BGP Flapping Issue Case Study

目录

硬件平台

GSR

软件版本

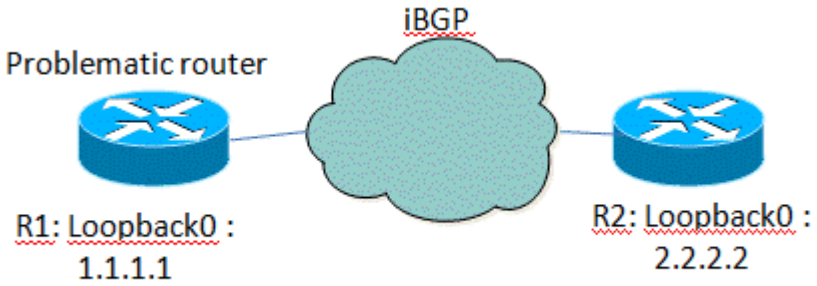
IOS 12.0(32)SY8

案例介绍

用户发现两个IBGP邻居不停的flapping，有一个明显的规律是bgp 邻居建立起来之后经过大约5分钟的时间就会因为holdtimer超时而down掉，然后邻居又会马上建立起来。

问题分析思路

1. 用户环境
   两个IBGP邻居用loopback0建立IBGP邻居,中间经过多台路由器。示意拓扑如下：



2. 问题的规律

从log中我们可以发现这个IBGP邻居断掉以及重建的规律。每次都是因为holdtimer超时，并且是因为对端收不到本端发出去keepalive报文。

```
Dec  6 13:28:36: %BGP-5-ADJCHANGE: neighbor 2.2.2.2 Up
Dec  6 13:33:55: %BGP-3-NOTIFICATION: received from neighbor 2.2.2.2 4/0 (holdtime expired) 0
bytes
Dec  6 13:33:55: %BGP-5-ADJCHANGE: neighbor 2.2.2.2 Down BGP Notification received
Dec  6 13:34:22: %BGP-5-ADJCHANGE: neighbor 2.2.2.2 Up
```

```
Dec  6 13:39:37: %BGP-3-NOTIFICATION: received from neighbor 2.2.2.2 4/0 (holdtime expired) 0
bytes
Dec  6 13:39:37: %BGP-5-ADJCHANGE: neighbor 2.2.2.2 Down BGP Notification received
```

3. 因为bgp邻居每次断掉之后就会很快的重建起来，IP路由应该没有问题。(1)现在我们要理解bgp邻居建立起来之后会做什么。发送路由更新，也就是update报文。(2)对于bgp来说，邻居的维护使用keepalive报文，但是如果有路由更新，发送update报文的话，此update报文就具有keepalive报文的功能，路由器就不用再专门发送一个bgp keepalive报文。 综合以上两点我们可以怀疑是bgp update报文对端没有收到。
4. 由于是第一个update报文，根据路由表容量的大小，我们可以知道这个报文会达到最大值，我们知道bgp 报文最大可以达到4096个字节，同时也受限于TCP 对端的MSS大小。
   MSS（Maximum Segment Size ）是指一个host能够接受最大TCP datagram的大小，这个值是在tcp option字段中协商得到的。我们可以通过以下命令知道。如下这个命令输出的 *Datagrams (max data segment is 4394 bytes)*。

```
R1#  show ip bgp vpnv all neighbors 2.2.2.2
BGP neighbor is 2.2.2.2,  remote AS 65350, internal link
 Description: To_ R2
 Member of peer-group NXVRRgroup for session parameters
  BGP version 4, remote router ID 202.100.126.219
  BGP state = Established, up for 00:00:56
Last read 00:00:51, last write 00:00:56, hold time is 180, keepalive interval is 60
seconds
  Neighbor capabilities:
    Route refresh: advertised and received(new)
    Four-octets ASN Capability: advertised and received
    Address family VPNv4 Unicast: advertised and received
  Message statistics:
    InQ depth is 0
    OutQ depth is 0
                         Sent       Rcvd
    Opens:                 35         35
    Notifications:          2         28
    Updates:           935784        467
    Keepalives:        133137     147643
    Route Refresh:          0          1
    Total:             1068931     148175
  Default minimum time between advertisement runs is 0 seconds

 For address family: VPNv4 Unicast
  BGP table version 1316545, neighbor version 0/0
 Output queue size : 0
  Index 3, Offset 0, Mask 0x8
  Route-Reflector Client
  Member of update-group 3
  NXVRRgroup peer-group member
  NEXT_HOP is always this router
                                 Sent       Rcvd
  Prefix activity:               ----       ----
    Prefixes Current:            3591        184 (Consumes 12512 bytes)
    Prefixes Total:                 0        184
    Implicit Withdraw:              0          0
    Explicit Withdraw:              0          0
    Used as bestpath:             n/a         46
    Used as multipath:            n/a          0

                                   Outbound    Inbound
  Local Policy Denied Prefixes:    --------    -------
    Total:                                0          0
  Number of NLRIs in the update sent: max 0, min 0

  Address tracking is enabled, the RIB does have a route to 2.2.2.2
```

```
  Connections established 35; dropped 34
  Last reset 00:01:17, due to BGP Notification received, hold time expired
Connection state is ESTAB, I/O status: 1, unread input bytes: 0
Mininum incoming TTL 0, Outgoing TTL 255
Local host: 1.1.1.1, Local port: 179
Foreign host: 2.2.2.2, Foreign port: 24434

Enqueued packets for retransmit: 0, input: 0  mis-ordered: 0 (0 bytes)

Event Timers (current time is 0x74A9E3D88):
Timer          Starts    Wakeups            Next
Retrans             2          0             0x0
TimeWait            0          0             0x0
AckHold             4          3             0x0
SendWnd             0          0             0x0
KeepAlive           0          0             0x0
GiveUp              0          0             0x0
PmtuAger            0          0             0x0
DeadWait            0          0             0x0

iss: 1432533502  snduna: 1432533575  sndnxt: 1432533575     sndwnd:  65463
irs: 4098882880  rcvnxt: 4098886860  rcvwnd:      61556  delrcvwnd:   3979

SRTT: 836 ms, RTTO: 3946 ms, RTV: 1137 ms, KRTT: 0 ms
minRTT: 0 ms, maxRTT: 300 ms, ACK hold: 200 ms
Flags: passive open, nagle, path mtu capable, gen tcbs,
  SACK option permitted

Datagrams (max data segment is 4394 bytes):
Rcvd: 8 (out of order: 0), with data: 6, total data bytes: 3979
Sent: 5 (retransmit: 0, fastretransmit: 0), with data: 1, total data bytes: 72
```

5. 所以我们可以用带DF的ping来测试此路径中是否能允许此update报文通过，因为bgp update报文在路径中不能被分片。

```
R1#ping
Protocol [ip]:
Target IP address: 2.2.2.2
Repeat count [5]:
Datagram size [100]: 2200 //datagram2200
Timeout in seconds [2]:
Extended commands [n]: y
Source address or interface: loopback0
Type of service [0]:
Set DF bit in IP header? [no]: yes
Validate reply data? [no]:
Data pattern [0xABCD]:
Loose, Strict, Record, Timestamp, Verbose[none]:
Sweep range of sizes [n]:
Type escape sequence to abort.
Sending 5, 2200-byte ICMP Echos to 2.2.2.2, timeout is 2 seconds:
.....
Success rate is 0 percent (0/5)
```

6. 至此我们可以知道是路径中某台路由器的接口mtu较小导致。

问题总结

此问题原因是因为客户的IGP环境发生了改变，某一台路由器的流量出口选择了一条备份链路，但是此链路接口mtu很小，导致bgp update报文在此被堵塞而造成holdtimer超时。

经验总结

1. 对于bgp 的troubleshooting，因为bgp 是基于tcp报文的，所以不仅仅是ip层面的，也有可能是也要基于tcp层面进行分析。
2. BGP neighbor发出的notification而导致的邻居down掉，我们都可以基于给出的 error code 和subcode(holdtimer 超时是4/0)来知道down掉的原因。具体的code的分类和意思可以参考RFC4271.

相关命令

Show ip bgp *