使用BGP“慢对等体”功能解决慢对等体问题

目录

简介

本文档介绍如何使用边界网关协议(BGP)慢速对等体功能解决慢速对等体问题。该功能可识别BGP更新组中的慢速对等体，并可将慢速对等体永久或临时移出更新组。

背景信息

本节概述慢速对等体功能和更新组的使用。

更新组

更新组中使用慢速对等体功能。更新组是一种动态方法，用于将BGP对等体与相同的出站策略进行分组。更新组的优点是使用组策略对消息进行一次格式化，然后将其复制并传输到组的其他成员。此方法比单独格式化每个对等体的BGP更新的需要更有效。

实施此方法时，如果出站策略更改，则对等组会按更新组更改。根据地址系列(AF)形成更新组。

以下是AF IPv4单播的不同更新组中两个BGP对等体的示例，但AF VPNv4的相同更新组：

```
R2#show ip bgp update-group
BGP version 4 update-group 1, external, Address Family: IPv4 Unicast
 Has 1 member (* indicates the members currently being sent updates):
```

```
  10.1.3.4

BGP version 4 update-group 2, external, Address Family: IPv4 Unicast
 Has 1 member (* indicates the members currently being sent updates):
  10.1.2.3

R2#show ip bgp vpnv4 all update-group
BGP version 4 update-group 1, external, Address Family: VPNv4 Unicast
 Has 2 members (* indicates the members currently being sent updates):
  10.1.2.3         10.1.3.4
```

随着更新组中包含的BGP对等体数量的增加，更新组变得更高效。通常，内部BGP(iBGP)对等体具有相同的出站策略。对于iBGP，路由反射器(RR)可以有许多iBGP对等体；因此，它将具有大型更新组。提供商边缘(PE)路由器可以在一个虚拟/路由转发(VRF)中拥有许多指向客户边缘(CE)路由器的外部BGP(eBGP)对等体。 PE路由器可以具有大型更新组，也可用于VRF接口上与CE路由器的对等。

问题

慢速对等体是一个对等体，它无法跟上路由器在更新组中长时间（按分钟顺序）生成BGP更新消息的速率。原因可能是持续的网络问题。网络原因可能是丢包和/或加载链路，或BGP会话的吞吐量问题。此外，BGP对等体在CPU方面可能负载过重，无法以所需速度为TCP连接提供服务。

慢速对等体会影响完整更新组的BGP融合。如果一个BGP对等体速度较慢，则会导致整个更新组速度减慢。结果是，其他更新组成员的收敛速度也会较慢。因此，应解决该问题。

您可以识别慢速对等体并将其移出更新组。为完成此任务，您可以更改该BGP对等体的出站策略；但是，这是手动任务。您必须首先确定速度较慢的对等体，然后将其移出更新组。慢速对等体功能可以自动执行此操作，因此无需用户干预。

解决方案

慢对等体功能有三个部分：

- 检测慢速对等体
- 将慢速对等体移动到慢速更新组
- 慢速对等体的恢复（将恢复的对等体移回其原始更新组）

以下各节将对这些过程进行更详细的说明。

检测

慢速对等体功能可检测更新组中慢速对等体。每个更新组都有缓存队列，格式化的BGP更新在传输之前暂时存储在该队列中。

以下是此类更新组缓存的示例：

```
R2#show ip bgp replication
```

```
                                                                  Current    Next
Index  Members          Leader       MsgFmt    MsgRepl     Csize    Version Version
   1        1        10.1.1.1            0          0    0/100          6/0
   2        3        10.1.2.3            2          6    0/1000         6/0
   3        1        10.1.2.6            3          0    0/100          6/0
```

缓存的大小是动态计算的，具体取决于：

- 更新组中的对等体数
- 已安装的系统内存
- 更新组中的对等体类型
- AF的类型

当一个对等体（慢速对等体）没有像其他成员那样快速确认BGP消息时，等待传输的格式化BGP更新数可以在一个更新组中构建。达到缓存限制后，组不再有配额来排队新消息。在缓存减少之前（直到慢速对等体确认某些消息之前），不能格式化新消息。 这禁止BGP对等体，并且不允许它向组的更快成员发送新消息（更新或撤消）。因此，这会减慢更新组中所有对等体的收敛。

为了使慢速对等体功能识别慢速对等体，它指BGP更新时间戳和对等TCP参数。

默认情况下禁用慢速对等体检测。要启用慢速对等体检测，请使用以下方法之一：

- 为BGP进程启用功能（可以从AF/VRF配置）：

  bgp slow-peer detection [threshold

  注意：阈值的范围为120到3,600秒，默认值为300秒。
- 启用每个对等体的功能：

  neighbor {

- 通过对等策略模板启用该功能：

  slow-peer detection [threshold < seconds >]

  [no] slow-peer detection

当检测到慢速对等体时，将显示类似于此的系统日志消息：

%BGP-5-SLOWPEER_DETECT: Neighbor IPv4 Unicast 10.1.6.7 has been detected as a slow peer.

可以输入以下show命令以查看慢速对等体：

- show ip bgp summary slow
- show ip bgp neighbors slow
- show ip bgp update-group summary slow

以下是使用slow关键字时的show命令输出示例：

```
R2#show ip bgp update-group summary slow
Summary for  Update-group 1, Address Family IPv4 Unicast
Summary for  Update-group 2, Address Family IPv4 Unicast
Summary for  Update-group 3, Address Family IPv4 Unicast
Summary for  Update-group 4, Address Family IPv4 Unicast
BGP router identifier 10.1.6.2, local AS number 2
BGP table version is 966013, main routing table version 966013
BGP main update table version 966013
50000 network entries using 6050000 bytes of memory
50000 path entries using 2600000 bytes of memory
5001/5000 BGP path/bestpath attribute entries using 700140 bytes of memory
5000 BGP AS-PATH entries using 183632 bytes of memory
0 BGP route-map cache entries using 0 bytes of memory
0 BGP filter-list cache entries using 0 bytes of memory
BGP using 9533772 total bytes of memory
BGP activity 208847/158847 prefixes, 508006/458006 paths, scan interval 60 secs
Neighbor        V    AS MsgRcvd MsgSent   TblVer  InQ OutQ Up/Down  State/PfxRcd
10.1.6.7        4     7     165   50309        0    0  100 00:10:35        0
```

如输出所示，对等体10.1.6.7是AF IPv4单播的慢对等体。其他AF不显示任何慢对等体。

要验证检测计时器当前是否运行及其值，请输入以下命令：

```
R2#show ip bgp update-group
BGP version 4 update-group 3, external, Address Family: IPv4 Unicast
 BGP Update version : 116013/0, messages 164 queue 164, not converged
 Private AS number removed from updates to this neighbor
 Update messages formatted 5948, replicated 11589
 Number of NLRIs in the update sent: max 249, min 1
 Minimum time between advertisement runs is 30 seconds
 Slow-peer detection timer (expires in 111 seconds)
  Has 3 members (* indicates the members currently being sent updates):
  10.1.4.5         10.1.5.6         10.1.6.7
```

如示例输出所示，检测计时器已启动。当更新组缓存已满时，检测计时器开始。

在本示例中，您可以看到检测到慢速对等体，但只有在慢速对等体检测计时器到期后，它才会从更新组中移出：

```
R2#show ip bgp update-group
âÂ¦
BGP version 4 update-group 3, external, Address Family: IPv4 Unicast
 BGP Update version : 516013/566013, messages 357 queue 357, not converged
 Private AS number removed from updates to this neighbor
 Update messages formatted 27044, replicated 53645
 Number of NLRIs in the update sent: max 249, min 0
 Minimum time between advertisement runs is 30 seconds
 Slow-peer detection timer (expires in 20 seconds)
```

```
 Has 3 members (* indicates the members currently being sent updates)
 (1 dynamically detected as slow):

  *10.1.4.5        *10.1.5.6         10.1.6.7
```

慢速对等体识别

如果未启用慢速对等体检测功能，则必须手动识别慢速对等体。首先，检查更新组中对等体的表版本和输出队列：

```
R2#show ip bgp update-group 3 summary
Summary for  Update-group 3, Address Family IPv4 Unicast
BGP router identifier 10.1.6.2, local AS number 2
BGP table version is 552583, main routing table version 552583
BGP main update table version 552583
37870 network entries using 4582270 bytes of memory
37870 path entries using 1969240 bytes of memory
5002/3788 BGP path/bestpath attribute entries using 700280 bytes of memory
5001 BGP AS-PATH entries using 183656 bytes of memory
0 BGP route-map cache entries using 0 bytes of memory
0 BGP filter-list cache entries using 0 bytes of memory
BGP using 7435446 total bytes of memory
BGP activity 158847/108847 prefixes, 295876/258006 paths, scan interval 60 secs
Neighbor        V    AS MsgRcvd MsgSent   TblVer  InQ OutQ Up/Down  State/PfxRcd
10.1.4.5        4     5      77   26840   516013    0    0 01:07:12        0
10.1.5.6        4     6      69   26833   516013    0    0 01:00:30        0
10.1.6.7        4     7      79   26761   516013    0  194 00:45:42        0
```

在本示例中，验证对等体的表版本(TblVer)是否赶上了主BGP表版本，或者是否始终落后。其次，检查一个或多个输出队列值非常高的对等体。它们很可能是速度较慢的同行。

当您查看可疑的慢速BGP对等体时，请考虑以下问题（在BGP会话的两端）：

- 上次写作是多*久前*的？
- keepalive是否在限制中？
- 输出队列是否高？
- SRTT/*RTTO高*吗？
- 重传次数是否增加？
- 是否存在排队的重传数据包？
- TCP发*送窗*口是很低还是零？

示例如下：

```
R2#show ip bgp neighbors 10.1.6.7
BGP neighbor is 10.1.6.7,  remote AS 7, external link
Member of peer-group group3 for session parameters
 BGP version 4, remote router ID 10.1.6.7
 BGP state = Established, up for 00:56:09
 Last read 00:00:43, last write 00:00:17, hold time is 180, keepalive interval
```

```
is 60 seconds
 Keepalives are temporarily in throttle due to closed TCP window
 Neighbor capabilities:
   Route refresh: advertised and received(new)
   Address family IPv4 Unicast:
advertised and received
 Message statistics
   InQ depth is 0
   OutQ depth is 0    Partial message pending
                        Sent       Rcvd
   Opens:                  5          4
   Notifications:          0          0
   Updates:            29004          0
   Keepalives:             0       1426
   Route Refresh:          0          0
   Total:              30336       1431
 Default minimum time between advertisement runs is 30 seconds
For address family: IPv4 Unicast
 BGP table version 250001, neighbor version 200001/250001
 Output queue size : 410
 Index 3, Offset 0, Mask 0x8
 3 update-group member
 group3 peer-group member
 Inbound soft reconfiguration allowed
 Private AS number removed from updates to this neighbor
 Inbound path policy configured
 Route map for incoming advertisements is eBGP-in
                               Sent       Rcvd
 Prefix activity:              ----       ----
   Prefixes Current:           2596          0
   Prefixes Total:           102624          0
   Implicit Withdraw:            28          0
   Explicit Withdraw:        100000          0
   Used as bestpath:            n/a          0
   Used as multipath:           n/a          0
                                 Outbound    Inbound
 Local Policy Denied Prefixes:    --------    -------
   Total:                                0          0
 Maximum prefixes allowed 20000
 Threshold for warning message 80%, restart interval 300 min
 Number of NLRIs in the update sent: max 249, min 0
 Last detected as dynamic slow peer: never
Dynamic slow peer recovered: never
 Oldest update message was formatted: 00:02:24
 Address tracking is enabled, the RIB does have a route to 10.1.6.7
 Connections established 4; dropped 3
 Last reset 00:57:39, due to User reset
 Transport(tcp) path-mtu-discovery is enabled
Connection state is ESTAB, I/O status: 1, unread input bytes: 0
Connection is ECN Disabled
Mininum incoming TTL 0, Outgoing TTL 1
Local host: 10.1.6.2, Local port: 20298
Foreign host: 10.1.6.7, Foreign port: 179
Connection tableid (VRF): 0

Enqueued packets for retransmit: 15, input: 0  mis-ordered: 0 (0 bytes)
Event Timers (current time is 0x4A63D14):
Timer          Starts    Wakeups            Next
Retrans           697         29       0x4A6590C
TimeWait            0          0             0x0
AckHold            64         63             0x0
SendWnd             0          0             0x0
KeepAlive           0          0             0x0
GiveUp              0          0             0x0
```

```
PmtuAger          128        127       0x4A64CB7
DeadWait            0          0             0x0
Linger              0          0             0x0

iss:  130287252  snduna:  131516888  sndnxt:  131532233     sndwnd:  16384
irs: 1184181084  rcvnxt: 1184182346  rcvwnd:      15123  delrcvwnd:   1261

SRTT: 20122 ms, RTTO: 20440 ms, RTV: 318 ms, KRTT: 0 ms
minRTT: 20028 ms, maxRTT: 20796 ms, ACK hold: 200 ms
Status Flags: none
Option Flags: nagle, path mtu capable, higher precendence

Datagrams (max data segment is 1460 bytes):
Rcvd: 922 (out of order: 0), with data&colon; 65, total data bytes: 1261
Sent: 1463 (retransmit: 29 fastretransmit: 1),with data&colon; 1391, total
data bytes: 1245129
```

移动

本节介绍在各种场景中慢速对等体功能的移动过程。

不使用慢速对等项功能的移动

无需慢速对等体功能，即可手动将慢速对等*体移*动到新更新组中。

在慢速对等体功能可用之前，您需要识别慢速对等体，然后手动将其移出更新组。完成此操作时，将更改该BGP对等体的出站策略。此出站策略必须与使用的任何其他策略不同，因为必须确保慢速对等体不会移动到当前存在的另一个更新组（并将问题移到该更新组）。 您可以应用的最佳更改是不影响实际策略的更改。例如，您可以更改对等体（在特定AF下）的最小路由通告间隔(MRAI)。

以下示例显示当慢速对等体功能不可用时慢速对等体的手动移动：

```
RR1#debug ip bgp groups
BGP groups debugging is on

RR1(config)#router bgp 1
RR1(config-router)#address-family vpnv4
RR1(config-router-af)#neighbor 10.100.1.3 advertisement-interval 3

BGP-DYN(4): 10.100.1.3 cannot join update-group 1 due to an advertisement-interval
mismatch
BGP(4): Scheduling withdraws and update-group membership change for 10.100.1.3
BGP(4): Resetting 10.100.1.3's version for its transition out of update-group 1
BGP-DYN(4): 10.100.1.3 cannot join update-group 1 due to an advertisement-interval
mismatch
BGP-DYN(4): Removing 10.100.1.3 from update-group 1
BGP-DYN(4): 10.100.1.3 cannot join update-group 1 due to an advertisement-interval
mismatch
BGP-DYN(4): Created update-group 0 from neighbor 10.100.1.3
BGP-DYN(4): Adding 10.100.1.3 to update-group 0
```

静态慢对等移动

要将一个对等体从更新组移动到新更新组，可以将其配置为静态*慢速*对等体。如果有多个慢速对等体，则具有相同出站策略的静态慢速对等体将放入同一*慢速更*新组。

要静态移动慢速对等体，可以使用以下命令对其进行配置：

- 启用每个邻居或每个对等组的静态对等体移动：

  [no] neighbor {

- 通过对等策略模板启用静态对等体移动：

  [no] slow-peer split-update-group static

动态慢速对等移动

默认情况下，慢速对等体移动处于禁用状态。要启用慢速对等移动，可以通过以下方法之一对其进行配置：

- 为BGP进程启用慢速对等移动：

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

  [no] bgp slow-peer split-update-group dynamic

  注意：可以从地址系列/拓*扑/VRF视图配置此*配置。
- 启用每个对等体的慢速对等移动：

  neighbor {

- 通过对等策略模板启用慢速对等体移动：

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

  [no] slow-peer split-update-group dynamic

  注意：permanent关键字表示慢速对等体不会自动恢复。在这种情况下，您可以通过一个clear命令将恢复的慢速对等体移回其原始更新组。

静态慢速对等体和动态慢速对等体位于同一慢速对等体更新组中。在本示例中，您可以在慢速更新组中看到一个慢速对等体：

```
R2#show ip bgp update-group
âÂ¦
BGP version 4 update-group 4, external, Address Family: IPv4 Unicast
 BGP Update version : 0/566013, messages 100 queue 100, not converged
 Slow update group
  Private AS number removed from updates to this neighbor
 Update messages formatted 2497, replicated 0
 Number of NLRIs in the update sent: max 10, min 1
```

```
 Minimum time between advertisement runs is 30 seconds
 Has 1 member (* indicates the members currently being sent updates)
 (1 dynamically detected as slow):
 *10.1.6.7
```

恢复

一旦确认慢速对等体不再是慢速对等体（它将迎接），就可以在其原始更新组（与出站策略匹配）下重新分组。 当慢速对等体更新组收敛时，恢复计时器开始。当恢复计时器过期时，慢速对等体将移回常规更新组。

注意：要查看与检测/恢复计时器相关*的行为*，请输入**debug ip bgp updates events命**令。

当慢速对等体移回原始更新组（这表示恢复）时，将显示类似以下的系统日志消息：

```
%BGP-5-SLOWPEER_RECOVER: Slow peer IPv4 Unicast 10.1.6.7 has recovered.
```

要验证恢复计时器当前是否运行以及值，请输入以下命令：

```
R2#show ip bgp update-group
BGP version 4 update-group 1, external, Address Family: IPv4 Unicast
 BGP Update version : 165973/0, messages 0 queue 0, converged
 Route map for outgoing advertisements is dummy
 Update messages formatted 0, replicated 0
 Number of NLRIs in the update sent: max 0, min 0
 Minimum time between advertisement runs is 30 seconds
 Slow-peer recovery timer (expires in 16 seconds)
  Has 1 member (* indicates the members currently being sent updates):
  10.1.1.1
```

在本例中，值为**16秒的恢复计时器**表示，可能较慢的对等体可能会在16秒后移回其原始更新组。

在本例中，您可以看到已从慢对等状态恢复的对等体：

```
R2#show ip bgp neighbor 10.1.6.7

BGP neighbor is 10.1.6.7,  remote AS 7, external link
Member of peer-group group3 for session parameters
 BGP version 4, remote router ID 10.1.6.7
âÂ¦
3 update-group member
 group3 peer-group member
âÂ¦
Number of NLRIs in the update sent: max 249, min 0
 Last detected as dynamic slow peer: 00:12:49
  Dynamic slow peer recovered: 00:01:57
  Oldest update message was formatted: 00:00:55
```

清除慢速对等体状态

使用以下命令可以手动清除慢速对等体状态：

- clear ip bgp * slow

- clear ip bgp *AF {unicast|multicast}* <AS number> slow
- clear ip bgp *AF {unicast|multicast}* peer-group <group-name> slow
- clear ip bgp <neighbor-address> slow
- clear bgp *AF {unicast|multicast}* * slow
- clear bgp *AF {unicast|multicast}* <AS number> slow
- clear bgp *AF {unicast|multicast}* peer-group <group-name> slow
- clear bgp *AF {unicast|multicast}* <neighbor-address> slow

注意：使用这些命令时，请用实际**地址**系列替换AF。

使用这些命令，对等体将移回原始更新组。

输入**show ip bgp internal**命令以查看慢速对等体检测和移动设置：

```
R2#show ip bgp internal
Time left for bestpath timer: 593 secs
Address-family IPv4 Unicast, Mode : RW
   Table Versions : Current 622091, RIB 622091
   Start time : 00:00:01.168    Time elapsed 01:21:56.740
   First Peer up in : 00:00:07    Exited Read-Only in : 00:02:16
   Done with Install in : 00:02:26    Last Update-done in : never
   0 updates expanded
   Attribute list queue size: 0
   Slow-peer detection is enabled  Threshold is 300 seconds
   Slow-peer split-update-group dynamic is enabled
   BGP Nexthop scan:-
       penalty: 0, Time since last run: never,  Next due in: none
       Max runtime : 0 ms Latest runtime : 0 ms Scan count: 0
   BGP General Scan:-
       Max runtime : 14572 ms Latest runtime : 14572 ms Scan count: 78
   BGP future scanner version: 79
   BGP scanner version: 0
```

注意：总之，BGP慢速对等体是一种功能，可检测BGP更新组中的慢速对等体，并允许通过将慢速对等体移出更新组来加快BGP融合。