使用BGP「慢速對等體」功能解決慢速對等體問題

目錄

簡介

本文說明如何使用邊界閘道通訊協定(BGP)*慢速對等體*功能解決慢速對等體問題，此功能會識別BGP更新組中的慢速對等體，並可將慢速對等體永久或暫時移出更新組。

背景資訊

本節概述了慢速對等體功能和更新組的使用。

更新組

更新組中使用慢速對等體功能。更新組是一種動態方法，用於將具有相同出站策略的BGP對等體分組。更新組的好處在於，使用組策略格式化消息一次，然後複製這些消息並傳輸到組的其他成員。此方法比分別為每個對等體格式化BGP更新的需要更有效率。

實施此方法時，如果出站策略發生更改，對等體組會根據更新組進行更改。更新組按照地址系列(AF)形成。

以下範例顯示兩個BGP對等體位於不同的AF IPv4單播更新組中，但具有相同的AF VPNv4更新組：

```
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
```

隨著更新組中包含的BGP對等體數量的增加，更新組將變得更加高效。通常，內部BGP(iBGP)對等體具有相同的傳出策略。若是iBGP，路由反射器(RR)可以有許多iBGP對等點；因此，它將具有大型更新組。提供者邊緣(PE)路由器可以在一個虛擬/路由轉送(VRF)中擁有多個針對客戶邊緣(CE)路由器的外部BGP(eBGP)對等體。 PE路由器可以有大型更新組，也可以用於在VRF介面上與CE路由器對等的情況。

問題

慢速對等體是指無法跟上路由器在更新組中長時間內（以分鐘為單位）生成BGP更新消息的速率。原因可能是持續存在的網路問題。網路原因可能是丟包和/或載入的鏈路，或者BGP會話的吞吐量問題。此外，BGP對等體可能在CPU方面負載過重，無法以所需的速度為TCP連線提供服務。

較慢的對等體會影響完整更新組的BGP收斂。如果一個BGP對等體變慢，將導致整個更新組變慢。結果是，其他更新組成員的收斂速度也會較慢。因此，應解決此問題。

您可以識別慢速對等體並將其從更新組中移出。為了完成此任務，您可以變更該BGP對等體的出站策略；但是，這是一個手動任務。必須首先識別速度慢的對等體，然後將其從更新組中移出。慢速對等功能可以自動完成此操作，因此不需要使用者干預。

解決方案

慢速對等功能有三個部分：

- 檢測慢速對等體
- 將慢速對等體移動到慢速更新組
- 慢速對等體的恢復（將恢復的對等體移回其原始更新組）

這些過程將在後續章節中進一步詳述。

檢測

慢速對等體功能檢測更新組中的慢速對等體。每個更新組都有一個快取隊列，格式化的BGP更新會在傳輸之前暫時儲存。

以下是此類更新組快取的示例：

```
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
```

快取記憶體的大小是動態計算的，取決於：

- 更新組中的對等體數
- 已安裝的系統記憶體
- 更新組中的對等體型別
- AF的型別

當一個對等體（慢速對等體）無法像其他成員那樣快速確認BGP消息時，等待傳輸的格式化的BGP更新數可以構建在一個更新組中。達到快取限制時，組沒有更多的配額來排隊新郵件。在減少快取之前（直到慢速對等體確認某些消息之前），無法格式化新消息。 這禁止BGP對等體，並且不允許它向群組的較快成員傳送新消息（更新或撤消）。因此，這會減慢更新組中所有對等體的收斂速度。

為了讓慢速對等體功能識別慢速對等體，它引用BGP更新時間戳和對等TCP引數。

預設情況下禁用慢速對等體檢測。若要啟用慢速對等體檢測，請使用以下方法之一：

- 為BGP進程啟用功能（可以從AF/VRF配置）：

  bgp slow-peer detection [threshold

  附註：閾值範圍為120到3,600秒，預設值為300秒。
- 啟用每個對等點的功能：

  neighbor {

- 通過對等策略模板啟用該功能：

  slow-peer detection [threshold < seconds >]

  [no] slow-peer detection

當檢測到慢速對等體時，會顯示類似以下內容的系統日誌消息：

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

您可以輸入以下show命令來檢視慢速對等路由器：

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

以下是使用slow關鍵字時的show命令輸出範例：

```
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
```

如輸出所示，對等10.1.6.7是AF IPv4單播的較慢對等。其他AF未顯示任何慢速對等體。

若要確認偵測計時器目前是否執行及其值，請輸入以下命令：

```
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
```

如示例輸出所示，檢測計時器已啟動。檢測計時器在更新組快取已滿時啟動。

在此範例中，您可以看到偵測到較慢的對等體，但只有在較慢的對等體偵測計時器到期後，它才會從更新群組移出：

```
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
```

慢速對等體識別

如果未啟用慢速對等體檢測功能，則必須手動識別慢速對等體。首先，檢查更新組中對等體的表版本和輸出隊列：

```
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
```

在本例中，驗證對等體的表版本(TblVer)是否曾經趕上主BGP表版本，或者它是否始終落後。第二，檢查一個或多個輸出隊列值非常高的對等體。這些可能是慢速同行。

當您檢視疑似慢速BGP對等點時，請考慮以下問題（在BGP作業階段的兩端）：

- 最後一次寫作*是多久前*
- keepalive啟動了嗎？
- 輸出隊列是否高？
- *SRTT*/*RTO是否高*?
- 重新傳輸的數量是否增加？
- 是否存在任何排隊的重新傳輸資料包？
- TCP發*送視窗*是否非常低或為零？

以下是範例：

```
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
```

移動

本節介紹各種方案中慢速對等項功能的移動過程。

無慢速對等功能的移動

慢速對等體可以手動移動到新的更新組*中*，而無需慢速對等體功能。

在慢速對等體功能可用之前，您需要識別慢速對等體，然後手動將其從更新組中移出。完成此操作後，將對該BGP對等體的出站策略進行更改。此出站策略必須與所使用的任何其他策略不同，因為必須確保慢速對等體不會移動到當前存在的另一個更新組（並將問題移動到該更新組）。 可以應用的最佳更改是不影響實際策略的更改。例如，您可以更改對等點的最小路由通告間隔(MRAI)（在特定AF下）。

以下範例顯示當慢速對等體功能不可用時慢速對等體的手動移動：

```
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
```

靜態慢速對等移動

若要將一個對等體從更新組移動到新的更新組，可以將其配置為*靜態*慢對等體。如果存在多個慢速對等體，則具有相同出站策略的靜態慢速對等體將被置於同一*慢速*更新組中。

為了靜態移動慢速對等體，可以使用以下命令進行配置：

- 啟用每個鄰居或每個對等組的靜態對等移動：

```
[no] neighbor {
```

- 通過對等策略模板啟用靜態對等移動：

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

動態慢速對等移動

預設情況下禁用慢速對等移動。為了啟用慢速對等體移動，可以通過以下方法之一進行配置：

- 為BGP進程啟用慢速對等移動：

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

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

附註：可從*address-family/topology/VRF*檢視中配置此項。

- 啟用每個對等體的慢速對等體移動：

```
neighbor {
```

- 通過對等策略模板啟用慢速對等移動：

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

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

附註：permanent關鍵字表示慢速對等體不會自動恢復。在這種情況下，您可以通過其中一個clear命令，將已恢復慢速對等體移回其原始更新組。

靜態慢速對等體和動態慢速對等體位於同一個慢速對等體更新組中。在此範例中，您可以看到慢速更新群組中的一個慢速對等點：

```
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
```

復原

一旦確認慢速對等體不再為慢速對等體（它趕上），即可在其原始更新組（與出站策略匹配）下重組慢速對等體。 當慢速對等更新組收斂時，恢復計時器啟動。恢復計時器到期時，慢速對等體將移回常規更新組。

附註：若要檢視與偵測/復原計時*器相關的*行為，請輸入debug ip bgp updates events命令。

當較慢的對等體移回原始更新組（這意味著恢復）時，會顯示類似以下內容的系統日誌消息：

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

若要確認復原計時器目前是否執行且值是否執行，請輸入以下命令：

```
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，值為16秒的recovery timer表示可能較慢的對等體可能會在16秒後移回其原始更新組。

在此範例中，您可以看到已從慢速對等體狀態中復原的對等體：

```
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
```

清除慢速對等體狀態

可以使用以下命令手動清除慢速對等體狀態：

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

附註：使用這些命令時，請將AF替換為實際地址系列。

使用這些命令後，對等體將移回原始更新組。

輸入show ip bgp internal命令以檢視慢速對等體偵測和移動設定：

```
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
```

附註：總而言之，BGP慢速對等體是一項功能，可檢測BGP更新組中的慢速對等體，並允許隨著慢速對等體從更新組中移出而實現更快的BGP收斂。