BGP Neighbor Flaps with MTU Troubleshooting TechNote

Document ID: 116377

Contributed by Amer Ibrahimovic, Mani Ganesan, and Jay Kulkarni,
Cisco TAC Engineers.
Jul 16, 2013

Contents

Introduction

This document describes how to determine if internal or external Border Gateway Protocol (BGP) neighbor flaps are caused by maximum transmission unit (MTU) issues.

Prerequisites

Ensure you complete these tasks on both BGP routers before you complete the procedures in this document:

- Check the BGP configuration.
- Verify that the BGP neighbor is reachable via Internet Control Message Protocol (ICMP) and no drops are observed.
- Verify that the connected interface used to peer BGP is not oversubscribed and does not have any input/output drops or errors.
- Check the CPU and memory utilization.

Problem

BGP neighbors form; however, at the time of prefix exchange, the BGP state drops and the logs generate missing BGP hello keepalives or the other peer terminates the session.

Complete these steps in order to determine if the MTU causes the BGP neighbors to flap:

1. Use the below commands in order to check which neighbor is affected and the connected interface on both BGP routers. If the peering address is a loopback address, check the connected interface through which the loopback is reachable. Also, check for the BGP OutQ on both peering routers. The consistent non−zero OutQ is a strong indication that updates do not reach the peer due to an MTU issue in the path.

```
Router#show ip bgp summ | in InQ|10.10.10.2
Neighbor      V   AS MsgRcvd MsgSent  TblVer  InQ OutQ Up/Down  State/PfxRcd
10.10.10.2    4   3     64      62       3     0    0   00:00:3       2

Router#show ip route 10.10.10.2
Routing entry for 10.10.10.0/24
  Known via "connected", distance 0, metric 0 (connected, via interface)
  Routing Descriptor Blocks:
  * directly connected, via GigabitEthernet1/0
      Route metric is 0, traffic share count is 1
```

2. Check the interface MTU on both sides:

   ```
   Router#show ip int g1/0 | i MTU
     MTU is 1500 bytes
   Router#
   ```

3. Confirm the TCP agreed max data segment for both BGP speakers:

   ```
   Router#show ip bgp neigh 20.20.20.2 | inc segment
   Datagrams (max data segment is 1460 bytes):
   Router#
   ```

   In the example above, 1460 is correct as 20 bytes is assigned to the TCP header and another 20 to the IP header.

4. Confirm if BGP used *path−mtu is enabled*:

   ```
   Router#show ip bgp neigh 10.10.10.2 | in tcp
     Transport(tcp) path−mtu−discovery is enabled
   Router#
   ```

5. Ping the BGP peer with max interface MTU and DF (Don't Fragment) bit set:

   ```
   Router#ping 10.10.10.2 size 1500 df

   Type escape sequence to abort.
   Sending 5, 1500−byte ICMP Echos to 10.10.10.2, timeout is 2 seconds:
   Packet sent with the DF bit set
   .....
   Success rate is 0 percent (0/5)
   ```

6. Decrease the ICMP size value in order to determine the maximum MTU size that can be used:

   ```
   ping 10.10.10.2 size 1300 df
   ```

Solution

Here are some possible causes:

- The interface MTU on both routers do not match.
- The interface MTU on both routers match, but the Layer 2 domain over which the BGP session is formed does not match.
- Path MTU discovery determined the incorrect max datasize for the TCP BGP session.
- The BGP Path Maximum Transmission Unit Discovery (PMTUD) could be failing due to PMTUD ICMP packets blocked (firewal or ACL)

Here are possible ways to resolve MTU issues:

1. The interface MTU on both routers should be the same; run the ***show ip int | in MTU*** command in order to check the current MTU settings.

2. If the interface MTU on both routers are correct (for example, 1500) but the ping tests with DF bit set do not exceed 1300, then the Layer 2 domain on which the affected BGP session is formed might include inconsistent MTU configurations. Check each Layer 2 interface MTU. Correct the Layer 2 interface MTU in order to resolve the issue.

3. If you are unable to check/change the Layer 2 domain, you can set the ***ip tcp mss*** global command to lesser value like 1000, which will force all locally originated TCP max data segment sessions (which includes BGP) to 1000. For more information on this command, refer to the ip tcp mss section of the *Cisco IOS IP Application Services Command Reference*.

   In addition, you can use the ***ip tcp adjust−mss*** command in order to troubleshoot further; this command is configured at the interface level and affects all TCP sessions. For more information on this command, refer to the ip tcp adjust−mss section of the *Cisco IOS IP Application Services Command Reference*.

4. (*Optional*) The BGP Path Maximum Transmission Unit Discovery (PMTUD) might not generate the correct maximum data size. You can disable it globally or per neighbor in order to confirm if this is the cause. When BGP PMTUD is disabled, the BGP Maximum Segment Size (MSS) defaults to 536 as defined in RFC 879.

   For information on how to disable PMTUD, refer to the Configuring BGP Support for TCP Path MTU Discovery per Session section of the *Cisco IOS BGP Configuration Guide*.

   For more information on PMTUD, refer to What Is PMTUD?

Updated: Jul 16, 2013 Document ID: 116377