Troubleshooting BGP Convergence Issues

Vinit Jain (CCIE# 22854)

BRKRST-3320 Preview

Cisco live!

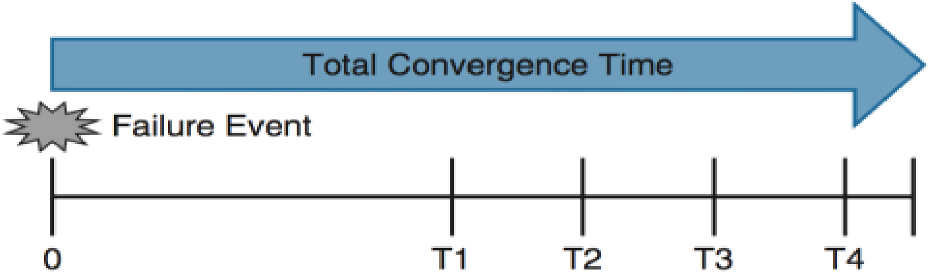
Agenda

- Introduction
- What is Convergence? Convergence in terms of BGP
- Understanding Update Groups
    - Update Groups in Cisco IOS
    - Update Groups in IOS XR
- BGP slow-peer

Scenario 3 - Troubleshooting BGP Convergence

Problem Description

- BGP Table is getting updated slowly
- Traffic loss (Traffic Black-Hole) is experienced
- High CPU



Troubleshooting BGP Convergence

What is convergence in terms of BGP?

- Establish sessions with a number of peers
- Locally generate all the BGP path (either via network command, redistribute static/connected/IGP), and/or from other component for other address family
  - e.g. MVPN from multicast, L2VPN from l2vpn mgr, EVPN from evpn mgr, etc.
- Send and receive multiple BGP tables (different BGP address-families) to/from each peer
- Upon receive all the paths from peers, do the best path calculation to find the best path (and/or multi path, additional-path, backup path, etc.)

Troubleshooting BGP Convergence

What is convergence in terms of BGP?

- If import/export is involved, the import/export of all kind of variations
  - VRF import, AF import, global import, MVPN import, EVPN import, etc.
- Install the best path into multiple routing table
  - Default RIB or VRF, IPv4/IPv6
- For other address family, pass the path calculation result to different lower layer components like step 2 (mvpn, evpn, l2vpn, etc.)

Troubleshooting BGP Convergence

Dimensional Factors

- Number of peers
- Number of address-families
- Number of path/prefix per address-family
- Link speed of individual interface, individual peer
- Different update group settings and topology
- Complexity of attribute creation / parsing for each address-family

Complex Routing Policy – IOS XR

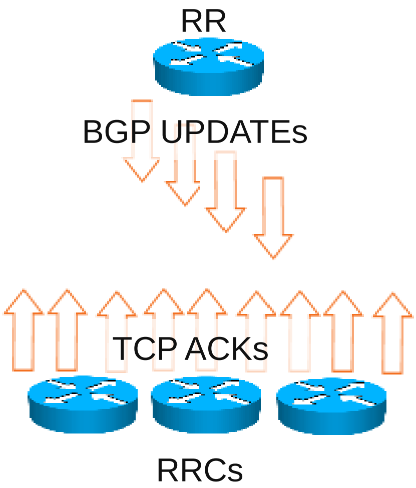
```
as-path-set match-ases
  ios-regex '^(.*65531)$',
  ios-regex '^(.*65532)$',
  ios-regex '^(.*65533)$',
     <snip>
prefix-set K1-routes
  10.170.53.0/24
end-set
prefix-set K2-routes
  10.147.4.0/24
end-set
prefix-set K3-routes
  198.168.44.0/23,
  198.168.46.0/24
end-set
```

```
route-policy Inbound-ROUTES
  if destination in K1-routes then
    pass
  elseif destination in K2-routes then
    pass
  elseif destination in K3-routes then
    pass
else
    drop
  endif
end-policy
!
router bgp 65530
neighbor-group IGW
  remote-as 65530
address-family ipv4 unicast
route-policy Inbound-ROUTES in
```

Convergence

Dropping TCP Acks

- Primarily an issue on RRs (Route Reflectors) with
  - One or two interfaces connecting to the core
  - Hundreds of RRCs (Route Reflector Clients)
- RR sends out tons of UPDATES to RRCs
- RRCs send TCP ACKs
- RR core facing interface(s) receive huge wave of TCP ACKs



Convergence

Dropping TCP Acks

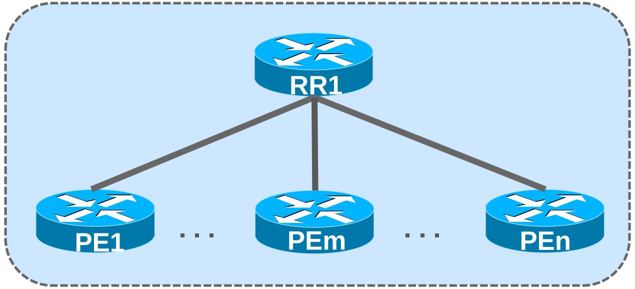
- Interface input queue fills up…TCP ACKs are dropped ☹
  - Each time a TCP packet is dropped, the session goes into slow start
  - It takes a good deal of time for a TCP session to come out of slow start
- Increase the input queue
  - **hold-queue 1000 in**
- If you still see drops increase to 4096

BGP Update Groups

Troubleshooting BGP Convergence

Update Groups

- Update Group is a collection of peers with identical outbound policy.
- Helps in improving IBGP convergence
  - Update messages are formatted and replicated to all the peers
- A Master is selected in the update group, which is updated first in the group
- Based on the message formatted for the master / Leader, all the peers are then replicated with the same formatted message
  - The message formatting only happens once.



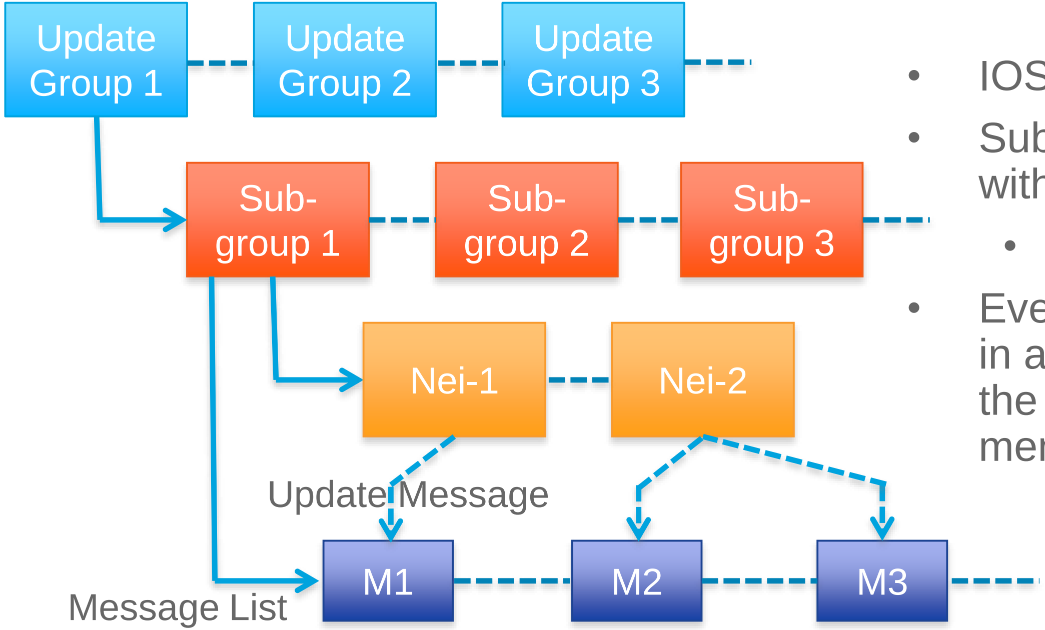
Troubleshooting BGP Convergence

Update Groups

```
R1#show bgp ipv4 unicast update-group
BGP version 4 update-group 2, internal, Address Family: IPv4 Unicast
  BGP Update version : 7/0, messages 0, active RGs: 1
  Route-Reflector Client
  Route map for outgoing advertisements is dummy
  Topology: global, highest version: 7, tail marker: 7
  Format state: Current working (OK, last not in list)
                Refresh blocked (not in list, last not in list)
  Update messages formatted 4, replicated 15, current 0, refresh 0, limit
1000
  Number of NLRIs in the update sent: max 1, min 0
  Minimum time between advertisement runs is 0 seconds
  Has 4 members:
   10.1.12.2        10.1.13.2*        10.1.14.2        10.1.15.2
```

Troubleshooting BGP Convergence

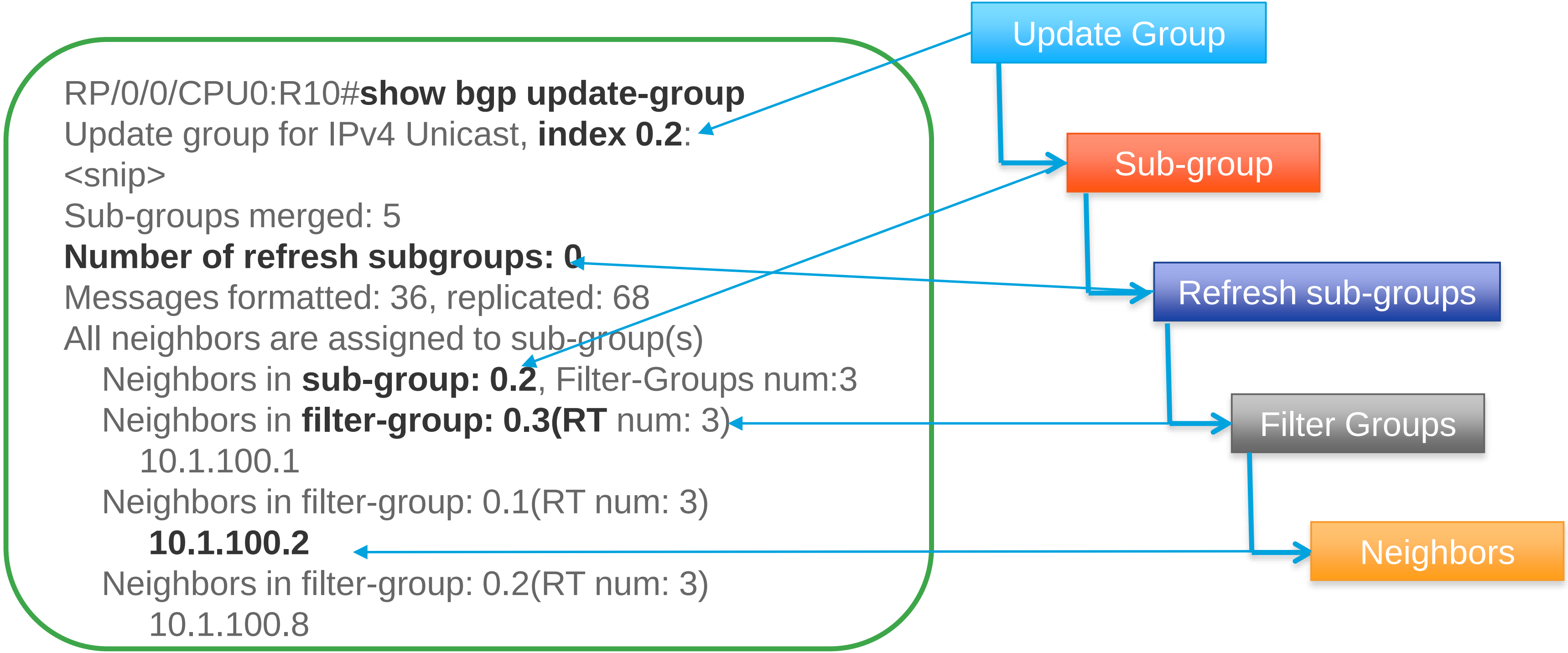
Update Groups on IOS XR



- IOS XR have hierarchical update groups
- Sub-Groups are subset of neighbors within an update Group
  - Neighbors running at same pace
- Even a newly configured neighbor is put in a separate sub-group till it reaches the same table version as other members

Troubleshooting BGP Convergence

Update Groups on IOS XR



```
RP/0/0/CPU0:R10#show bgp update-group
Update group for IPv4 Unicast, index 0.2:
<snip>
Sub-groups merged: 5
Number of refresh subgroups: 0
Messages formatted: 36, replicated: 68
All neighbors are assigned to sub-group(s)
   Neighbors in sub-group: 0.2, Filter-Groups num:3
   Neighbors in filter-group: 0.3(RT num: 3)
      10.1.100.1
   Neighbors in filter-group: 0.1(RT num: 3)
       10.1.100.2
   Neighbors in filter-group: 0.2(RT num: 3)
       10.1.100.8
```

Verify TCP Stats – IOS XR

```
RP/0/8/CPU0:R10#show tcp brief | include 10.1.102.2
0x10146a20 0x60000000  0  0  10.1.102.1:62233  10.1.102.2:179 ESTAB
```

```
RP/0/8/CPU0:R10#show tcp stat pcb 0x10146a20 location 0/8/CPU0
==================================================================
 Statistics for PCB 0x10146a20, vrfid 0x60000000
Send:    0 bytes received from application
        <snip>
         0 packets failed getting queued to network (v4/v6 IO)
         0 packets failed getting queued to network (NetIO)
Rcvd:    722 packets received from network
         380 packets queued to application
         0 packets failed queuing to application
```

Verify TCP NSR Stats – IOS XR

```
RP/0/8/CPU0:R10#show tcp nsr statistics pcb 0x10146a20
PCB 0x10146a20
Number of times NSR went up: 1
Number of times NSR went down: 0
Number of times NSR was disabled: 0
Number of times switch-over occured : 0
IACK RX Message Statistics:
  Number of iACKs dropped because SSO is not up              : 0
  Number of stale iACKs dropped                              : 0
  Number of iACKs not held because of an immediate match     : 0
TX Messsage Statistics:
    Data transfer messages:
        Sent 118347, Dropped 0, Data (Total/Avg.) 2249329/19
                <SNIP>
```

Troubleshooting BGP Convergence – IOS XR

Show bgp all all convergence

```
RP/0/0/CPU0:R10# show bgp all all convergence
Address Family: IPv4 Unicast
==========================================
Converged.
All received routes in RIB, all neighbors updated.
All neighbors have empty write queues.

Address Family: VPNv4 Unicast
==========================
Not converged.
Received routes may not be entered in RIB.
One or more neighbors may need updating.
```

Not converged – implies that there are BGP neighbors that for which the replication has not completed yet

Troubleshooting BGP Convergence – IOS XR

Verifying Performance Statistics

```
RP/0/0/CPU0:R10#show bgp ipv4 unicast update-group 0.2 performance-
statistics
Update group for IPv4 Unicast, index 0.2:
  <snip>
  Messages formatted: 0, replicated: 0
  All neighbors are assigned to sub-group(s)
    Neighbors in sub-group: 0.1, Filter-Groups num:1
     Neighbors in filter-group: 0.1(RT num: 0)
      10.1.102.2    10.1.103.2    10.1.104.2    10.1.105.2
  Updates generated for 0 prefixes in 10 calls(best-external:0)
             (time spent: 10.000 secs)
 <snip>
```

Verify the time spent in generating and replicated the updates

BGP Convergence – NX-OS

Show bgp convergence detail

```
R20# show bgp convergence detail vrf all
Global settings:
BGP start time 5 day(s), 13:55:45 ago
Config processing completed 0.119865 after start
BGP out of wait mode 0.119888 after start
LDP convergence not required
Convergence to ULIB not required
Information for VRF default
Initial-bestpath timeout: 300 sec, configured 0 sec
BGP update-delay-always is not enabled
First peer up 00:09:18 after start
Bestpath timer not running
 Contd…
```

Troubleshooting BGP Convergence – NX-OS

Show bgp convergence detail

```
Contd…
IPv4 Unicast:
   First bestpath signalled 00:00:27 after start
   First bestpath completed 00:00:27 after start
   Convergence to URIB sent 00:00:27 after start
   Peer convergence after start:
    10.1.202.2          (EOR after bestpath)
    10.1.203.2          (EOR after bestpath)
    10.1.204.2          (EOR after bestpath)
    10.1.205.2          (EOR after bestpath)
```

If bestpath is received before EOR or peer fails to send EOR marker, it can lead to traffic loss

Troubleshooting BGP Convergence – NX-OS

Enable Debugging using Filters

```
debug bgp events updates rib brib import
debug logfile bgp
debug-filter bgp vrf vpn1
debug-filter bgp address-family ipv4 unicast
debug-filter bgp neighbor 10.1.202.2
debug-filter bgp prefix 192.168.2.2/32
```

Troubleshooting BGP Convergence – NX-OS

When Route is not downloaded into URIB

When the route is not download into URIB, it may not be a problem with BGP.

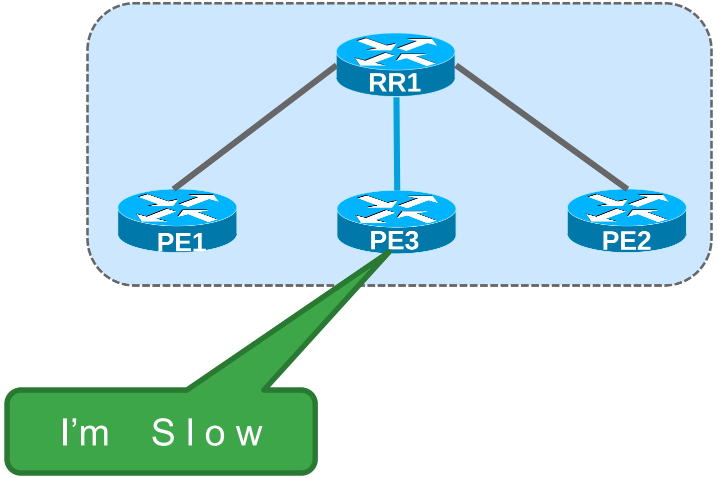
- Show routing internal event-history ufdm
- Show routing internal event-history ufdm-summary
- Show routing internal event-history recursive

BGP Slow Peer

Scenario 4 – BGP Slow Peer

Problem Description

- Customer reports updates not getting across all PE routers
- Caused due to:
  - RR's sending updates with high speed
  - Slow processing peers
- Symptoms
  - High CPU due to BGP
  - Updates not replicated to all peers
  - Router reloads



BGP OutQ & Cache Size

- OutQ column should show very high OutQ value
- Should be reaching the maximum cache size for that update-group

```
Router# show ip bgp vpnv4 all summary
..
Neighbor        V    AS MsgRcvd MsgSent    TblVer  InQ OutQ Up/Down  State/PfxRcd
12.123.67.97    4   109      42   87065         0    0 1000 00:10:00        0
12.122.78.19    4   109      42   87391         0    0  674 00:10:00        0

Router# show ip bgp vpnv4 all replication
                                                                    Current Next
Index  Members          Leader       MsgFmt    MsgRepl     Csize    Version Version
    1      348    12.123.67.97   1726595727 1938155978  999/1000 1012333000/1012351142
    2        2    12.122.78.19     79434677   79398843     0/200  1012351503/1012351503
    3        1   199.37.187.24            0          0     0/100          0/0
    4        2   12.122.78.249     79219618   97412908     0/200  1012351504/1012351504
```

TCP sndwnd

```
Router#show ip bgp neighbor 10.1.0.1
..
iss: 3662804973  snduna: 3668039487  sndnxt: 3668039487     sndwnd:      0
irs: 1935123434  rcvnxt: 1935222998  rcvwnd:      16003  delrcvwnd:    381

SRTT: 300 ms, RTTO: 303 ms, RTV: 3 ms, KRTT: 0 ms
minRTT: 0 ms, maxRTT: 512 ms, ACK hold: 200 ms
Status Flags: passive open, gen tcbs
Option Flags: nagle, path mtu capable
```

- Check for send window (sndwnd) and receive window (rcvwnd) using “show ip bgp neighbor <x.x.x.x>”
- For the TCP session for which outQ is high, we might notice that sndwd is very low or zero.
- On the remote end, we should see the rcvwnd value is very low or zero.

Solution - Static Slow peer

- The manual knob to flag a peer as slow will create a separate update group for the peer.
- The advantage - there is a limit to the overhead that this feature will create.
- The drawback - slow member update group will have to progress at the pace of the slowest of the slow peers.

```
neighbor {<nbr-addr>/<peer-grp-name>} slow-peer split-update-group static
```

- This command will manually mark a neighbor as slow peer.
- The peer will be part of slow update group.

Solution - Dynamic Slow peer

- IOS BGP will monitor the transmission speeds of the peers.
- A peer will have to be exhibiting slowness for several minutes to be flagged.
- Log message for when a slow peer is detected/recovered

```
bgp slow-peer detection [threshold <seconds>]

neighbor {<nbr-addr>/<peer-grp-name>} slow-peer detection [threshold < seconds >]
```

- The threshold defines "the threshold time in seconds" to detect a peer as slow peer.
- The range is 120 seconds to 3600 seconds. Default is 300 seconds.

Solution - Slow peer protection

- Depends on Dynamic Slow Peer feature
- When a slow peer recovery is detected (the peer has converged), the peer will be moved back to its original group

```
bgp slow-peer split-update-group dynamic [permanent]

neighbor {<nbr-addr>/<peer-grp-name>} slow-peer split-update-group dynamic [permanent]
```

- When “permanent” is not configured, the “slow peer” will be moved to its regular original update group, after it becomes regular peer (converges).
- If “permanent” is configured, the peer will not be moved to its original update group automatically

Syslog Messages

- The below log message will be generated when a peer is detected as dynamic slow peer.

```
"bgp neighbor %s in af %d is detected as slow-peer"
```

- The below log message will be generated when a “slow-peer” recovers.

```
"slow bgp peer %s in af %d has recovered"
```

BGP Slow Peer - Commands

- Show Commands

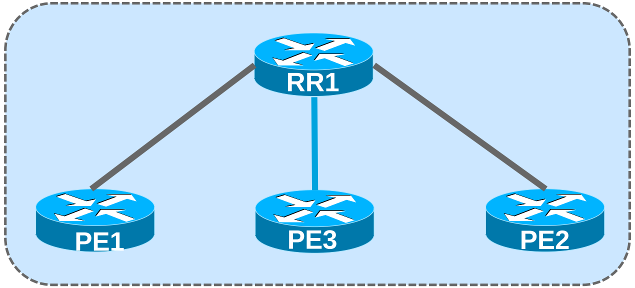
```
show ip bgp [AF/scope/topo] update-group summary slow

show ip bgp [AF/scope/topo] summary slow

show ip bgp [AF/scope/topo] neighbor slow
```

- Clear Commands

```
Clear [ip] bgp <nbr-addr> slow

Clear [ip] bgp peer-group <group-name> slow

Clear [ip] bgp af * slow

Clear ip bgp * slow
```

TAC Case Example - 6

BGP Slow Peer

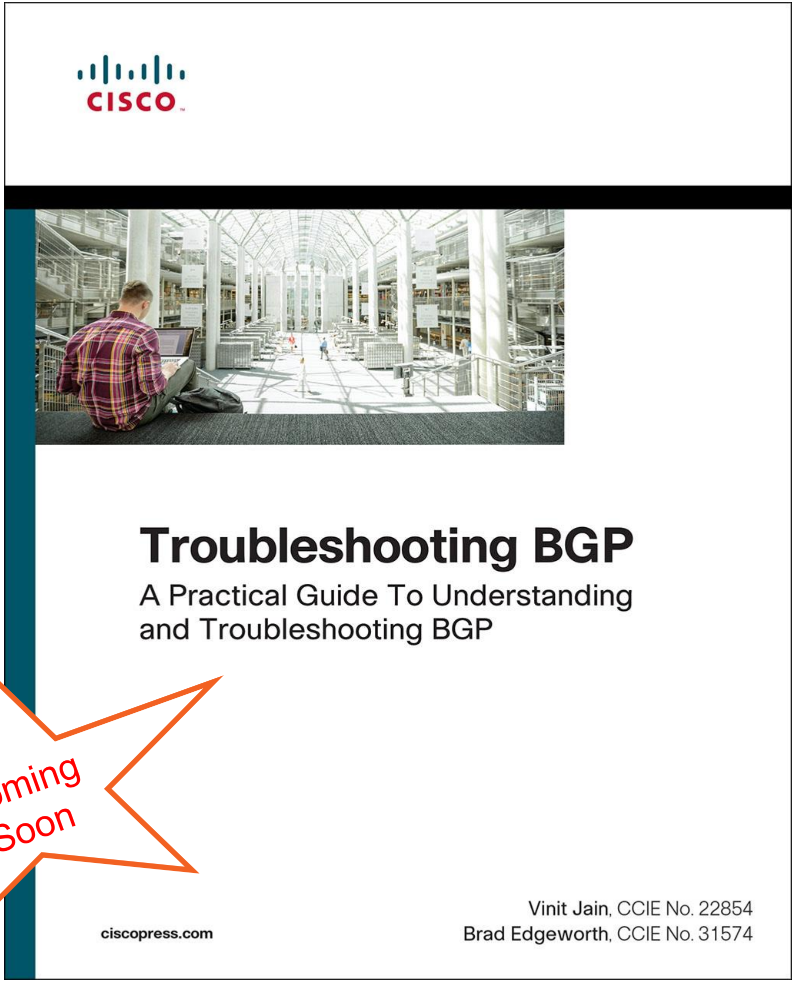


- Customer reported routes were stuck in BGP RR.
- Their end-customer removed service from one of their locations but the routes are still seen on their RR and other locations
- Soft clearing the neighborship temporarily resolved the problem but reoccurred again after sometime

```
RR1# show ip bgp vpnv4 all replication
                                                                    Current    Next
Index  Members          Leader      MsgFmt    MsgRepl     Csize    Version Version
    1      150     216.156.3.10   274950548  650809652 2000/2000  421492656/421493582
    2        5      65.106.7.100   41049479  204232170     0/500   421493582/0
    5        1   66.239.189.212    16143960   16143960     0/100   421491282/421493582
```

Resolution

- Two neighbors were identified to be showing slow peer symptoms
- Customer’s RR router didn’t had the slow peer capability in the IOS they were running
- Two workarounds / solutions:
    - Create a separate outbound policy for slow peers.
    - Use the "neighbor <ip> advertisement-interval <interval>".
        - Default for internal neighbors is 5 sec and for external is 30 seconds.



CISCO

Troubleshooting BGP

A Practical Guide To Understanding
and Troubleshooting BGP

ciscopress.com

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

Coming
Soon

Thank you

We’re ready. Are you?