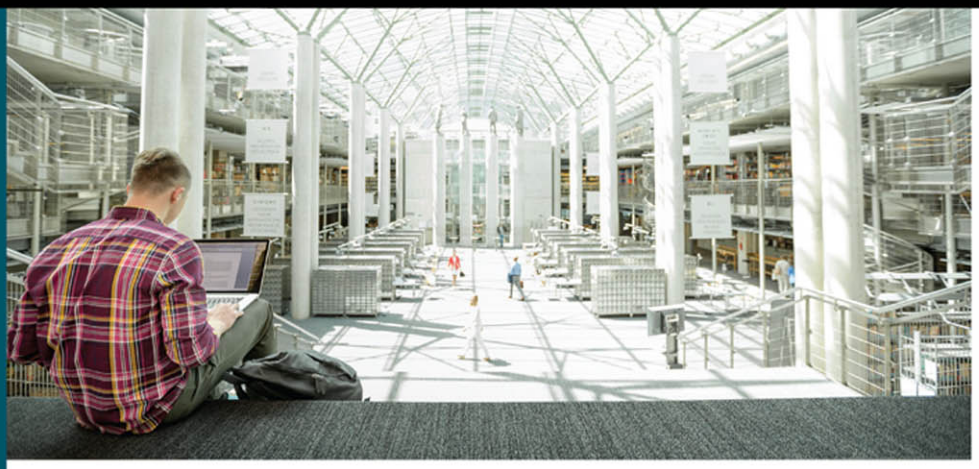

Troubleshooting BGP

A Practical Guide to Understanding and Troubleshooting BGP

Vinit Jain, CCIE No. 22854
Brad Edgeworth, CCIE No. 31574

ciscopress.com

```
  Refresh Epoch 7
  Local, (Received from a RR-client), (received & used)
    10.1.35.1 (metric 31) from 192.168.3.3 (192.168.3.3)
      Origin incomplete, metric 0, localpref 100, valid, internal, best
      rx pathid: 0, tx pathid: 0x0
```

```
! Output after 10.1.35.1 becomes unreachable
R2# show bgp ipv4 unicast 192.168.5.5
BGP routing table entry for 192.168.5.5/32, version 66
Paths: (2 available, best #2, table default)
  Advertised to update-groups:
     2
  Refresh Epoch 9
  Local, (Received from a RR-client), (received & used)
    10.1.45.1 (metric 41) from 192.168.4.4 (192.168.4.4)
      Origin incomplete, metric 0, localpref 100, valid, internal
      rx pathid: 0, tx pathid: 0
  Refresh Epoch 7
  Local, (Received from a RR-client), (received & used)
    10.1.35.1 (metric 2) from 192.168.3.3 (192.168.3.3)
      Origin incomplete, metric 0, localpref 100, valid, internal, best
      rx pathid: 0, tx pathid: 0x0
```

Similar issues can be seen when there is a summarized route in IP RIB. These kind of issues can be resolved using the BGP Selective Next-Hop Tracking feature, which is discussed later in the chapter.

BGP Next-Hop Tracking

The BGP NHT feature was designed to overcome some of the challenges of the BGP Scanner process and convergence issues that are seen with default or summarized routes in RIB. BGP NHT feature is designed on Address Tracking Filter (ATF) infrastructure, which enables event-driven NEXT_HOP reachability updates. ATF provides a scalable event-driven model for dealing with IP RIB changes. A standalone component, ATF allows for selective monitoring of IP RIB updates for registered prefixes. The following actions are taken after ATF tracks all prefixes that are registered:

- Notify client about route changes.
- Client, in this case BGP, is responsible for taking action based on ATF notifications.

For event-driven NEXT_HOP reachability, BGP registers its next-hops from all available paths with ATF during the best-path calculation. ATF maintains all the information

regarding the next-hops and tracks any IP RIB events in a dependency database. When a route for BGP next-hop changes (add/modify/delete) in the RIB, ATF notifies BGP about the change of its next-hop. Upon notification, BGP triggers a table walk to compute the best path for each prefix. This table walk is known as a lightweight "BGP Scanner" run. Figure 5-6 illustrates the interaction between BGP, ATF, and the IP RIB.

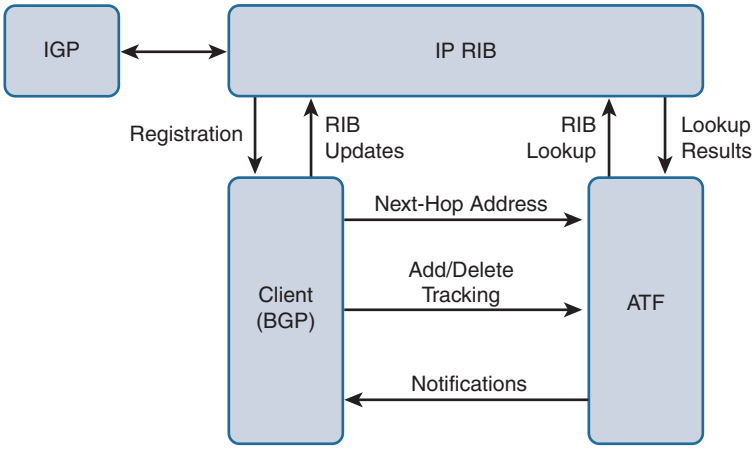

Figure 5-6 *Interaction Between BGP, ATF, and IP RIB*

Because BGP is capable of holding a large number of prefixes, and each prefix can have multiple paths, reacting to each notification from ATF could be costly for BGP. For this reason, the table walk is scheduled after a calculated time.

BGP NHT feature is enabled by default in almost all releases where the feature is supported, unless it is explicitly disabled on the router using the command **no bgp nexthop trigger enable** under the BGP address-family on Cisco IOS. BGP NHT cannot be disabled on IOS XR and NX-OS platforms. On Cisco IOS devices, a list of next-hops that are registered with ATF is viewed using the hidden command **show ip bgp attr nexthop**. Example 5-10 displays the output of all next-hops registered with ATF on the route-reflector router in Figure 5-5.

Example 5-10 *Next-Hops Registered with ATF*

```
R2# show ip bgp attr nexthop
 Next-Hop            Metric      Address-Family   Table-id rib-filter
 10.1.13.3           0           ipv4 unicast     0        0x10B624C3
 10.1.45.5           3           ipv4 unicast     0        0x10B62400
 10.1.35.5           2           ipv4 unicast     0        0x10B62493
 10.1.14.4           0           ipv4 unicast     0        0x10B62433
```

After an ATF notification is received, BGP waits 5 seconds (by default) before triggering the next-hop scan. This timer is called the NHT trigger delay. The NHT trigger delay

is changed for an address-family by using the command **bgp nexthop trigger delay** *seconds* on Cisco IOS. IOS XR and NX-OS platforms categorize the delay timers differently. The RIB notifications are classified based on the severity—critical and noncritical. The delay timers on IOS XR are configured by using the address-family command **nexthop trigger-delay** [**critical** | **non-critical**] *seconds*. NX-OS only provides command-line interface (CLI) to configure a trigger delay just for critical notifications using the address-family command **nexthop trigger-delay critical** *seconds*.

Note The trigger delay is 5 seconds by default in almost all releases, except for 12.0(30)S or earlier (it is 1 second).

Selective Next-Hop Tracking

BGP NHT overcomes the problem faced because of periodic BGP scan by introducing the event-driven quick scan paradigm, but it still does not resolve the inconsistencies caused by default route or summarized route present in the RIB. To overcome these problems, a new enhancement was introduced in BGP NHT called the BGP Selective Next-Hop Tracking or BGP Selective Next-Hop Route Filtering.

A route map is used during best-path calculation and is applied on the routes in IP RIB that cover the next-hop of BGP prefixes. If the route to the next-hop fails the route-map evaluation during a BGP NHT scan triggered by a notification from ATF, the route to the next-hop is marked as unreachable. Selective Next-Hop Tracking is configured per address-family; this allows for different route maps to be applied for next-hop routes in different address-families.

Selective NHT is enabled using the command **bgp nexthop route-map** *route-map-name* on Cisco IOS software. Examine the Selective NHT feature configuration in Example 5-11.

Example 5-11 *Selective NHT Configuration*

```
router bgp 100
address-family ipv4 unicast
 bgp nexthop route-map Loop32
!
ip prefix-list le-31 seq 5 per 0.0.0.0/0 le 31
!
route-map Loop32 deny 10
 match ip address prefix-list le-31
!
route-map Loop32 permit 20
```

On Cisco IOS XR, use the command **nexthop route-policy** *route-policy-name* to implement the Selective Next-Hop Tracking feature.

Note The IOS XR implementation for BGP Next-Hop Tracking and Selective Next-Hop Tracking supports every address-family identifier (AFI)/subaddress-family identifier (SAFI), whereas IOS supports only IPv4/VPNv4 and VPNv6.

Slow Convergence due to Advertisement Interval

BGP neighbor advertisement interval or MRAI causes delays in update generation if set to a higher value configured manually. It is a good practice to have the same MRAI timer at both ends of the neighbor and also across different platforms. Cisco IOS has advertisement interval of 0 seconds for IBGP as well as EBGP session in a VRF and 30 seconds for EBGP session, whereas IOS XR and NX-OS has the advertisement interval of 0 seconds for both IBGP and EBGP sessions. Thus if there is an IOS router and an IOS XR router both having EBGP sessions, then IOS router advertises any update after the MRAI timer has passed, which is 30 seconds, whereas IOS XR advertises the update immediately. The higher advertisement interval can cause slow convergence because the updates are not replicated before the MRAI timer expires.

The advertisement interval or MRAI value is modified using the command **neighbor** *ip-address* **advertisement-interval** *time-in-seconds* on Cisco IOS and the command **advertisement-interval** *time-in-seconds* under the neighbor configuration mode on IOS XR. The MRAI value is not configurable on the NX-OS platform.

Computing and Installing New Path

By default, BGP always selects only one best path (assuming BGP multipath is not configured). In case of failure of the best path, BGP has to go through the path selection process again to compute the alternative best path. This takes time and thus impacts convergence time. Also, features such as BGP NHT help improve the convergence time by providing fast reaction to IGP events, but that is still not significant because it depends on the total number of prefixes to be processed for best-path selection. With the BGP multipath feature, equal cost paths can be used for both redundancy and faster failover.

The question is, what happens when there are unequal cost paths available in the network? Can the backup path be programmed in the forwarding engine (that is, Cisco Express Forwarding in case of Cisco devices) of the router? And how can multiple paths be received across a route reflector? There have been new enhancements in past few years in this area with which all these problems can be resolved. Features such as BGP Best External, BGP Add Path, and BGP Prefix Independent Convergence (PIC) are a few of the features that are answers to all these questions. These features highly enhance the convergence for BGP and provide high availability in the network. Backup paths are now precomputed and installed in the forwarding engine. Thus in case of any failure, the traffic switches to the installed backup path.

Note BGP PIC and other High Availability topics are covered in Chapter 14, "BGP High Availability."

Troubleshooting BGP Convergence on IOS XR

BGP convergence troubleshooting techniques that are specific to IOS XR are covered. When troubleshooting IOS XR BGP convergence issues, the first thing you need to do is to verify whether the issue is seen right after the router boots or if it's seen after a link or protocol flap, if fast convergence features are enabled, availability of an alternate path, and how fast the other supporting infrastructure functions (Label Switch Database (LSD), RIB, forwarding information base (FIB), and the like) are updated. It is also important to verify if any route policy is implemented or a route map on the peer router is attached to the BGP peer. Other possible scenarios that could lead to slow convergence on IOS XR could be a memory leak condition on the BGP process or by any critical application running in the system. This may cause the BGP to either converge slowly or hinder the BGP process from processing messages in a timely manner.

Verifying Convergence During Initial Bring Up

Verification needs to be performed when the router is rebooted, the BGP process has restarted or crashed, or BGP is configured for the first time. The following actions confirm convergence:

- **BGP process state:** Ensure that the BGP process is in Run state when it is started or restarted. If it is in a different state or is continuously restarting, then it needs further investigation and a Technical Assistance Center (TAC) case has to be opened.
- **bgp update-delay configuration:** Sometimes the BGP update-delay parameter is modified as part of the design requirements. BGP stays in Read-Only (RO) mode until the update-delay time elapses, unless it receives End of Row (EoR) from all peers.
- **Non-Stop Routing (NSR):** If NSR is configured, a Stateful Switchover (SSO) is achieved by using the **nsr process-failures switchover** configuration knob. Verify the state using the command **show redundancy** on the IOS XR platforms.
- **Verify BGP process performance statistics:** Check BGP performance statistics, such as when the first BGP peer was established, what time it moved out of RO mode, and so on. This is checked by using the command **show bgp process**. When a new session is established and when the router begins exchanging OPEN messages, the router enters into BGP RO mode.
- **performance-statistics detail:** This command is AFI/SAFI aware and should be checked for relevant AFI/SAFI.

Example 5-12 demonstrates the verification of the preceding points to troubleshoot any BGP convergence issue on the router.

Example 5-12 *Troubleshooting Convergence on IOS XR*

```
RP/0/0/CPU0:R10# show process bgp | include Process state
           Process state: Run

! Verifying bgp update-delay configuration
RP/0/0/CPU0:R10# show run router bgp | include update-delay
 bgp update-delay 360

! Verifying BGP in RO mode or Normal mode
RP/0/0/CPU0:R10# show bgp process detail | include State
State: Normal mode.

RP/0/0/CPU0:R10# show bgp process performance-statistics detail | begin First nei
First neighbor established:  Oct 13 23:40:05
Entered DO_BESTPATH mode:    Oct 13 23:46:09
Entered DO_IMPORT mode:      Oct 13 23:46:09
Entered DO_RIBUPD mode:      Oct 13 23:46:09
Entered Normal mode:         Oct 13 23:46:09
Latest UPDATE sent:          Oct 13 23:46:14
```

Note The difference between the first established neighbor and update sent should be approximately the same as the update-delay time.

Verifying BGP Reconvergence in Steady State Network

To troubleshoot BGP reconvergence in steady state, always wait for the BGP application to be in RW (read/write) or Normal mode. After the BGP is in any of these states, check the router to see whether it exhibits any of the following symptoms:

- Check whether the router is not exceeding maximum allowed BGP peers and prefixes.
- Check for memory leaks by BGP application or any critical process.
- Check for constant link/peer flaps and troubleshoot based on that.
- Check whether the slow convergence is noticed for a particular peer or multiple peers. Compare the peers converging at a slower pace to the ones converging faster.
- Check for any inefficiently configured route policy.
- Ensure whether Path MTU Discovery is enabled.
- Check for any issues with RIB/FIB/BCDL infrastructure that are adding to the convergence delay