrastructure and Layer 2 connectivity is shared between dual route processors. SSO also enables the maintenance of Layer 2 connectivity across a route processor switchover.
Routing Information Base	The aggregate collection of all routing information on a router. The Routing Information Base (RIB) may contain multiple references to different IP destinations.
Forwarding Information Base	The Forwarding Information Base (FIB) is an optimized routing table formed by examining the RIB and selecting only the best paths to any particular IP destination. If load-sharing or load-balancing is enabled, multiple best paths may be selected.
Cisco Express Forwarding	Cisco Express Forwarding (CEF) is a further optimization of the FIB, which allows very fast switching of IP packets. Distributed CEF is a variation of CEF that runs on line cards.
Route Selection Process	The process by which Border Gateway Protocol (BGP) selects the best routes to particular destinations using all available information obtained from peers. Also known as Best-path Selection.
Convergence	The point at which every router on the network has received and processed all routing information from its peer routers.
Autonomous System	Technically, a group of routers under a common administrative control. Practically, a group of routers that share a commonly configured autonomous system number in their BGP configuration.
eBGP	BGP peering connections between routers in different autonomous systems
iBGP	BGP peering connections between routers in the same autonomous system
Interior Gateway Protocol (IGP)	A protocol—typically OSPF, IS-IS or EIGRP—run within an autonomous system to provide next-hop reachability information.



1.0 Overview

Cisco Nonstop Forwarding with Stateful Switchover (NSF with SSO) is a Cisco innovation for routers with dual route processors. Cisco NSF with SSO allows a router that has experienced hardware or software failure of an active route processor to maintain data link layer connections and continue forwarding packets during the switchover to the standby route processor. This forwarding can continue despite lost routing protocol peering arrangements with other routers. Routing information is recovered dynamically, in the background, while packet forwarding proceeds uninterrupted.

Cisco NSF for BGP is a combination of internal system modifications to the various NSF-capable hardware platforms, and external enhancements to the BGP-4 protocol. The modifications to the BGP protocol (BGP Graceful Restart) have been submitted to the Internet Engineering Task Force (IETF):

<http://www.ietf.org/internet-drafts/draft-ietf-idr-restart-06.txt>¹

This document will detail specific changes to the BGP protocol. In addition, it will explore common deployment scenarios for BGP NSF and the basic troubleshooting techniques available to analyze the functionality of this new technology.

2.0 Benefits of Cisco Nonstop Forwarding

In pursuit of higher revenues and profitability, Service Providers and Enterprises are increasingly putting more mission-critical, time sensitive services on their IP infrastructure. One of the key challenges in this business model is achieving and delivering high network availability. This network availability is measured and billed appropriately via a Service Level Agreement (SLA); therefore, users must address the following issues:

- Increase network and node availability during planned or unplanned software restarts, peer resets, and/or hardware (e.g., Route Processor [RP]) changes. In other words, negatively impacting events must be minimized, and maintenance windows decreased.
- Minimize topology changes seen in the network. Topological changes could cause route flapping, consuming expensive CPU cycles on the router, and producing packet jitter and undesirable traffic patterns through sub-optimal routing.
- Reduce Capital Expenditures (CapEx) and Operational Expenditures (OpEx) in deploying a highly available network.

Cisco NSF addresses these issues by:

- Providing transparent RP switchover during a hardware or software fault on a router
- Masking the impact of any failure by localizing the associated topology changes so they do not cascade throughout the entire network
- Reducing CapEx and OpEx by providing intelligent unattended failover within a single router.
- Maintaining original capital investment by enabling features on existing Cisco platforms and route processors

1. Internet drafts are frequently updated. If the URL above is not found, it is recommended that the reader browse to <http://www.ietf.org/internet-drafts/> and search for "idr-restart" to obtain the latest version of the draft.



3.0 Protocol Enhancements

This section provides an examination of how BGP Graceful Restart works. For a more complete discussion, refer to the IETF Internet draft at <http://www.ietf.org/internet-drafts/draft-ietf-idr-restart-06.txt>.

Cisco routers that support dual RPs and Cisco NSF with SSO can maintain Layer 2 data link connections and sufficient forwarding information to continue processing packets during a RP switchover. The ability to maintain Layer 2 connections during such an event is referred to as Cisco SSO.² The ability to continue forwarding packets is implemented by using the existing information in the Forwarding Information Base (FIB) and the Cisco Express Forwarding (CEF) tables until the routing protocols can reconverge. This document will refer to a router going through an RP switchover as the Restarting Router.

However, all of these innovations would be for naught if routers that were peered with the router that performs the switchover (hereafter, the Peer Routers) did not continue to forward packets to it. In order for a Peer Router to continue packet forwarding, several conditions must be met:

- The Restarting Router and the Peer Router must each agree to support BGP Graceful Restart
- The Peer Router must not prematurely declare the Restarting Router as unavailable
- The Peer Router must not communicate any state change in the Restarting Router to any of its peers. This avoids the network-wide detrimental effect on performance associated with the sudden failure of a router
- The Peer Router must send BGP updates to help the restarting NSF router reacquire its BGP Routing Information Base (RIB)
- The Peer Router must signal the completion of the initial routing update by sending the End-of-RIB marker (discussed below)
- In the interim (before the Restarting Router has reacquired the routing information), the Peer Router must mark any routes associated with the Restarting Router as “stale”, but continue to use those routes for packet forwarding

The protocol modifications begin when the initial BGP connection is established. Both the NSF-capable router and its peer indicate their understanding of the BGP Graceful Restart mechanism by exchanging a new BGP capability (#64) during the initial BGP OPEN that establishes the session.

Note that the Peer Router will send Capability 64, regardless of whether it is capable of restartability. Capability 64 does not alone indicate restartability. It can indicate that the router in question has implemented the BGP enhancements specified in the IETF draft. Thus, a Cisco 7200 Series Router that is BGP graceful-restart configured will still advertise Capability 64 to its peers, even though it does not support dual RPs and cannot restart BGP.

Additionally, the NSF-capable router will provide a list of Address Family Identifiers (AFI) and Subsequent Address Family Identifiers (SAFI) for which it has the capability to maintain forwarding state across a BGP Restart. The AFI and SAFI indicate different types of protocols for which BGP can carry information. This would include protocol support, including IPv4, IPv6, MPLS, and Unicast/Multicast routing.

Figure 1 illustrates the significant fields within the new Capabilities 64 exchange, and provides a brief discussion of their usage. Table 1 offers more detailed information about each field.

2. Cisco SSO actually provides capabilities beyond the maintenance of Layer 2 connectivity. It also manages state of all of the supporting platform and infrastructure. Maintaining Layer 2 connectivity is simply the most outwardly visible of the services it provides.



Figure 1
Format of BGP Graceful Restart Capability 64

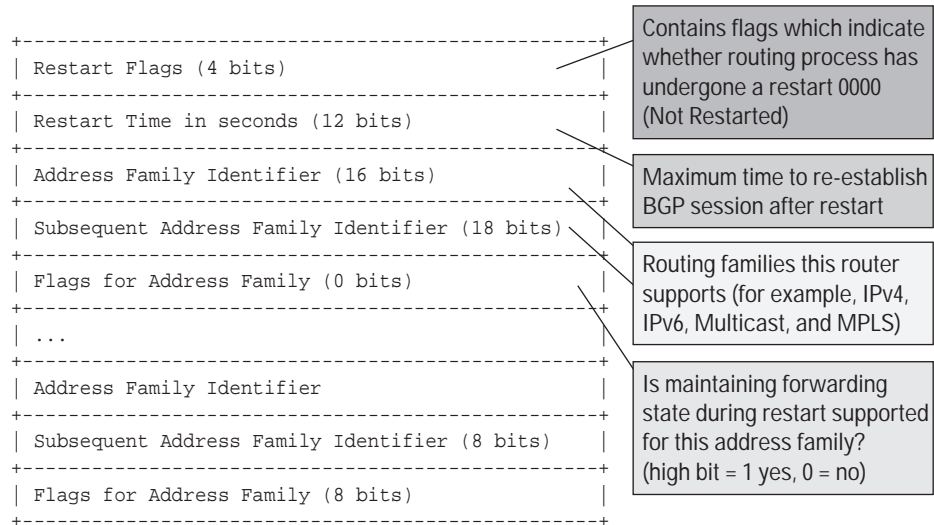


Table 1 BGP Capability 64 Fields

Restart Flags	<p>The high-bit of this field is significant; all other bits must be set to 0. If the high-bit is set, it indicates that the router sending this capability has restarted the BGP process. Possible values are:</p> <ul style="list-style-type: none"> • 0000—Not restarted (normal operation) • 1000—Restarted (BGP Graceful Restart in effect) <p>Both the restartable router and its peer(s) will have all bits set to 0 during normal operation. On restart of BGP, the Restarting Router will set the high-bit to 1.</p>
Restart Time	<p>Each peer, in its Capability 64 exchange, will advertise a restart time in seconds. This represents the 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 considered a failure, and normal BGP recovery procedures take effect.</p> <p>There is no requirement that the restart timers match on BGP peers, although it is generally recommended.</p>
AFI/SAFI	<p>Describes the IP-based protocol(s) that are enabled on this router. The SAFI provides more specific information about the protocol. For example, the AFI could contain a code representing IPv4, and the SAFI would indicate whether the related IPv4 was unicast or multicast.</p>
Address Family Flags	<p>Indicates whether forwarding state is maintained for this AFI/SAFI. This is indicated by setting the high-bit within the field:</p> <ul style="list-style-type: none"> • 00000000—AFI/SAFI forwarding not maintained • 10000000—AFI/SAFI forwarding maintained

When the NSF-capable router performs a route processor switchover, the TCP connection to the Peer Router is cleared; a Peer Router that does not support BGP restart then clears all routes associated with the Restarting Router and no longer forwards packets to it. With BGP Graceful Restart, the Peer Router marks all routes to the Restarting Router as stale, but continues to use them for packet forwarding, based upon the knowledge that the Restarting Router will re-establish the BGP session shortly and that it maintains the capability to forward packets in the interim.



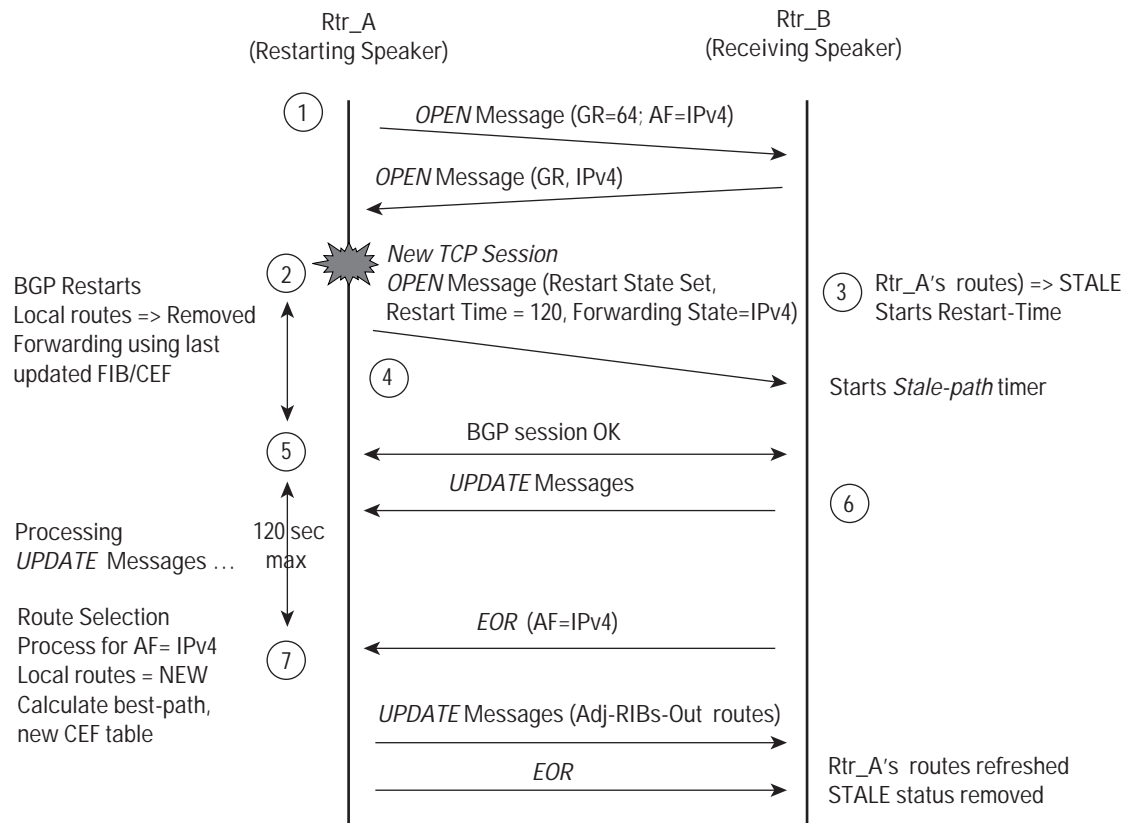
When the Restarting Router's newly active RP opens the new BGP session, it will again send the Graceful Restart capability (#64). However, this time, the restart bit in the Restart Flags portion of the capability exchange will be set. This notifies the Peer Routers that the restart of the BGP process on the Restarting Router caused the disconnect/reconnect.

While continuing to forward packets, the Peer Router refreshes the Restarting Router with any relevant BGP updates. The Peer Router indicates completion of this process by sending an End-of-RIB (EOR) marker. The EOR marker for IPv4 is a BGP update message that is of the minimum length—23 bytes. The EOR does not contain any routes to be added or withdrawn. Essentially, it is an “empty” update, whose sole purpose is to indicate that all available routes have been sent. The EOR marker helps speed convergence, because it allows the router to begin best-path selection as quickly as possible, without waiting for the timer to expire.³

Once the Restarting Router has received all available routes from each peer, it can conduct best-path selection, and send any updates to its Peer Routers. The Restarting Router will also use the EOR to indicate the completion of this process.

Figure 2 provides a graphical representation of this process.

Figure 2
BGP Graceful Restart Procedures



3. Although the EOR marker was introduced as part of the BGP Graceful Restart IETF draft, it is worth noting that it has applicability outside of the context of Graceful Restart. The ability to signal the completion of initial BGP route updates is a generally useful enhancement, and can be used to speed BGP network convergence even if BGP Graceful Restart is not concurrently deployed.



Consider the step-by-step protocol exchange to clarify the implementation of End of RIB (*EOR*) and Graceful Restart (*GR*). The goal is to restart a BGP session without the Restarting Router's peers redirecting traffic around the Restarting Router.

1. The BGP process of Router A (RTR_A) BGP begins, and it establishes a peering relationship with router B (RTR_B). It sends an *OPEN* message to router B, and the *OPEN* message includes the *Graceful Restart* Capability (Code 64) and Address Family of IPv4, Subsequent Address Family of unicast. Because router B also supports *GR*, it also sends an acknowledgement via its own *OPEN* Message, which contains *GR=64* and *AF=IPv4*.
2. An RP switchover occurs and Router A's BGP process restarts on the newly active RP. Router A does not have a routing information base on this RP, and must reacquire it from its Peer Routers. Router A will continue to forward IP packets destined for (or through) Router B using the last updated FIB and CEF table.
3. When the Receiving router (Router B) detects that the TCP session between it and Restarting Router is cleared, it immediately marks routes learned from the Restarting Router as STALE. Router B only marks routes learned from Router A as STALE; If B had other peers, then the routes learned from those peers would remain in the UP state. Router B also initializes a *Restart-timer* for the Restarting Router. The default setting for this timer is 120 seconds. The Restart-Timer is the amount of time that a Receiving router will wait for an *OPEN* message from the Restarting Router. A Receiving router will remove all STALE routes unless it receives an *OPEN* message from the Restarting Router within the specified *Restart-time*. Once router B receives router A's *OPEN* message, the Restart-timer is reset. During this time, Routers A and B continue to forward traffic using the last updated CEF table.
4. Router A's BGP process has initialized. It will now attempt to re-establish a BGP session with router B. It first establishes a new TCP session, and then it sends an *OPEN* message (Restart State bit set, Restart Time= n, and Forwarding State= IPv4). By default, Restart time is 120 seconds and it is also configurable. When Router B receives this *OPEN* message, it resets its own Restart-timer and starts a *Stale-path* timer. *Stale-path*, by default, is 360 seconds and is configurable.
5. Both routers successfully re-establish their session. At this point, if Router B recognizes that the Forwarding State in Router A's *OPEN* message is not set for IPv4, it immediately removes any STALE routes, which it had learned from the Restarting Speaker, and re-computes its routing database. (Normally, the Forwarding State will be set for IPv4)
6. Router B will begin to send *UPDATE* messages to Router A. These messages contain IP prefix information, and Router A will process them accordingly. Until an *EOR* indication is received from all peers (or the *bgp* update-delay timer expires), Router A will not start the BGP Route Selection Process. A new routing information database is available after the Route Selection Process is finished and the CEF information is updated accordingly. Router A starts an update-delay timer and waits up to 120 seconds to receive *EOR* from all of its NSF-peers.
7. Once Router A has received *EOR* from all its peers, it will begin the BGP Route Selection Process. Once this process is complete, it will begin to send *UPDATE* messages, which contain prefix information, to router B. Router A concludes this process by sending an *EOR* indication to Router B so that B, in turn, can start its Route Selection Process. Once Router B receives an *EOR* from A, and it has completed its Route Selection Process, then any STALE entries in BGP will be refreshed with newer information or removed from the BGP RIB and FIB. Router B is now converged. While Router B waits for an *EOR*, it also monitors *stalepath-time*. If the timer expires, all STALE routes will be removed and "normal" BGP processes will be started.



4.0 Router Preparation and Network Configuration:

In order to ensure a successful migration to a Graceful Restart-capable router, there are a few important principles to consider:

The router must have compatible RPs installed. In addition, care should be taken when mixing RP types:

- Cisco 12000 Series Internet Router: GRP and GRP-B RPs can be used together. If using a PRP on this router, it must be paired with another PRP.
- Cisco 10000 Series Internet Router: two PRE-1s must be used. The original PRE for this router is not supported for purposes of Cisco NSF with SSO.
- Cisco 7500: RSP-2 and RSP-4 can be used in combination. RSP-8 and RSP-16 can also be used in combination. However, an RSP-8 or RSP-16 cannot be mixed with an RSP-2 or an RSP-4.
- For all RP types on all supported platforms, the active and standby RPs must have the same amount of memory

A wide variety of line cards support Cisco NSF with SSO, but—for optimum performance of BGP Graceful Restart—every card in the router chassis should support Cisco SSO. For a list of supported line cards, please visit: <http://www.cisco.com/en/US/products/sw/iosswrel/ps1829/1221748>

Cisco SSO may not be supported on any line card not specifically listed in the aforementioned document. In this case, that specific line card will operate in RPR+ mode. At the time of the RP switchover, the dCEF table on the card will be cleared. This will cause Cisco NSF to destinations reachable through that card to fail.

Subsequent releases of Cisco IOS Software provide additional hardware support for Cisco SSO on specific line cards. Please check the release notes for later releases of Cisco IOS Software to determine if support for a particular line card may be available.

The referenced document also supplies detailed instructions on enabling SSO on the platforms that support Cisco NSF. Cisco SSO is an absolute requirement for enabling Cisco NSF; it will not work unless both are concurrently enabled.

On the Cisco 12000 Series Internet Router, there is a method to validate whether all line cards within a chassis are supported. Load a software image enabled with Cisco NSF with SSO and then issue the command “*show redundancy mode-supported*”. Each card in the chassis will be listed, and indicate the highest level of system redundancy it supports (RPR, RPR-Plus, Cisco SSO).

To achieve the full benefit of Cisco NSF with SSO, all line cards should support Cisco SSO. Furthermore, depending on platform, Distributed Cisco Express Forwarding (dCEF) must be enabled for the line cards in order for NSF to work.

The correct software image must be loaded on the flash disks of both route processors. Currently, mixing software versions between the active and standby router processors is not supported—even if both software images support Cisco NSF with SSO.

The software boot image in bootflash should also be upgraded and should correspond to the software image being loaded on the RP.



BGP Graceful Restart is configured under the global “*router bgp*” configuration command. The most basic configuration is “*bgp graceful-restart*”

```
Router(config-route)# [no] bgp graceful-restart
Router(config-route)# [no] bgp graceful-restart restart-time n
Router(config-route)# [no] bgp update-delay n
Router(config-route)# [no] bgp graceful-restart stalepath-time n
```

The “*bgp graceful-restart*” command must be entered on the Cisco NSF-capable router, and also must be entered on any NSF-aware peer that will be participating in Graceful Restart. Graceful Restart is not enabled by default, and must be explicitly configured on both the Restarting Router and all Peer Routers.

The “*bgp graceful-restart restart-time n*” is the 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.

The “*bgp update-delay n*” command may be entered on the 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, and *n* is always measured in 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.

The “*bgp graceful-restart stalepath-time n*” command may be entered on the NSF-aware peer(s) of the restarting router. This timer 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.

5.0 Deployment

There are many different variations on design and deployment of BGP networks. However, to simplify matters, it is easier to think about BGP design in terms of router functionality: What does a particular router need to accomplish given its particular placement within the network? There are three basic types of routers within a BGP network:

- Inter-AS routers run a combination of eBGP and iBGP to connect different autonomous systems. There are many variations to this: edge routers that connect Enterprise customers to the Service Provider network, Internet peering points that connect Service Provider autonomous systems together, edge routers that exist on the boundary of a BGP confederation sub-AS. (see RFC 3065). Yet, the functionality of each of these routers is identical from the Cisco NSF perspective.
- Intra-AS routers exist in the distribution layer or core of an individual AS. These routers run only iBGP and interact only with routers within their own autonomous system. Any knowledge they have of the world outside of their AS is communicated to them via Inter-AS routers
- Route Reflectors (RRs) act as aggregation and distribution points for BGP routing information. Intra-AS routers report BGP routing information to the RRs and receive information from them. RRs increase the scalability of a BGP network by removing the restriction that all iBGP peers must be fully-meshed. The two most common deployment scenarios for RRs are



- Centralized RRs: The Route Reflectors exist at the core of the BGP network, roughly equidistant from all the other routers in the AS. Each router in the AS forms a BGP session with this RR. Frequently, there will be redundant RRs in this configuration
- Distributed RRs: In this design, some subset of routers within an AS will be administratively grouped and have a local RR to which each router will form a BGP session. These RRs subsequently form BGP sessions to other RRs in other regions, or a meshed connection to other RR as well as Intra-AS routers in the core. A typical example of this type of configuration would be a Service Provider that has local RRs in each of its Points of Presence (PoPs).

5.1 Deployment Scenarios

Figure 3
NSF Deployment —Inter-AS Peers

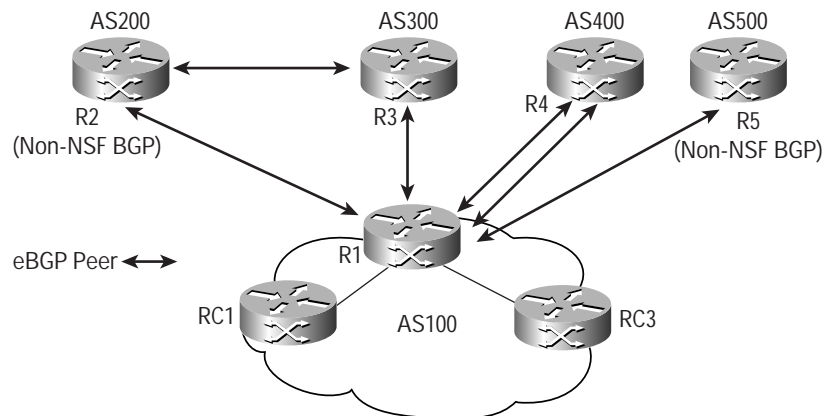


Figure 3 illustrates an eBGP deployment with peers in several different autonomous systems. R1, R2, R3, R4 and R5 are all eBGP peers. Furthermore, the connections R1-to-R4 are multi-homed and are peering to the loopback interface address. R1, R3, and R4 recognize each other as NSF-aware peers because each sends and acknowledges each other's GR capability. On the other hand, the BGP sessions between R1-R2 and R1-R5 are non-NSF, meaning they do not signal or acknowledge the GR capability. The following occurs when R1 goes through a BGP restart:

- R2 detects failure of the TCP/BGP session established over the R1-R2 link and will attempt to route around it. As a result of this re-computation, R2 will take the R2-R3/R3-R1 links to access AS100.
- R1 (the Restarting Router) continues to forward packets destined to AS200 along the R1-to-R2 link, because its CEF table remains intact after switchover. We now have an example of asymmetric routing that can occur when there is a mixture of NSF-aware and non-NSF-aware peers. Although asymmetric routing is not a desirable condition and may result in some packet loss, it is still preferable to the network disruption that would have ensued had R1 completely reinitialized.
- R3 and R4 will not flush routes that they had previously learned from R1. R1 should continue to forward IP packets between R3 and R4, using its last updated CEF Table.
- R3 and R4 should continue to forward IP packets to R1 using their last updated CEF table.



- R5 is non-NSF-aware; as such, it will lose the BGP session to R1 and initialize the BGP session from scratch. R1 will continue to forward packets bound for AS500 through R5, but there is no return path for the traffic. There will be packet loss until R5 successfully reconverges with R1.
- There is an exception to this rule, if R5 has a static default route pointing to R1 as the next-hop. In this case, R5 was only using BGP so it could advertise its routes into the R1 BGP table. The R5 routes are preserved at R1 and R5 only needs a default route, so there should be no packet loss.

Figure 4
NSF—Inter-AS peers

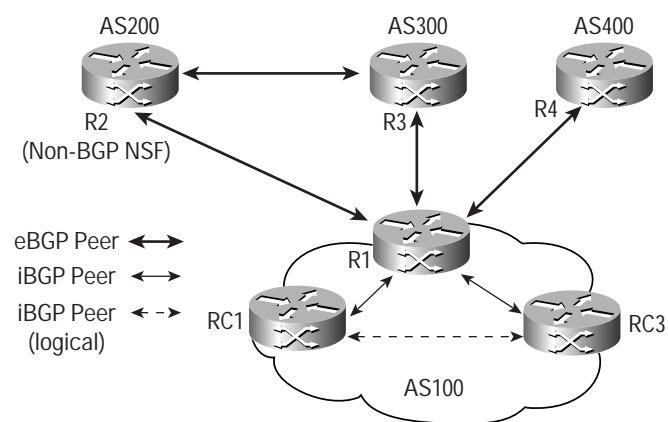


Figure 4 illustrates that R1, RC1 and RC3 are all BGP NSF-aware. Additionally, R1 is NSF-capable—meaning that it supports BGP restartability. Should R1 restart BGP, there should be little-to-no packet loss within AS100. Because R1 is also an Inter-AS router, there may be some packet loss before R2 reconverges to route traffic destined to AS100 via R3. This process was described in the previous section.

Note that there is an important deployment consideration in this scenario. In this topology, it is very common to be running an IGP protocol (i.e.: OSPF or IS-IS) to provide next-hop reachability within AS100. There is interdependence between BGP and the selected IGP protocol. During best-path calculation, BGP knows the IP address of the router advertising certain destination prefixes. However, it relies on the information from the IGP to determine the next-hop to reach that advertising router.

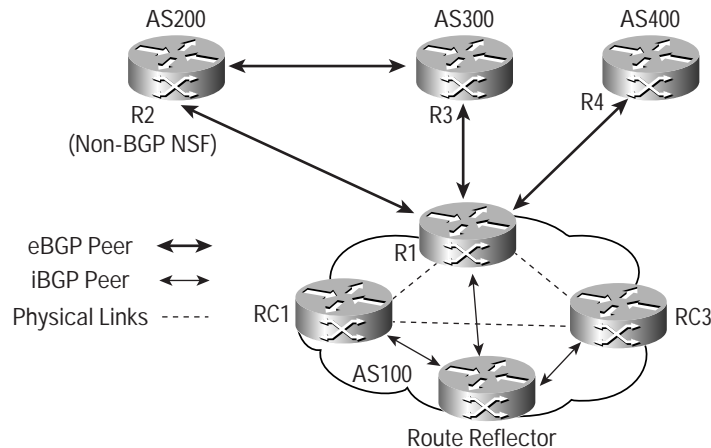
Because BGP Graceful Restart can alter the timing of BGP convergence, situations can potentially occur where BGP is ready to conduct best-path selection, but the IGP has not yet converged. Therefore, some destination prefixes could exist in BGP, but cannot be added to the CEF table because a path to the advertising router has not been calculated by the IGP yet. This could result in packet loss. See section 6.5 of this paper for an example of this situation.

Therefore, it is strongly recommended that NSF for IS-IS or OSPF should be configured in addition to BGP Graceful Restart. To learn more about NSF for both OSPF and IS-IS, review Globally Resilient IP deployment papers: <http://www.cisco.com/warp/public/732/Tech/grip/tech.shtml>

Figure 5



Cisco NSF—Inter-AS with Route Reflectors



This topology demonstrates that R1, RC1, RC2 and the Route Reflector (RR) are NSF-aware. RR is deployed as a control plane to reduce the requirements for a full iBGP mesh. Thus, it is not in the forwarding path, but forms iBGP peering arrangements with R1, RC1 and RC2. It is also assumed that a flavor of IGP NSF is implemented in this topology.

- When R1 restarts its BGP, it relies on the existing CEF table and continues to forward packets destined to (or through) RC1 and RC2.
- Meanwhile, the only peering arrangement that R1 has is with the Route Reflector. It has no direct peering with RC1 or RC2.
- Because RR is NSF-aware, it masks the fact that R1 has restarted BGP from propagating to RC1 or RC2. RC1 and RC2 continue to forward through R1.

A more interesting variation on this occurs if the Route Reflector is actually NSF-capable and restarted its BGP process.

- When the Route Reflector restarts BGP, all the clients will keep routing information, which was reflected by RR. None of the clients will switch to a backup RR (assuming a backup RR is available).
- Some special considerations must be made when using an NSF-capable Route Reflector
 - It is anticipated that a Route Reflector will have more BGP peers and a larger aggregate collection of BGP data than other routers in the AS; due to these conditions, best-path selection may take longer to complete during a switchover.
 - Users must balance the requirement to provide uninterrupted packet forwarding and routing stability to the network versus the likelihood of a significant routing change before convergence is complete. During this interim period, Cisco NSF uses the CEF table and not BGP routing information to forward packets.
 - Assuming that the decision has been made to use Cisco NSF on the Route Reflector, another configuration adjustment may be required. If it is anticipated that the entire process of reconvergence will exceed 360 seconds, then the default value of the “*bgp graceful-restart stalepath-time xxx*” may need to be adjusted on all of the peers of the route reflector. The value for stale-path should be adjusted to equal the expected convergence time (in seconds) plus an additional buffer zone of 30-60 seconds to account for variances in convergence time based on changing network conditions.



The decision of whether to use BGP Graceful Restart on a Route Reflector is a complex one and depends largely on network operations. Users must consider the key trade-offs in this decision:

- Is there an alternate availability strategy (i.e.: backup RR); if so, does it provide acceptable failover time?
- How long does it take for the restarting Route Reflector to reconverge, so its Peer Routers can begin to base forwarding decisions on fresh information?
- What is the likelihood that there will be other significant BGP routing changes that occur while the Route Reflector is reconverging?

While these questions are posed in the context of a decision to use Cisco NSF with SSO on a Route Reflector, they are also good general questions that should help in determining where and how to deploy Cisco NSF with SSO.

6.0 Troubleshooting

Because the protocol changes associated with Nonstop Forwarding for BGP are well defined and documented, troubleshooting Cisco NSF with SSO becomes a matter of assuring that the correct protocol exchanges occur at the correct times. The various Cisco IOS Software “show” and “debug” commands have been modified to provide this type of information.

6.1 Validating BGP Graceful Restart

As described in the “Protocol Enhancements” section of this paper, Cisco NSF / BGP begins with, and depends upon, an agreement between the BGP peers that they have implemented the protocol extensions in the IETF Graceful-Restart draft. Examining the output of the *show ip bgp neighbor X.X.X.X* command can validate this negotiation.

Figure 6
sh ip bgp neighbor

```
Router#sh ip bgp neighbors 192.168.3.3
BGP neighbor is 192.168.3.3, remote AS 4230, internal link
  BGP version 4, remote router ID 200.218.71.10
  BGP state = Established, up for 00:03:08
  Last read 00:00:08, hold time is 180, keepalive interval is 60 seconds
  Neighbor capabilities:
    Route refresh: advertised and received(new)
    Address family IPv4 Unicast: advertised and received
    Address family VPNv4 Unicast: advertised and received
    Address family IPv4 Multicast: advertised and received
    Graceful Restart Capabilty: advertised and received
    Remote Restart timer is 120 seconds
    Address families preserved by peer:
      IPv4 Unicast
```

Figure 6 illustrates that the *show ip bgp neighbor* command now documents the successful negotiation of the Graceful Restart Capability. In order for Cisco NSF to work, the Restarting Router and its peer must each have implemented the BGP protocol enhancements. It is thus vital that the capability be both “advertised and received”.



This output should be the same, regardless of whether it is queried on the NSF-capable router or its NSF-aware peer. Note that the advertisement of the BGP graceful-restart capability does not specify that a particular router is NSF-capable; it only signifies that the router understands the protocol changes specified in the IETF draft.

The value of the Remote Restart timer is listed. Although the local and remote BGP peers are permitted to have different restart timers, it is wise to be suspicious of a significant variance between the values. The restart timer sets an upper bound on the amount of time it will take to re-establish the TCP connection and renegotiate the graceful-restart capability in a new BGP OPEN message. Setting this value too low on either of the peers makes it unlikely that the OPEN can be re-established in time, and will lead to the failure of the Cisco NSF process.⁴ BGP will revert to “normal” recovery procedures in this type of case.

Refer to the following line, “*Address Families preserved by peer*” for information about what Address Families (i.e.: types of traffic) for which the Peer Router can maintain forwarding state during the progression of an RP switchover. IPv4 is currently the only supported Address Family. Future extensions to Cisco NSF will add support for other Address Families (or Subsequent Address Families), including Multicast, IPv6 and MPLS when these protocols are made NSF-capable.

6.2 Validating Stale Routes on an NSF-aware Peer

One of the agreed-upon capabilities of the NSF-aware Peer Router is the use of “stale” routes (the last known good routes before the restart) to continue packet forwarding while the Restarting Router is dynamically reacquiring routing information in the background (see previous section for details). This capability can be queried with the *show ip bgp* command.

Figure 7
sh ip bgp on NSF-aware peer

```
ip9-75b# show ip bgp
BGP table version is 209, local router ID is 11.11.11.11
Status codes: s suppressed, d damped, h history, * valid, > best, i - internal   S Stale
Origin codes: i - IGP, e - EGP, ? - incomplete
   Network          Next Hop        Metric LocPrf  Weight  Path
*> 11.0.0.0         0.0.0.0         0           0    32768  i
*> S170.10.10.0/24  180.10.10.3     0           0    200 101e
*> S180.10.10.0/24  180.10.10.3     0           0    200 101e
*> S190.10.10.0/24  180.10.10.3     5           0    200 101e
```

In Figure 7, the NSF-aware router has marked the routes for the 170, 180 and 190 networks as “stale” (as signified by the capital “S” beside the route itself). This behavior is expected. Only the routes reachable through the Restarting Router will be “stale”. All other routes should appear as normal.

4. Note that setting the restart timer value too high can also have some undesirable consequences. Except for directly-connected peers, an NSF-aware router cannot distinguish between a Restarting Router and a router that has been removed from the network for other administrative purposes. In the latter case, BGP reconvergence will not start until the restart timer on the NSF-aware router has expired. Thus, setting the restart timer to an unnecessarily high value may delay convergence during “normal” operation.



If the specific routes through the Restarting Router are not marked “S”, one of several explanations might apply:

- If the route is missing from *sh ip bgp*, then Cisco NSF has failed for some reason and “normal” BGP reconvergence is taking place. Possible causes:
 - The Peer Router is not NSF-aware. See 6.1, above
 - The restart timer has expired
 - The stale-path timer has expired
- If the route is present, but is not marked “S”, then one of the following might apply:
 - The Cisco NSF-aware router has not detected the restart on the NSF-capable router yet
 - The Cisco NSF process is complete, and the NSF-aware router has replaced the stale route with a fresh one
 - If the next-hop is not pointing to the Restarting Router, then Cisco NSF has failed and “normal” BGP recovery procedures have occurred, and the network has reconverged around the Restarting Router

6.3 Debugging Output on the NSF-aware Peer

As with almost all Cisco-supported protocols, BGP NSF supports debugging commands that can monitor the progression of events as a route processor switchover occurs. Table 2 illustrates the output of the commands *debug ip bgp events* and *debug ip bgp updates* on the NSF-aware peer of a Restarting Router at the time of a switchover.

Note: The use of the *debug ip bgp updates* command can be extremely CPU-intensive, and must be used with great care. This command should not be used on a production router.

Table 2 Debugging output on NSF-aware peer

debug ip bgp events debug ip bgp updates	
Line 1	%BGP-5-ADJCHANGE: neighbor 170.10.10.3 Down NSF peer closed the session
Line 2	BGPNSF: Marked nbr 170.10.10.3 for nsf processing. Restart timer started for 120 secs
Line 3	BGPNSF state: 170.10.10.3 went from nsf_in_progress to nsf_marked_stale
Line 4	BGPNSF: Building graceful restart capability for 170.10.10.3
Line 5	BGPNSF: 170.10.10.3 SessionRestart Timer stopped. Stalepath Timer started
Line 6	BGPNSF: 170.10.10.3 Peer has restarted. Restart Time: 120
Line 7	%BGP-5-ADJCHANGE: neighbor 170.10.10.3 Up
Line 8	BGP: 170.10.10.3 initial update completed
Line 9	BGPNSF(0): Send End-of-RIB for AF IPv4 Unicast to nbr 170.10.10.3
Line 10	BGP(0): 170.10.10.3 rcvd UPDATE w/ attr: nexthop 170.10.10.3, origin i, path 200 201
Line 11	BGP(0): 170.10.10.3 rcvd 11.11.0.0/16...duplicate ignored
Line 12	BGPNSF: nbr 170.10.10.3 sent End-of-RIB marker for IPv4 Unicast
Line 13	End-of-RIB was expected from 170.10.10.3
Line 14	BGPNSF: got EOR for all AFs. Stop stalepath timer for 170.10.10.3.



Table 2 shows the normal progression of events during a route processor switchover from the point-of-view of the BGP peer. The most relevant events are:

- Line 1: The actual TCP connection from the Restarting Router is cleared. Although this looks like an error message, it is expected behavior and it is necessary in order to initiate the Cisco NSF process. The clearing of the TCP session does not cause a clearing of the forwarding entries on the NSF-aware peer.
- Lines 5-7: A new TCP connection arrives from the Restarting Router. Because this TCP connection has arrived before the local restart timer has expired, this timer is stopped and the stale-path timer is started.⁵
- Line 9: After completion of the initial BGP updates, the Cisco NSF-aware peer sends End-of-RIB to the Restarting Router. After receiving EOR from all of its peers, the Restarting Router can begin best-path selection.
- Line 12: The Restarting Router has completed best-path selection, so it sends all of its BGP updates to its peers, and sends End-of-RIB when complete.
- Line 14: The NSF-aware peer has received all updates, and can therefore stop the stalepath timer, do its own best-path selection and update the FIB and CEF table. Any stale routes that remain in the FIB and CEF tables are cleared.⁶

The debugging output can be used to isolate problems in the sequence of events that occur during a route processor switchover.

6.4 Debugging Output on the Restarting Router

Debugging on the router that is actually undergoing a restart is more challenging, because the commands must be entered on the console of the Standby route processor. The method used to access the Standby console varies between platforms. On the Cisco 12000, you can attach to the Standby console using the *attach <slot-number-of-standby-console>* command. On the Cisco 10000, you can access the Standby console via the console port. However, by default, this console is disabled. It must be explicitly enabled from configuration mode using the *hidden standby console enable* command:

On the Cisco 7500 Series Router, the standby console can be accessed directly by simply attaching a terminal to it. No special configuration is required.

```
c10K (config)# redundancy main-cpu submode
c10K(config-r-mc)# standby console enable
```

Important Note: Configuration changes on the standby console can cause significant negative results, and may cause future switchovers to fail. The standby console should never be used for configuration; all configuration changes should be performed on the primary console.

Additionally, Cisco recommends that any debugging performed on the standby console be executed under the direction of Cisco Technical Assistance Center personnel only.

5. The order of these lines is a bit confusing, because the events happen concurrently. The lines are listed in the order in which the debugging output displays them. However, it might be easier to visualize the progression of events if line 8 preceded line 6.

6. Why would there still be stale routes in the FIB/CEF table after switchover? If network paths that were available at the time of switchover became unavailable during the reconvergence process, their stale entries are not refreshed. They must be purged from the FIB and CEF tables.



Once the user has obtained access to the standby console, debugging output can then be enabled. The commands *debug ip bgp events* and *debug ip bgp updates* are recommended for analyzing BGP/NSF problems. Table 3 provides the output of these commands. Lines that hold especial interest have been rendered in bold.

Note: The use of the *debug ip bgp updates* command can be extremely CPU-intensive, and must be used with great care. This command should not be used on a production router.

Table 3 Debugging Output on the Restarting Router

debug ip bgp events debug ip bgp updates	
Line 1	REDUNDANCY-5-PEER_MONITOR_EVENT: secondary received a switchover(raisevent=peer_redundancy_state_change(5))
Line 2	BGPNSF: building graceful restart capability for 170.10.10.2
Line 3	BGPNSF: 170.10.10.2 OPEN has graceful restart capability. Length = 6
Line 4	BGPNSF: 170.10.10.2 peer has not restarted. Restart time: 120
Line 5	%BGP-5-ADJCHANGE: neighbor 170.10.10.2 up
Line 6	BGP(0): Delaying initial update for up to 120 seconds
Line 7	BGP(0): 170.10.10.2 rcvd UPDATE w/ attr: nexthop 170.10.10.2, origin i path 101 100
Line 8	BGP(0): 170.10.10.2 rcvd 190.10.10.0/24
Line 9	BGPNSF: nbr 170.10.10.2 sent End-of-RIB marker for IPv4 Unicast
Line 10	End-of-RIB was awaited from 170.10.10.2
Line 11	BGPNSF: 180.10.10.4 Peer has not restarted. Restart Time: 120
Line 12	BGPNSF: 180.10.10.4 Address family IPv4 Unicast is preserved
Line 13	%BGP-5-ADJCHANGE: neighbor 180.10.10.4 Up
Line 14	BGP(0): 180.10.10.4 rcvd UPDATE w/ attr: nexthop 180.10.10.4, origin i, metric 0, path 201
Line 15	BGP(0): 180.10.10.4 rcvd 11.11.0.0/16
Line 16	BGPNSF: nbr 180.10.10.4 sent End-of-RIB marker for IPv4 Unicast
Line 17	End-of-RIB was awaited from 180.10.10.4
Line 18	BGP: compute bestpath for address family 0
Line 19	BGP(0): Revise route installing 11.11.0.0/16 -> 180.10.10.4 to main IP table
Line 20	BGP(0): Revise route installing 190.10.10.0/24 -> 170.10.10.2 to main IP tableBGP:
Line 21	Compute bestpath for address family 1
Line 22	BGP: compute bestpath for address family 2
Line 23	BGPNSF: Listeners notified of convergence
Line 24	BGP(0): 170.10.10.2 computing updates, afi 0, neighbor version 0, table version 4, starting at 0.0.0.0
Line 25	BGP(0): 170.10.10.2 send UPDATE (format) 11.11.0.0/16, next 170.10.10.3, metric 0, path 201



Table 3 Debugging Output on the Restarting Router

debug ip bgp events debug ip bgp updates	
Line 26	BGP: 170.10.10.2 initial update completed
Line 27	BGPNSF: Send End-of-RIB for AF IPv4 Unicast to nbr 170.10.10.2
Line 28	BGPNSF(0): Send End-of-RIB for AF IPv4 Unicast to nbr 180.10.10.4

- Line 1: The standby route processor receives notification that a state change has occurred, and it will become the new active route processor.
- Line 3-5: TCP connectivity and a new BGP OPEN message are exchanged with the NSF-aware peer. In this case, the Restarting Router has two different peers, and the second peer connects in lines 11-13.
- Line 9: The Restarting Router begins receiving BGP updates from its peer. The completion of the initial updates is signaled by receipt of the End-of-RIB marker in line 9.
- Line 16: The second peer has completed its initial update and sends End-of-RIB
- Line 18: The Restarting Router has all updates from all peers, so it can begin best-path calculation now (and only now).
- Line 23: Processes internal to the router that have registered for notification of convergence from BGP are alerted. Assuming any other routing protocols are converged, the FIB and CEF are both refreshed at this point with any new information. Additionally any stale paths remaining in the FIB or CEF are purged (see 6.3).
- Line 24: Routes to be distributed to peers are calculated.
- Lines 27-28: End-of-RIB sent to both peers, meaning that all available updates have been sent. Peers can start their own best-path calculation. The network is effectively reconverged. Note that packet forwarding has continued throughout this entire process.

6.5 Debugging Timing Problems in BGP Graceful Restart

From a technical perspective, failure scenarios are always interesting to investigate. Table 4 shows the (edited) debugging output of a failed BGP/NSF switchover. The symptom reported is that approximately fifteen seconds after switchover, traffic begins to drop. What could be wrong?

Table 4

Line 1	Jun 19 14:16:40.546: %BGP-5-ADJCHANGE: neighbor 30.0.1.1 Down NSF peer closed the session
Line 2	Jun 19 14:16:59.942: BGP: 30.0.1.1 rcv OPEN, version 4
Line 3	Jun 19 14:16:59.942: BGPNSF: 30.0.1.1 SessionRestart Timer stopped. Stalepath Timer started
Line 4	Jun 19 14:16:59.942: BGP: 30.0.1.1 OPEN has CAPABILITY code: 64, > >> length 6
Line 5	Jun 19 14:16:59.942: BGPNSF: 30.0.1.1 OPEN has Graceful Restart capability. Length = 6
Line 6	Jun 19 14:16:59.942: BGPNSF: 30.0.1.1 Peer has restarted. Restart Time : 120
Line 7	Jun 19 14:16:59.942: BGPNSF: 30.0.1.1 Address family IPv4 Unicast is preserved



Table 4

Line 8	BGPNSF state: 30.0.1.1 went from nsf_marked_stale to sf_delete_stale_af
Line 9	Jun 19 14:17:18.857: BGPNSF state: 30.0.1.1 went from nsf_delete_stale_af to nsf_delete_stale
Line 10	Jun 19 14:17:18.961: BGPNSF state: 30.0.1.1 went from nsf_delete_stale to nsf_not_active
Line 11	Jun 19 14:17:18.961: RT: del 220.0.0.0 via 30.1.1.1, bgp metric [20/0] Jun 19 14:17:18.961: RT: delete network route to 220.0.0.0 Jun 19 14:17:18.961: RT: del 220.0.1.0 via 30.1.2.1, bgp metric [20/0] Jun 19 14:17:18.961: RT: delete network route to 220.0.1.0 Jun 19 14:17:18.961: RT: del 220.0.2.0 via 30.1.3.1, bgp metric [20/0] Jun 19 14:17:18.961: RT: delete network route to 220.0.2.0 Jun 19 14:17:18.961: RT: del 220.0.3.0 via 30.1.4.1, bgp metric [20/0] Jun 19 14:17:18.961: RT: delete network route to 220.0.3.0 Jun 19 14:17:18.961: RT: del 220.0.4.0 via 30.1.5.1, bgp metric [20/0] Jun 19 14:17:18.961: RT: delete network route to 220.0.4.0 Jun 19 14:17:18.961: RT: del 220.0.5.0 via 30.1.6.1, bgp metric [20/0] Jun 19 14:17:18.961: RT: delete network route to 220.0.5.0 Jun 19 14:17:18.961: RT: del 220.0.6.0 via 30.1.7.1, bgp metric [20/0] Jun 19 14:17:18.961: RT: delete network route to 220.0.6.0 Jun 19 14:17:18.961: RT: del 220.0.7.0 via 30.1.8.1, bgp metric [20/0] Jun 19 14:17:18.961: RT: delete network route to 220.0.7.0 Jun 19 14:17:18.961: RT: del 220.0.8.0 via 30.1.9.1, bgp metric [20/0] Jun 19 14:17:18.961: RT: delete network route to 220.0.8.0 Jun 19 14:17:18.961: RT: del 220.0.9.0 via 30.1.10.1, bgp metric [20/0] Jun 19 14:17:18.961: RT: delete network route to 220.0.9.0 ... Jun 19 14:17:18.997: RT: del 220.0.174.0 via 30.5.35.1, bgp metric [20/0]
Line 12	Jun 19 14:18:11.293: RT: add 220.0.105.0/24 via 30.4.1.1, bgp metric [20/0] Jun 19 14:18:11.293: RT: add 220.0.106.0/24 via 30.4.2.1, bgp metric [20/0] Jun 19 14:18:11.293: RT: add 220.0.107.0/24 via 30.4.3.1, bgp metric [20/0] Jun 19 14:18:11.293: RT: add 220.0.108.0/24 via 30.4.4.1, bgp metric [20/0]

Problem background: This trace is taken on the NSF-aware peer of a Restarting Router. The debugging of the problem is especially challenging because we only have the output of the *debug ip bgp* and *debug ip bgp updates* commands. There is no record of the output of *debug ip bgp event*, so the user must intuit some of the missing information.

The problem clearly occurred in line 11, where various network routes are deleted, subsequently causing the packet loss. Some of those same routes are added back into the BGP routing table less than one minute later.

There are two possible explanations, based on existing knowledge of the BGP/Cisco NSF process. BGP/Cisco NSF will remove routes under two conditions:

- The stale-path timer expires, or
- Stale routes are removed once End-of-RIB has been received from the peer and best-path calculation is complete

Which case is this?



In line 3 of the debugging output, the stale-path timer is started at 14:16:59.942. The first route is cleared at 14:17:18.961. The stale-path timer might be set for only 19 seconds; while this is unlikely, we should check the stale-path timer on the local router. The configuration shows that it is set to the default of 360 seconds—so this is clearly not the problem.

The other possibility is that the stale routes are removed once the local router has completed best-path calculation. There are several theories that explain why this might occur:

1. The routes were never sent by the remote router
2. The routes were sent, but were filtered at the local router
3. The routes are available, but something keeps them from being entered in the FIB and CEF table.

The first two options are unlikely, because the correct routes do get into the FIB and CEF tables eventually (line 12). Option three is the only remaining possibility—the routes are present, but cannot be entered into the FIB/CEF tables. In order for BGP to calculate a best-path and put an entry in the CEF table, it needs to know a) The address of the advertising router and b) the path (next-hop) to the advertising router. These will usually be available through the Interior Gateway Protocol (i.e.: OSPF, ISIS) running within the domain.

The eventual resolution of this problem was that the IGP was not NSF-capable, so that when the initial best-path calculation was performed (line 11) all of the necessary information to put the entry in the FIB/CEF table was not available. Once the IGP converged, a second best-path calculation was run (line 12) and the routes could be added at that time. Unfortunately, traffic was lost in the interim.

This example is instructive, because it illustrates the dependencies between the routing protocols, and the effect they may have on BGP/Cisco NSF. (This is an atypical problem, because BGP will usually take longer to converge than any IGP protocol). Some possible solutions or workarounds for the problem might be:

- Use static routes to assign the next-hop gateways for advertising routers. This is a good solution, but it might not scale if the number of advertising routers is very large.
- Use the BGP *next-hop-self* command to have the edge routers advertise the routes they are bringing into the autonomous system using themselves as the next-hop. This will work for a meshed environment, but will not work with BGP route-reflectors (the route-reflectors are not the actual next-hop for the target networks).
- Make the IGP NSF-capable as well. This is most likely the best solution. OSPF and IS-IS are the IGP's that currently support NSF-capabilities.

7.0 Software and Platform

NSF BGP for IPv4 unicast requires the following:

- Cisco IOS Software Release 12.0(22)S or later.
- Dual Route Processor: Active and Standby, on Restarting Router.
- Same Cisco NSF-capable image on both Route Processors.
- Cisco 7500 Series Router, Cisco 10000 or 12000 Series Internet Router
- NSF aware peers:
 - Cisco IOS Software Release 12.0(22)S or later
 - Cisco IOS Software Release 12.2(15)T or later
 - Any vendor supporting the BGP Graceful Restart protocol extensions

8.0 Conclusion

This document has examined the protocol changes associated with Cisco NSF for BGP, as well as some common deployment scenarios and methods of troubleshooting these BGP enhancements.

During the last few years, network availability has become an increasingly important topic for both Service Providers and Enterprises. The networking industry has responded with many enhancements, including Graceful Restart. As users and vendors gain more deployment experience with such enhancements, further refinements will be made to the protocols themselves and the way in which networks are deployed.



Corporate Headquarters
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134-1706
USA
www.cisco.com
Tel: 408 526-4000
800 553-NETS (6387)
Fax: 408 526-4100

European Headquarters
Cisco Systems International BV
Haarlerbergpark
Haarlerbergweg 13-19
1101 CH Amsterdam
The Netherlands
www-europe.cisco.com
Tel: 31 0 20 357 1000
Fax: 31 0 20 357 1100

Americas Headquarters
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134-1706
USA
www.cisco.com
Tel: 408 526-7660
Fax: 408 527-0883

Asia Pacific Headquarters
Cisco Systems, Inc.
Capital Tower
168 Robinson Road
#22-01 to #29-01
Singapore 068912
www.cisco.com
Tel: +65 6317 7777
Fax: +65 6317 7799

Cisco Systems has more than 200 offices in the following countries and regions. Addresses, phone numbers, and fax numbers are listed on the Cisco Web site at www.cisco.com/go/offices

Argentina • Australia • Austria • Belgium • Brazil • Bulgaria • Canada • Chile • China PRC • Colombia • Costa Rica • Croatia
Czech Republic • Denmark • Dubai, UAE • Finland • France • Germany • Greece • Hong Kong SAR • Hungary • India • Indonesia • Ireland
Israel • Italy • Japan • Korea • Luxembourg • Malaysia • Mexico • The Netherlands • New Zealand • Norway • Peru • Philippines • Poland
Portugal • Puerto Rico • Romania • Russia • Saudi Arabia • Scotland • Singapore • Slovakia • Slovenia • South Africa • Spain • Sweden
Switzerland • Taiwan • Thailand • Turkey • Ukraine • United Kingdom • United States • Venezuela • Vietnam • Zimbabwe

All contents are Copyright © 1992-2003 Cisco Systems, Inc. All rights reserved. Cisco, Cisco IOS, Cisco Systems, and the Cisco Systems logo are registered trademarks or trademarks of Cisco Systems, Inc. and/or its affiliates in the U.S. and certain other countries.

All other trademarks mentioned in this document or Web site are the property of their respective owners. The use of the word partner does not imply a partnership relationship between Cisco and any other company.
(0303R) 203031.A/ETMG_04/03