iBGP PE-CE功能的IOS实施

目录

简介

本文档介绍如何在Cisco IOS®中实现提供商边缘(PE)和客户边缘(CE)之间的内部边界网关协议(iBGP)功能。

背景信息

在新的iBGP PE-CE功能之前，PE和CE(因此在PE路由器上的虚拟路由和转发(VRF)接口上)之间的iBGP不受正式支持。一个例外是多VRF CE(VRF-Lite)设置中VRF接口上的iBGP。部署此功能的动机是：

- 客户希望在VRF的多个站点上拥有一个自治系统编号(ASN)，而无需使用as-override部署外部边界网关协议(eBGP)。
- 客户希望向CE路由器提供内部路由反射，就像服务提供商(SP)核心是一个透明路由反射器(RR)一样。

通过此功能，VRF的站点可以具有与SP核心相同的ASN。但是，如果VRF站点的ASN与SP核心的ASN不同，则使用本地自治系统(AS)功能可以使其显示相同。

实施iBGP PE-CE

要使此功能正常工作，以下是两个主要部分：

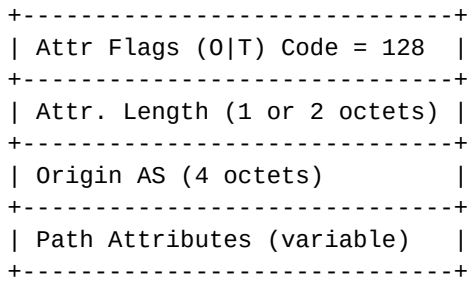
- BGP协议中添加了新属性ATTR_SET，以便以透明方式在SP核心中传输VPN BGP属性。
- 使PE路由器成为指向VRF中CE路由器的iBGP会话的RR，并成为指向VPNv4邻居（其他PE路由器或RR）的RR。

新的ATTR_SET属性允许SP以透明方式传送客户的所有BGP属性，并且不干扰SP属性和BGP策略。此类属性包括集群列表、本地首选项、社区等。

BGP客户路由属性

ATTR_SET是用于传送SP客户的VPN BGP属性的新BGP属性。它是可选的可传递属性。在此属性中，可以携带BGP更新消息（MP_REACH和MP_UNREACH属性除外）中的所有客户BGP属性。

ATTR_SET属性具有以下格式：



属性标志是常规BGP属性标志（请参阅RFC 4271）。 属性长度指示属性长度是一个二进制八位数还是两个二进制八位数。Origin AS字段的目的是防止源自一个AS的一个路由泄漏到另一个AS，而无需正确操作AS_PATH。可变长度路径属性字段传送VPN BGP属性，这些属性必须在SP核心中传送。

在出口PE路由器上，VPN BGP属性被推送到此属性中。在入口PE路由器上，在BGP前缀发送到CE路由器之前，这些属性会从属性中弹出。此属性提供SP网络和客户VPN之间的BGP属性隔离，反之亦然。例如，SP路由反射集群列表属性在VPN网络中未被看到和考虑。但是，VPN路由反射集群列表属性在SP网络中未被看到和考虑。

查看图1，查看客户BGP前缀在SP网络中的传播。

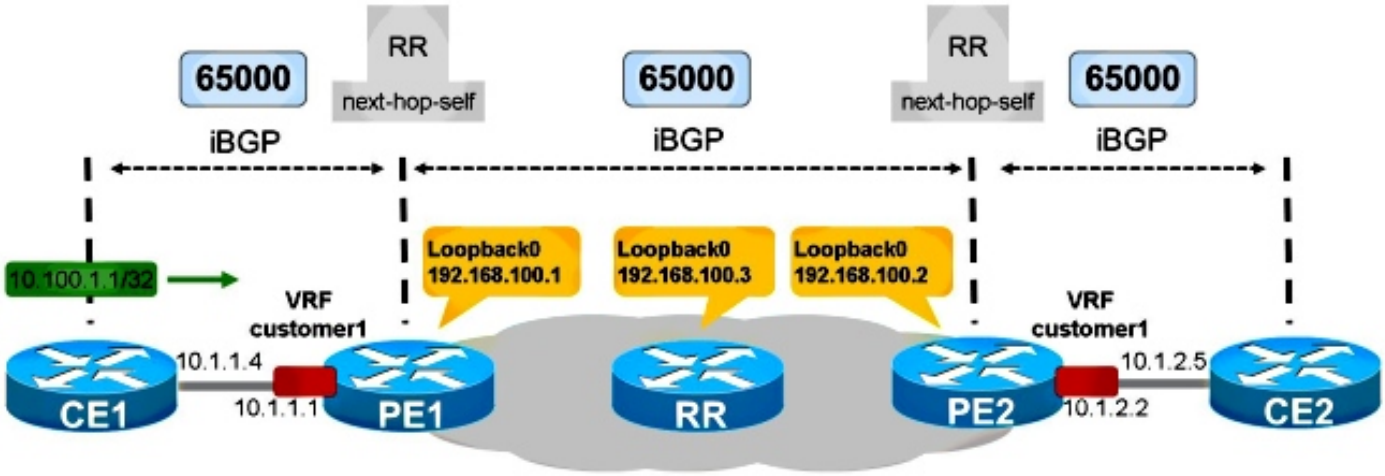


图 1

CE1和CE2与SP网络位于同一AS中：65000. PE1为CE1配置了iBGP。PE1反映前缀10.100.1.1/32通往SP网络中RR的路径。RR像往常一样反映指向PE路由器的iBGP路径。PE2反映通向CE2的路径。

要使此功能正常工作，您必须：

- 在具有iBGP PE-CE功能支持的PE1和PE2上有代码
- 配置PE1和PE2，以便对指向各自CE路由器的BGP会话执行路由反射
- 在PE路由器上为指向其CE路由器的BGP会话提供下一跳自身
- 确保每个VPN站点使用不同的路由区分器(RD)

配置

请参见图 1。

以下是PE1和PE2所需的配置：

```
PE1

vrf definition customer1
rd 65000:1
route-target export 1:1
route-target import 1:1
!
address-family ipv4
exit-address-family

router bgp 65000
bgp log-neighbor-changes
neighbor 192.168.100.3 remote-as 65000
neighbor 192.168.100.3 update-source Loopback0
!
address-family vpnv4
```

```
 neighbor 192.168.100.3 activate
 neighbor 192.168.100.3 send-community extended
exit-address-family
!
address-family ipv4 vrf customer1
 neighbor 10.1.1.4 remote-as 65000
 neighbor 10.1.1.4 activate
 neighbor 10.1.1.4 internal-vpn-client
 neighbor 10.1.1.4 route-reflector-client
 neighbor 10.1.1.4 next-hop-self
exit-address-family
```

PE2

```
vrf definition customer1
 rd 65000:2
 route-target export 1:1
 route-target import 1:1
 !
 address-family ipv4
 exit-address-family

router bgp 65000
 bgp log-neighbor-changes
 neighbor 192.168.100.3 remote-as 65000
 neighbor 192.168.100.3 update-source Loopback0
 !
 address-family vpnv4
 neighbor 192.168.100.3 activate
 neighbor 192.168.100.3 send-community extended
 exit-address-family
 !
 address-family ipv4 vrf customer1
 neighbor 10.1.2.5 remote-as 65000
 neighbor 10.1.2.5 activate
 neighbor 10.1.2.5 internal-vpn-client
 neighbor 10.1.2.5 route-reflector-client
 neighbor 10.1.2.5 next-hop-self
 exit-address-family
```

注意：如果PE没有CE邻居的**neighbor <internal-CE> internal-vpn-client**命令，它不会将前缀从CE传播到SP RRs/PE路由器。

注意：如果PE不是VRF中的RR，则它不会将前缀从RR/PE路由器传播到CE路由器。

新命令

有一个新命令neighbor <internal-CE> internal-vpn-client，可使此功能正常工作。必须在PE路由器上仅为指向CE路由器的iBGP会话配置它。

注意：iBGP PE-CE Multi-VRF CE(VRF-Lite)功能仍受支持，无需**neighbor <internal-CE> internal-vpn-client**命令。

注意：配置**neighbor <internal-CE> internal-vpn-client** 命令后，**neighbor <internal-CE> route-reflector-client**和**neighbor <internal-CE> next-hop-self**命令也会自动放入配置中。当删除**neighbor <internal-CE> route-reflector-client**和**neighbor <internal-CE> next-hop-self**命令（或

两者）之一并执行重新加载时，这些命令将自动重新放入配置中。

ATTR_SET的详细介绍

请参见图 1。

这是CE1通告的前缀：

```
CE1#show bgp ipv4 unicast 10.100.1.1/32
BGP routing table entry for 10.100.1.1/32, version 2
Paths: (1 available, best #1, table default)
 Advertised to update-groups:
    4
 Refresh Epoch 1
 Local
   0.0.0.0 from 0.0.0.0 (10.100.1.1)
     Origin IGP, metric 0, localpref 100, weight 32768, valid, sourced, local, best
     rx pathid: 0, tx pathid: 0x0
```

当PE1从CE1收到BGP前缀10.100.1.1/32时，它将其存储两次：

```
PE1#show bgp vpnv4 unicast all 10.100.1.1/32
BGP routing table entry for 65000:1:10.100.1.1/32, version 21
Paths: (2 available, best #1, table customer1)
 Advertised to update-groups:
    5
 Refresh Epoch 1
 Local, (Received from ibgp-pece RR-client)
   10.1.1.4 (via vrf customer1) from 10.1.1.4 (10.100.1.1)
     Origin IGP, metric 0, localpref 200, valid, internal, best
     mpls labels in/out 18/nolabel
     rx pathid: 0, tx pathid: 0x0
 Refresh Epoch 1
 Local, (Received from ibgp-pece RR-client), (ibgp sourced)
   10.1.1.4 (via vrf customer1) from 10.1.1.4 (10.100.1.1)
     Origin IGP, localpref 100, valid, internal
     Extended Community: RT:1:1
     mpls labels in/out 18/nolabel
     rx pathid: 0, tx pathid: 0
```

第一条路径是PE1上的实际路径，因为它是从CE1接收的。

第二条路径是通告给RR/PE路由器的路径。上面标有源自ibgp的。它包含ATTR_SET属性。请注意，此路径附加了一个或多个路由目标(RT)。

PE1通告前缀，如下所示：

```
PE1#show bgp vpnv4 unicast all neighbors 192.168.100.3 advertised-routes
BGP table version is 7, local router ID is 192.168.100.1
Status codes: s suppressed, d damped, h history, * valid, > best, i - internal,
              r RIB-failure, S Stale, m multipath, b backup-path, f RT-Filter,
              x best-external, a additional-path, c RIB-compressed,
Origin codes: i - IGP, e - EGP, ? - incomplete
RPKI validation codes: V valid, I invalid, N Not found

    Network          Next Hop            Metric LocPrf Weight Path
```

```
Route Distinguisher: 65000:1 (default for vrf customer1)
 *>i 10.100.1.1/32    10.1.1.4                 0    200      0 i

Total number of prefixes 1
```

RR通过以下方式查看路径：

```
RR#show bgp vpnv4 un all 10.100.1.1/32
BGP routing table entry for 65000:1:10.100.1.1/32, version 10
Paths: (1 available, best #1, no table)
 Advertised to update-groups:
    3
 Refresh Epoch 1
 Local, (Received from a RR-client)
   192.168.100.1 (metric 11) (via default) from 192.168.100.1 (192.168.100.1)
     Origin IGP, localpref 100, valid, internal, best
     Extended Community: RT:1:1
     Originator: 10.100.1.1, Cluster list: 192.168.100.1
      ATTR_SET Attribute:
       Originator AS 65000
       Origin IGP
       Aspath
       Med 0
       LocalPref 200
       Cluster list
       192.168.100.1,
       Originator 10.100.1.1
     mpls labels in/out nolabel/18
     rx pathid: 0, tx pathid: 0x0
```

请注意，此VPNv4单播前缀在核心中的本地优先级是100。在ATTR_SET中，存储原始本地优先级200。但是，这对SP核心中的RR是透明的。

在PE2上，您会看到如下所示的前缀：

```
PE2#show bgp vpnv4 unicast all 10.100.1.1/32
BGP routing table entry for 65000:1:10.100.1.1/32, version 5
Paths: (1 available, best #1, no table)
 Not advertised to any peer
 Refresh Epoch 2
 Local
   192.168.100.1 (metric 21) (via default) from 192.168.100.3 (192.168.100.3)
     Origin IGP, localpref 100, valid, internal, best
     Extended Community: RT:1:1
     Originator: 10.100.1.1, Cluster list: 192.168.100.3, 192.168.100.1
     ATTR_SET Attribute:
       Originator AS 65000
       Origin IGP
       Aspath
       Med 0
       LocalPref 200
       Cluster list
       192.168.100.1,
       Originator 10.100.1.1
     mpls labels in/out nolabel/18
     rx pathid: 0, tx pathid: 0x0
BGP routing table entry for 65000:2:10.100.1.1/32, version 6
Paths: (1 available, best #1, table customer1)
 Advertised to update-groups:
    1
 Refresh Epoch 2
 Local, imported path from 65000:1:10.100.1.1/32 (global)
```

```
   192.168.100.1 (metric 21) (via default) from 192.168.100.3 (192.168.100.3)
     Origin IGP, metric 0, localpref 200, valid, internal, best
     Originator AS(ibgp-pece): 65000
     Originator: 10.100.1.1, Cluster list: 192.168.100.1
     mpls labels in/out nolabel/18
     rx pathid:0, tx pathid: 0x0
```

第一条路径是从RR(ATTR_SET)接收的路径。请注意，RD是65000:1，即源RD。第二条路径是从RD 65000:1的VRF表导入的路径。ATTR_SET已删除。

以下是CE2上看到的路径：

```
CE2#show bgp ipv4 unicast 10.100.1.1/32
BGP routing table entry for 10.100.1.1/32, version 10
Paths: (1 available, best #1, table default)
 Not advertised to any peer
 Refresh Epoch 1
 Local
   10.1.2.2 from 10.1.2.2 (192.168.100.2)
     Origin IGP, metric 0, localpref 200, valid, internal, best
     Originator: 10.100.1.1, Cluster list: 192.168.100.2, 192.168.100.1
     rx pathid: 0, tx pathid: 0x0
```

注意下一跳是10.1.2.2，即PE2。集群列表包含路由器PE1和PE2。这些路由器是VPN内部的重要路由器。SP RR(10.100.1.3)不在群集列表中。

SP网络中的VPN中保留了200的本地优先级。

debug bgp vpnv4 unicast updates命令显示在SP网络中传播的更新：

```
PE1#
BGP(4): Revise route installing 1 of 1 routes for 10.100.1.1/32 -> 10.1.1.4
(customer1) to customer1 IP table
BGP(4): 192.168.100.3 NEXT_HOP changed SELF for ibgp rr-client pe-ce net
65000:1:10.100.1.1/32,
BGP(4): 192.168.100.3 Net 65000:1:10.100.1.1/32 from ibgp-pece 10.1.1.4 format
ATTR_SET
BGP(4): (base) 192.168.100.3 send UPDATE (format) 65000:1:10.100.1.1/32, next
192.168.100.1, label 16, metric 0, path Local, extended community RT:1:1
BGP: 192.168.100.3 Next hop is our own address 192.168.100.1
BGP: 192.168.100.3 Route Reflector cluster loop; Received cluster-id 192.168.100.1
BGP: 192.168.100.3 RR in same cluster. Reflected update dropped

RR#
BGP(4): 192.168.100.1 rcvd UPDATE w/ attr: nexthop 192.168.100.1, origin i, localpref
100, originator 10.100.1.1, clusterlist 192.168.100.1, extended community RT:1:1,
[ATTR_SET attribute:  originator AS 65000, origin IGP, aspath , med 0, localpref 200,
cluster list 192.168.100.1 , originator 10.100.1.1]
BGP(4): 192.168.100.1 rcvd 65000:1:10.100.1.1/32, label 16
RT address family is not configured. Can't create RTC route
BGP(4): (base) 192.168.100.1 send UPDATE (format) 65000:1:10.100.1.1/32, next
192.168.100.1, label 16, metric 0, path Local, extended community RT:1:1

PE2#
BGP(4): 192.168.100.3 rcvd UPDATE w/ attr: nexthop 192.168.100.1, origin i, localpref
100, originator 10.100.1.1, clusterlist 192.168.100.3 192.168.100.1, extended community
RT:1:1, [ATTR_SET attribute:  originator AS 65000, origin IGP, aspath , med 0, localpref
200, cluster list 192.168.100.1 , originator 10.100.1.1]
BGP(4): 192.168.100.3 rcvd 65000:1:10.100.1.1/32, label 16
RT address family is not configured. Can't create RTC route
BGP(4): Revise route installing 1 of 1 routes for 10.100.1.1/32 -> 192.168.100.1
```

```
(customer1) to customer1 IP table
BGP(4): 10.1.2.5 NEXT_HOP is set to self for net 65000:2:10.100.1.1/32,
```

注意：PE1从RR收到自己的更新，然后将其丢弃。这是因为PE1和PE2在RR上处于同一更新组中。

注意：如果要转储完整的十六进制更新消息，请使用**detail关**键字进行**debug BGP updates**命令。

```
PE2#   debug bgp vpnv4 unicast updates detail
BGP updates debugging is on with detail for address family: VPNv4 Unicast

PE2#
BGP(4): 192.168.100.3 rcvd UPDATE w/ attr: nexthop 192.168.100.1, origin i,
localpref 100, originator 10.100.1.1, clusterlist 192.168.100.3 192.168.100.1,
extended community RT:1:1, [ATTR_SET attribute:  originator AS 65000, origin IGP,
aspath , med 0, localpref 200, cluster list 192.168.100.1 , originator 10.100.1.1]
BGP(4): 192.168.100.3 rcvd 65000:1:10.100.1.1/32, label 17
RT address family is not configured. Can't create RTC route
BGP: 192.168.100.3 rcv update length 125
BGP: 192.168.100.3 rcv update dump: FFFF FFFF FFFF FFFF FFFF FFFF FFFF FFFF
0090 0200 00
PE2#00 7980 0E21 0001 800C 0000 0000 0000 0000 C0A8 6401 0078 0001 1100 00FD E800
0000 010A 6401 0140 0101 0040 0200 4005 0400 0000 64C0 1008 0002 0001 0000 0001 800A
08C0 A864 03C0 A864 0180 0904 0A64 0101 C080 2700 00FD E840 0101 0040 0200 8004 0400
0000 0040 0504 0000 00C8 800A 04C0 A864 0180 0904 0A64 0101
BGP(4): Revise route installing 1 of 1 routes for 10.100.1.1/32 -> 192.168.100.1
(customer1) to customer1 IP table
BGP(4): 10.1.2.5 NEXT_HOP is set to self for net 65000:2:10.100.1.1/32,
```

下一跳处理

必须在PE路由器上配置Next-hop-self以实现此功能。原因是下一跳通常使用iBGP传输不变。但是，这里有两个独立的网络：VPN网络和SP网络，它们运行单独的内部网关协议(IGP)。 因此，IGP度量无法轻松进行比较并用于两个网络之间的最佳路径计算。RFC 6368选择的方法是对指向CE的iBGP会话强制使用next-hop-self，这一方法将共同避免之前描述的问题。优点是VRF站点可以使用此方法运行不同的IGP。

RD

RFC 6368提到，建议同一VPN的不同VRF站点使用不同（唯一）的RD。在Cisco IOS中，此功能是必需的。

iBGP PE-CE功能与Local-AS

请参阅图2。VPN customer1有ASN 65001。

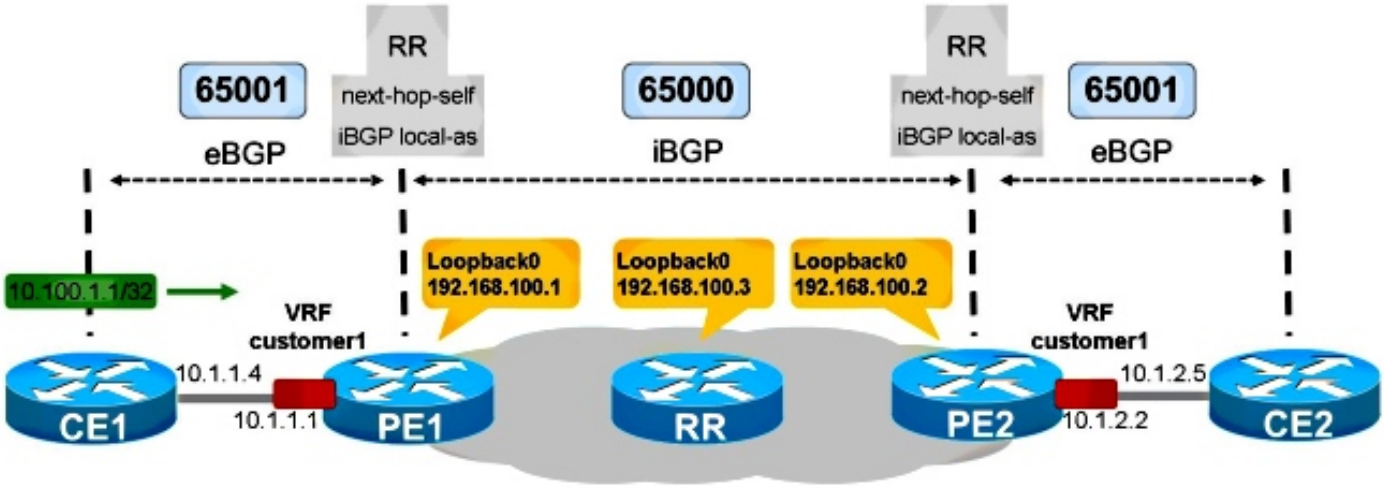


图 2

CE1在AS 65001中。为了从PE1的角度使此内部BGP，它需要iBGP local-as功能。

```
CE1

router bgp 65001
 bgp log-neighbor-changes
 network 10.100.1.1 mask 255.255.255.255
 neighbor 10.1.1.1 remote-as 65001

PE1

router bgp 65000
 bgp log-neighbor-changes
 neighbor 192.168.100.3 remote-as 65000
 neighbor 192.168.100.3 update-source Loopback0
 !
 address-family vpnv4
 neighbor 192.168.100.3 activate
 neighbor 192.168.100.3 send-community extended
 exit-address-family
 !
 address-family ipv4 vrf customer1
 neighbor 10.1.1.4 remote-as 65001
 neighbor 10.1.1.4 local-as 65001
 neighbor 10.1.1.4 activate
 neighbor 10.1.1.4 internal-vpn-client
 neighbor 10.1.1.4 route-reflector-client
 neighbor 10.1.1.4 next-hop-self
 exit-address-family
```

PE2和CE2的配置类似。

PE1看到BGP前缀，如下所示：

```
PE1#show bgp vpnv4 unicast all 10.100.1.1/32
BGP routing table entry for 65000:1:10.100.1.1/32, version 41
Paths: (2 available, best #1, table customer1)
 Advertised to update-groups:
    5
 Refresh Epoch 1
 Local, (Received from ibgp-pece RR-client)
```

```
   10.1.1.4 (via vrf customer1) from 10.1.1.4 (10.100.1.1)
     Origin IGP, metric 0, localpref 200, valid, internal, best
     mpls labels in/out 18/nolabel
     rx pathid: 0, tx pathid: 0x0
 Refresh Epoch 1
 Local, (Received from ibgp-pece RR-client), (ibgp sourced)
   10.1.1.4 (via vrf customer1) from 10.1.1.4 (10.100.1.1)
     Origin IGP, localpref 100, valid, internal
     Extended Community: RT:1:1
     mpls labels in/out 18/nolabel
     rx pathid: 0, tx pathid: 0
```

前缀是内部BGP。

PE2看到以下内容：

```
PE2#show bgp vpnv4 unicast all 10.100.1.1/32
BGP routing table entry for 65000:1:10.100.1.1/32, version 33
Paths: (1 available, best #1, no table)
 Not advertised to any peer
 Refresh Epoch 5
 Local
   192.168.100.1 (metric 21) (via default) from 192.168.100.3 (192.168.100.3)
     Origin IGP, localpref 100, valid, internal, best
     Extended Community: RT:1:1
     Originator: 10.100.1.1, Cluster list: 192.168.100.3, 192.168.100.1
     ATTR_SET Attribute:
       Originator AS 65001
       Origin IGP
       Aspath
       Med 0
       LocalPref 200
       Cluster list
       192.168.100.1,
       Originator 10.100.1.1
     mpls labels in/out nolabel/18
     rx pathid: 0, tx pathid: 0x0
BGP routing table entry for 65000:2:10.100.1.1/32, version 34
Paths: (1 available, best #1, table customer1)
 Advertised to update-groups:
    5
 Refresh Epoch 2
 Local, imported path from 65000:1:10.100.1.1/32 (global)
   192.168.100.1 (metric 21) (via default) from 192.168.100.3 (192.168.100.3)
     Origin IGP, metric 0, localpref 200, valid, internal, best
     Originator AS(ibgp-pece): 65001
     Originator: 10.100.1.1, Cluster list: 192.168.100.1
     mpls labels in/out nolabel/18
     rx pathid: 0, tx pathid: 0x0
```

发起方AS是65001，是从PE2向CE2发送前缀时使用的AS。因此，AS将保留，本例中的本地首选项也是如此。

```
CE2#show bgp ipv4 unicast 10.100.1.1/32
BGP routing table entry for 10.100.1.1/32, version 3
Paths: (1 available, best #1, table default)
 Not advertised to any peer
 Refresh Epoch 1
 Local
   10.1.2.2 from 10.1.2.2 (192.168.100.2)
     Origin IGP, metric 0, localpref 200, valid, internal, best
     Originator: 10.100.1.1, Cluster list: 192.168.100.2, 192.168.100.1
```

```
    rx pathid: 0, tx pathid: 0x0
```

您将看**到Local**而不是AS路径。这意味着它是源自AS 65001的内部BGP路由，也是路由器CE2的已配置ASN。所有BGP属性都取自ATTR_SET属性。这符合下一节中案例1的规则。

不同VRF站点之间路由交换的规则

ATTR_SET包含源VRF的发起方AS。当远程PE在将前缀发送到CE路由器之前删除ATTR_SET时，会检查此始发AS。

第 1 种情况：如果始发AS与CE路由器的已配置AS匹配，则当PE将路径导入目标VRF时，BGP属性从ATTR_SET属性中获取。

第 2 种情况：如果始发AS与CE路由器的已配置AS不匹配，则构建路径的属性集如下所示：

1. 路径属性设置为ATTR_SET属性中包含的属性。

2. iBGP特定属性被丢弃（LOCAL_PREF、ORIGINATOR和CLUSTER_LIST）。

3. ATTR_SET属性中包含的**源AS**编号在AS_PATH前面，并遵循应用于源和目标AS之间外部BGP对等的规则。

4. 如果与VRF关联的自治系统与VPN提供商自治系统相同，且VPN路由的AS_PATH属性不为空，则VRF路由的AS_PATH属性应预置到该属性之前。

   请参阅图3。CE1和PE1具有AS 65000，并配置了iBGP PE-CE功能。CE2具有ASN 65001。这意味着PE2和CE2之间有eBGP。

图 3

PE2会看到如下路由：

```
PE2#show bgp vpnv4 unicast all 10.100.1.1/32
BGP routing table entry for 65000:1:10.100.1.1/32, version 43
Paths: (1 available, best #1, no table)
 Not advertised to any peer
 Refresh Epoch 6
 Local
   192.168.100.1 (metric 21) (via default) from 192.168.100.3 (192.168.100.3)
     Origin IGP, localpref 100, valid, internal, best
     Extended Community: RT:1:1
     Originator: 10.100.1.1, Cluster list: 192.168.100.3, 192.168.100.1
     ATTR_SET Attribute:
       Originator AS 65000
       Origin IGP
       Aspath
       Med 0
       LocalPref 200
       Cluster list
       192.168.100.1,
       Originator 10.100.1.1
```

```
     mpls labels in/out nolabel/17
     rx pathid: 0, tx pathid: 0x0
BGP routing table entry for 65000:2:10.100.1.1/32, version 44
Paths: (1 available, best #1, table customer1)
 Advertised to update-groups:
    6
 Refresh Epoch 6
 Local, imported path from 65000:1:10.100.1.1/32 (global)
   192.168.100.1 (metric 21) (via default) from 192.168.100.3 (192.168.100.3)
     Origin IGP, metric 0, localpref 200, valid, internal, best
     Originator AS(ibgp-pece): 65000
     Originator: 10.100.1.1, Cluster list: 192.168.100.1
     mpls labels in/out nolabel/17
     rx pathid: 0, tx pathid: 0x0
```

这是CE2上显示的前缀：

```
CE2#show bgp ipv4 unicast 10.100.1.1/32
BGP routing table entry for 10.100.1.1/32, version 5
Paths: (1 available, best #1, table default)
 Not advertised to any peer
 Refresh Epoch 1
 65000
   10.1.2.2 from 10.1.2.2 (192.168.100.2)
     Origin IGP, localpref 100, valid, external, best
     rx pathid: 0, tx pathid: 0x0
```

这是案例2。ATTR_SET属性中包含的**源AS** 编号由PE2预置到AS_PATH前面，并遵循适用于源和目标AS之间eBGP对等的规则。当PE2创建要通告给CE2的路由时，PE2会忽略iBGP特定属性。因此，本地首选项为100，而不是200（如ATTR_SET属性所示）。

CE-to-CE VRF-Lite反射

请参阅图 4。

图 4

图4显示了连接到PE1的另一台CE路由器CE3。CE1和CE3都连接到同一VRF实例上的PE1:customer1。这意味着CE1和CE3是PE1的多VRF CE路由器（也称为VRF-Lite）。PE1在将前缀从CE1通告给CE3时将自身作为下一跳。如果不需要此行为，可以配置**neighbor 10.1.3.6 next-hop-unchanged**。要配置此配置，必须删除PE1上的**neighbor 10.1.3.6 next-hop-self。然后，CE3发现来自CE1且CE1的路由将成为这些BGP前缀的下一跳。**为了实现此目的，您需要CE3的路由表中这些BGP下一跳的路由。您需要动态路由协议(IGP)或CE1、PE1和CE3上的静态路由，以确保路由器为彼此的下一跳IP地址提供路由。但是，此配置存在问题。

PE1上的配置是：

```
router bgp 65000
 !
 address-family ipv4 vrf customer1
 neighbor 10.1.1.4 remote-as 65000
 neighbor 10.1.1.4 activate
 neighbor 10.1.1.4 internal-vpn-client
 neighbor 10.1.1.4 route-reflector-client
 neighbor 10.1.1.4 next-hop-self
 neighbor 10.1.3.6 remote-as 65000
 neighbor 10.1.3.6 activate
```

```
 neighbor 10.1.3.6 internal-vpn-client
 neighbor 10.1.3.6 route-reflector-client
 neighbor 10.1.3.6 next-hop-unchanged
 exit-address-family
```

在CE3上，CE1的前缀可以正常显示：

```
CE3#show bgp ipv4 unicast 10.100.1.1
BGP routing table entry for 10.100.1.1/32, version 9
Paths: (1 available, best #1, table default)
 Not advertised to any peer
 Refresh Epoch 1
 Local
   10.1.1.4 from 10.1.3.1 (192.168.100.1)
     Origin IGP, metric 0, localpref 200, valid, internal, best
     Originator: 10.100.1.1, Cluster list: 192.168.100.1
     rx pathid: 0, tx pathid: 0x0
```

但是，CE3上会看到来自CE2的前缀，如下所示：

```
CE3#show bgp ipv4 unicast 10.100.1.2
BGP routing table entry for 10.100.1.2/32, version 0
Paths: (1 available, no best path)
 Not advertised to any peer
 Refresh Epoch 1
 Local
   192.168.100.2 (inaccessible) from 10.1.3.1 (192.168.100.1)
     Origin IGP, metric 0, localpref 100, valid, internal
     Originator: 10.100.1.2, Cluster list: 192.168.100.1, 192.168.100.2
     rx pathid: 0, tx pathid: 0
```

BGP下一跳是192.168.100.2，即PE2的环回IP地址。PE1在将前缀10.100.1.2/32通告给CE3时没有将BGP下一跳重写到自己。这使CE3上的此前缀不可用。

因此，如果iBGP PE-CE功能在MPLS-VPN和iBGP VRF-Lite中混合使用，则必须确保在PE路由器上始终具有next-hop-self。

当PE路由器是RR时，您无法保留下一跳，该RR反映PE上本地VRF接口上从一个CE到另一个CE的iBGP路由。在MPLS VPN网络中运行iBGP PE-CE时，必须对指向CE路由器的iBGP会话**使用internal-vpn-client。**当PE路由器上的VRF中有多个本地CE时，必须为这些BGP对**等体保留**next-hop-self。

您可以查看路由映射，以便将从其他PE路由器接收的前缀的下一跳设置为自己，但不能将来自其他本地连接CE路由器的反射前缀设置为自己。但是，当前不支持在出站路由映射中将下一跳设置为自己。该配置如下所示：

```
router bgp 65000

 address-family ipv4 vrf customer1
 neighbor 10.1.1.4 remote-as 65000
 neighbor 10.1.1.4 activate
 neighbor 10.1.1.4 internal-vpn-client
 neighbor 10.1.1.4 route-reflector-client
 neighbor 10.1.1.4 next-hop-self
 neighbor 10.1.3.6 remote-as 65000
 neighbor 10.1.3.6 activate
 neighbor 10.1.3.6 internal-vpn-client
 neighbor 10.1.3.6 route-reflector-client
 neighbor 10.1.3.6 route-map NH-setting out
```

```
 exit-address-family

ip prefix-list PE-loopbacks seq 10 permit 192.168.100.0/24 ge 32
!

route-map NH-setting permit 10
 description set next-hop to self for prefixes from other PE routers
 match ip route-source prefix-list PE-loopbacks
 set ip next-hop self
!

route-map NH-setting permit 20
 description advertise prefixes with next-hop other than the prefix-list in
route-map entry 10 above
!
```

但是，这不受支持：

```
PE1(config)#route-map NH-setting permit 10
PE1(config-route-map)# set ip next-hop self
% "NH-setting" used as BGP outbound route-map, set use own IP/IPv6 address for the nexthop not
supported
```

PE路由器上较旧的Cisco IOS

如果PE1运行缺少iBGP PE-CE功能的旧版Cisco IOS软件，则PE1从不将自己设置为反射的iBGP前缀的下一跳。这意味着从CE1(10.100.1.1)到CE2（通过PE1）的反射BGP前缀(10.100.1.1/32)将CE1(10.1.1.4)用作下一跳。

```
CE3#show bgp ipv4 unicast 10.100.1.1
BGP routing table entry for 10.100.1.1/32, version 32
Paths: (1 available, best #1, table default)
 Not advertised to any peer
 Refresh Epoch 1
 Local
   10.1.1.4 from 10.1.3.1 (192.168.100.1)
     Origin IGP, metric 0, localpref 200, valid, internal, best
     Originator: 10.100.1.1, Cluster list: 192.168.100.1
     rx pathid: 0, tx pathid: 0x0
```

来自CE2(10.100.1.2/32)的前缀被视为下一跳，因为PE1也不对此前缀执行next-hop-self:

```
CE3#show bgp ipv4 unicast 10.100.1.2
BGP routing table entry for 10.100.1.2/32, version 0
Paths: (1 available, no best path)
 Not advertised to any peer
 Refresh Epoch 1
 Local
   192.168.100.2 (inaccessible) from 10.1.3.1 (192.168.100.1)
     Origin IGP, localpref 100, valid, internal
     Originator: 10.100.1.2, Cluster list: 192.168.100.1, 192.168.100.3, 192.168.100.2
     ATTR_SET Attribute:
       Originator AS 65000
       Origin IGP
       Aspath
       Med 0
       LocalPref 100
       Cluster list
       192.168.100.2,
```

```
      Originator 10.100.1.2
    rx pathid: 0, tx pathid: 0
```

要使iBGP PE-CE功能正常工作，启用该功能的VPN的所有PE路由器必须具有支持该功能的代码并启用该功能。

VRF上eBGP的下一跳自己

请参阅图 5。

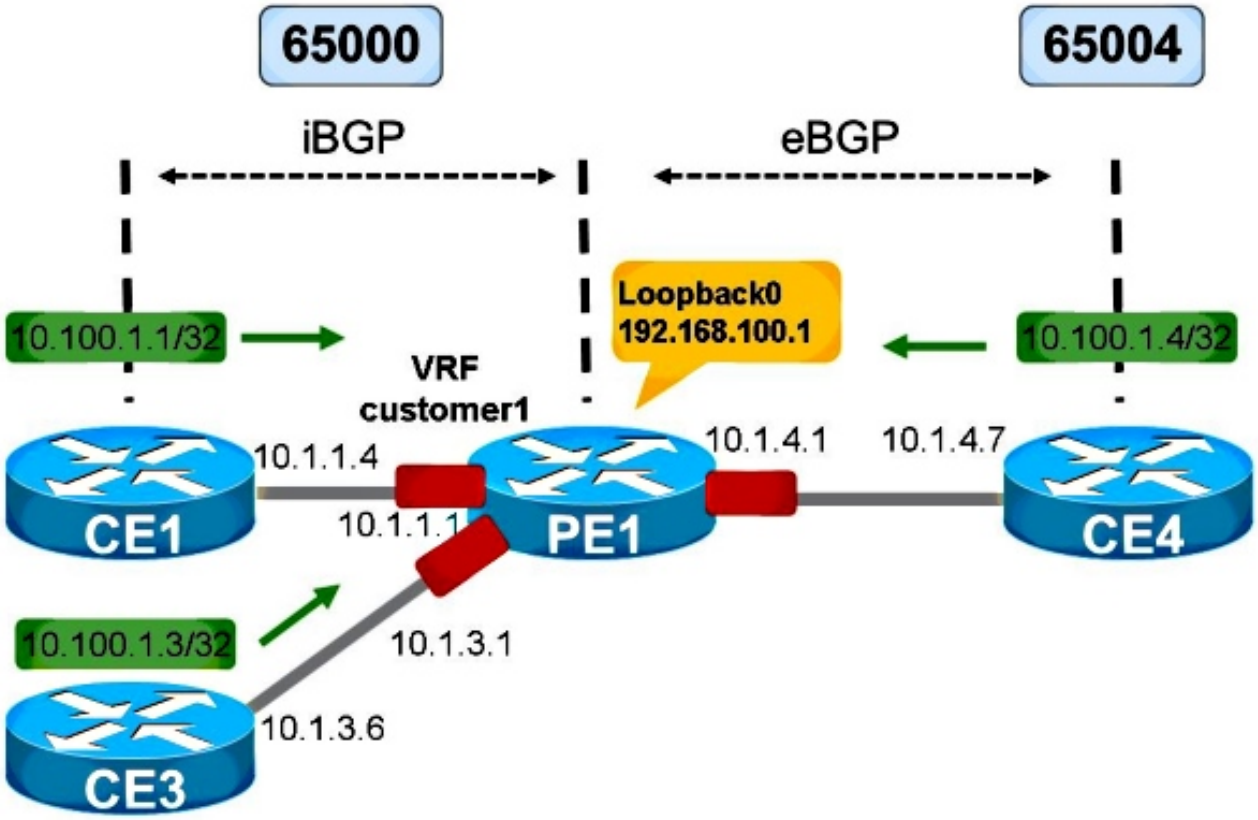


图 5

图5显示了VRF-Lite设置。从PE1到CE4的会话是eBGP。从PE1到CE3的会话仍为iBGP。

对于eBGP前缀，当下一跳向VRF上的iBGP邻居通告前缀时，始终将其设置为自。无论VRF上指向iBGP邻居的会话是否设置了下一跳自身，都是如此。

在图5中，CE3将来自CE4的前缀视为下一跳，而PE1为下一跳。

```
CE3#show bgp ipv4 unicast 10.100.1.4
BGP routing table entry for 10.100.1.4/32, version 103
Paths: (1 available, best #1, table default)
 Not advertised to any peer
 Refresh Epoch 1
 65004
   10.1.3.1 from 10.1.3.1 (192.168.100.1)
     Origin IGP, metric 0, localpref 100, valid, internal, best
     rx pathid: 0, tx pathid: 0x0
```

当PE1上的next-hop-self通向CE3或不通向CE3时，会发生这种情况。

如果PE1上指向CE3和CE4的接口不在VRF中，而是在全局环境中，则指向CE3的下一跳自身会有所不同。

在通往CE3的PE1上没有next-hop-self，您将看到：

```
PE1#show bgp vrf customer1 vpnv4 unicast neighbors 10.1.3.6
BGP neighbor is 10.1.3.6,  vrf customer1,  remote AS 65000, internal link
...
 For address family: VPNv4 Unicast
 Translates address family IPv4 Unicast for VRF customer1
 Session: 10.1.3.6
 BGP table version 1, neighbor version 1/0
 Output queue size : 0
 Index 12, Advertise bit 0
 Route-Reflector Client
 12 update-group member
 Slow-peer detection is disabled
 Slow-peer split-update-group dynamic is disabled
 Interface associated: (none)
```

尽管隐式启用了next-hop-self，但输出并未指明这一点。

在通往CE3的PE1上，使用next-hop-self，您将看到：

```
PE1#show bgp vrf customer1 vpnv4 unicast neighbors 10.1.3.6
BGP neighbor is 10.1.3.6,  vrf customer1,  remote AS 65000, internal link
..
 For address family: VPNv4 Unicast
...
 NEXT_HOP is always this router for eBGP paths
```

但是，如果指向CE3和CE4的接口处于全局情景中，则在未配置next-hop-self时，来自CE4的前缀的下一跳是CE4本身：

```
CE3#show bgp ipv4 unicast 10.100.1.4
BGP routing table entry for 10.100.1.4/32, version 124
Paths: (1 available, best #1, table default)
 Not advertised to any peer
 Refresh Epoch 1
 65004
   10.1.4.7 from 10.1.3.1 (192.168.100.1)
     Origin IGP, metric 0, localpref 100, valid, internal, best
     rx pathid: 0, tx pathid: 0x0
```

对于指向CE3的PE1上的下一跳自身：

```
CE3#show bgp ipv4 unicast 10.100.1.4
BGP routing table entry for 10.100.1.4/32, version 125
Paths: (1 available, best #1, table default)
 Not advertised to any peer
 Refresh Epoch 1
 65004
   10.1.3.1 from 10.1.3.1 (192.168.100.1)
     Origin IGP, metric 0, localpref 100, valid, internal, best
     rx pathid: 0, tx pathid: 0x0
```

这是基于RFC 4364完成的。

如果不想为通过VRF接口指向iBGP会话的eBGP前缀设置next-hop-self，则必须配置next-hop-

unchanged。仅Cisco Bug ID CSCuj11720支持[此功能](#)。

```
router bgp 65000
...
 address-family ipv4 vrf customer1
 neighbor 10.1.1.4 remote-as 65000
 neighbor 10.1.1.4 activate
 neighbor 10.1.1.4 route-reflector-client
 neighbor 10.1.3.6 remote-as 65000
 neighbor 10.1.3.6 activate
 neighbor 10.1.3.6 route-reflector-client
 neighbor 10.1.3.6 next-hop-unchanged
 neighbor 10.1.4.7 remote-as 65004
 neighbor 10.1.4.7 activate
 exit-address-family
```

现在，CE3将CE4视为CE4通告的前缀的下一跳：

```
CE3#show bgp ipv4 unicast 10.100.1.4
BGP routing table entry for 10.100.1.4/32, version 130
Paths: (1 available, best #1, table default)
 Not advertised to any peer
 Refresh Epoch 3
 65004
   10.1.4.7 from 10.1.3.1 (192.168.100.1)
     Origin IGP, metric 0, localpref 100, valid, internal, best
     rx pathid: 0, tx pathid: 0x0
```

如果尝试在Cisco Bug ID **CSCuj11720之前为Cisco IOS代码上指向CE3的iBGP会话配置next-hop-unchanged**关键字，则会遇到以下错误：

```
PE1(config-router-af)# neighbor 10.1.3.6 next-hop-unchanged
%BGP: Can propagate the nexthop only to multi-hop EBGP neighbor
```

在Cisco Bug ID [CSCuj11720](#)之后，**next-hop-unchanged**关键字对多跳eBGP邻居和iBGP VRF-Lite邻居有效。