BGP Graceful Restart-FS for FRR community

Author:

Biswajit Sadhu

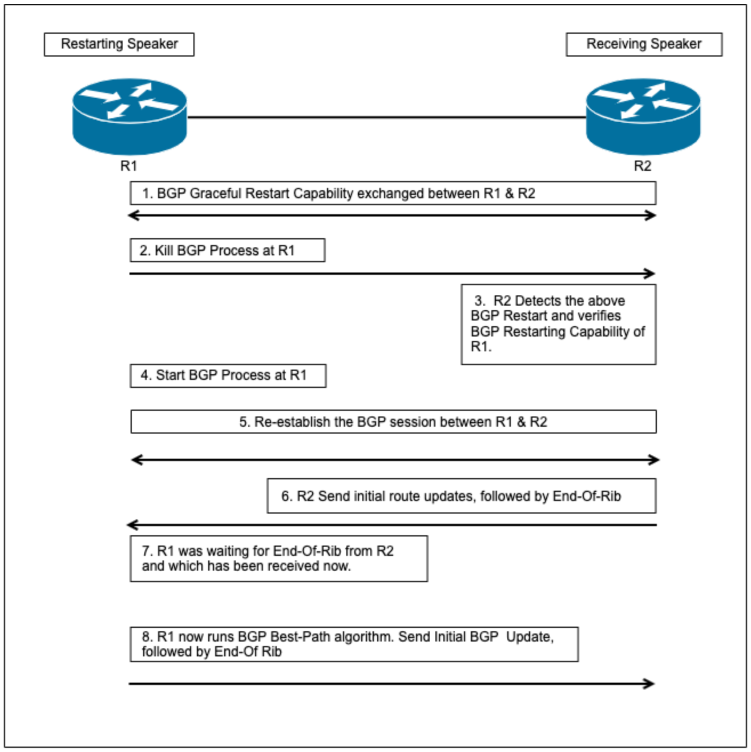
sadhub@vmware.com

Functional Specification

BGP graceful restart functionality as defined in RFC 4724 (https://tools.ietf.org/html/rfc4724) defines the mechanisms that allows BGP speaker to continue to forward data packets along known routes while the routing protocol information is being restored. Goal is to provide functionality as below:

1. Ability to enable and disable graceful restart helper mode functionality at peer level.
2. Defer best path selection post restart with graceful restart enabled.
3. Retention of routes in zebra while BGP is restarting when graceful restart enabled.

Overview of Graceful restart (RFC 4724)


Restarting Speaker
Receiving Speaker
R1
R2
1. BGP Graceful Restart Capability exchanged between R1 & R2
2. Kill BGP Process at R1
3. R2 Detects the above BGP Restart and verifies BGP Restarting Capability of R1.
4. Start BGP Process at R1
5. Re-establish the BGP session between R1 & R2
6. R2 Send initial route updates, followed by End-Of-Rib
7. R1 was waiting for End-Of-Rib from R2 and which has been received now.
8. R1 now runs BGP Best-Path algorithm. Send Initial BGP Update, followed by End-Of Rib

BGP Graceful Restart

Usually, when BGP on a router restarts, all the BGP peers detect that the session went down and then came up. This "down/up" transition results in a "routing flap" and causes BGP route re-computation, generation of BGP routing updates, and unnecessary churn to the forwarding tables. The following functionality is provided by graceful restart:

1) The feature allows the restarting router to indicate to the helping peer the routes it can preserve in its forwarding plane during control plane restart by sending graceful restart capability in the OPEN message sent during session establishment.
2) The feature allows helping router to advertise to all other peers the routes received from the restarting router which are preserved in the forwarding plane of the restarting router during control plane restart.

End-of-RIB (EOR) message

An UPDATE message with no reachable Network Layer Reachability Information (NLRI) and empty withdrawn NLRI is specified as the End-of-RIB marker that can be used by a BGP speaker to indicate to its peer the completion of the initial routing update after the session is established.

For the IPv4 unicast address family, the End-of-RIB marker is an UPDATE message with the minimum length. For any other address family, it is an UPDATE message that contains only the MP_UNREACH_NLRI attribute with no withdrawn routes for that <AFI, SAFI>.

Although the End-of-RIB marker is specified for the purpose of BGP graceful restart, it is noted that the generation of such a marker upon completion of the initial update would be useful for routing convergence in general, and thus the practice is recommended.

Route Selection Deferral Timer

Specifies the time the restarting router defers the route selection process after restart.

Restarting Router

The usage of route election deferral timer is specified in https://tools.ietf.org/html/rfc4724#section-4.1

Once the session between the Restarting Speaker and the Receiving Speaker is re-established, the Restarting Speaker will receive and process BGP messages from its peers. However, it MUST defer route selection for an address family until it either

1. Receives the End-of-RIB marker from all its peers (excluding the ones with the "Restart State" bit set in the received capability and excluding the ones that do not advertise the graceful restart capability).

2. The Selection_Deferral_Timer timeout.

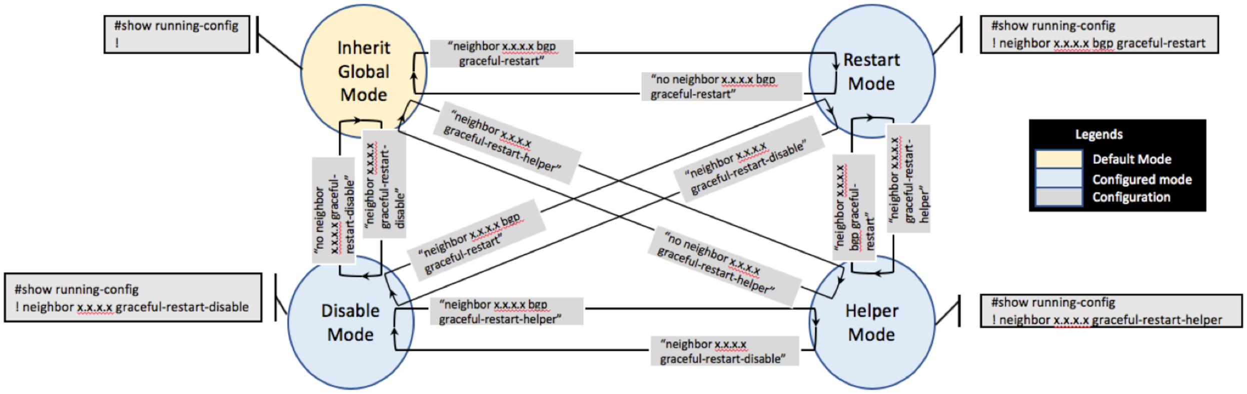
Functional Description

Per Peer Graceful Restart

Support for global graceful restart and helper disable mode will be added. This will ensure that router is not in graceful restart mode or in helper mode. Per peer graceful restart will be implemented with configuration CLI where for a given peer GR/Helper/Disable mode can be enabled and disabled as desired. Changes will be made to existing GR and helper functionality to disable it when required and also to enable only for a configured peer.

Global Mode

A new configuration CLI will be introduced for the same. With the new set of CLI's graceful restart mode change will be as shown below. By default it router will stay in helper mode.



#show running-config
!
Inherit Global Mode
"neighbor x.x.x.x bgp graceful-restart"
"no neighbor x.x.x.x bgp graceful-restart"
Restart Mode
#show running-config
! neighbor x.x.x.x bgp graceful-restart
"neighbor x.x.x.x graceful-restart-helper"
"neighbor x.x.x.x graceful-restart-disable"
Legends
Default Mode
Configured mode
Configuration
"no neighbor x.x.x.x graceful-restart-disable"
"neighbor x.x.x.x graceful-restart-disable"
"neighbor x.x.x.x bgp graceful-restart"
"no neighbor x.x.x.x graceful-restart-helper"
"neighbor x.x.x.x bgp graceful-restart"
"neighbor x.x.x.x graceful-restart-helper"
#show running-config
! neighbor x.x.x.x graceful-restart-disable
Disable Mode
"neighbor x.x.x.x bgp graceful-restart-helper"
"neighbor x.x.x.x graceful-restart-disable"
Helper Mode
#show running-config
! neighbor x.x.x.x graceful-restart-helper

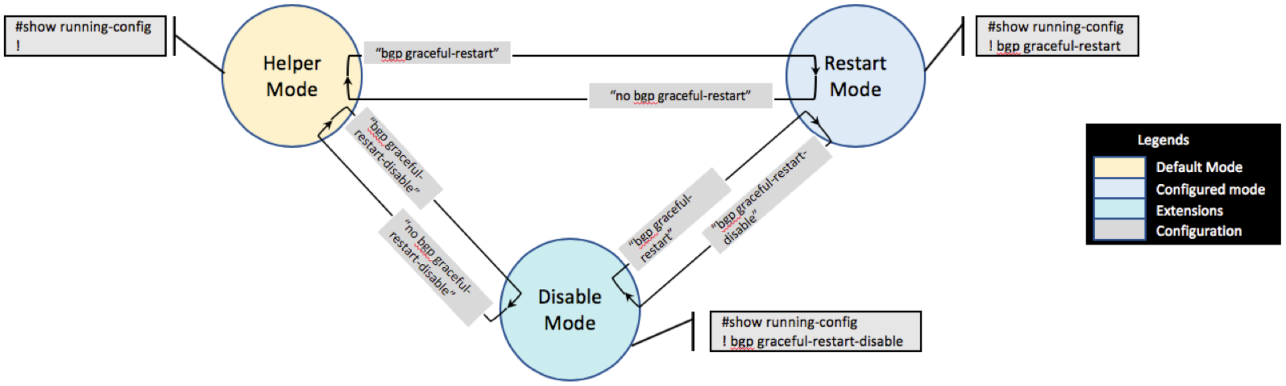
The following CLI commands will be implemented to disable helper mode and graceful restart at the global level

Command	Description	Comments
router> **bgp graceful-restart disable**	Disable graceful restart and helper mode	This is new command, the functionality is modified to disable both graceful restart and helper mode

router> **no bgp graceful-restart disable**	Enable default behavior (global helper mode)	This is new command, the functionality is to revert to default behavior (helper mode)

Peer Mode

Few configuration commands under neighbor will be added to enable/disable GR feature as desired for a given peer. Mode changes with these new set of commands under peer is as show below. By default given peer will always inherit global restart mode.



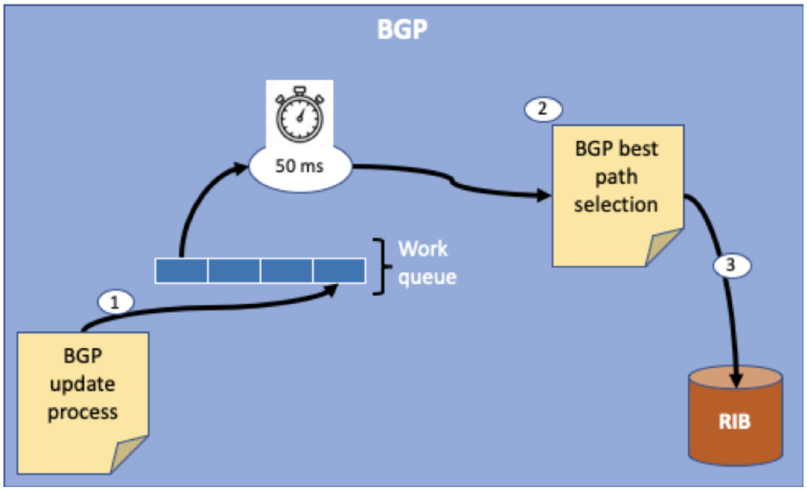
Below are the set of commands that will be added to support per peer graceful restart feature.

Command	Description	Comments
router> **neighbor A.B.C.D graceful-restart**	Enable graceful restart for the peer	This is new command, the functionality is to enable both graceful restart and helper mode for the peer
router> **no neighbor A.B.C.D graceful-restart**	Disable graceful restart for the peer	This is new command, the functionality is to disable both graceful restart and helper mode for the peer. The peer will inherit global configuration
router> **neighbor A.B.C.D graceful-restart-helper**	Enable helper mode for peer	This is new command, the functionality is to enable helper mode for the peer
router> **no neighbor A.B.C.D graceful-restart-helper**	Disable helper mode for peer	This is new command, disables the helper mode for peer. The peer will inherit global configuration
router> **neighbor A.B.C.D graceful-restart-disable**	Disable grace restart and helper for peer	This is new command, the functionality is to enable helper mode for the peer
router> **no neighbor A.B.C.D graceful-restart-disable**	Enable Inherit mode for peer	This is new command, disables the helper mode for peer. The peer will inherit global configuration

Deferral Timer

Deferral timer will get affect only when BGP has come up after restart and is in GR mode. If GR is not enabled and BGP receives update message, BGP process all the routes and keeps adding it to work queue in BGP.

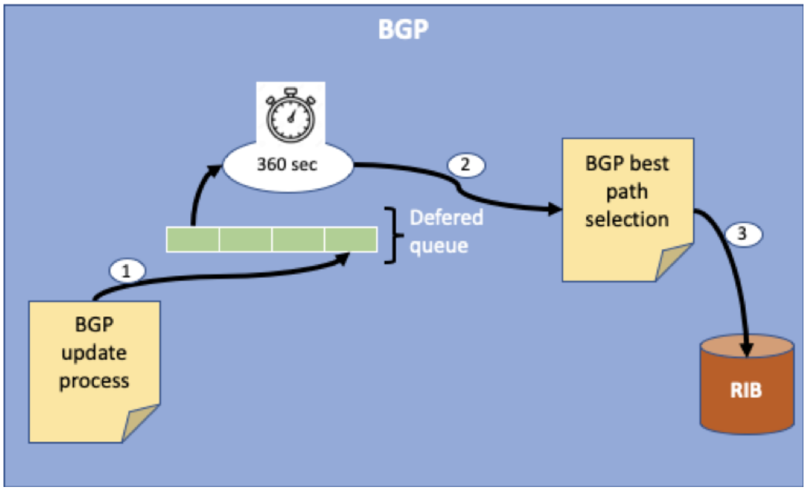
Work queue timers are as aggressive as 50ms and it would not be deferred for seconds. Sequence diagram are as show below.



1. BGP receives an update process each route and puts that in to work queue.
2. After 50ms work queue processing starts and sends each of the route for best path selection.
3. Post best path selection BGP puts selected route in RIB.

When GR is enabled and BGP restarts. Post restart BGP helpers (peers) start sending updates to BGP. BGP will process updates and defers processing of routes till deferred timer expiry.

Deferred time will give enough time for restarting router to receive routes from peer and consolidate them and run the best path selection at once for routes received from all peers. Sequence diagram for deferral timer is as show below.



1. BGP receives an update process each route and puts that in to deferred queue.
2. After deferral timer expiry, deferred queue processing starts and sends each of the route for best path selection.
3. Post best path selection BGP puts selected route in RIB.

4.

The selection deferral timer will not be configurable in product and default value will be used (360 secs). For debugging in FRR below CLI will be added for deferral timer.

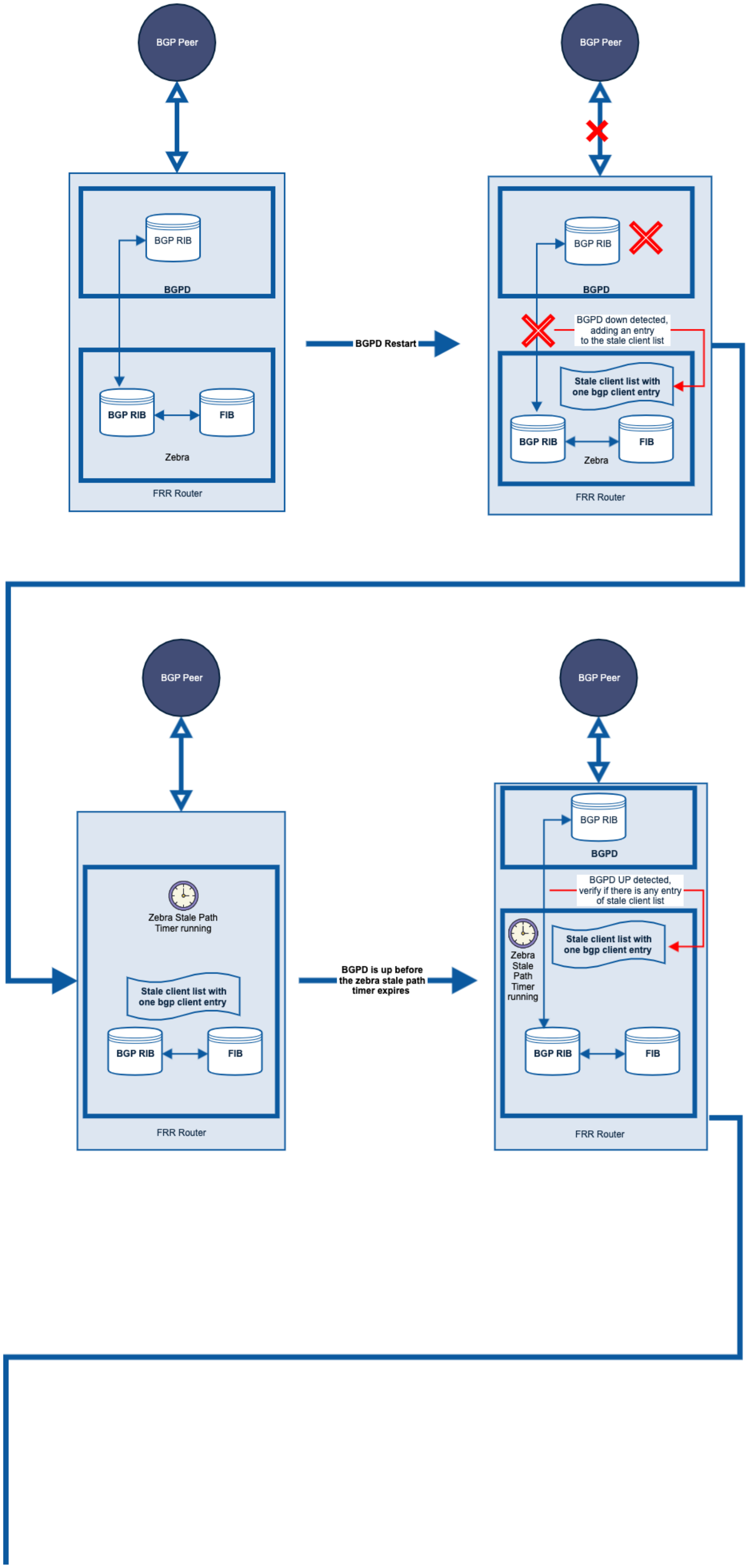
Command	**Description**	**Comments**

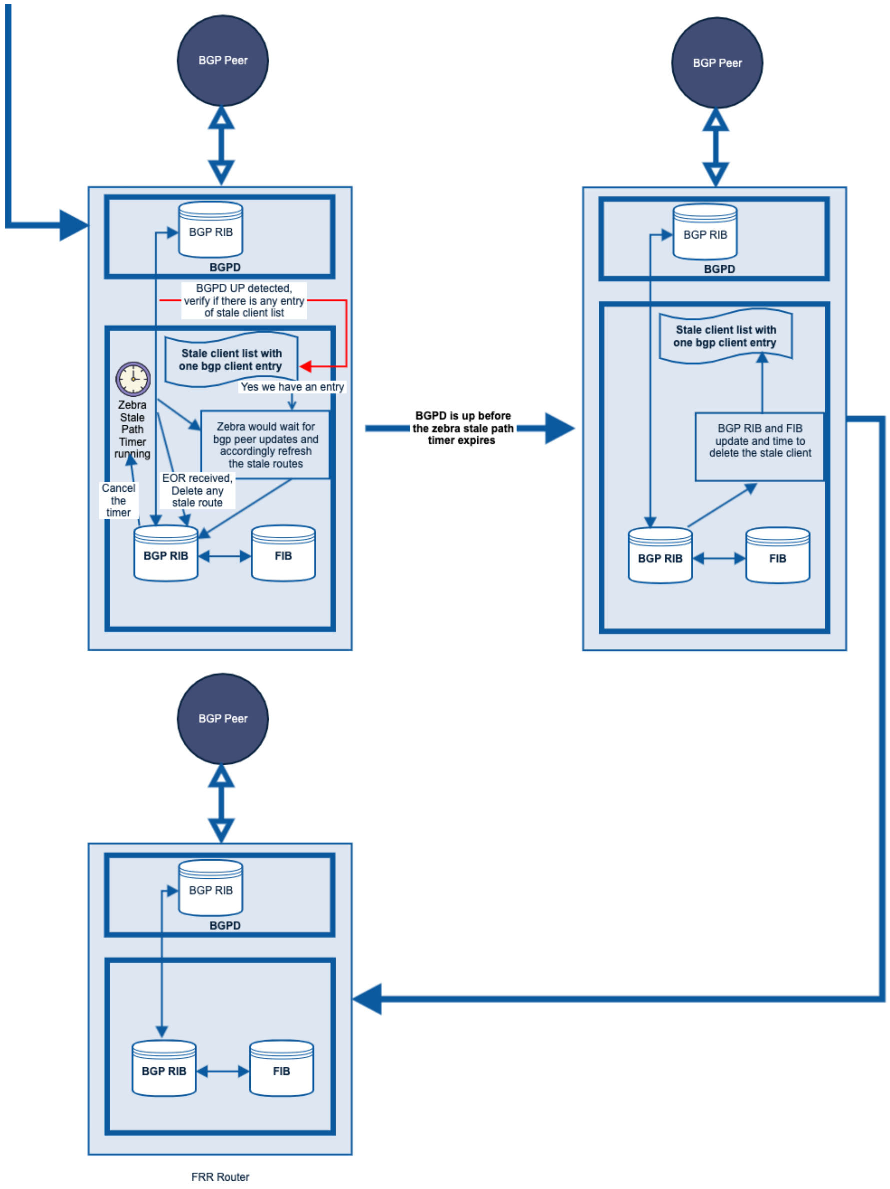
router> **bgp graceful-restart select-defer-time (0-3600)**	Set deferral time to value specified	This is new command, will set deferral time to value specified.
router> **no bgp graceful-restart select-defer-time (0-3600)**	Set deferral time to default value.	This is new command, will set deferral time to default value 3600 seconds.
router> **bgp graceful-restart rib-stale-time (1-3600)**	Specify the stale route removal timer in RIB	This is new command, will set the time for which stale routes are kept in RIB
router > **no bgp graceful-restart rib-stale-time (1-3600)**	Set the tale route removal timer in RIB	This is new command, will set rib stale time to default value (500 sec)

Route Retention in ZEBRA During BGP Restart

Support for route retention in zebra when BGP goes for graceful restart will be implemented. To support this feature below changes will be implemented in BGP and Zebra.

1. BGP graceful restart when enabled needs to send capabilities with GR enable and stale time value for which routes needs to be retained.
2. BGP restarts then Zebra will mark all the routes received from BGP as stale.
3. BGP comes back up and starts installing all the routes that it learns from its peers.
4. Zebra unmarks all the routes which it received from BGP as not stale and any stale route not updated by the bgp peer would be deleted after the End-Of-Rib is received.
5. In case the End-Of-Rib is not received. Zebra waits till the stale timer expires and remove all the stale entries.


BGP Peer
BGP RIB
BGPD
BGP RIB
FIB
Zebra
FRR Router
BGPD Restart
BGP Peer
BGP RIB
BGPD
BGPD down detected, adding an entry to the stale client list
Stale client list with one bgp client entry
BGP RIB
FIB
Zebra
FRR Router
BGP Peer
Zebra Stale Path Timer running
Stale client list with one bgp client entry
BGP RIB
FIB
FRR Router
BGPD is up before the zebra stale path timer expires
BGP Peer
BGP RIB
BGPD
BGPD UP detected, verify if there is any entry of stale client list
Zebra Stale Path Timer running
Stale client list with one bgp client entry
BGP RIB
FIB
FRR Router


BGP Peer
BGP RIB
BGPD
BGPD UP detected, verify if there is any entry of stale client list
Stale client list with one bgp client entry
Yes we have an entry
Zebra Stale Path Timer running
Zebra would wait for bgp peer updates and accordingly refresh the stale routes
Cancel the timer
EOR received, Delete any stale route
BGP RIB
FIB
BGPD is up before the zebra stale path timer expires
BGP Peer
BGP RIB
BGPD
Stale client list with one bgp client entry
BGP RIB and FIB update and time to delete the stale client
BGP RIB
FIB
BGP Peer
BGP RIB
BGPD
BGP RIB
FIB
FRR Router

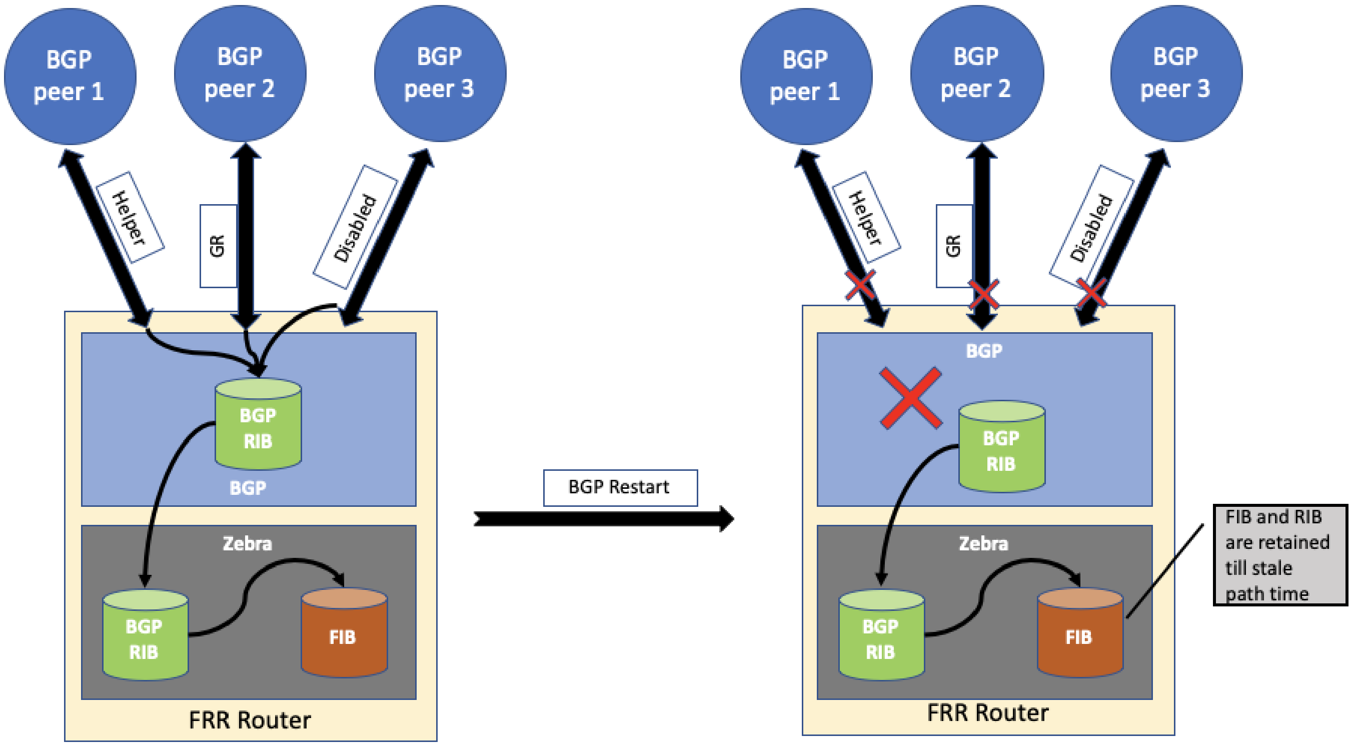
Caveats

Per Peer GR Behaviour

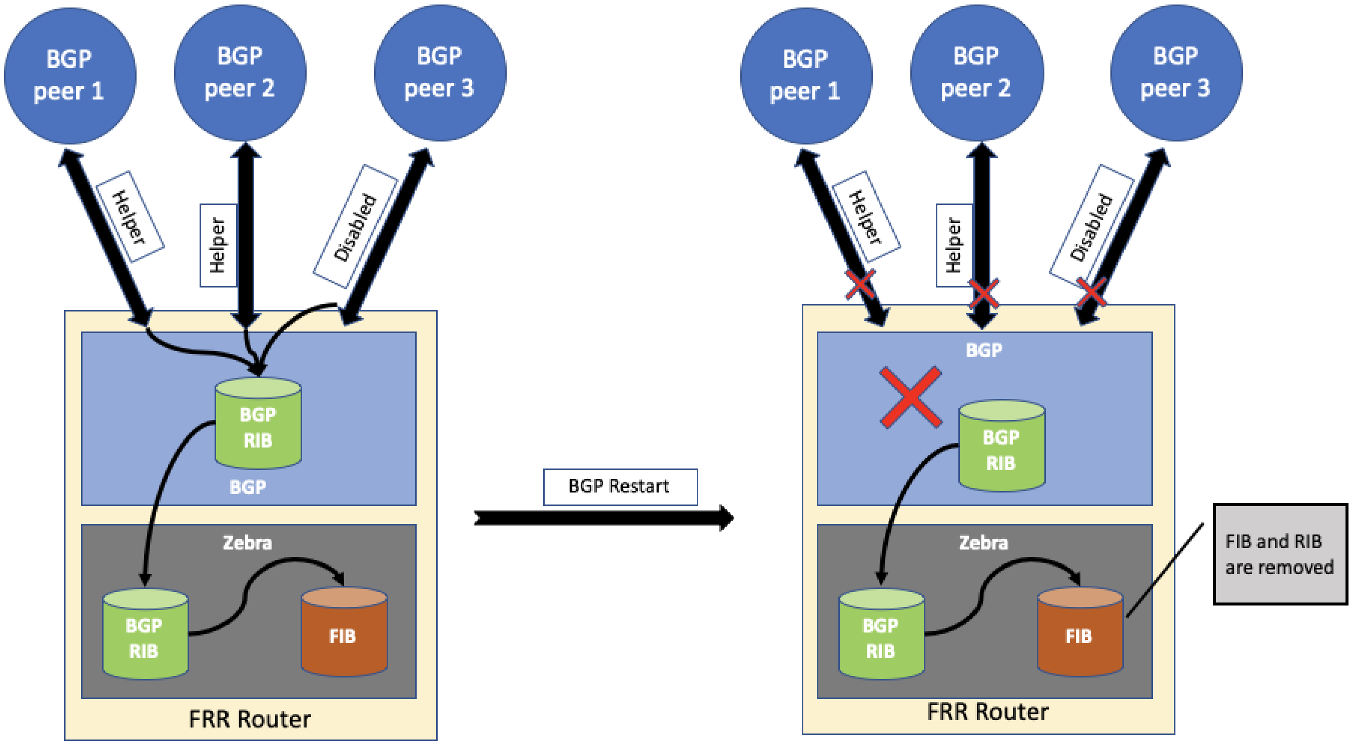
Route retention is for FIB. In FIB there is no notion of peer's route, it would be consolidated for a best path received from multiple peer. With BGP per peer functionality only some of the peers will have BGP GR enabled and some of them may be in disabled or helper mode. In this case when BGP goes for graceful restart then Zebra will retain all the routes received by BGP.

Retaining routes for transient time for peers from whom GR was not enabled will not have any issues as BGP will eventually converge and remove all the stale routes if there is a change in the network while BGP was restarting.

When per peer is configured and at least one of the peer has GR enabled. In this case BGP will send zebra with GR capabilities and zebra will retain routes for all BGP received routes as show below.



When per peer is configured none of the peer has GR enabled. In this case BGP will not send zebra with GR capabilities and zebra will flush all BGP received routes.



References

RFC : https://tools.ietf.org/html/rfc4724

https://www.cisco.com/en/US/docs/ios-xml/ios/iproute_bgp/configuration/15-1sg/irg-grace-restart-neighbor.html

Glossary

FRR – Free Range Routing Stack

BGP - Border Gateway Protocol

EBGP - Exterior Border Gateway Protocol

IBGP - Interior Border Gateway Protocol

RIB - Routing Information Base

FIB - Forwarding Information Base

PLR - Provider Logical Router or Tier 0 router

SR - Service Router

DR - Distributed Router

AS - Autonomous System

TOR – Top of the Rack Switch