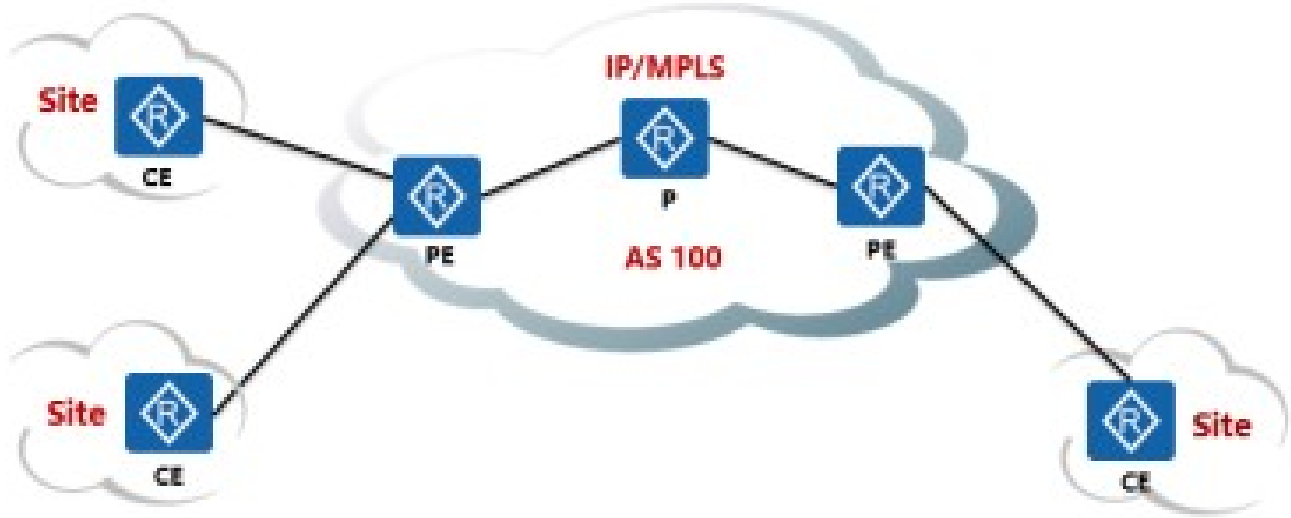
MPLS VPN 跨域

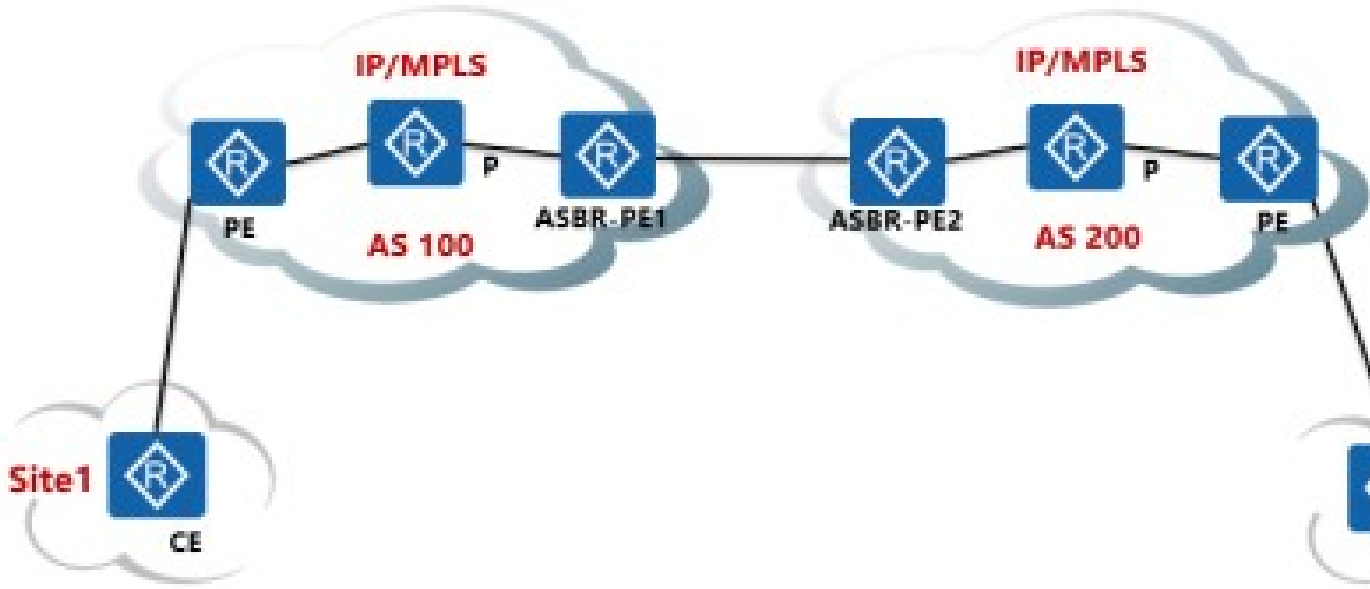
随着 MPLS VPN 解决方案的广泛应用，服务的终端用户的规格和范围也在增长，在一个企业内部的站点数目越来越大，某个地理位置与另外一个服务提供商相连的需求变得非常的普遍，例如国内运营商的不同城域网之间，或相互协作的运营商的骨干网之间都存在着跨越不同自治域的情况。
一般的 MPLS VPN 体系结构都是在一个自治系统 AS（Autonomous System）内运行，任何 VPN 的路由信息都是只能在一个 AS 内按需扩散，没有提供 AS 内的 VPN 信息向其他 AS 扩散的功能。因此，为了支持运营商之间的 VPN 路由信息交换，就需要扩展现有的协议和修改 MPLS VPN 体系框架，提供一个不同于基本的 MPLS VPN 体系结构所提供的互连模型——跨域（Inter-AS）的 MPLS VPN，以便可以穿过运营商间的链路来发布路由前缀和标签信息。

传统 MPLS BGP VPN 的解决方案

一般的 MPLS VPN 体系结构都是在一个自治系统内运行，任何 VPN 的路由信息都是只能在一个自治系统内按需扩散，没有提供自治系统内的 VPN 信息向其他自治系统扩散的功能。



基于 MPLS 的 VPN 可以将私有网络的不同 Site 连接起来，形成一个统一的网络，基于 MPLS 的 VPN 还支持对不同 VPN 间的互通控制。CE（Customer Edge）是用户边缘设备；PE（Provider Edge）是服务商边缘路由器，位于骨干网络；P（Provider），是服务提供商网络中的骨干路由器，不与 CE 直接相连。



如果同一 VPN 的两个站点位于不同的 AS，那么普通的 MPLS BGP VPN 是否还合适于业务的部署？
答案是否定的。因为此时连接 VPN 的 PE 路由器已经无法简单地建立 IBGP 邻居关系了，或是与 RR 建立邻居关系。因此，需要一些手段通过建立 EBGP 邻居关系来传递 VPNv4 路由。

为了支持不同 AS 之间的 VPN 路由信息交换，就需要扩展现有的协议和修改 MPLS VPN 体系框架，提供一个不同于基本的 MPLS VPN 体系结构所提供的互连模型——跨域（Inter-AS）的 MPLS VPN，以便可以穿过 AS 间的链路来发布路由前缀和标签信息。

三种跨域 VPN 解决方案

跨域 VPN-OptionA 方式：
需要跨域的 VPN 在 ASBR 间通过专用的接口管理自己的 VPN 路由，也称为 VRF-to-VRF；
背对背：ASBR 互为对方的 CE ，发布普通的 IPv4 路由，AS 内两层标签，AS 间无标签
优点：配置简单，ASBR 之间不需要运行 MPLS，也不需要为跨域进行特殊配置。
不足：可扩展性差，ASBR 需要管理所有 VPN 路由，为每个 VPN 创建 VPN 实例。
场景：拓扑结构简单，跨域的 VPN 数量比较少的情况

通过 eBGP 邻居关系把 VPNv4 路由转变成普通 IPv4 路由从一个 AS 传递到另一个 AS

跨域 VPN-OptionB 方式：
ASBR 间通过 MP-EBGP 发布 VPN-IPv4 路由，也称为 EBGP redistribution of labeled VPN-IPv4 routes
EBGP 单跳 ，AS 内两层标签，AS 间一层标签
优点：不受 ASBR 之间互连链路数目的限制，采用两层标签方式传递路由，实现和维护都比较简单
不足：VPN 的路由信息是通过 AS 之间的 ASBR 来保存和扩散的，当 VPN 路由较多时，ASBR 负担重
ASBR 设备既要负责 IPv4 报文的路由控制，又要负责 VPNv4 报文的路由控制，还要承担数据转发
任务，因此设备的负担比较重，容易成为网络中性能的瓶颈。
AS 内 PE 只有 VPNv4 邻居，无 BGP 单播邻居

跨域 VPN-OptionC 方式：
PE 或 RR 间通过 Multi-hop MP-EBGP 发布标签 VPN-IPv4 路由，也称为 Multihop EBGP redistribution of labeled VPN-IPv4 routes

EBGP 多跳，方案一：AS 内三层标签，AS 间两层标签　方案二：AS 内两层标签，AS 间两层标签

AS 间有 BGP 单播邻居，要为单播路由分发标签（默认只为 VPNv4 路由发标签）

方案一，ASBR 在收到对端 ASBR 发来的 BGP 标签路由后，需要配置策略产生一个新的标签并发布给 AS 内的 PE 或者 RR 设备，以建立一条完整的 BGP LSP。是非对称结构
BGP 去分发标签
在 ASBR1 和 ASBR2 之间使用 MP-BGP 来分配标签，而 ASBR 和 PE 之间，也使用 MP-BGP 来分配标签。实际上，相当于 LDP 对 MP-BGP 分配的标签进行了封装。
这种实现方式不需要把 BGP 路由引入 IGP，但 ASBR 需要和所有的 PE 建立 BGP 邻居，并和 PE 之间使用 MP-BGP 来分配 ipv4 的路由标签。

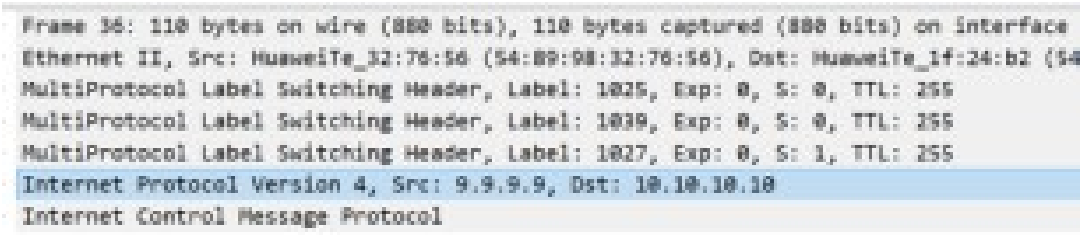
方案二中，ASBR 需要配置 MPLS 触发为 BGP 标签路由分发标签，因此在 AS 内的 PE 或 RR 上可以看到去往对端 PE 或 RR 的 LDP LSP，而非 BGP LSP。该方案需要把 BGP 的路由表引入 IGP，可能会使得 IGP 过大
BGP 引入 IGP ,由 LDP 去分发标签

优点：ASBR 负担轻，ASBR 上不保存也不通告 VPN-IPv4 路由，中间域设备可以不支持 MPLS VPN，更适合在跨越多个 AS 时使用，适应于大型环境。
不足：维护一条端到端的 PE 或 RR 连接管理代价较大。
需要在域之间互相通告环回接口的路由，造成所谓的路由泄漏问题。同时由于需要建立跨域的 LSP，在管理上带来较大的麻烦。另外，由于跨域同一 VPN 的所有 PE 之间都要建立 eBGP 连接，存在严

重的扩展性问题。

三层标签：
VPN label（私网）
BGP LSP（RR 到 ASBR，用于在两 PE 或 RR 间交换数据）
Tunnel LSP LDP 协议分配（公网）

底层标签是由对端 PE 分配的与 VPN 路由相关联的 VPN 标签，
中间的标签是 ASBR 分配的与去往对端 PE 的路由相关联的标签，
外层标签则是与去往下一跳 ASBR 的路由相关联的标签。



问题一：本端的 PE，无法为对端 AS 内的 PE 的 vpn 路由分配标签。不同 AS 往往属于不同运营商，两个运营商之间的 vpn 独立管理，通常无法互通。两个运营商之间的 PE 就更无法交互 vpn 路由了。

问题二：端到端的外层标签如何建立。运营商之间通常不进行 IGP 的互通，即便运行 LDP，也无法分标签；所以普通 L3VPN 的外层标签交换在 ASBR 的位置就会被中断。

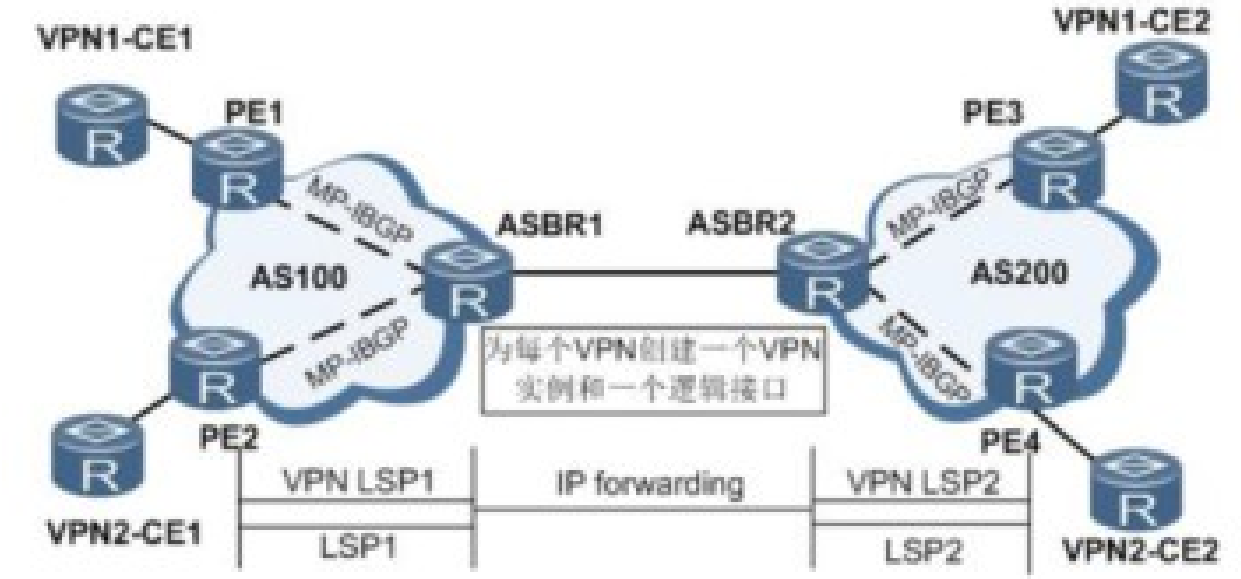
Option C 方式的思想就是在跨域的情况下，PE 之间仍然可以像域内那样，在 PE 和 PE 之间建立直接的 BGP 邻居，交换 VPNv4 路由信息，这样就不需要中间设备再保存、维护和扩散 VPN 路由信息。唯一不同的是域内是 MP-iBGP 邻居，跨域是 MP-eBGP 邻居。

方案二
1 在 R1、R2、R3 上配置 IGP（例如 ospf），使之互通（R3 和 R4 之间的链路不在 IGP 中发布）；
2 在 R1、R3、R4、R6 上配置 BGP，R1 和 R3，R4 和 R6 分别建立 iBGP 邻居，R3 和 R4 建立 eBGP 邻居；
3 在 R3 上发布 R1 的 loopback 地址至 BGP 进程中，在 R4 上发布 R6 的 loopback 地址至 BGP 进程，使得 R1 和 R6 可以互通；
4 在 R1、R6 上配置 vpn-instance；R1 和 R6 建立 vpnv4 的 EBGP 邻居；
5 使 ASBR 之间开始使用 BGP 交互 ipv4 的路由。
6 在 R3 和 R4 上使能 LDP 对标签路由分配标签，使得标签能够衔接起来。
7 在 ASBR（R3、R4）上，把 BGP 路由引入 IGP。
该实现方式的缺陷：需要把 BGP 的路由表引入 IGP，可能会使得 IGP 过大。

=========

OptionA

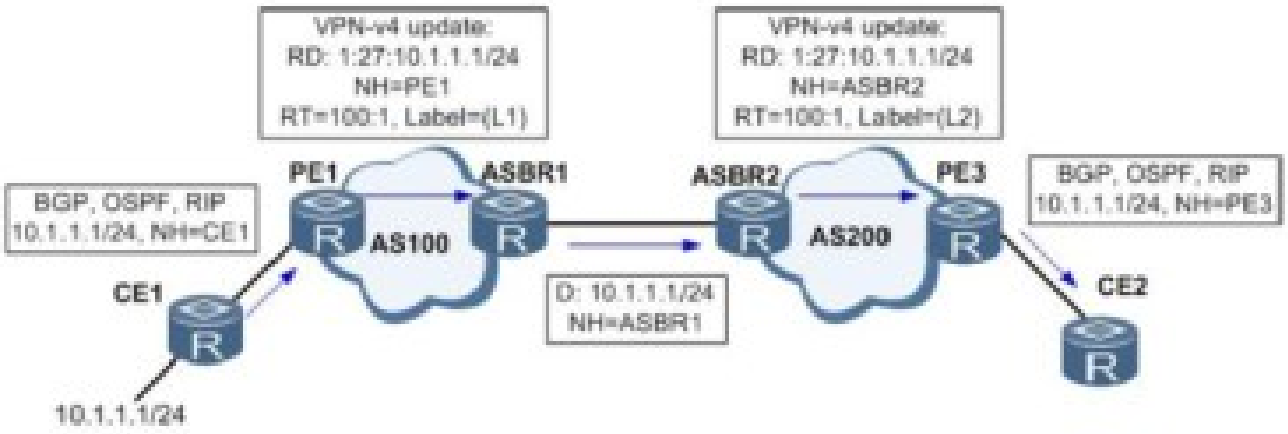
跨域 VPN-OptionA 是基本 BGP/MPLS IP VPN 在跨域环境下的应用，ASBR 之间不需要运行 MPLS，也不需要为跨域进行特殊配置。这种方式下，两个 AS 的边界 ASBR 直接相连，ASBR 同时也是各自所在自治系统的 PE。两个 ASBR 都把对端 ASBR 看作自己的 CE 设备，将会为每一个 VPN 创建 VPN 实例，使用 EBGP 方式向对端发布 IPv4 路由。
在图 1 中，对于 AS100 的 ASBR1 来说，AS200 的 ASBR2 只是它的一台 CE 设备；同样，对于 ASBR2，ASBR1 也只是一台接入的 CE 设备。图中，VPNLSP（Label Switched Path）表示私网隧道，LSP 表示公网隧道。



OptionA 的路由发布

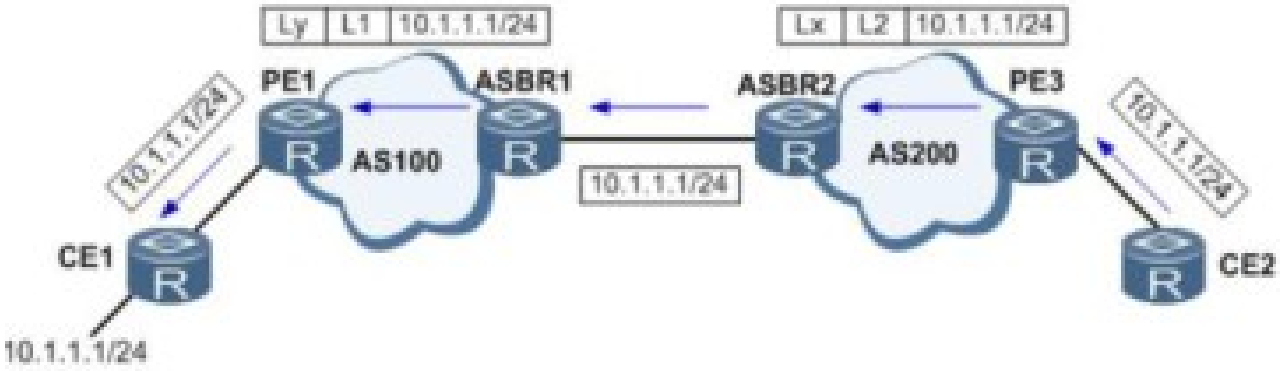
在 PE 和 ASBR 之间运行 MP-IBGP 协议交换 VPN-IPv4 路由信息。两个 ASBR 之间可以运行普通的 PE-CE 路由协议（BGP 或 IGP 多实例）或静态路由来交互 VPN 信息；但这是不同 AS 之间的交互，建议使用 EBGP。

例如 CE1 将目的地址为 10.1.1.1/24 的路由发布给 CE2，其流程如图 2 所示。其中，D 表示目的地址，NH 表示下一跳，L1 和 L2 表示所携带的私网标签。图中省略了公网 IGP 路由和标签的分配。



OptionA 的报文转发

以 LSP 为公网隧道的报文转发流程如图 3 所示。其中，L1 和 L2 表示私网标签，
Lx 和 Ly 表示公网外层隧道标签。

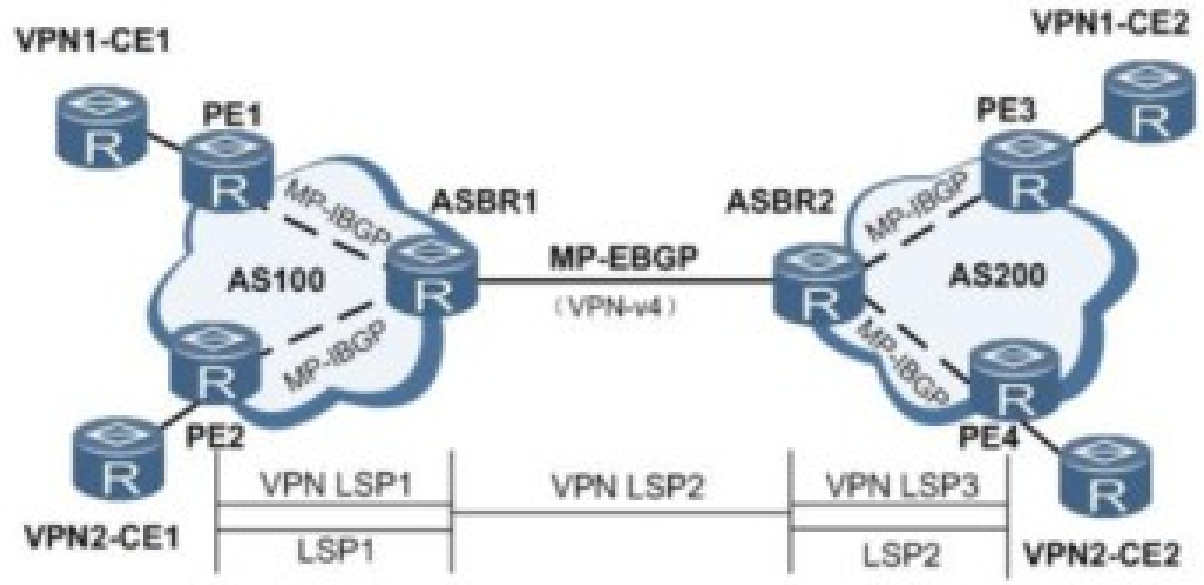


OptionA 的特点

1 优点：配置简单由于 ASBR 之间不需要运行 MPLS，也不需要为跨域进行特殊配置。
2 缺点：可扩展性差：由于 ASBR 需要管理所有 VPN 路由，为每个 VPN 创建 VPN 实例。这将导致 ASBR 上的 VPN-IPv4 路由数量过大。并且，由于 ASBR 间是普通的 IP 转发，要求为每个跨域的 VPN 使用不同的接口，从而提高了对 PE 设备的要求。如果跨越多个自治域，中间域必须支持 VPN 业务，不仅配置量大，而且对中间域影响大。在需要跨域的 VPN 数量比较少的情况，可以优先考虑使用。

OptionB

跨域 VPN-OptionB 中，两个 ASBR 通过 MP-EBGP 交换它们从各自 AS 的 PE 设备接收的标签 VPN-IPv4 路由。图中，VPN LSP 表示私网隧道，LSP 表示公网隧道。



跨域 VPN-OptionB 方案中，ASBR 接收本域内和域外传过来的所有跨域 VPN-IPv4 路由，再把 VPN-IPv4 路由发布出去。但 MPLS VPN 的基本实现中，PE 上只保存与本地 VPN 实例的 VPN Target 相匹配的 VPN 路由。通过对标签 VPN-IPv4 路由进行特殊处理，让 ASBR 不进行 VPN Target 匹配把收到的 VPN 路由全部保存下来，而不管本地是否有和它匹配的 VPN 实例。

bgp 100

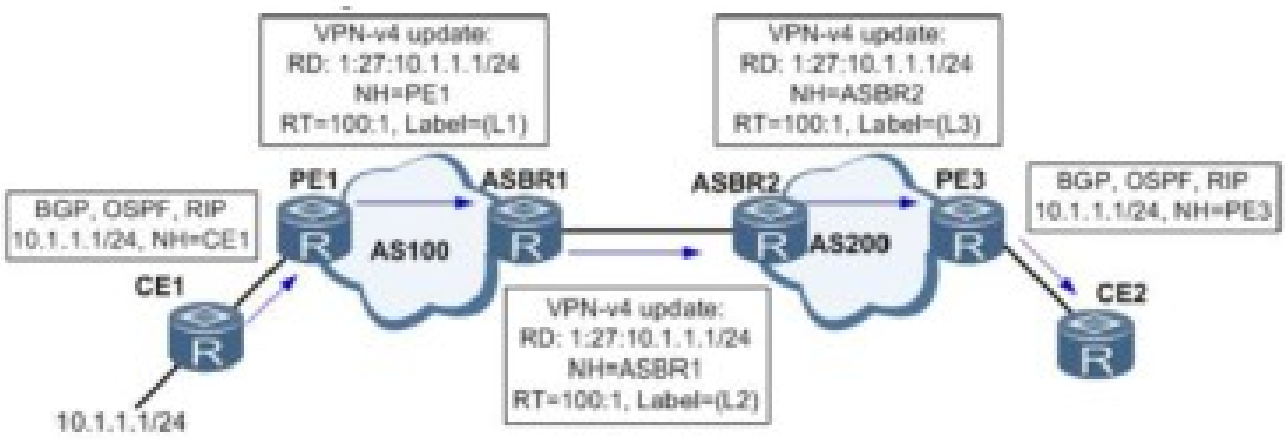
ipv4-family vpnv4

undo policy vpn-target

这种方案的优点是所有的流量都经过 ASBR 转发，使流量具有良好的可控性，但 ASBR 的负担重。可以同时使用 BGP 路由策略（如对 RT 的过滤），使 ASBR 上只保存部分 VPN-IPv4 路由。

OptionB 的路由发布

CE1 将 10.1.1.1/24 的路由发布给 CE2。NH 表示下一跳，L1、L2 和 L3 表示所携带的私网标签。



具体过程如下：

1 .CE1 通过 BGP、OSPF 或 RIP 方式将路由发布给 AS100 内的 PE1。

2 .AS100 内的 PE1 先通过 MP-IBGP 方式把标签 VPNv4 路由发布给 AS100 的 ASBR1，或发布给路由反射器 RR（Route Reflector），由 RR 反射给 ASBR1。

3 .ASBR1 通过 MP-EBGP 方式把标签 VPNv4 路由发布给 ASBR2。由于 MP-EBGP 在传递路由时，需要改变路由的下一跳，ASBR1 向外发布时给这些 VPNv4 路由信息分配新标签。

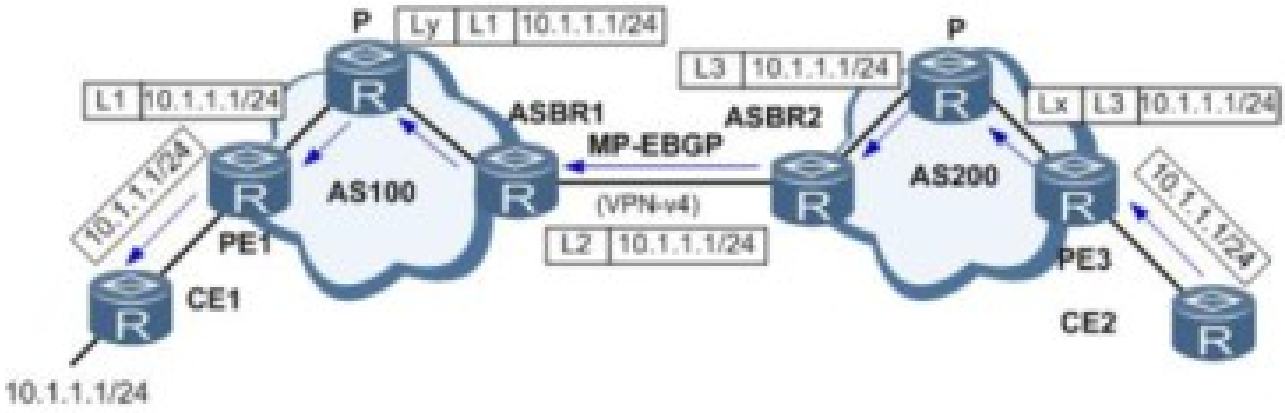
4 .ASBR2 通过 MP-IBGP 方式把标签 VPNv4 路由发布给 AS200 内的 PE3，或发布给 RR，由 RR 反射给 PE3。当 ASBR2 向域内的 MP-IBGP 对等体发布路由时，将下一跳改为自己。

5 .AS200 内的 PE3 通过 BGP、OSPF 或 RIP 方式将路由发布给 CE2。在 ASBR1 和 ASBR2 上都对 VPNv4 路由交换内层标签，域间的标签由 BGP 携带，因此 ASBR 之间不需要运行 LDP（Label Distribution Protocol）或 RSVP（Resource Reservation Protocol）等协议。

OptionB 的报文转发

在跨域 VPN-OptionB 方式的报文转发中，在两个 ASBR 上都要对 VPN 的 LSP 做一次交换。以 LSP 为公网隧道的报文转发流程如图

6 所示。其中，L1、L2 和 L3 表示私网标签。Lx 和 Ly 表示公网外层隧道标签。



OptionB 的特点

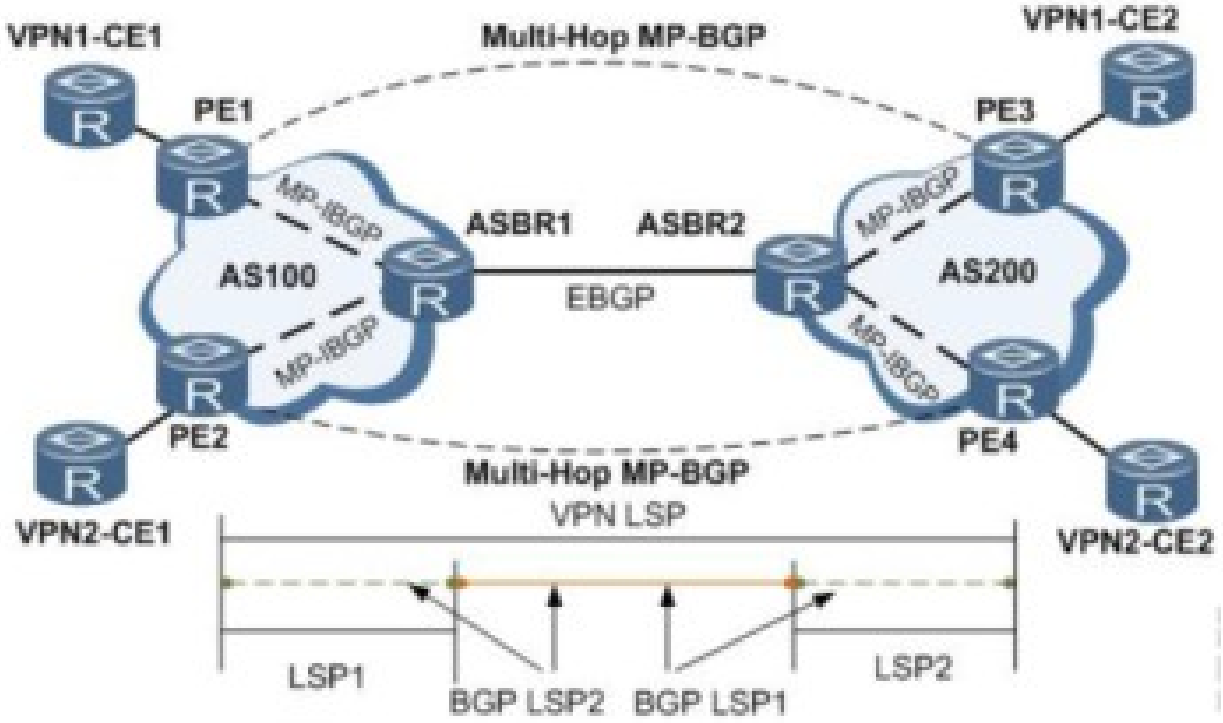
1 优点：不受 ASBR 之间互连链路数目的限制。
2 缺点：VPN 的路由信息是通过 AS 之间的 ASBR 来保存和扩散的，当 VPN 路由较多时，ASBR 负担重，容易成为故障点。因此在 MP-EBGP 方案中，需要维护 VPN 路由信息的 ASBR 一般不再负责公网 IP 转发。

OptionC

前面介绍的两种方式都能够满足跨域 VPN 的组网需求，但这两种方式也都需要 ASBR 参与 VPN-IPv4 路由的维护和发布。当每个 AS 都有大量的 VPN 路由需要交换时，ASBR 就很可能阻碍网络进一步的扩展。

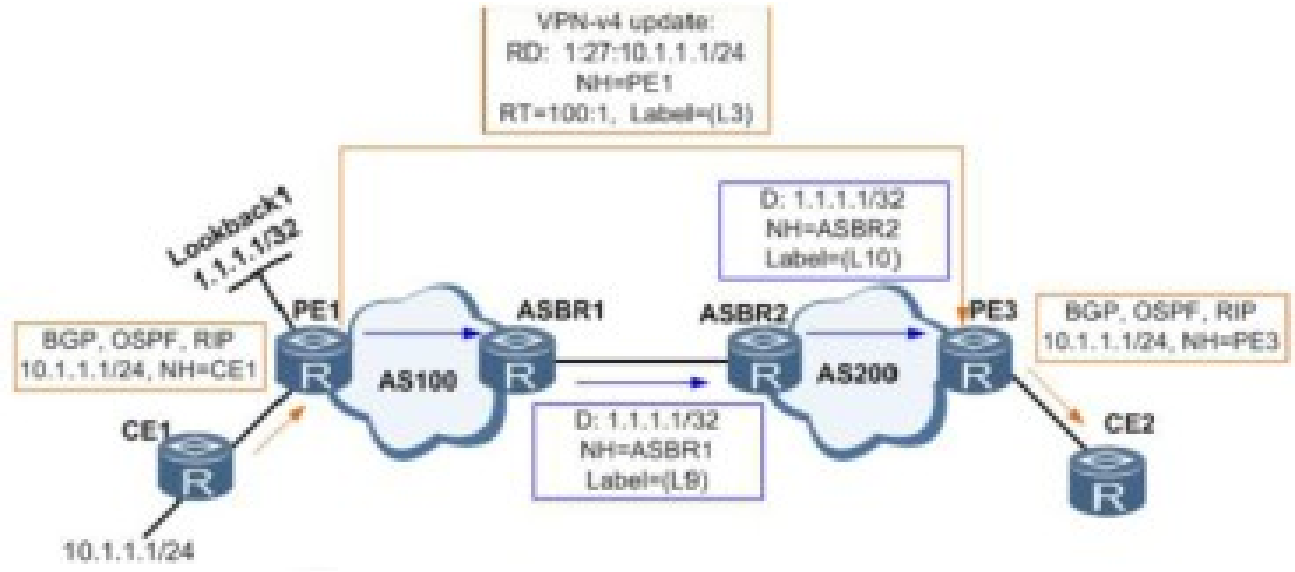
解决上述问题的方案是：ASBR 不维护或发布 VPN-IPv4 路由，PE 之间直接交换 VPN-IPv4 路由。
1 ASBR 通过 MP-IBGP 向各自 AS 内的 PE 设备发布标签 IPv4 路由，并将到达本 AS 内 PE 的标签 IPv4 路由通告给它在对端 AS 的 ASBR 对等体，过渡 AS 中的 ASBR 也通告带标签的 IPv4 路由。

这样，在入口 PE 和出口 PE 之间建立一条 LSP；
2 不同 AS 的 PE 之间建立 Multihop 方式的 EBGP 连接，交换 VPN-IPv4 路由；
3. ASBR 上不保存 VPN-IPv4 路由，相互之间也不通告 VPN-IPv4 路由。
下图为跨域 VPN-OptionC 的组网图，其中，VPN LSP 表示私网隧道，LSP 表示公网隧道。BGP LSP 主要作用是两个 PE 之间相互交换 Loopback 信息，由两部分组成，例如图中从 PE1 到 PE3 方向建立 BGP LSP1，PE3 到 PE1 方向建立 BGP LSP2。



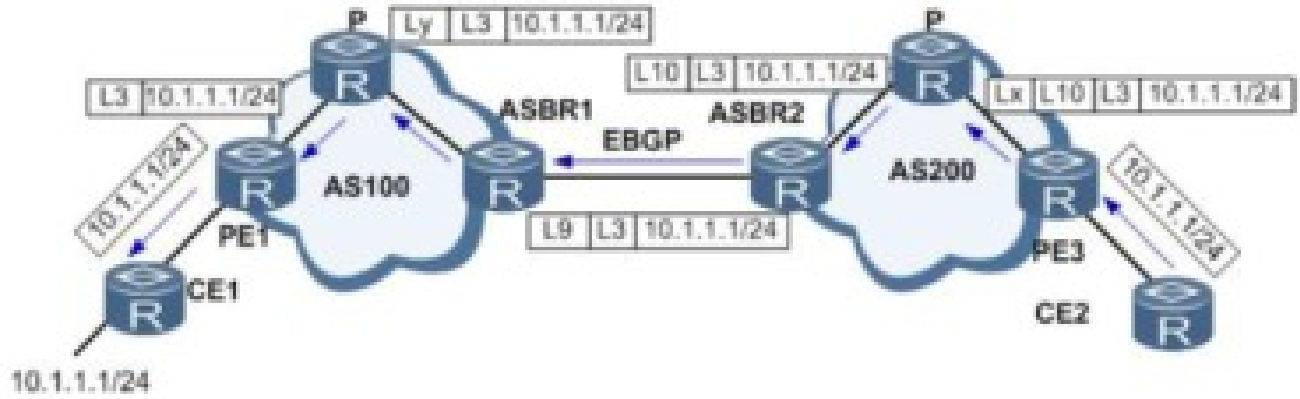
OptionC 的路由发布

跨域 VPN-OptionC 关键实现是公网跨域隧道的建立。例如在 CE1 中有一条 10.1.1.1/24 的路由信息，其发布流程如图 9 所示。D 表示目的地址，NH 表示下一跳，L3 表示所携带的私网标签，L9、L10 表示 BGP LSP 的标签。图中省略了公网 IGP 路由和标签的分配。



OptionC 的报文转发

以 LSP 为公网隧道的报文转发流程如图所示。其中，L3 表示私网标签，L10 和 L9 表示 BGP LSP 的标签，Lx 和 Ly 表示域内公网外层隧道标签。



报文从 PE3 向 PE1 转发时，需要在 PE3 上打上三层标签，分别为 VPN 路由的标签、BGP LSP 的标签和公网 LSP 的标签。到 ASBR 2 时，只剩下两层标签，分别是 VPN 的路由标签和 BGP LSP 的标签；进入 ASBR1 后，BGP LSP 终结，之后就是普通的 MPLS VP N 的转发流程。

OptionC 的特点

1 优点：VPN 路由在入口 PE 和出口 PE 之间直接交换，不需要中间设备的保存和转发。VPN 的路由信息只出现在 PE 设备上，而 P 和 ASBR 只负责报文的转发，使得中间域的设备可以不支持 MPLS VPN 业务，只需支持 MPLS 转发，ASBR 设备不再成为性能瓶颈。因此跨域 VPN-OptionC 更适合在跨越多个 AS 时使用。更适合支持 MPLS VPN 的负载分担。
2 缺点：:维护一条端到端的 PE 连接管理代价较大。

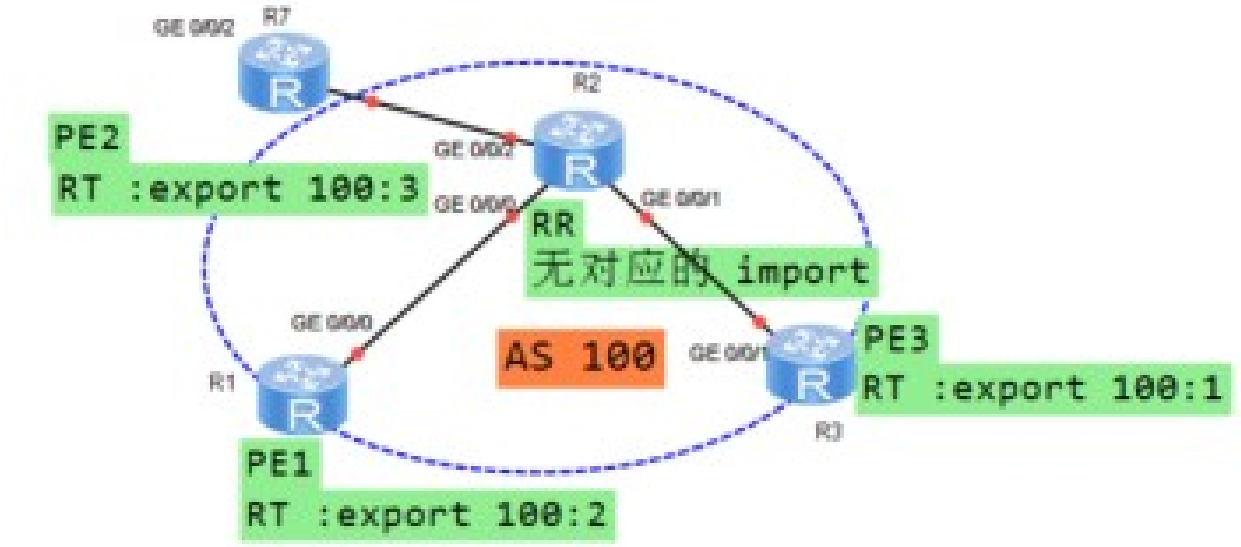
==========

不对 RT 过滤路由

配置“undo policy vpn-target”，即 RR 上不用 RT 来过滤路由。用来取消对接收的 VPN 路由或者标签块进行 VPN-Target 过滤。

在 BGP/MPLS IP VPN 组网中，VPN-Target 属性用来对接收到的 VPN 路由或者标签块进行过滤。如果不配置 VPN-Target，则会丢弃接收到的 VPN 路由或者标签块。
但在如下场景的特定设备上：
1.BGP/MPLS IP VPN 骨干网上部署的反射器 RR。
2.BGP/MPLS IP VPN 跨域 OptionB 方式中的 ASBR（不兼做 PE）。

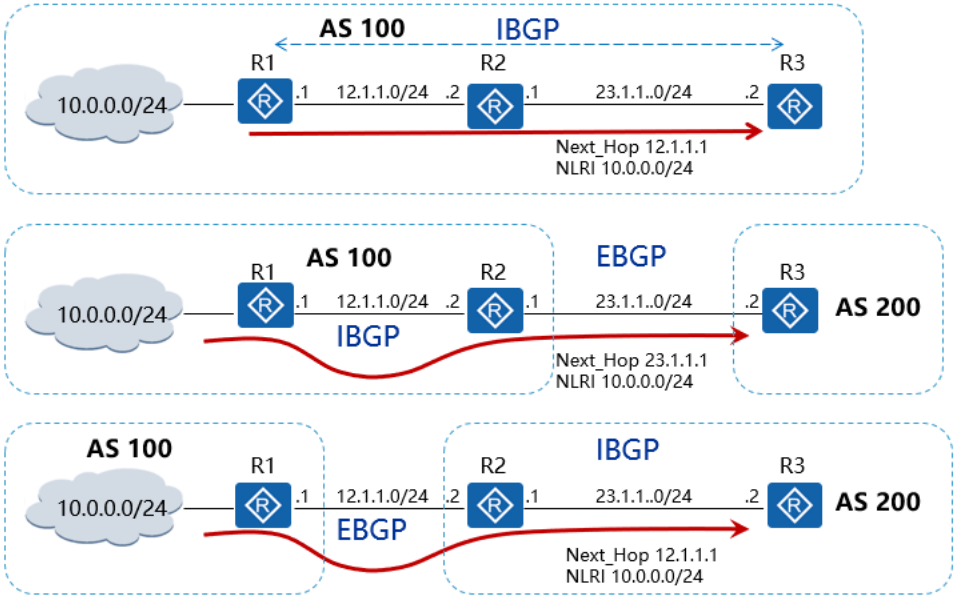
不会在其上创建 VPN，也就不配置 VPN-Target，这样会造成 RR 或者 ASBR 上不会保存 VPN 路由或者标签块。但 RR 或者 ASBR 又需要保存所有 PE 发来的 VPN 路由或者标签块，为解决这个问题，需要在 RR 或者 ASBR 上配置 undo policy vpn-target 命令，不对 VPN 路由或者标签块进行 VPN-Target 过滤。



RR 上没有相应的 VPN instance 的 import 与之相对应，
RR 就不会接收这 3 个 PE 发送的 VPNv4 路由，需要在 RR 上配 undo policy vpn-target

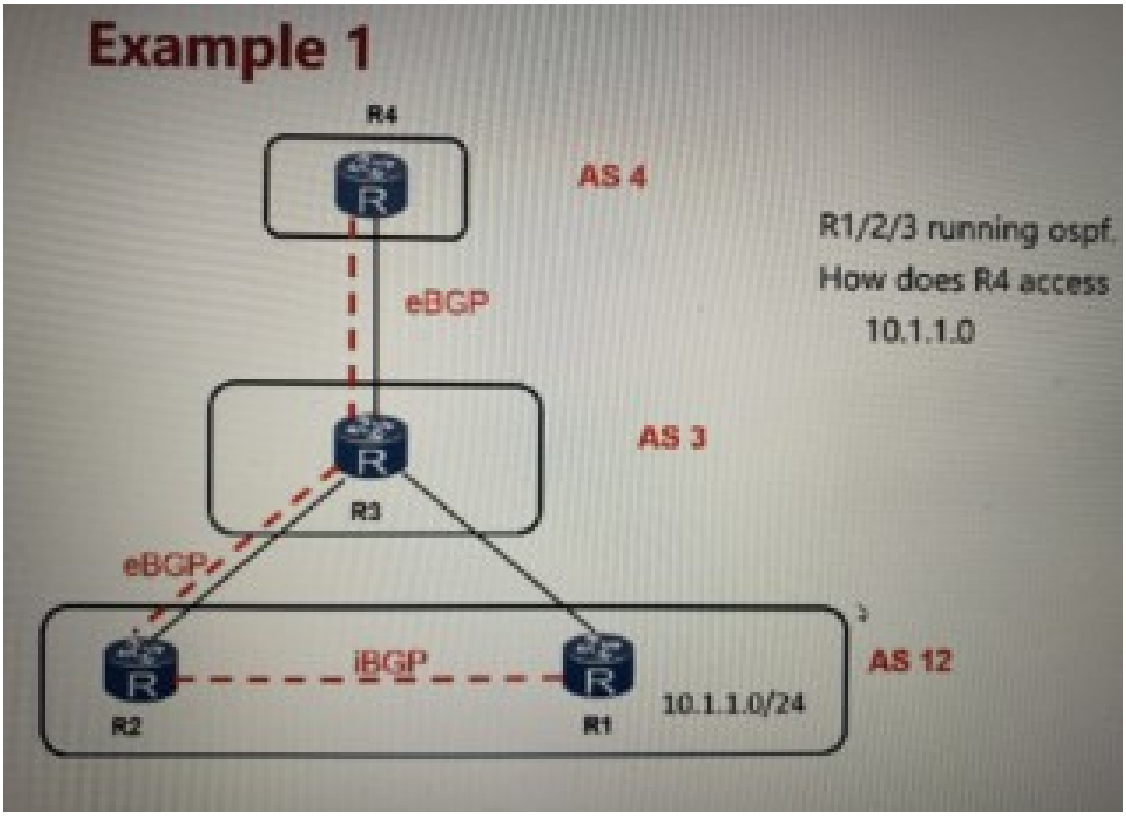
不改变下一跳

Next_Hop 属性记录了路由的下一跳信息。BGP 的下一跳属性和 IGP 的有所不同，不一定就是邻居设备的 IP 地址。通常情况下，Next_Hop 属性遵循下面的规则：



1 BGP Speaker 将本地始发路由发布给 IBGP 对等体时，会把该路

由信息的下一跳属性设置为本地与对端建立 BGP 邻居关系的接口地址。
2 BGP Speaker 在向 EBGP 对等体发布某条路由时，会把该路由信息的下一跳属性设置为本地与对端建立 BGP 邻居关系的接口地址。
3 BGP Speaker 在向 IBGP 对等体发布从 EBGP 对等体学来的路由时，并不改变该路由信息的下一跳属性。 要设置 peer 1.1.1.1 next-hop-local ,改变下一跳



bgp 100
peer 192.168.23.3 next-hop-invariable

peer next-hop-invariable 命令配置不同 AS 域的 PE 向 EBGP 对等体发布路由时不改变下一跳；向 IBGP 对等体发布引入的 IGP 路由时使用 IGP 路由的下一跳地址。

RR 之间在 vpnv4 视图下建立 MP-EBGP 邻居，向对端传递路由时不改变下一跳。
即对端 PE 学到的 VPNv4 路由的下一跳是本端 PE。RR 与本端 PE 建立 vpnv4 邻居，RR 向本端 PE 传递路由时不改变下一跳，即本端 PE 学到的 VPNv4 路由的下一跳是对端 PE。本端 PE 只与本端 RR 建立 VPNv4 邻居，本端 RR 与对端 RR 建立 VPNv4 邻居，实现了跨域 VPN 路由的传递。
保证对端 PE 可以迭代路由，用于流量传输时，通往本端 PE 的 BGP LSP。