采用 MP-BGP EVPN 控制平面的 VXLAN 网络

设计指南

目录

简介

虚拟可扩展局域网 (VXLAN) 是网络虚拟化的重叠技术。它在共享的第 3 层底层基础设施网络上提供第 2 层扩展，方法是在 IP 用户数据报协议（IP/UDP 中 的 MAC）隧道封装中使用 MAC 地址。在重叠网络中获得第 2 层扩展的目的是克服物理服务器机架和地理位置边界的局限，并获得在数据中心内和不同数据中心之间放置工作负载的灵活性。

初始的 IETF VXLAN 标准 (RFC 7348) 定义了一个基于组播、不采用控制平面的“泛洪和学习”VXLAN。它对远程 VXLAN 隧道终端 (VTEP) 对等体发现和远程终端主机学习依靠数据驱动的“泛洪和学习”行为。重叠广播、未知单播和组播流量封装到组播 VXLAN 数据包并通过底层组播转发传输到远程 VTEP 交换机。此类部署中的泛洪可能给解决方案的可扩展性带来挑战。在底层网络中启用组播功能的要求也会带来挑战，因为某些组织不希望在其数据中心或广域网网络中启用组播。

要克服 RFC 7348 中定义的“泛洪和学习”VXLAN 的局限，组织可以将多协议边界网关协议以太网虚拟专用网络 (MP-BGP EVPN) 用作 VXLAN 控制平面。MP-BGP EVPN 已由 IETF 定义为 VXLAN 重叠的基于标准的控制平面。MP-BGP EVPN 控制平面提供基于协议的 VTEP 对等体发现和终端主机可达性信息分发，允许适合私有和公共云的更具扩展性的 VXLAN 重叠网络设计。MP-BGP EVPN 控制平面引入了一组功能，这些功能可减少或消除重叠网络中的流量泛洪，并为西-东和南-北流量启用最优转发。

本文档讨论了 MP-BGP EVPN 的功能和配置，并介绍了使用 MP-BGP EVPN 的典型 VXLAN 重叠网络设计。

本文档不讨论 VXLAN 的基础知识、基于组播的“泛洪和学习”模式中的 VXLAN 或相关的网络设计方案。有关 VXLAN 和采用基于组播的“泛洪和学习”模式的 VXLAN 的详细信息，请参阅以下文档：

- VXLAN 概述：Cisco Nexus® 9000 系列交换机：http://www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/white-paper-c11-729383.html。
- Cisco Nexus 9300 平台交换机的 VXLAN 设计：http://www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/white-paper-c11-732453.html。

本文档假定读者已掌握有关 BGP、MP-BGP 以及 BGP 和多协议标签交换 (BGP/MPLS) IP VPN 的知识。有关详细信息，请参阅以下 IETF RFC 文档：

- RFC 4271 - 边界网关协议 4 (BGP-4)：https://tools.ietf.org/html/rfc4271
- RFC 4760 - BGP-4 的多协议扩展：https://tools.ietf.org/html/rfc4760
- RFC 4364 - BGP/MPLS IP VPN：https://tools.ietf.org/html/rfc4364#page-15

MP-BGP EVPN 控制平面：概述

MP-BGP EVPN 是基于行业标准的 VXLAN 控制协议。在 EVPN 之前，VXLAN 重叠网络在“泛洪和学习”模式下运行。在此模式下，终端主机信息学习和 VTEP 发现均由数据平面驱动，在 VTEP 之间没有控制协议分发终端主机可达性信息。MP-BGP EVPN 更改此模型。它为远程 VTEP 后面的终端主机引入了控制平面学习。它提供控制平面和数据平面的分离，并为 VXLAN 重叠网络中的第 2 层和第 3 层转发提供统一的控制平面。

MP-BGP EVPN 控制平面提供以下主要优势：

- MP-BGP EVPN 协议基于行业标准，允许多供应商互操作性。
- 它启用终端主机第 2 层和第 3 层可达性信息的控制面板学习，使组织能够构建更强大、更具扩展性的 VXLAN 重叠网络。
- 它使用已面世十年的 MP-BGP VPN 技术支持可扩展的多租户 VXLAN 重叠网络。

- EVPN 地址系列包含第 2 层和第 3 层可达性信息，从而在 VXLAN 重叠网络中提供集成的桥接和路由。
- 它通过基于协议的主机 MAC/IP 路由分发和本地 VTEP 上的地址解析协议 (ARP) 抑制，最大限度地减少网络泛洪。
- 它为东-西和北-南流量提供最优的转发，并通过分布式任播功能支持工作负载移动性。
- 它提供 VTEP 对等体发现和身份验证，可在 VXLAN 重叠网络中减少欺诈 VTEP 的风险。
- 它为在第 2 层构建双活多宿主提供机制。

MP-BGP EVPN 控制平面的软件和硬件支持

根据设备在 MP-BGP EVPN VXLAN 网络中所起的作用，对于采用 MP-BGP EVPN 控制平面的 VXLAN 网络，它可能只需支持控制平面功能，或需要同时支持控制平面和数据平面功能。

运行 MP-BGP EVPN 的 IP 传输设备

IP 传输设备在底层网络中提供 IP 路由。通过运行 MP-BGP EVPN 协议，它们成为 VXLAN 控制平面的一部分，并在其 MP-BGP EVPN 对等体之间分发 MP-BGP EVPN 路由。设备可能是 MP-iBGP EVPN 对等体、路由反射器或 MP 外部 BGP (MP-eBGP) EVPN 对等体。其操作系统软件需要支持 MP-BGP EVPN，以便可以了解 MP-BGP EVPN 更新，并使用标准定义的结构将其分配给其他 MP-BGP EVPN 对等体。对于数据转发，IP 传输设备仅根据 VXLAN 封装的数据包的外部 IP 地址执行 IP 路由。它们不需要支持 VXLAN 数据封装和解封功能。

运行 MP-BGP EVPN 的 VTEP

运行 MP-BGP EVPN 的 VTEP 需要同时支持控制平面和数据平面功能。在控制平面中，它们启动 MP-BGP EVPN 路由以通告其本地主机。它们从其对等体接收 MP-BGP EVPN 更新，并在其转发表中安装 EVPN 路由。对于数据转发，它们在 VXLAN 中封装用户流量并通过 IP 底层网络发送用户流量。它们按反方向从其他 VTEP 接收 VXLAN 封装的流量，将其解封，然后采用本地以太网封装方式将流量转发到主机。

需要为不同的网络角色选择正确的交换机平台。对于 IP 传输设备，软件需要支持 MP-EVPN 控制平面，但是硬件不需要支持 VXLAN 数据平面功能。对于 VTEP，交换机需要同时支持控制平面和数据平面功能。

VXLAN 间的路由

MP-BGP EVPN 控制平面提供集成的路由和桥接，方法是为 VXLAN 重叠网络上的终端主机分发第 2 层和第 3 层可达性信息。不同子网中的主机之间的通信需要进行 VXLAN 间的路由。BGP EVPN 通过以主机 IP 地址路由或 IP 地址前缀的形式分发第 3 层可达性信息来启用此通信。在数据平面中，VTEP 需要支持 IP 地址路由查找并根据查找结果执行 VXLAN 封装。此功能称为 VXLAN 路由功能。并非所有交换机硬件平台均支持 VXLAN 路由，因此会影响硬件平台的选择。

Cisco Nexus 9000 系列交换机上的 MP-BGP EVPN VXLAN 支持

VXLAN 的 MP-BGP EVPN 控制平面已引入 Cisco Nexus 9000 系列交换机的 Cisco® NX-OS 软件版本 7.0(3)I1(1)。软件功能也将实施在其他 Cisco Nexus 交换机平台（例如 Cisco Nexus 7000 系列交换机）的思科 NX-OS 软件系列中。

在思科 NX-OS 7.0(3)I1(1) 中，Cisco Nexus 9300 平台交换机支持 MP-BGP EVPN 控制平面功能和 VTEP 数据平面功能。Cisco Nexus 9500 平台交换机支持 MP-BGP EVPN 控制平面功能。VTEP 数据平面功能将在思科 NX-OS 7.0(3)I1(1) 的一个维护版本中添加到 Cisco Nexus 9500 平台交换机。Cisco Nexus 9300 和 9500 平台均支持硬件中的 VXLAN 间路由。

虽然许多 MP-BGP EVPN 功能和本文档中的设计讨论独立于平台，但是鉴于 Cisco Nexus 9000 系列是支持该协议的第一款交换机平台，因此所举示例均基于 Cisco Nexus 9000 系列。

MP-BGP EVPN 中的多租户

作为现有 MP-BGP 的扩展，MP-BGP EVPN 使用虚拟路由和转发 (VRF) 结构继承对 VPN 多租户的支持。在 MP-BGP EVPN 中，多个租户可以共存和共享共同的 IP 传输网络，同时在 VXLAN 重叠网络中拥有各自的单独 VPN。

在 EVPN VXLAN 重叠网络中，VXLAN 网络标识符 (VNI) 定义第 2 层域并实施第 2 层分段，方法是不允许第 2 层流量流经 VNI 边界。同样，VXLAN 租户之间的第 3 层分段的实现方法是应用第 3 层 VRF 技术并实施租户之间的路由隔离（通过使用映射到每个 VRF 实例的单独第 3 层 VNI）。每个租户都有其自己的 VRF 路由实例。特定租户的 VNI 的 IP 子网在将第 3 层路由域与其他租户分离的同一第 3 层 VRF 实例中。

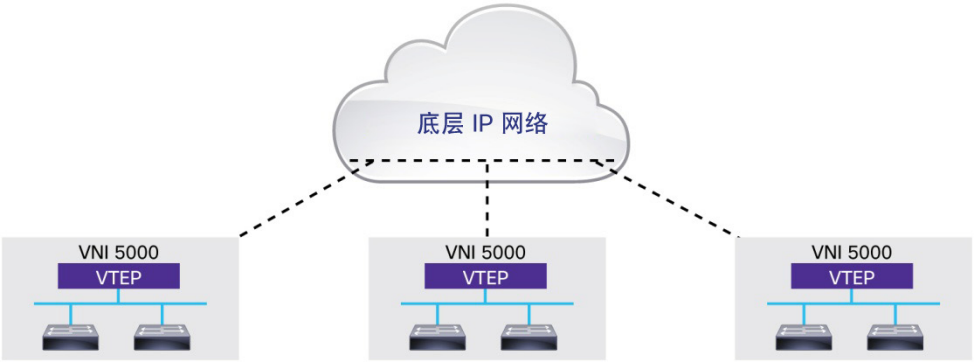
与基于组播的“泛洪和学习”VXLAN 和不具有多租户功能的其他第 2 层扩展技术相比，内置的多租户支持是 MP-BGP EVPN VXLAN 的一项优势。它使 VXLAN 技术更适合于使用多租户模型部署的云网络。

MP-BGP EVPN NLRI 和 L2VPN EVPN 地址系列

与其他网络路由控制协议一样，MP-BGP EVPN 旨在分发网络的网络层可达性信息 (NLRI)。EVPN NLRI 的一项独特功能是它包括位于 EVPN VXLAN 重叠网络的终端主机的第 2 层和第 3 层可达性信息。换句话说，它将通告 EVPN VXLAN 终端主机的 MAC 和 IP 地址。此功能构成 VXLAN 集成路由和桥接支持的基准。

需要分配第 2 层 MAC 地址，因为 VXLAN 是第 2 层扩展技术。不同于传统的 VLAN（局限在网络中的特定位置并保持在第 2 层和第 3 层边界内），VNI 是重叠网络中的一个虚拟的第 2 层网段。但是，从底层网络角度来看，它可以跨多个非邻近的站点，到达底层基础设施的第 2 层和第 3 层边界外（图 1）。同一 VNI 中的终端主机之间的流量需要在重叠网络中桥接，这意味着特定 VNI 中的 VTEP 设备需要了解此 VNI 中的终端主机的其他 MAC 地址。通过 BGP EVPN 分配 MAC 地址可以减少或消除 VXLAN 中的未知单播泛洪。

图 1. 底层 IP 网络中的 VNI



第 3 层主机 IP 地址通过 MP-BGP EVPN 通告，以便 VXLAN 间的流量可以通过一个最优路径路由到目标终端主机。对于需要路由到目标终端主机的 VXLAN 间流量，基于主机的 IP 路由可以提供至目标主机确切位置的最优转发路径。

MP-BGP EVPN 还可以通告 VNI 的 IP 子网前缀路由。在缺少主机 IP 路由时，前缀路由可以用于将流量路由到目标主机：例如在 VTEP 尚未通过 MP-BGP 学习主机 IP 路由时。如果子网需要可路由并为 VXLAN 网络外部所知，则 VTEP 还可以将前缀路由通告到 VXLAN 网络之外。

EVPN NLRI 使用 BGP 多协议扩展包含在 BGP 中，采用名为第 2 层 VPN (L2VPN) EVPN 的新地址系列。EVPN 的 L2VPN EVPN 地址系列类似基于 BGP MPLS 的 IP VPN (RFC 4364) 中的 VPNv4 地址系列，使用路由标识符 (RD) 保持不同 VRF 实例中的相同路由之间的唯一性，并使用路由目标 (RT) 定义用于确定如何通告路由以及由不同 VRF 实例共享路由的策略。

路由标识符是一个 8 位二进制数字，用于区分不同的路由组（一个 VRF 实例）。这个唯一数字作为每个路由的前缀，以便在相同的路由用于多个不同的 VRF 实例时，BGP 可以将其视为不同的路由。在与 MP-BGP 对等体交换 EVPN 路由时，路由标识符与路由一起通过 MP-BGP 传输。

路由目标可以应用于 VRF 实例，以控制此实例与其他 VRF 实例之间的路由导入和导出。路由的 route-target 属性以 BGP 扩展社区属性的形式分配，因此运行 MP-BGP EVPN 的设备上的 BDP 配置必须启用，才能够生成或处理扩展社区属性。

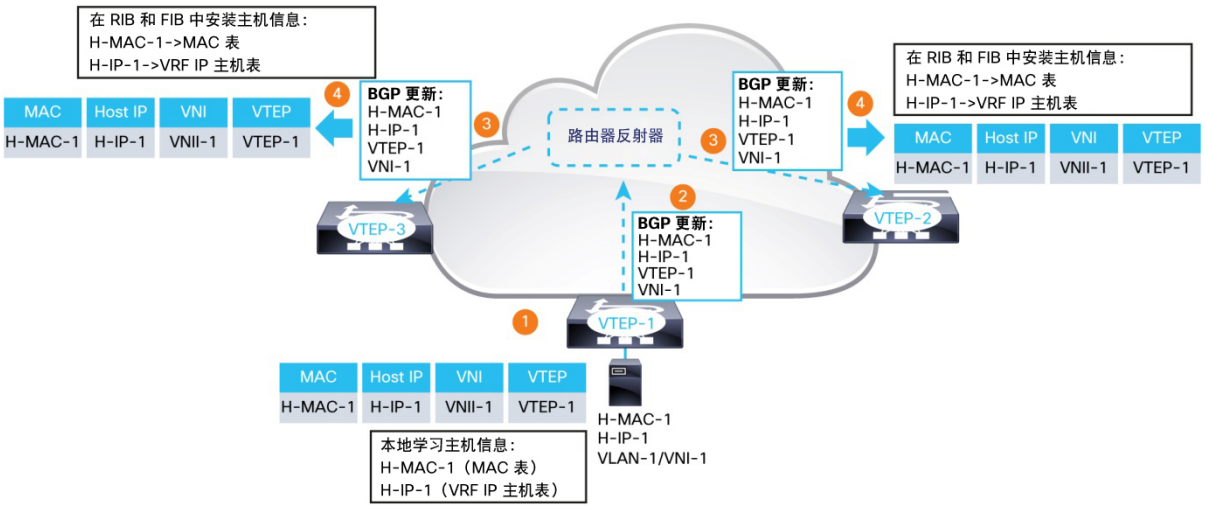
在思科 NX-OS 实施中，可以自动生成 BGP 路由标识符和路由目标以便于配置。BGP 路由标识符可以从 VTEP 交换机的 VNI 和 BGP 路由器 ID 自动得出，并且 BGP 路由目标可以自动生成为 BGP AS: VNI。或者，您也可以手动配置 BGP 路由标识符和路由目标。如果网络中的所有 MP-BGP EVPN VTEP 是 Cisco Nexus 交换机平台，则建议的方法是使用自动生成的 route-distinguisher 和 route-target 值。如果多个供应商的 VTEP 设备进行互操作，则建议的方法是手动配置值以避免由供应商实施的差异导致的问题。对于 VTEP 位于不同域的 eBGP 部署方案，必须手动分配 BGP 路由目标。

MP-BGP EVPN 控制平面提供集成的路由和桥接

MP-BGP EVPN 控制平面提供集成的路由和桥接，方法是为位于 VXLAN 重叠网络上的终端主机分发第 2 层和第 3 层可达性信息。每个 VTEP 执行本地学习以从其本地连接主机获取 MAC 和 IP 地址信息，然后通过 MP-BGP EVPN 控制平面分发此信息。通过 MP-BGP 控制平面远程学习连接到远程 VTEP 的主机。这种方法减少了终端主机学习的网络泛洪，并且可以更好地控制终端主机可达性信息分发。

图 2 显示了使用路由反射器的 MP-iBGP EVPN 中的终端主机 NLRI 学习和分发示例。

图 2. MP-BGP EVPN 主机 NLRI 学习和分发



本地主机学习

通过本地学习，MP-BGP EVPN 中的 VTEP 学习本地连接终端主机的 MAC 地址和 IP 地址。此学习可以基于本地数据平面，使用标准的以太网和 IP 学习过程，例如从传入以太网帧进行的源 MAC 地址学习，以及在主机为 VTEP 上的网关 IP 地址发送无故 ARP (GARP) 和反向 ARP (RARP) 数据包或 ARP 请求时进行的 IP 地址学习。或者，可以通过使用控制平面或通过 VTEP 与本地主机之间的管理平面集成来实现学习。

EVPN 路由通告和远程主机学习

在学习本地主机 MAC 和 IP 地址后，VTEP 在 MP-BGP EVPN 控制平面中通告主机信息，以便可以将此信息分发到其他 VTEP。此方法使 EVPN VTEP 能够在 MP-BGP EVPN 控制平面中学习远程终端主机。

通过 L2VPN EVPN 地址系列通告 EVPN 路由。BGP L2VPN EVPN 路由包括以下信息：

- Rd：路由标识符
- MAC 地址长度：6 个字节
- MAC 地址：主机 MAC 地址
- IP 地址长度：32 或 128
- IP 地址：主机 IP 地址（IPv4 或 IPv6）
- L2 VNI：终端主机所属的桥域的 VNI
- L3 VNI：与租户 VRF 路由实例相关联的 VNI

MP-BGP EVPN 使用 BGP 扩展社区属性在 EVPN 路由中传输导出的路由目标。当 EVPN VTEP 接收某个 EVPN 路由时，它会将接收的路由中的 route-target 属性与其本地配置的 route-target 导入策略进行比较，以决定是要导入还是忽略该路由。此方法使用已面世十年的 MP-BGP VPN 技术 (RFC 4364)，并提供可扩展的多租户功能，在该功能中，不在本地具有 VRF 的节点不会导入相应的路由。可通过使用 BGP 结构（例如路由目标约束的路由分配 [RFC 4684]）进一步增强 VPN 扩展。

当 VTEP 交换机为其本地学习的终端主机发起 MP-BGP EVPN 路由时，它将其自己的 VTEP 地址用作 BGP 下一跳。此 BGP 下一跳必须在网络中的路由分发过程中保持不变，因为在转发重叠网络的数据包时，远程 VTEP 必须学习作为 VXLAN 封装下一跳的原始 VTEP 地址。

对于用于通过底层网络将封装的 VXLAN 数据包路由到出口 VTEP 的所有 VTEP 地址，底层网络均提供 IP 可达性。底层网络中的网络设备只需保持 VTEP 地址的路由信息。它们无需学习 EVPN 路由。此方法简化了底层网络运营并提高了其稳定性和可扩展性。

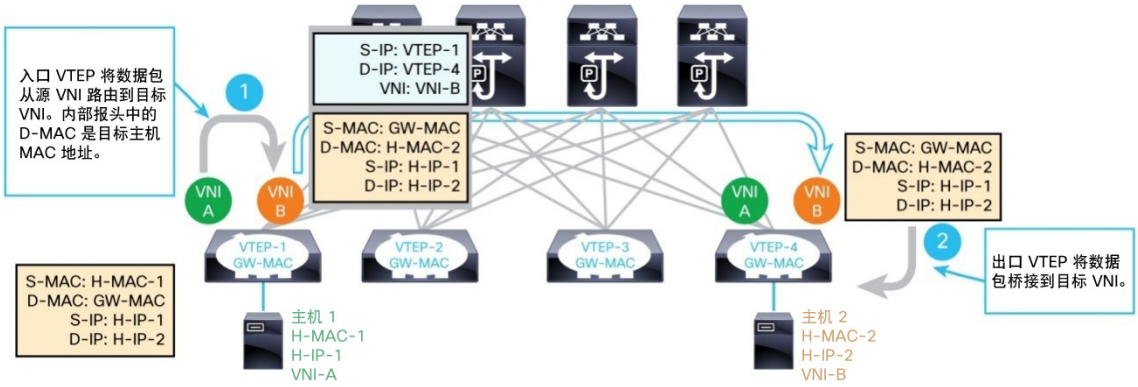
对称和非对称集成路由和桥接

IETF EVPN 草案定义了两个集成的路由和桥接 (IRB) 语义：非对称 IRB 和对称 IRB。Cisco Nexus 交换机平台的思科 NX-OS 实施对称 IRB 以获得其可扩展性优势，以及简化的第 2 层和第 3 层多租户支持。

非对称 IRB

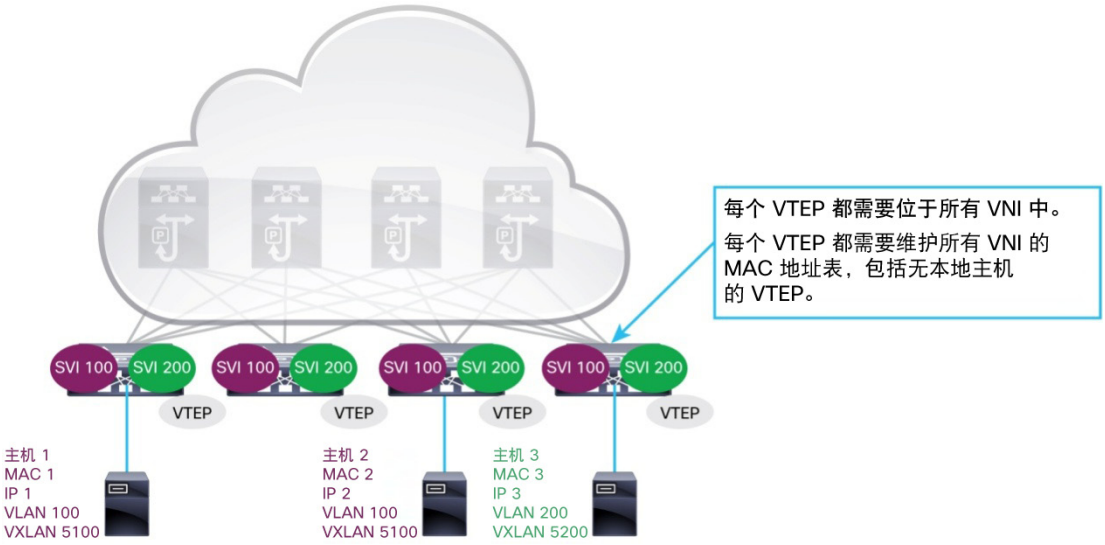
对于非对称 IRB，入口 VTEP 执行第 2 层桥接和第 3 层路由查找，而出口 VTEP 仅执行第 2 层桥接查找。如图 3 所示，对于非对称 IRB，当数据包在两个 VNI 之间传输时，入口 VTEP 将数据包从源 VNI 路由到目标 VNI。出口 VTEP 将数据包桥接到目标 VNI 内的目标端口。

图 3. 非对称 IRB 的 VXLAN 路由



非对称 IRB 要求使用第 2 层和第 3 层转发的源和目标 VNI 配置入口 VTEP。从根本上讲，这要求使用 VXLAN 网络中的所有 VNI 配置每个 VTEP，并学习连接到那些 VNI 的所有终端主机的 ARP 条目和 MAC 地址。随着终端主机的密度和/或重叠网络中的 VXLAN VNI 数量增加，此行为可能会导致可扩展性问题。

图 4. 非对称 IRB 中的 VTEP VNI 成员资格



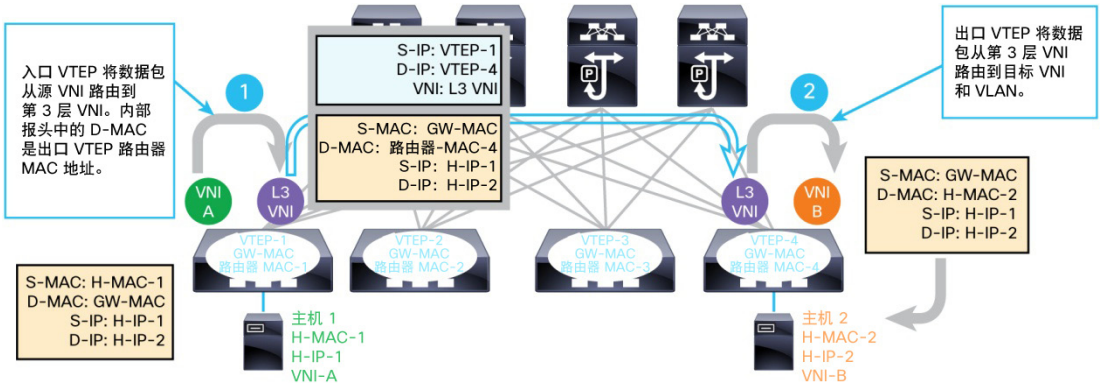
对称 IRB

对于对称 IRB，入口和出口 VTEP 执行第 2 层和第 3层 查找。对称 IRB 引入一些新的逻辑结构：

- **第 3 层 VNI：**每个租户 VRF 实例映射到网络中的唯一第 3 层 VNI。此映射需要在网络中的所有 VTEP 上保持一致。所有 VXLAN 间的路由流量使用 VXLAN 报头中的第 3 层 VNI 封装，并为接收的 VTEP 提供 VRF 情景。接收的 VTEP 使用此 VNI 确定转发内部 IP 数据包时需要处于的 VRF 情景。此 VNI 为在数据平面中实施第 3 层分段提供基准。
- **VTEP 路由器 MAC 地址：**每个 VTEP 均有唯一的系统 MAC 地址，其他 VTEP 可以将其用于 VNI 间的路由。此 MAC 地址在此处称为路由器 MAC 地址。路由器 MAC 地址用作路由的 VXLAN 数据包的内部目标 MAC 地址。

如图 5 所示，在将数据包从 VNI A 发送到 VNI B 时，入口 VTEP 将数据包路由到第 3 层 VNI。它会将内部目标 MAC 地址改写为出口 VTEP 的路由器 MAC 地址，并将 VXLAN 报头中的第 3 层 VNI 编码。在出口 VTEP 接收封装的 VXLAN 数据包后，它首先通过删除 VXLAN 报头解封数据包。然后它查看内部数据包报头。由于内部数据包报头中的目标 MAC 地址是其自己的 MAC 地址，因此它执行第 3 层路由查找。VXLAN 报头中的第 3 层 VNI 提供执行此路由查找时所处的 VRF 情景。

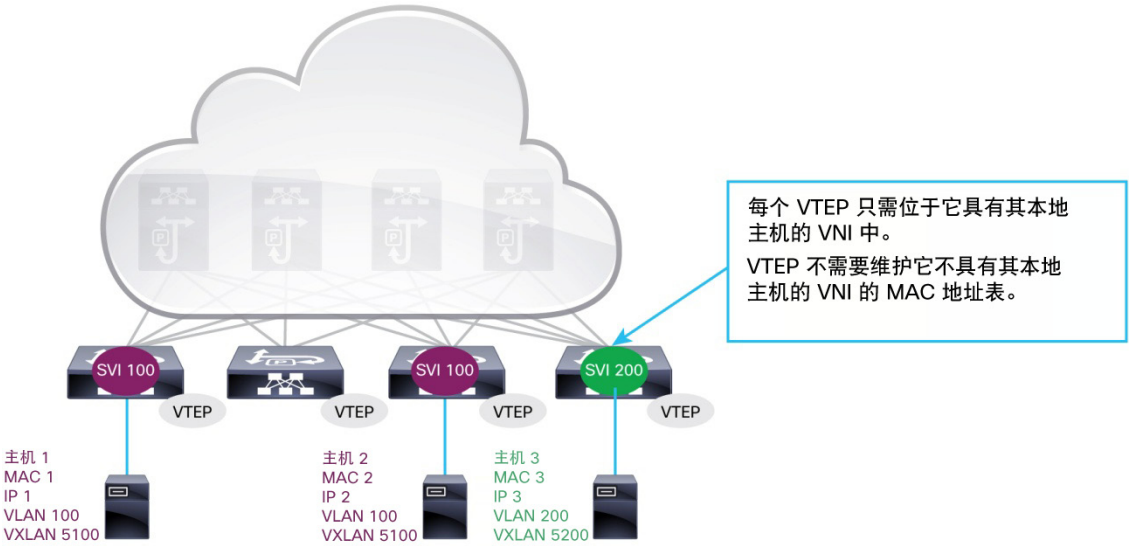
图 5. 对称 IRB 的 VXLAN 路由



对称 IRB 的优势

对于对称 IRB，入口 VTEP 不需要知道 VNI 间的路由的目标 VNI。因此，对于连接到它不具有本地主机的出口 VNI 的远程主机，VTEP 不需要学习和保持 MAC 地址信息（图 6）。这种方法使得可以更好地利用 VTEP 上的 MAC 地址表和 ARP 邻接关系。例如，在图 6中，并非 VNI-B 中的所有主机 MAC 地址和 ARP 邻接关系均需要存在于 VTEP-1。因此，与使用非对称 IRB 相对，路由和桥接更具可扩展性。思科 NX-OS 实施对称 IRB 以实现最优学习和扩展。

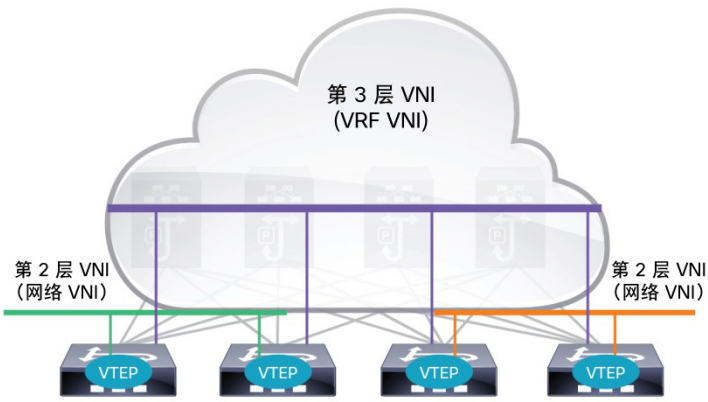
图 6. 对称 IRB 的 VTEP VNI 成员资格



桥域的 VNI 和 IP VRF 实例

EVPN VXLAN 租户可以有多个第 2 层网络，每个网络都有相应的 VNI。这些第 2 层网络是重叠网络中的桥域。与其相关联的 VNI 通常称为第 2 层 (L2) VNI。如果需要 VXLAN 间的路由，每个租户还需要对称 IRB 的第 3 层 (L3) VNI。虽然 VTEP 可以具有 VXLAN EVPN 中的所有第 2 层 VNI 或其子集，但是它必须将第 3 层 VNI 用于 VXLAN 间的路由。EVPN 中的所有 VTEP 必须具有相同的第 3 层 VNI（图 7）。

图 7. 桥域的 VNI 和 IP VRF 实例

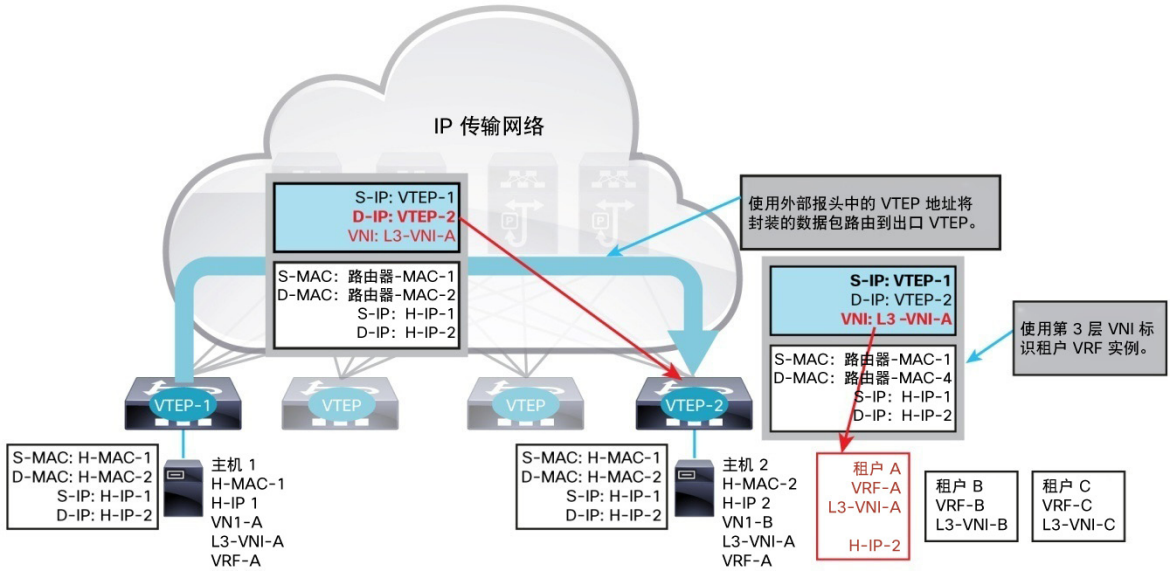



```
vlan 200
  vn-segment 20000
vlan 201
  vn-segment 20100

vlan 3900
  name l3-vni-vlan-for-tenant-1
  vn-segment 39000

vrf context evpn-tenant-1
  vni 39000
  rd auto
  address-family ipv4 unicast
    route-target import 39000:39000
    route-target export 39000:39000
    route-target both auto evpn

interface Vlan200
  no shutdown
  vrf member evpn-tenant-1
  ip address 20.0.0.1/24
  fabric forwarding mode anycast-gateway

interface Vlan201
  no shutdown
  vrf member evpn-tenant-1
  ip address 20.1.0.1/24
  fabric forwarding mode anycast-gateway

interface Vlan3900
  description l3-vni-for-tenant-1-routing
  no shutdown
  vrf member evpn-tenant-1
  ip forward
```

当 EVPN VTEP 为它从其本地终端主机接收的数据包执行转发查找和 VXLAN 封装时，它使用第 2 层 VNI 或 VXLAN 报头中的第 3 层 VNI，具体取决于是需要桥接还是路由数据包。如果原始数据包报头中的目标 MAC 地址不属于本地 VTEP，则本地 VTEP 执行第 2 层查找并将数据包桥接到与源主机位于同一第 2 层 VNI 的目标终端主机。本地 VTEP 在 VXLAN 报头中嵌入此 第 2 层 VNI。在这种情况下，源和目标主机位于同一第 2 层广播域。如果目标 MAC 地址属于本地 VTEP 交换机，换句话说，如果本地 VTEP 是源主机的 IP 网关，并且源主机和目标主机位于不同的 IP 子网，则数据包将由本地 VTEP 进行路由。在这种情况下，它执行第 3 层路由查找。然后它使用 VXLAN 报头中的第 3 层 VNI 封装数据包，并将内部目标 MAC 地址改写为远程 VTEP 的路由器 MAC 地址。在接收封装的 VXLAN 数据包后，远程 VTEP 会根据内部 IP 报头执行另一个路由查找，因为接收的数据包中的内部目标 MAC 地址属于远程 VTEP 本身。

VXLAN 数据包的外部 IP 报头中的目标 VTEP 地址标识了底层网络中的目标主机位置。VXLAN 数据包根据外部目标通过底层网络路由至出口 VTEP。在数据包到达出口 VTEP 后，会检查 VXLAN 报头中的 VNI 以确定应在其中桥接数据包的 VLAN 或应将数据包路由到的租户 VRF 实例。在后一种情况下，使用第 3 层 VNI 对 VXLAN 报头进行编码。第 3 层 VNI 与租户 VRF 路由实例相关联，因此，出口 VTEP 可以将路由的 VXLAN 数据包直接映射到适当的租户路由实例。图 8 显示了对称 IRB 中的此转发概念。此方法使多租户功能更容易支持第 2 层和第 3 层分段。

图 8. 对称 IRB 路由的 VXLAN 数据包转发



MP-BGP EVPN 中的 VTEP 对等体发现和身份验证

在 MP-BGP EVPN 之前，VXLAN 不具有基于控制协议的 VTEP 对等体发现机制或用于对 VTEP 对等体进行身份验证的方法。这些局限在现实的 VXLAN 部署中会带来重大的安全风险，因为它们允许将欺诈的 VTEP 轻松插入到 VNI 段以发送或接收 VXLAN 流量。

使用 MP-BGP EVPN 控制平面，VTEP 设备首先需要与其他 VTEP 或与内部 BGP (iBGP) 路由反射器建立 BGP 邻居邻接关系。除了终端主机 NLRI 的 BGP 更新外，VTEP 通过 BGP 交换关于自身的以下信息：

- 第 3 层 VNI
- VTEP 地址
- 路由器 MAC 地址

一旦 VTEP 从远程 VTEP BGP 邻居接收 BGP EVPN 路由更新，它就会将 VTEP 地址从该路由通告添加到 VTEP 对等体列表。然后，此 VTEP 对等体列表用作有效 VTEP 对等体的白名单。不在此白名单上的 VTEP 被视为无效或未经授权的来源。来自这些无效 VTEP 的 VXLAN 封装流量将被其他 VTEP 丢弃。

在数据平面转发中，BGP EVPN VTEP 仅接受来自白名单上的 VTEP 对等体的 VXLAN 封装数据包。因此，MP-BGP EVPN 引入基于协议的 VTEP 发现，并且能够将 VXLAN 重叠流量分配仅限于 BGP 学习的 VTEP。

除了促进 VTEP 对等体学习的 VTEP 地址外，BGP EVPN 路由还包含 VTEP 路由器 MAC 地址。每个 VTEP 均具有路由器 MAC 地址。一旦通过 MP-BGP 分配并由其他 VTEP 学习 VTEP 的路由器 MAC 地址，其他 VTEP 就会将其用作 VTEP 对等体的一个属性，以将 VXLAN 间路由的数据包封装到该 VTEP 对等体。路由器 MAC 地址设定为路由的 VXLAN 的内部目标 MAC 地址。

为获得额外的安全性，现有的 BGP 消息摘要 5 (MD5) 身份验证可以方便地应用于 BGP 邻居会话，以便交换机无法成为 BGP 邻居以交换 MP-BGP EVPN 路由，除非交换机使用预先配置的 MD5 三重数据加密标准 (3DES) 密钥成功对彼此进行身份验证。MP-BGP EVPN 中的 BGP 邻居身份验证采用与以前在 BGP 中支持的相同方法进行配置。示例如下所示：

在 VTEP-1 上

```
router bgp 100
  router-id 10.1.1.101
  log-neighbor-changes
  address-family ipv4 unicast
  address-family l2vpn evpn
  neighbor 10.1.1.102 remote-as 100
    password 3 a667d47acc18ea6b
    update-source loopback0
    address-family ipv4 unicast
      send-community both
    address-family l2vpn evpn
      send-community both
```

- 可以在命令行界面 (CLI) 上使用明文密码进行配置：**password ciscol23**。系统会在运行配置显示中自动将此密码更改为 3DES 加密的密码。
- 两个邻居需要具有完全相同的密码。

在 VTEP-2 上

```
router bgp 100
  router-id 10.1.1.102
  log-neighbor-changes
  address-family ipv4 unicast
  address-family l2vpn evpn
    retain route-target all
  neighbor 10.1.1.101 remote-as 100
    password 3 a667d47acc18ea6b
    update-source loopback0
    address-family ipv4 unicast
      send-community both
    address-family l2vpn evpn
      send-community both
```

以下示例显示了思科 NX-OS 中的 VNI 对等体状态和信息：

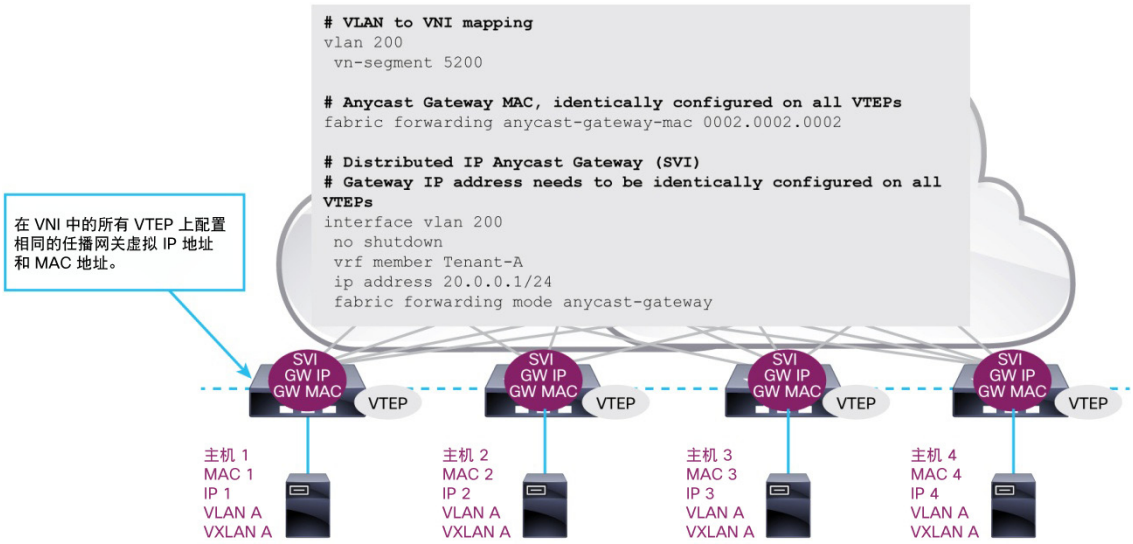
```
VTEP1-1# sh nve peers
Interface Peer-IP          State LearnType Uptime   Router-Mac
--------- ---------------  ----- --------- -------- -----------------
nve1      10.1.1.102        Up    CP        1w3d    6412.2574.6ae7
nve1      10.1.1.134        Up    CP        1w3d     7c69.f6df.e71f

VTEP-1#
```

```
VTEP-1# sh nve peers peer-ip 10.1.1.102 det
Details of nve Peers:
----------------------------------------
Peer-Ip: 10.1.1.102
    NVE Interface       : nve1
    Peer State          : Up
    Peer Uptime         : 1w3d
    Router-Mac          : 6412.2574.6ae7
    Peer First VNI      : 20100
    Configured VNIs     : 20000,20100,21000,21100,39000,39010
    Provision State     : add-complete
    Route-Update        : Yes
    Peer Flags          : DisableLearn
    Learnt CP VNIs      : 20000,20100
----------------------------------------
VTEP-1#
```

MP-BGP EVPN 中的分布式任播网关

在 MP-BGP EVPN 中，VNI 中的任何 VTEP 均可以成为其 IP 子网中的终端主机的分布式任播网关，方法是支持相同的虚拟网关 IP 地址和虚拟网关 MAC 地址（图 0）。借助 EVPN 中的任播网关，VNI 中的终端主机始终可以将此 VNI 的本地 VTEP 用作其默认网关，以将流量发送到其 IP 子网之外。此功能支持对来自 VXLAN 重叠网络的终端主机的北向流量进行最优的转发。分布式任播网关还在 VXLAN 重叠网络中提供无缝主机移动性的优势。由于在 VNI 内的所有 VTEP 上相同地调配网关 IP 和虚拟 MAC 地址，因此当终端主机从一个 VTEP 移至另一个 VTEP 时，它不需要发送另一个 ARP 请求以重新学习网关 MAC 地址。

图 9. MP-BGP EVPN 中的分布式任播网关



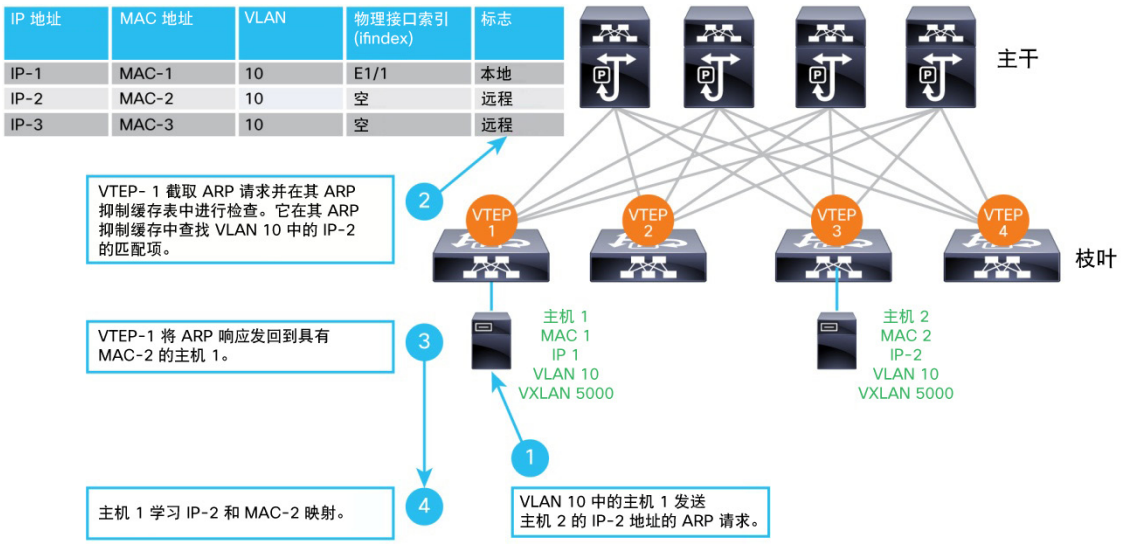
MP-BGP EVPN 中的 ARP 抑制

ARP 抑制是 MP-BGP EVPN 控制平面提供的一项增强功能，可减少由来自 ARP 请求的广播流量导致的网络泛洪。

在为 VNI 启用 ARP 抑制后，其 VTEP 各保存已知 IP 主机的一个 ARP 抑制缓存表及其在 VNI 网段中的关联 MAC 地址。如图 10 所示，当 VNI 中的一台终端主机为另一终端主机 IP 地址发送 ARP 请求时，其本地 VTEP 会截取 ARP 请求并在其 ARP 抑制缓存表中检查经过 ARP 处理的 IP 地址。如果它找到匹配项，则本地 VTEP 会代表远程终端主机发送一个 ARP 响应。本地主机在 ARP 响应中学习远程主机的 MAC 地址。如果本地 VTEP 未在其 ARP 抑制表中包含经过 ARP 处理的 IP 地址，则它会使 ARP 请求涌入到 VNI 中的其他 VTEP。对于发送给网络中的静默主机的初始 ARP 请求，此 ARP 泛洪可能发生。网络中的 VTEP 看不到来自静默主机的任何流量，直到另一台主机为其 IP 地址发送 ARP 请求，并且它发回 ARP 响应。在本地 VTEP 学习静默主机的 MAC 和 IP 地址后，信息通过 MP-BGP EVPN 控制平面分发给所有其他 VTEP。任何后续的 ARP 请求均不需要泛洪。

由于大多数终端主机会在上线后立即发送 GARP 或 RARP 请求以向网络宣告自己，因此本地 VTEP 将立即有机会学习终端主机的 MAC 和 IP 地址，并通过 MP-BGP EVPN 控制平面将此信息分发给其他 VTEP。因此，应当由 VTEP 通过执行本地学习或基于控制平面的远程学习，来学习 VXLAN EVPN 中的大多数活动的 IP 主机。因此，ARP 抑制可减少由主机 ARP 学习行为导致的网络泛洪。

图 10. MP-BGP EVPN 中的 ARP 抑制



MP-BGP EVPN VTEP 配置

本部分概述了用于配置 MP-BGP EVPN VTEP 的步骤。

第 1 步： 执行每台 VTEP 交换机的初始配置。

启用 VXLAN 和 MP-BGP EVPN 控制平面。

```
feature nv overlay
feature vn-segment-vlan-based
feature bgp
nv overlay evpn
```

- `feature nv overlay` — 启用 VXLAN。
- `feature vn-segment-vlan-based` — 启用基于 VLAN 的 VXLAN（当前是唯一的模式）。
- `feature bgp` — 启用 BGP。
- `nv overlay evpn` — 为 VXLAN 启用 EVPN 控制平面。

Other features may need to be enabled.

```
feature ospf
feature pim
feature interface-vlan
```

- `feature ospf` — 如果开放最短路径优先 (OSPF) 被选为底层内部网关协议 (ISP) 路由协议，则启用它。
- `feature pim` — 在底层网络中启用 IP 协议无关组播 (PIM) 路由。
- `feature interface-vlan` — 如果 VTEP 需要是 VXLAN VLAN 子网的 IP 网关和路由，则启用 VLAN 交换机虚拟接口 (SVI)。

第 2 步： 配置 EVPN 租户 VRF 实例。

以下示例显示了两个租户 VRF 实例的配置：

```
vrf context evpn-tenant-1
  vni 39000
  rd auto
  address-family ipv4 unicast
    route-target both auto
    route-target both auto evpn
```

创建 VXLAN 租户 VRF 实例。

为此租户 VRF 实例指定 VXLAN 路由的第 3 层 VNI。

定义 VRF 路由标识符。

在地址系列 ipv4 单播中定义 VRF 路由目标导入和导出策略。本示例为此 VRF 使用路由目标自动生成。

```
vrf context evpn-tenant-2
  vni 39010
  rd auto
  address-family ipv4 unicast
    route-target import 100:39010
    route-target import 100:39010 evpn
    route-target export 100:39010
    route-target export 100:39010 evpn
```

在前面的步骤后创建第二个租户 VRF 实例。该示例为路由目标导入和导出策略使用手动配置。

第 3 步： 为每个租户 VRF 实例创建第 3 层 VNI。

```
vlan 3900
  name l3-vni-vlan-for-tenant-1
  vn-segment 39000

interface Vlan3900
  description l3-vni-for-tenant-1-routing
  no shutdown
  vrf member evpn-tenant-1
  ip forward

vrf context evpn-tenant-1
  vni 39000
  rd auto
  address-family ipv4 unicast
    route-target import 39000:39000
    route-target export 39000:39000
    route-target both auto evpn
```

为第 3 层 VNI 创建 VLAN。为每个租户 VRF 路由实例创建一个第 3 层 VNI。

为第 3 层 VNI 创建 SVI。将此 SVI 放在租户 VRF 情景中。命令“ip forward”为 VNI IP 子网启用基于前缀的路由。必须执行此操作，才能完成至 VNI 网络中的静默主机的初始路由。

将第 3 层 VNI 与租户 VRF 路由实例相关联。

```
vlan 3901
  name l3-vni-vlan-for-tenant-2
  vn-segment 39010

interface Vlan3901
  description l3-vni-for-tenant-2-routing
  no shutdown
  vrf member evpn-tenant-2
  ip forward

vrf context evpn-tenant-2
  vni 39010
  rd auto
  address-family ipv4 unicast
    route-target import 39010:39010
    route-target export 39010:39010
    route-target both auto evpn
```

在前面的步骤后为第二个租户定义第 3 层 VNI。

第 4 步： 为第 2 层网络配置 EVPN 第 2 层 VNI。

此步骤涉及将 VLAN 映射到第 2 层 VNI 并定义其 EVPN 参数。

```
vlan 200
  vn-segment 20000
vlan 210
  vn-segment 21000
```

将 VLAN 映射至 VXLAN VNI。

```
evpn
  vni 20000 l2
    rd auto
    route-target import auto
    route-target export auto
  vni 21000 l2
    rd auto
    route-target import auto
    route-target export auto
```

在 EVPN 配置下，为每个第 2 层 VNI 定义路由标识符以及路由目标导入和导出策略。

第 5 步： 为第 2 层 VNI 配置 SVI，并在 SVI 下启用任播网关。

```
interface Vlan200
  no shutdown
  vrf member evpn-tenant-1
  ip address 20.1.1.1/8
  fabric forwarding mode anycast-gateway


interface Vlan210
  no shutdown
  vrf member evpn-tenant-1
  ip address 21.1.1.1/8
  fabric forwarding mode anycast-gateway
```

为第 2 层 VNI 创建 SVI。将其与租户 VRF 实例相关联。

此 VLAN 和 VNI 的所有 VTEP 应当与分布式 IP 网关具有相同的 SVI IP 地址。

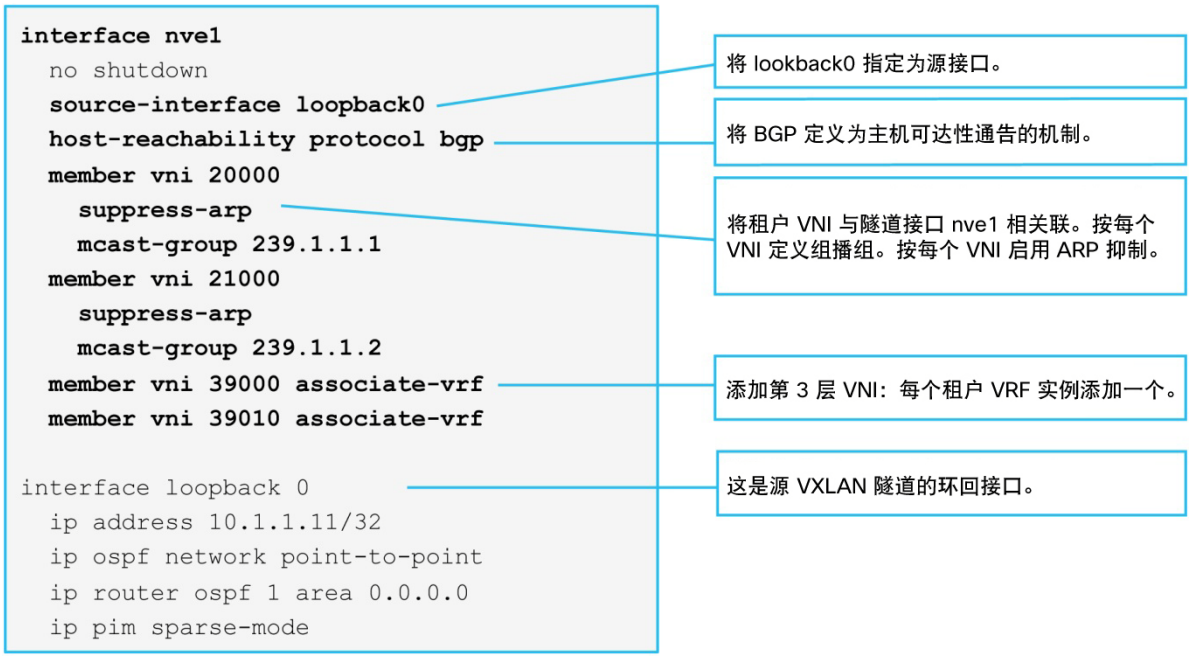
为此 VLAN 和 VNI 启用分布式任播网关。

第 6 步： 配置 EVPN 分布式任播网关。

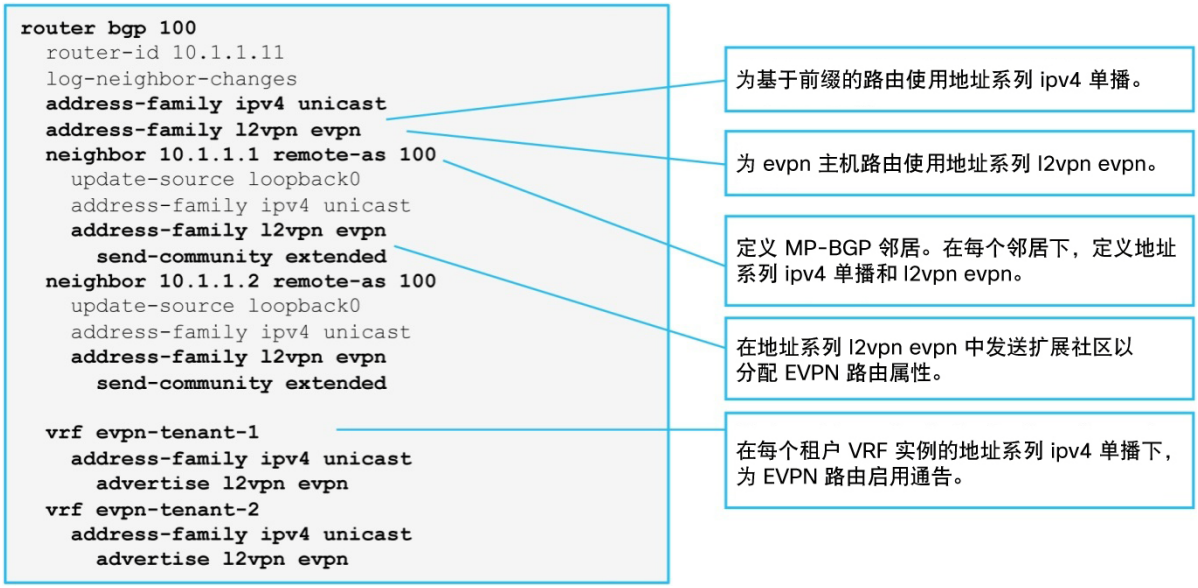
此步骤包括为每个 VTEP 配置任播网关虚拟 MAC 地址，以及为每个 VNI 配置任播网关 IP 地址。

EVPN 域中的所有 VTEP **必须**具有相同的任播网关虚拟 MAC 地址和相同的任播网关 IP 地址（对于这些 VTEP 作为默认 IP 网关的给定 VNI 而言）。

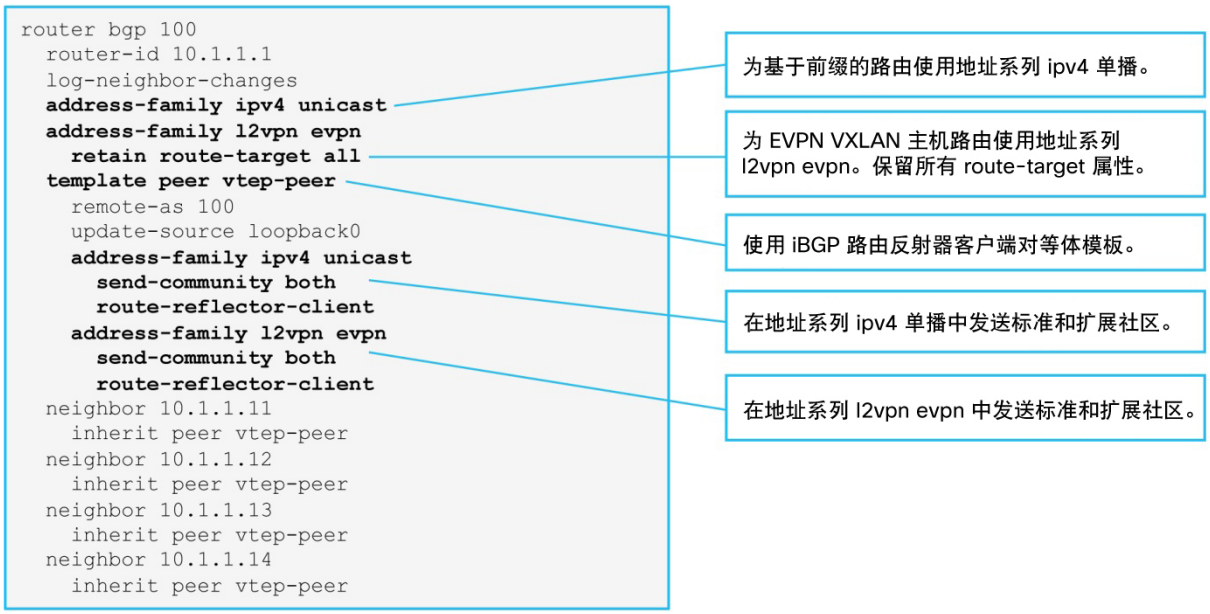
第 7 步： 配置 VXLAN 隧道接口 nve1，然后将第 2 层 VNI 和第 3 层 VNI 与其关联。



第 8 步： 在 VTEP 上配置 MP-BGP。



第 9 步： 配置 iBGP 路由反射器。



MP-BGP EVPN VXLAN 中的虚拟 Port-Channel VTEP

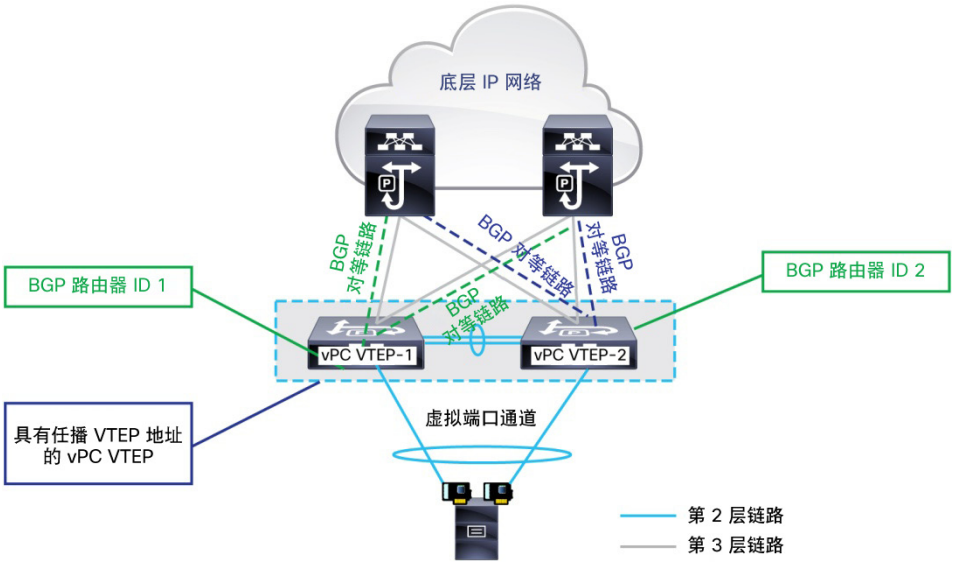
虚拟 Port-Channel (vPC) VTEP 将两项技术（vPC 和 VXLAN）结合起来，为 VTEP 提供设备级冗余。一对 vPC 交换机共享相同的 VTEP 地址，该地址通常被称为任播 VTEP 地址，并作为逻辑 VTEP。网络中的其他 VTEP 将这两台交换机看作具有任播 VTEP 地址的单台 VTEP。两台 vPC VTEP 交换机均启动并正常运行时，会以双活配置分担负载。如果一台 vPC 交换机发生故障，另一台交换机会接管整个流量负载，以便故障事件不会导致已连接到 vPC 对的设备丢失连接。

实施思科 NX-OS 中的 MP-BGP EVPN 控制平面是为了与 vPC VTEP 透明地协同工作。借助 MP-BGP EVPN 控制平面，vPC VTEP 继续作为具有任播 VTEP 地址的单个逻辑 VTEP 运行以便利用 VTEP 功能，但是从 MP-BGP 的角度看，它们作为两个单独的实体运行。它们具有不同的 BGP 路由器 ID，与 BGP 对等体单独建立 BGP 邻居邻接关系，并独立通告 EVPN 路由。在 EVPN 路由中，它们两者均将任播 VTEP 地址用作下一跳，以便远程 VTEP 可以使用学习到的 EVPN 路由并封装数据包（将任播 VTEP 地址用作所封装数据包的外部 IP 报头中的目标）。

EVPN vPC VTEP 配置

vPC VTEP 交换机配置为将环回接口上的辅助 IP 地址用作 VXLAN 隧道（接口 nve1）来源的 VTEP 地址。其余的 EVPN VXLAN 配置与标准的单个 VTEP 保持相同。两台交换机各自的 BGP 配置都需要具有独特的路由器 ID。图 11 说明了 MP-BGP EVPN vPC VTEP 的概念。在为 EVPN 路由构建 BGP 更新时，MP-BGP 将任播 VTEP 地址用作下一跳。

图 11. MP-BGP EVPN vPC VTEP



下面显示了一个 vPC VTEP 配置示例。

vPC VTEP-1 配置

```
interface nve1
 no shutdown
  source-interface loopback0
  host-reachability protocol bgp
  member vni 20000
    suppress-arp
    mcast-group 239.1.1.1
  member vni 20100
    suppress-arp
    mcast-group 239.1.1.2
```

这是 VXLAN 隧道

接口 loopback0 是 VXLAN 隧道的来源。

```
  member vni 39000 associate-vrf

interface loopback0
  ip address 10.1.1.13/32
  ip address 10.1.1.134/32 secondary
  ip ospf network point-to-point
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode

router bgp 100
  router-id 10.1.1.13
  log-neighbor-changes
  address-family ipv4 unicast
  address-family l2vpn evpn
  neighbor 10.1.1.1 remote-as 100
    update-source loopback0
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended
  neighbor 10.1.1.2 remote-as 100
    update-source loopback0
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended
  vrf evpn-tenant-1
    address-family ipv4 unicast
      advertise l2vpn evpn
evpn
  vni 20000 l2
    rd auto
    route-target import auto
    route-target export auto
  vni 20100 l2
    rd auto
    route-target import auto
    route-target export auto
vrf context evpn-tenant-1
  rd auto
  address-family ipv4 unicast
    route-target import 39000:39000
    route-target export 39000:39000
    route-target both auto evpn

n9396-vPC-VTEP-1#
```

此辅助 IP 地址用作任播 VTEP 地址。需要使用完全相同的任播 VTEP 地址配置两台 vPC VTEP。

BGP 实例都有其自己的路由器 ID：10.1.1.13

vPC VTEP-2 配置

```
interface nve1
  no shutdown
  source-interface loopback0
  host-reachability protocol bgp
  member vni 20000
    suppress-arp
    mcast-group 239.1.1.1
  member vni 20100
    suppress-arp
    mcast-group 239.1.1.2

  member vni 39010 associate-vrf


interface loopback0
  ip address 10.1.1.14/32
  ip address 10.1.1.134/32 secondary
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode

router bgp 100
  router-id 10.1.1.14
  log-neighbor-changes
  address-family ipv4 unicast
  address-family l2vpn evpn
  neighbor 10.1.1.1 remote-as 100
    update-source loopback0
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended
  neighbor 10.1.1.2 remote-as 100
    update-source loopback0
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended
  vrf evpn-tenant-1
    address-family ipv4 unicast
      advertise l2vpn evpn
evpn
  vni 20000 l2
    rd auto
    route-target import auto
```

这是 VXLAN 隧道接口。

接口 loopback0 是 VXLAN 隧道的来源。

此辅助 IP 地址用作任播 VTEP 地址。两台 vPC VTEP 要求配置完全相同的任播VTEP地址。

BGP 实例都有其自己的路由器 ID：10.1.1.14

```
    route-target export auto
  vni 20100 l2
    rd auto
    route-target import auto
    route-target export auto
vrf context evpn-tenant-1
  rd auto
  address-family ipv4 unicast
    route-target import 39000:39000
    route-target export 39000:39000
    route-target both auto evpn


n9396-vPC-VTEP-2#
```

vPC VTEP MP-BGP 状态和 EVPN 路由更新

对于其 MP-BGP 邻居，vPC VTEP 显示为两个单独的邻居。以下示例显示了来自 vPC VTEP 的 BGP 邻居的 **show bgp l2vpn evpn summary** 输出：

```
spine-9508-1# sh bgp l2vpn evpn summary
BGP summary information for VRF default, address family L2VPN EVPN
BGP router identifier 10.1.1.1, local AS number 100
BGP table version is 75, L2VPN EVPN config peers 4, capable peers 4
13 network entries and 13 paths using 1716 bytes of memory
BGP attribute entries [12/1728], BGP AS path entries [0/0]
BGP community entries [0/0], BGP clusterlist entries [0/0]

Neighbor        V    AS MsgRcvd MsgSent   TblVer  InQ OutQ Up/Down  State/PfxRcd
10.1.1.11       4   100    8247    8262       75    0    0    5d17h 6
10.1.1.12       4   100    8254    8259       75    0    0    1d08h 3
10.1.1.13       4   100    8258    8409       75    0    0    1d16h 2
10.1.1.14       4   100    8257    8455       75    0    0    1d16h 2
```

两台 vPC VTEP 显示为两个单独的 BGP 邻居。

两台 vPC VTEP 通告 EVPN 路由，同一任播 VTEP 地址作为 BGP 下一跳。下面显示了来自两台 vPC VTEP 的路由通告的示例。

在 VTEP-1 上

```
n9396-vPC-VTEP-1# sh bgp l2vpn evpn neighbors 10.1.1.1 advertised-routes

Peer 10.1.1.1 routes for address family L2VPN EVPN:
BGP table version is 94, local router ID is 10.1.1.13
Status: s-suppressed, x-deleted, S-stale, d-dampened, h-history, *-valid, >-best
Path type: i-internal, e-external, c-confed, l-local, a-aggregate, r-redist, I-
injected
```

```
Origin codes: i - IGP, e - EGP, ? - incomplete, | - multipath, & - backup

  Network      Next Hop      Metric   LocPrf   Weight Path
Route Distinguisher: 10.1.1.11:32967

Route Distinguisher: 10.1.1.11:32968

Route Distinguisher: 10.1.1.11:32977

Route Distinguisher: 10.1.1.12:2

Route Distinguisher: 10.1.1.12:6

Route Distinguisher: 10.1.1.13:32967    (L2VNI 20000)
*>l[2]:[0]:[0]:[48]:[0000.1330.e586]:[0]:[0.0.0.0]/216
                     10.1.1.134                          100      32768 i
*>l[2]:[0]:[0]:[48]:[0000.1330.e586]:[32]:[20.0.0.98]/272
                     10.1.1.134                          100      32768 i

Route Distinguisher: 10.1.1.13:32977    (L2VNI 21000)

Route Distinguisher: 10.1.1.14:32967

Route Distinguisher: 10.1.1.13:3    (L3VNI 39000)

n9396-vPC-VTEP-1#
```

下一跳是任播 VTEP 地址 10.1.1.134。

在 VTEP-2 上

```
n9396-vPC-VTEP-2# sh bgp l2vpn evpn neighbors 10.1.1.1 advertised-routes

Peer 10.1.1.1 routes for address family L2VPN EVPN:
BGP table version is 117, local router ID is 10.1.1.14
Status: s-suppressed, x-deleted, S-stale, d-dampened, h-history, *-valid, >-best
Path type: i-internal, e-external, c-confed, l-local, a-aggregate, r-redist, I-
injected
Origin codes: i - IGP, e - EGP, ? - incomplete, | - multipath, & - backup

   Network            Next Hop            Metric     LocPrf     Weight Path
Route Distinguisher: 10.1.1.11:32967

Route Distinguisher: 10.1.1.11:32968

Route Distinguisher: 10.1.1.11:32977

Route Distinguisher: 10.1.1.12:2
```

```
Route Distinguisher: 10.1.1.12:6

Route Distinguisher: 10.1.1.13:32967

Route Distinguisher: 10.1.1.14:32967    (L2VNI 20000)
*>l[2]:[0]:[0]:[48]:[0000.1330.e586]:[0]:[0.0.0.0]/216
                     10.1.1.134                          100      32768 i
*>l[2]:[0]:[0]:[48]:[0000.1330.e586]:[32]:[20.0.0.98]/272
                     10.1.1.134                          100      32768 i

Route Distinguisher: 10.1.1.14:32977    (L2VNI 21000)

Route Distinguisher: 10.1.1.14:3    (L3VNI 39000)

n9396-vPC-VTEP-2#
```

下一跳是任播 VTEP 地址 10.1.1.134。

在其他 VTEP 上学习 EVPN 路由，将任播 VTEP 作为下一跳。以下代码片段来自远程 VTEP 上的 **show bgp l2vpn evpn** 输出，该输出属于上例中通告的相同路由：

```
Route Distinguisher: 10.1.1.14:32967
* i[2]:[0]:[0]:[48]:[0000.1330.e586]:[0]:[0.0.0.0]/216
                     10.1.1.134                          100          0 i
*>i                  10.1.1.134                          100          0 i
*>i[2]:[0]:[0]:[48]:[0000.1330.e586]:[32]:[20.0.0.98]/272
                     10.1.1.134                          100          0 i
* i                  10.1.1.134                          100          0 i
```

MP-BGP EVPN VXLAN 交换矩阵设计

在部署新的可扩展数据中心网络时，越来越多的组织考虑采用两层的主干-枝叶交换矩阵架构（图 12）。两层的交换矩阵设计提供了网络发展所需的灵活性，以满足应用不断增长对连接密度和转发容量的需求。交换矩阵作为第 3 层网络运行，以利用现有第 3 层路由协议（例如开放最短路径优先 [OSPF]、BGP 和中间系统到中间系统 [IS-IS]）经过验证的稳定性和可扩展性。

图 12. 两层的主干-枝叶交换矩阵架构

对于第 3 层交换矩阵，第 2 层域包含在每台枝叶交换机下。对于假定在计算节点之间存在直接的第 2 层邻接关系的应用，这种模式可能限制工作负载放置。可以部署 VXLAN 以将第 2 层域扩展到第 3 层交换矩阵，从而实现工作负载放置的灵活性。本部分讨论了 VXLAN 交换矩阵的一些典型设计方案（将 MP-BGP EVPN 控制平面用于路由分配和多租户支持）。

MP-BGP EVPN 是 BGP 中的新地址系列，并且使用 BGP 中独立于该地址系列的机制。它不要求使用 iBGP 或 eBGP。这种灵活性使组织更容易从其当前的数据中心 BGP 设计过渡到 MP-BGP EVPN VXLAN 设计。该方法还在 BGP 自治系统编号 (ASN) 的分配方面提供了灵活性。本部分讨论了 MP-iBGP EVPN 设计和 MP-eBGP EVPN 设计。

采用 MP-iBGP EVPN 的 VXLAN 交换矩阵

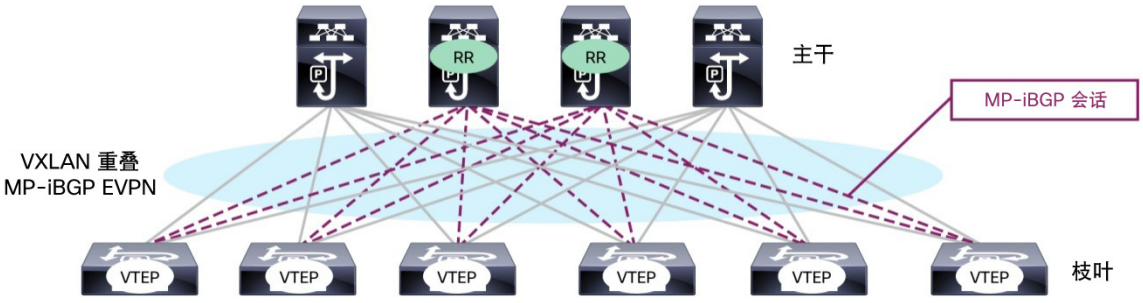
采用 MP-iBGP EVPN 设计时，所有 MP-BGP 扬声器均在同一 BGP 自治系统中。为了简化 iBGP 对等拓扑，iBGP 路由反射器通常部署在网络中。可以部署所选的 IGP 路由协议，以便在底层网络中为 VTEP 地址提供 IP 可达性。根据软件功能和可扩展性，可以将 iBGP 路由反射器部署在主干层或枝叶层上，也可以将其部署在专用的设备中以提高可扩展性。

主干层上的 MP-iBGP 路由反射器

在此设计中，枝叶交换机是 VTEP 设备。它们运行 MP-iBGP，并与主干交换机上运行的一对路由反射器对接。此设计要求所选的主干设备具有 MP-BGP EVPN 软件功能，但是它们不需要是 VTEP。

图 13 显示了在主干层上使用 iBGP 路由反射器 (RR) 的 MP-iBGP EVPN VXLAN 交换矩阵示例。在此设计中，每个 VTEP 枝叶有两个 iBGP 邻居，它们是两个主干 BGP 路由反射器。每个主干 BGP 路由反射器均将所有 VTEP 枝叶节点作为路由反射器客户端，并反射这些 VTEP 枝叶节点的 EVPN 路由。

图 13. 在主干层上使用路由反射器的 MP-iBGP EVPN VXLAN 交换矩阵设计



以下示例显示了此设计中的 VTEP 枝叶节点的 MP-iBGP 配置：

```
n9396-vtep-1# sh run bgp

!Command: show running-config bgp
!Time: Fri Jan 23 07:38:48 2015

version 7.0(3)I1(1)
feature bgp

router bgp 100
  router-id 10.1.1.11
  log-neighbor-changes
  address-family ipv4 unicast
  address-family l2vpn evpn
  neighbor 10.1.1.1 remote-as 100
    update-source loopback0
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended
  neighbor 10.1.1.2 remote-as 100
    update-source loopback0
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended

  vrf evpn-tenant-1
    address-family ipv4 unicast
      advertise l2vpn evpn
evpn
  vni 20000 l2
    rd auto
    route-target import auto
    route-target export auto
  vni 20100 l2
```

将两个主干 BGP 路由反射器配置为两个 iBGP 邻居。在每个邻居下，在地址系列 l2vpn evpn 中发送扩展社区。EVPN 路由使用扩展社区包含 EVPN 属性。

向地址系列 ipv4 单播通告 EVPN 路由。此步骤是可选的。如果此 VTEP 路由到某个外部设备（如广域网边缘路由器），并且需要将 EVPN 路由分发到外部，则需要执行此步骤。

```
    rd auto
    route-target import auto
    route-target export auto
  vni 21000 l2
    rd auto
    route-target import auto
    route-target export auto
  vni 21100 l2
    rd auto
    route-target import auto
    route-target export auto
vrf context evpn-tenant-1
  rd auto
  address-family ipv4 unicast
    route-target import 39000:39000
    route-target export 39000:39000
    route-target both auto evpn
vrf context evpn-tenant-2
  rd auto
  address-family ipv4 unicast
    route-target import 39010:39010
    route-target export 39010:39010
    route-target both auto evpn


n9396-vtep-1#
```

以下示例显示了主干 BGP 路由反射器的 MP-iBGP 配置：

```
feature bgp
nv overlay evpn
router bgp 100
  router-id 10.1.1.1
  log-neighbor-changes
  address-family ipv4 unicast
  address-family l2vpn evpn
    retain route-target all

 template peer vtep-peer
    remote-as 100
    update-source loopback0
    address-family ipv4 unicast
      send-community both
      route-reflector-client
    address-family l2vpn evpn
      send-community both
      route-reflector-client

  neighbor 10.1.1.11
    inherit peer vtep-peer
  neighbor 10.1.1.12
    inherit peer vtep-peer
  neighbor 10.1.1.13
    inherit peer vtep-peer
  neighbor 10.1.1.14
    inherit peer vtep-peer
```

启用 MP-BGP l2vpn evpn。

为 VXLAN EVPN 路由使用地址系列 l2vpn evpn。在从一个 iBGP 路由反射器客户端向其他客户端通告 EVPN 路由时，必须保留 route-target 属性。必须满足此要求，其他路由反射器客户端才能接收路由。

使用 iBGP RR 客户端对等体模板。

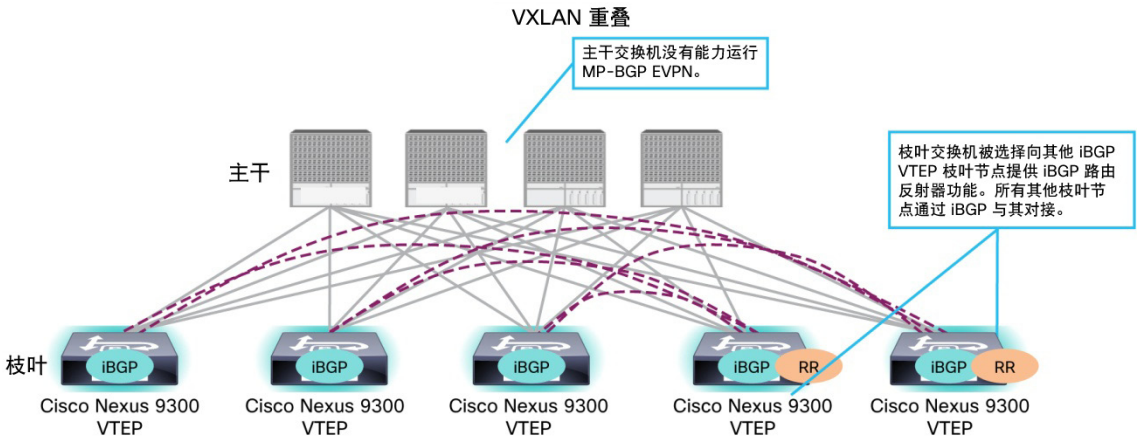
在地址系列 l2vpn evpn 中发送标准和扩展社区。

VTEP 枝叶节点是 iBGP 路由反射器客户端。

枝叶层上的 MP-iBGP 路由反射器

在主干层上部署 BGP 路由发射器是 MP-iBGP EVPN 的一种直观设计。它要求所选的主干设备支持 MP-iBGP EVPN 协议的软件功能，以便它们可以处理和分配 EVPN 路由的 MP-iBGP 更新。如果主干设备没有能力运行 MP-BGP EVPN，则 BGP 路由发射器功能需要迁移到枝叶层，该层的枝叶交换机支持 MP-BGP EVPN 和 VTEP 功能（图 14）。

图 14. 在枝叶层上使用 BGP 路由发射器功能的 MP-iBGP EVPN 交换矩阵设计



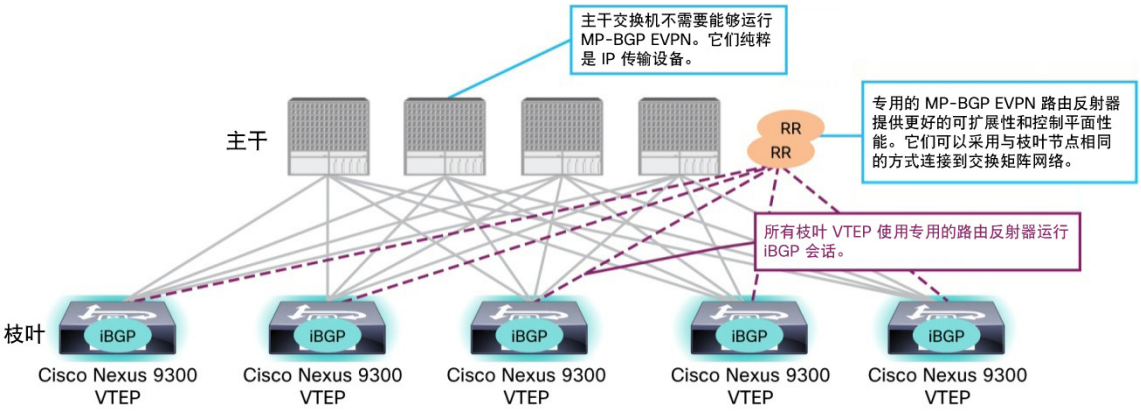
在此设计中，主干交换机完全不参与 MP-BGP EVPN 控制平面。它们运行底层网络路由协议，以确定 VTEP 地址和 iBGP 对等地址（如果它们与 VTEP 地址不相同：例如在 vPC VTEP 上）的 IP 可达性。

具有专用路由反射器的 MP-iBGP

EVPN 中的 MP-iBGP 路由反射器与标准的 iBGP 路由反射器具有相同的作用，即在 iBGP 对等体之间反射 BGP 更新，以便它们不需要构成全网状的 iBGP 对等拓扑。此方法可极大地简化 iBGP 拓扑，并使协议更具可扩展性。由于路由反射器功能完全是控制平面功能，因此 BGP 路由反射器不需要在数据平面转发路径中。使用此功能可在路由反射器部署和平台选择方面具有极大的灵活性。

可扩展设计的一种方案是将专用的设备用作数据路径之外的路由反射器（图 15）。所选的设备需要支持 MP-BGP EVPN，并且必须具有快速收敛所需的适当 BGP 控制平面可扩展性和计算能力。使用专用的路由反射器可消除主干层中的 MP-BGP EVPN 功能要求。它还消除了 VTEP 枝叶节点的负担，使其不必在执行数据转发之外还运行 BGP 路由反射器功能。虽然 VTEP 枝叶节点在逻辑上与路由反射器具有直接的 iBGP 邻居邻接关系，但是路由反射器可以采用与枝叶节点相同的方式在物理上连接到 VXLAN 交换矩阵网络，并在 VTEP 枝叶与路由反射器之间进行 iBGP 会话，以便经过交换矩阵底层网络中的多跳（通常为 2）。需要应用路由注意事项，以便 VTEP 地址之间的底层数据路径不经过路由反射器。此要求可帮助确保路由反射器在数据转发路径之外。

图 15. 具有专用路由反射器的 MP-iBGP EVPN 设计



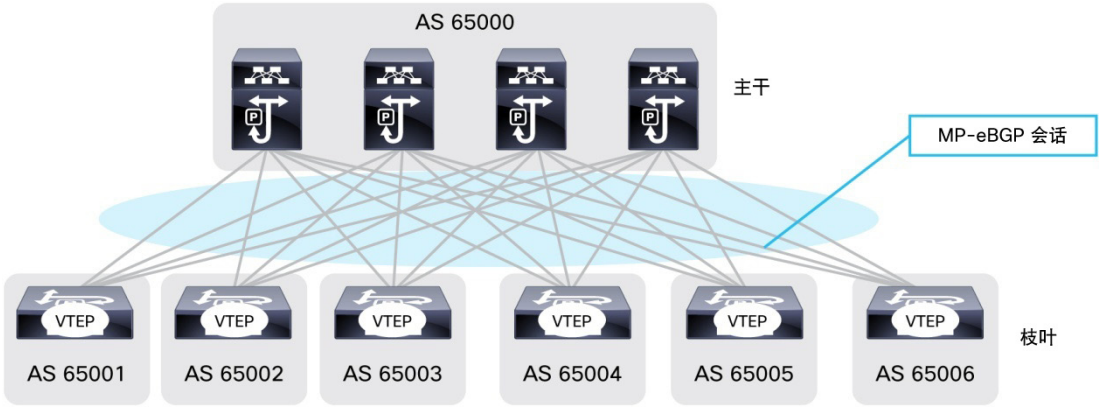
采用 MP-eBGP EVPN 的 VXLAN 交换矩阵

虽然 MP-iBGP EVPN 设计是常见做法，但是某些组织选择在枝叶与主干层之间运行 eBGP。MP-BGP EVPN 可以灵活地同时使用 iBGP 和 eBGP。使用 MP-eBGP 对等连接的 EVPN 是一种可行的设计方案。eBGP 设计为 BGP 自治系统 (AS) 分配提供了多个选项。图 16 显示了一种设计，每个 VTEP 枝叶在其各自独特的 BGP AS 中；图 17 显示了另一种设计，在该设计中，所有 VTEP 枝叶节点位于同一 AS 中，但是它们通过 eBGP 与主干交换机对接。

由于 MP-BGP EVPN 是 BGP 的扩展，因此它继承标准的 BGP 行为。在 MP-BGP EVPN 网络中，不需要某些默认行为。例如，当 BGP 路由器向 eBGP 对等体通告 BGP 路由时，默认情况下它将 BGP 下一跳更改为它自己的 IP 地址。在 MP-BGP EVPN 中，当 VTEP 启动 BGP 更新以通告其 EVPN 路由时，它会将其自己的 VTEP 地址用作 BGP 下一跳。该下一跳在逐跳的 BGP 路由分配中必须保持，以便其他 VTEP 可以接收将原始 VTEP 地址作为下一跳的 EVPN 路由，并且可以使用此路由在数据平面启动 VXLAN 隧道。

因此，需要配置主干交换机上的 eBGP，以使其不会更改 BGP 下一跳。BGP 路由器还可能在发送 eBGP 路由时修改 BGP 社区属性。在 MP-EVPN 中，此更改可能导致修改或删除 EVPN 路由中的 route-target 属性。因此，需要在中间 eBGP 对等体上应用其他配置，以帮助确保它们保留所有 route-target 属性。

图 16. VTEP 枝叶节点位于独特的自治系统中的 MP-eBGP EVPN VXLAN 交换矩阵



在此设计中，由于每个 VTEP 具有独特的 BGP AS，因此 NX-OS 中的 route-target 自动生成会导致同一 VNI 的 VTEP 上有不同的 route-target。建议手动配置导入和导出路由目标，以确保 VTEP 对于相同的第 3 层 VRF 实例和相同的 EVPN 第 3 层 VNI 具有相同的路由目标配置。

以下示例显示了主干交换机和 VTEP 枝叶的 MP-BGP 配置，如图 16 所示。主干交换机的 MP-BGP 配置包括对主干交换机应用出站策略，以使其不会更改 eBGP 路由下一跳。该示例还显示了第 3 层 VRF 实例和 EVPN 第 2 层 VNI 的 VTEP 枝叶的手动 route-target 配置。

```
[BGP configuration on a spine switch as in Figure 16 design]
route-map permit-all permit 10
  set ip next-hop unchanged

router bgp 65000
  router-id 10.1.1.1
  address-family ipv4 unicast
    redistribute direct route-map permitall
  address-family l2vpn evpn
    nexthop route-map permit-all
    retain route-target all
  neighbor 192.167.11.2 remote-as 65001
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended
      route-map permit-all out
  neighbor 192.168.12.2 remote-as 65002
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended
      route-map permit-all out
```

将下一跳策略设置为不更改下一跳属性。

在向 eBGP 对等体通告 EVPN BGP 路由时，保留路由及所有路由目标。

将出站策略设置为向此 eBGP 邻居通告所有路由。

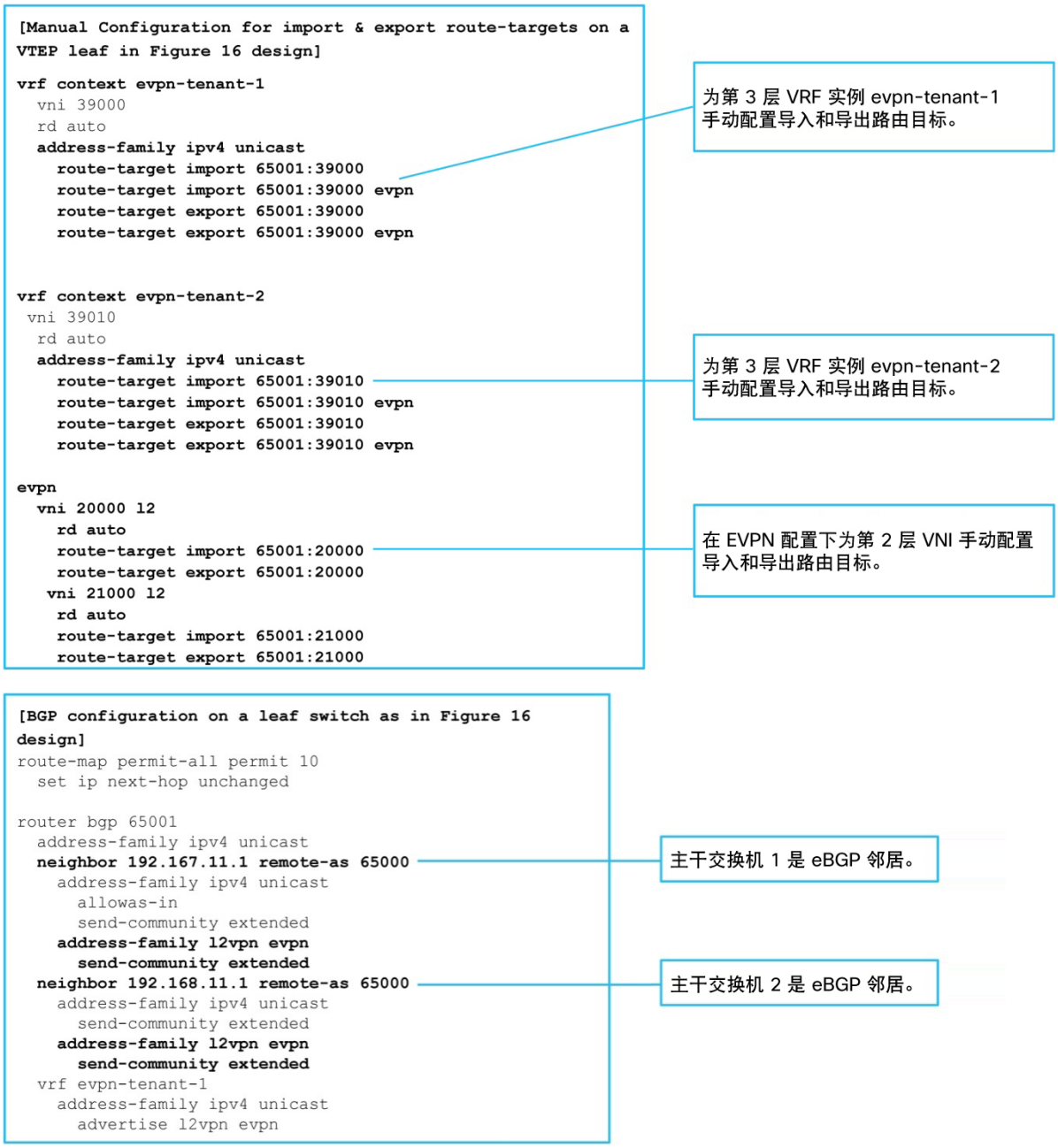
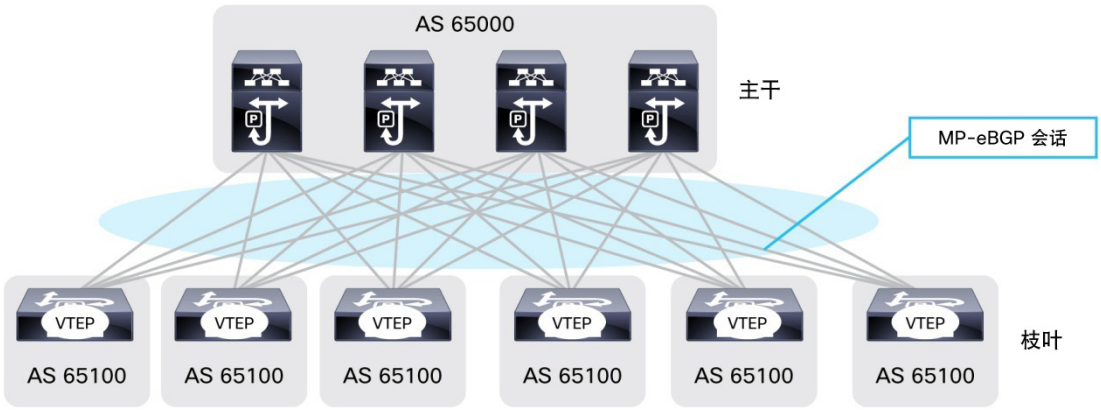


图 17 描绘了所有枝叶节点均在同一自治系统中的 MP-eBGP 设计，但是这些节点各自通过 MP-eBGP 与主干节点对接。

图 17. VTEP 枝叶节点均在同一自治系统中的 MP-eBGP 设计



以下示例显示了如图 17 所示的 VTEP 枝叶和主干交换机设计的配置。除了图 16 设计中的配置外，图 17 中的主干交换机需要禁用 **peer-as-check**，因为它们需要在两个位于同一 BGP 自治系统中的 eBGP 邻居之间传送 MP-BGP EVPN 路由。图 17 中的 VTEP 枝叶节点需要启用 **allowas-in**，以便它们从位于同一 BGP 自治系统中的其他 VTEP 接受 BGP 路由。在此设计中，由于所有 VTEP 枝叶位于同一 BGP 自治系统中，因此为第 3 层 VRF 实例和 EVPN 第 2 层 VNI 使用系统自动生成的导入和导出路由目标是合适的。

```
[BGP configuration on a spine switch as in Figure 17 design]
route-map permit-all permit 10
  set ip next-hop unchanged

router bgp 65000
  router-id 10.1.1.1
  address-family ipv4 unicast
    redistribute direct route-map permit-all
  address-family l2vpn evpn
    nexthop route-map permit-all
    retain route-target all
  neighbor 192.167.11.2 remote-as 65100
    address-family ipv4 unicast
    address-family l2vpn evpn
      disable-peer-as-check
      send-community extended
      route-map permit-all out
  neighbor 192.168.12.2 remote-as 65100
    address-family ipv4 unicast
    address-family l2vpn evpn
      disable-peer-as-check
      send-community extended
      route-map permit-all out
```

将下一跳策略设置为不更改下一跳属性。

在向 eBGP 对等体通告 EVPN BGP 路由时，保留所有 route-target 属性。

VTEP 枝叶是 eBGP 对等体。所有 VTEP 位于同一 BGP 自治系统中：AS 65100。

对此邻居禁用 peer-as-check。

将出站策略设置为向此 eBGP 邻居通告所有路由。

VTEP 枝叶是 eBGP 对等体。所有VTEP在同一BGP自治系统：AS 65100。

```
[BGP configuration on a leaf switch in Figure 17 design]
route-map permit-all permit 10
  set ip next-hop unchanged

router bgp 65001
  address-family ipv4 unicast
  neighbor 192.167.11.1 remote-as 65000
    address-family ipv4 unicast
      allowas-in
      send-community extended
    address-family l2vpn evpn
      allowas-in
      send-community extended
  neighbor 192.168.11.1 remote-as 65000
    address-family ipv4 unicast
      allowas-in
      send-community extended
    address-family l2vpn evpn
      allowas-in
      send-community extended
  vrf evpn-tenant-1
    address-family ipv4 unicast
      advertise l2vpn evpn
```

主干交换机 1 是 eBGP 邻居。

允许本地自治系统在此邻居的自治系统路径中的 BGP 路由。

主干交换机 2 是 eBGP 邻居。

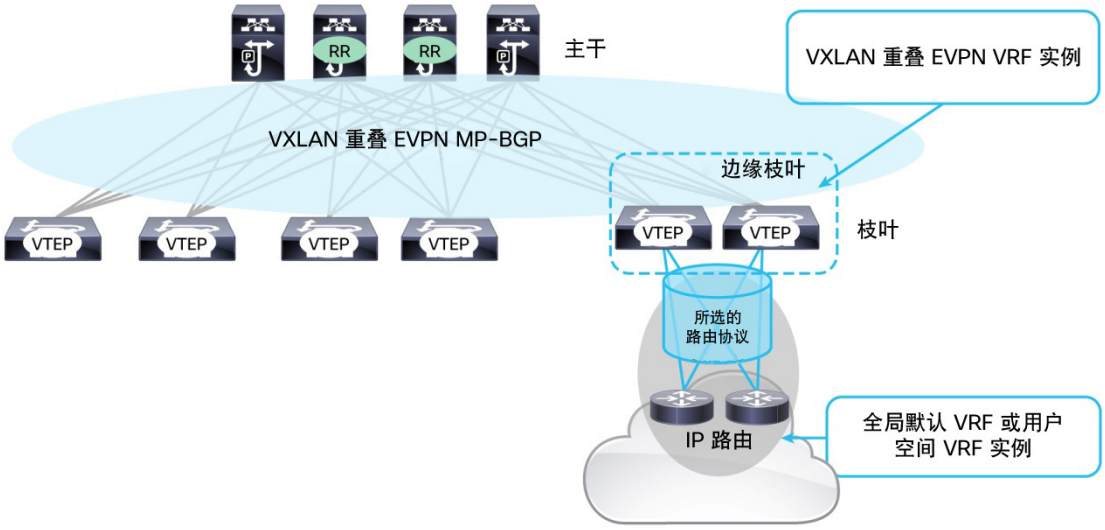
允许本地自治系统在此邻居的自治系统路径中的 BGP 路由。

MP-BGP EVPN VXLAN 的外部路由

在大多数组织中，数据中心未与网络的其余部分（包括园区网络、广域网和互联网）隔离。当 EVPN VXLAN 交换矩阵部署在数据中心中时，需要与位于 VXLAN 交换矩阵外部的这些网络保持连接。

采用标准的主干和枝叶交换矩阵架构时，可以通过使用边界枝叶节点连接到外部路由设备来实现外部连接。图 18 显示了具有一对边界枝叶交换机的此类设计。

图 18. 用于 MP-BGP EVPN VXLAN 交换矩阵的外部路由的边界枝叶交换机



边界枝叶交换机运行内部的 MP-BGP EVPN（其他 VTEP 在 VXLAN 交换矩阵中），并与这些 VTEP 交换 EVPN 路由。同时，它在租户 RRF 实例中运行正常的 IPv4 或 IPv6 单播路由（外部路由设备在外部）。路由协议可以是常规的 eBGP 或所选的任何 IGP。根据设计，MP-BGP EVPN 将在 IPv4 或 IPv6 单播地址系列中学习的 BGP 路由自动导入到 L2VPN EVPN 地址系列。

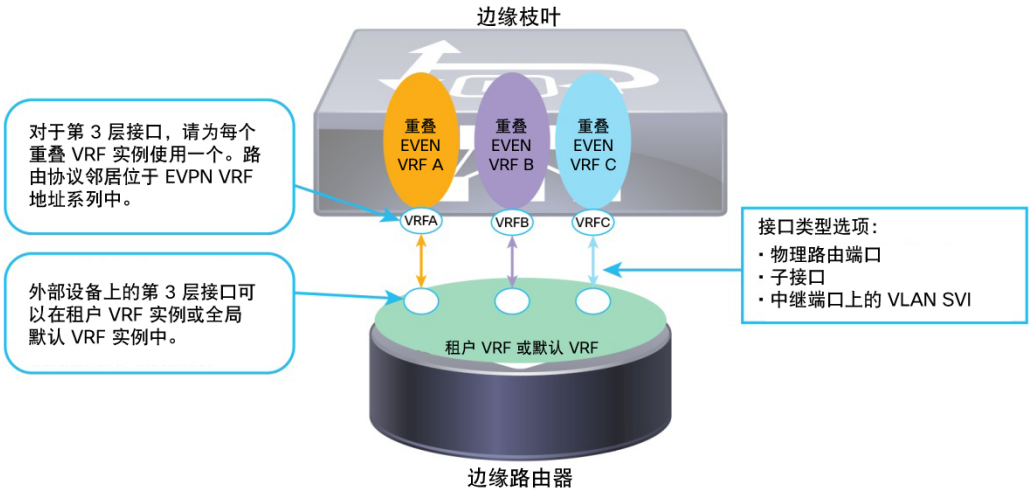
因此，在边界枝叶交换机学习外部路由后，它可以将这些路由作为 EVPN 路由向 EVPN 域通告，以便其他 VTEP 枝叶节点还可以学习用于发送出站流量的外部路由。边界枝叶交换机还可以配置为将在 L2VPN EVPN 地址系列中学习的 EVPN 路由发送到 IPv4 或 IPv6 单播地址系列，并向外部路由设备通告这些路由。因此，如果 VXLAN 交换矩阵中存在任何公共子网，可以向外部通告这些子网，以便从外部至这些公共子网的入站流量可以路由到 VXLAN 交换矩阵。

由于 MP-BGP EVPN 具有内置的多租户，因此 VXLAN 重叠网络中的第 3 层子网在租户 VRF 路由实例中。默认情况下，不同租户可以维护各自独立的第 3 层路由实例。因此，需要单独提供不同租户的外部路由。对于边界枝叶为其运行外部路由的每个租户 VRF 实例，边界枝叶需要具有连接外部的第 3 层接口（图 19）。

要将不同租户之间的此类第 3 层路由分段扩展到外部网络，外部路由器还可以将边界枝叶的第 3 层接口放置在租户 VRF 实例中。边界枝叶与外部路由器之间的路由会话将在 VRF-lite 的两侧运行。

在 VXLAN 边界枝叶上终止第 3 层分段的设计中，外部路由器可以在默认路由表中运行所有路由会话。在这种情况下，来自 VXLAN 交换矩阵中的不同租户路由实例的路由将合并到外部的同一默认路由表中。在此类型的设计中，由于租户实质上共享外部路由，因此 VXLAN 租户的 IP 地址不能重叠。

图 19. 采用多租户的 MP-BGP EVPN VXLAN 交换矩阵外部路由



VXLAN EVPN 边界枝叶与外部路由器之间的 eBGP 的配置示例

以下是 VXLAN 边界枝叶与外部路由器之间的 eBGP 路由的配置示例。eBGP 会话在边界枝叶上的租户 VRF 实例中，但是在外部路由器的默认路由表中，以便共享外部路由。

在边界枝叶上，BGP 配置为通告 VXLAN IP 子网前缀。默认情况下，BGP 通告 MP-BGP EVPN IP 主机路由。示例配置中应用路由过滤以阻止 /32 IP 主机路由，以便仅向外部路由器通告前缀路由。由于外部不需要入站流量的特定主机路由，因此使用该方法可以为外部路由提高路由器可扩展性。

在 VXLAN 边界枝叶上：

```
router bgp 100
  router-id 10.1.1.16
  log-neighbor-changes
  address-family ipv4 unicast
  address-family l2vpn evpn
  neighbor 10.1.1.1 remote-as 100
    update-source loopback0
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended
  neighbor 10.1.1.2 remote-as 100
    update-source loopback0
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended
  vrf evpn-tenant-1
    address-family ipv4 unicast
      network 20.0.0.0/24
    neighbor 30.10.1.2 remote-as 200
      address-family ipv4 unicast
       prefix-list outbound-no-hosts out
ip prefix-list outbound-no-hosts seq 5 deny 0.0.0.0/0 eq 32
ip prefix-list outbound-no-hosts seq 10 permit 0.0.0.0/0 le 32
```

eBGP 邻居位于外部。它在租户 VRF 路由实例的地址系列 ipv4 单播中。

为了提高可扩展性，请应用前缀列表以过滤掉 /32 IP 主机路由。仅向外部 eBGP 邻居通告前缀路由。

外部路由器的 BCP 配置：

```
router bgp 200
  router-id 10.1.1.18
  log-neighbor-changes
  address-family ipv4 unicast
    network 100.0.0.0/24
    network 100.0.1.0/24
  neighbor 30.10.1.1 remote-as 100
    address-family ipv4 unicast
```

在上例中，通过 VRF-lite eBGP 向外部路由器通告 VNI 子网路由 20.0.0.0/24，如下面的全局路由表所示：

```
N9372TX-2-ext# sh ip bgp 20.0.0.0/24
BGP routing table information for VRF default, address family IPv4 Unicast
BGP routing table entry for 20.0.0.0/24, version 36
Paths: (1 available, best #1)
Flags: (0x00001a) on xmit-list, is in urib, is best urib route

  Advertised path-id 1
  Path type: external, path is valid, is best path, no labeled nexthop
  AS-Path: 100 , path sourced external to AS
    30.10.1.1 (metric 0) from 30.10.1.1 (20.0.0.1)
      Origin IGP, MED not set, localpref 100, weight 0

  Path-id 1 not advertised to any peer
N9372TX-2-ext#
N9372TX-2-ext# sh ip route 20.0.0.0/24
IP Route Table for VRF "default"

20.0.0.0/24, ubest/mbest: 1/0
    *via 30.10.1.1, [20/0], 1w2d, bgp-200, external, tag 100
N9372TX-2-ext#
```

从外部路由器学习到的路由将由边界枝叶通过 MP-BGP EVPN 协议分配给 VXLAN 交换矩阵。以下示例显示了在内部 VTEP 上捕获外部路由。VTEP 通过路由反射器从边界枝叶学习外部路由。路由将通过 MP-BGP EVPN 分配。

```
n9396-vtep-1# sh vrf evpn-tenant-1 detail
VRF-Name: evpn-tenant-1, VRF-ID: 3, State: Up
    VPNID: unknown
    RD: 10.1.1.11:3
    VNI: 39000
    Max Routes: 0  Mid-Threshold: 0
    Table-ID: 0x80000003, AF: IPv6, Fwd-ID: 0x80000003, State: Up
    Table-ID: 0x00000003, AF: IPv4, Fwd-ID: 0x00000003, State: Up

n9396-vtep-1#

n9396-vtep-1# sh bgp l2vpn evpn rd 10.1.1.11:3 100.0.0.0
BGP routing table information for VRF default, address family L2VPN EVPN
Route Distinguisher: 10.1.1.11:3    (L3VNI 39000)
BGP routing table entry for [5]:[0]:[0]:[24]:[100.0.0.0]:[0.0.0.0]/224, version 324
Paths: (1 available, best #1)
Flags: (0x00001a) on xmit-list, is in l2rib/evpn

  Advertised path-id 1
  Path type: internal, path is valid, is best path, no labeled nexthop
             Imported from 10.1.1.16:3:[5]:[0]:[0]:[24]:[100.0.0.0]:[0.0.0.0]/120
  AS-Path: NONE, path sourced internal to AS
    10.1.1.16 (metric 3) from 10.1.1.1 (10.1.1.1)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 39000
      Extcommunity:  RT:100:39000 ENCAP:8 Router MAC:6412.2574.6ae7
      Originator: 10.1.1.16 Cluster list: 10.1.1.1

  Path-id 1 not advertised to any peer

n9396-vtep-1#
```

外部路由通过 EVPN 分发并导入到租户 VRF 实例。

```
n9396-vtep-1# sh ip bgp vrf evpn-tenant-1 100.0.0.0
BGP routing table information for VRF evpn-tenant-1, address family IPv4 Unicast
BGP routing table entry for 100.0.0.0/24, version 70
Paths: (1 available, best #1)
Flags: (0x08041a) on xmit-list, is in urib, is best urib route
  vpn: version 75, (0x100002) on xmit-list

  Advertised path-id 1, VPN AF advertised path-id 1
  Path type: internal, path is valid, is best path, no labeled nexthop
             Imported from unknown dest
  AS-Path: NONE, path sourced internal to AS
    10.1.1.16 (metric 3) from 10.1.1.1 (10.1.1.1)
      Origin IGP, MED not set, localpref 100, weight 0
      Received label 39000
      Extcommunity:  RT:100:39000 ENCAP:8 Router MAC:6412.2574.6ae7
      Originator: 10.1.1.16 Cluster list: 10.1.1.1

  VRF advertise information:
  Path-id 1 not advertised to any peer

  VPN AF advertise information:
  Path-id 1 not advertised to any peer
n9396-vtep-1#
n9396-vtep-1# sh ip route vrf evpn-tenant-1 100.0.0.0/24
IP Route Table for VRF "evpn-tenant-1"
'*' denotes best ucast next-hop
'**' denotes best mcast next-hop
'[x/y]' denotes [preference/metric]
'%<string>' in via output denotes VRF <string>

100.0.0.0/24, ubest/mbest: 1/0
    *via 10.1.1.16%default, [200/0], 01:01:14, bgp-100, internal, tag 100 (evpn)segid: 0x9858 tunnelid:
0xa010110 encap: 1

n9396-vtep-1#
```

这是外部路由。

下一跳是边界枝叶的 VTEP 地址。

租户是 VRF L3 VNI。

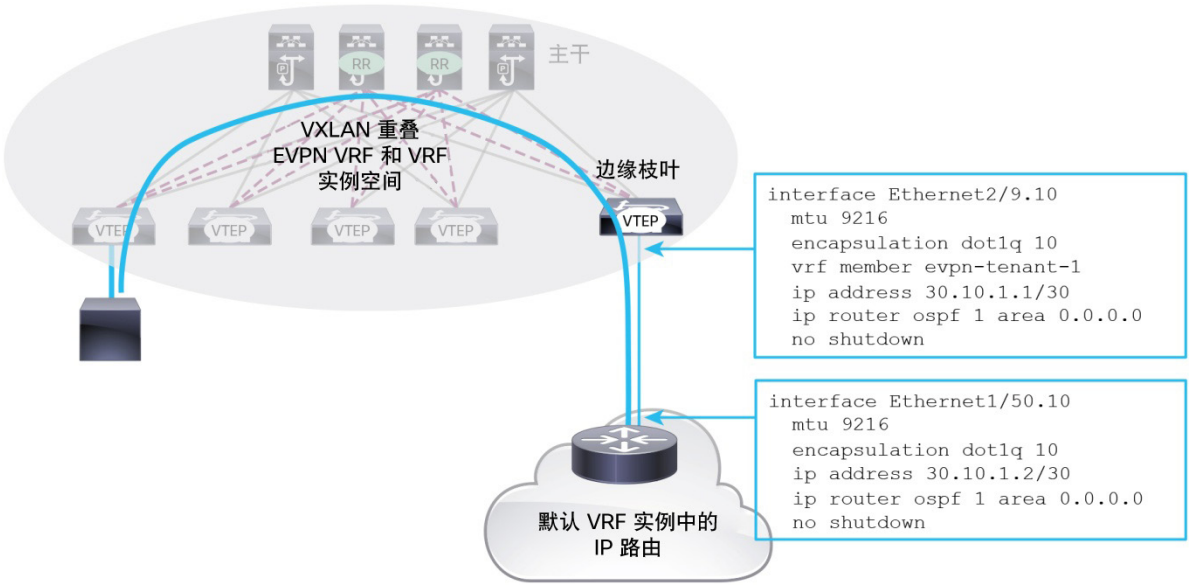
10.1.1.16 是边界枝叶的 BGP 路由器 ID。
10.1.1.1 是主干路由反射器。

这是 IBGP 路由。下一跳是边界枝叶的 VTEP 地址。

VXLAN EVPN 边界枝叶与外部路由器之间的 OSPF 的配置示例

图 20 中的示例将 OSPF 用作 EVPN VXLAN 边界枝叶上的外部路由协议，以便与外部交换路由。为了实现多租户，该示例使用子接口在边界枝叶与外部路由器之间进行路由。使用子接口，多个租户可以共享外部路由的相同物理链路，边界枝叶上的每个 VRF 路由实例有一个子接口。在本例中，外部路由器上的路由在默认 VRF 实例中。您还可以在外部设备上扩展租户 VRF 实例，方法是在外部设备上配置 VRF-Lite 子接口。

图 20. 采用 OSPF 的 EVPN VXLAN 外部路由



边界枝叶的相关配置如下所示：

```
ip prefix-list bgp-ospf-no-hosts seq 5 permit 0.0.0.0/0 eq 32
route-map permit-bgp-ospf deny 5
  match ip address prefix-list bgp-ospf-no-hosts
route-map permit-bgp-ospf permit 10
route-map permit-ospf-bgp permit 10

router ospf 1
  router-id 10.1.1.16
  vrf evpn-tenant-1
    redistribute bgp 100 route-map permit-bgp-ospf

router bgp 100
  router-id 10.1.1.16
  log-neighbor-changes
  address-family ipv4 unicast
  address-family l2vpn evpn
    retain route-target all
  neighbor 10.1.1.1 remote-as 100
    update-source loopback0
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended
  neighbor 10.1.1.2 remote-as 100
    update-source loopback0
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community extended
  vrf evpn-tenant-1
    address-family ipv4 unicast
      advertise l2vpn evpn
      redistribute ospf 1 route-map permit-ospf-bgp
```

将 BGP 路由重新分发到 OSPF。过滤掉/32 IP 主机路由。

当 BGP 路由器是自治系统边界路由器时，该路由器将在 l2vpn evpn 路由中修改路由目标。必须保留原始路由目标。

将 OSPF 重新分发到 BGR。向 L2VPN EVPN 通告重新分发的路由。

在此设计中，边界枝叶通过租户 VRF 实例中的 OSPF 学习外部路由。它将路由重新分发到 VRF 实例内的 MP-BGP，然后通过 MP-BGP L2VPN EVPN 向内部 VTEP 通告路由。

以下示例显示了边界枝叶的外部路由分发：

```
n9396-border-leaf# sh ip route 100.0.0.0/24 vrf evpn-tenant-1
IP Route Table for VRF "evpn-tenant-1"

100.0.0.0/24, ubest/mbest: 1/0
    *via 30.10.1.2, Eth2/9.10, [110/2], 01:43:07, ospf-1, intra

n9396-border-leaf# sh bgp l2vpn evpn 100.0.0.0 vrf evpn-tenant-1
BGP routing table information for VRF default, address family L2VPN EVPN
Route Distinguisher: 10.1.1.16:3    (L3VNI 39000)
BGP routing table entry for [5]:[0]:[0]:[24]:[100.0.0.0]:[0.0.0.0]/224, version 325
Paths: (1 available, best #1)
Flags: (0x00000a) on xmit-list, is not in l2rib/evpn

  Advertised path-id 1
  Path type: local, path is valid, is best path, no labeled nexthop
  AS-Path: NONE, path locally originated
    10.1.1.16 (metric 0) from 0.0.0.0 (10.1.1.16)
      Origin IGP, MED not set, localpref 100, weight 32768
      Received label 39000
      Extcommunity:  RT:100:39000

  Path-id 1 advertised to peers:
    10.1.1.1          10.1.1.2

n9396-border-leaf#
```

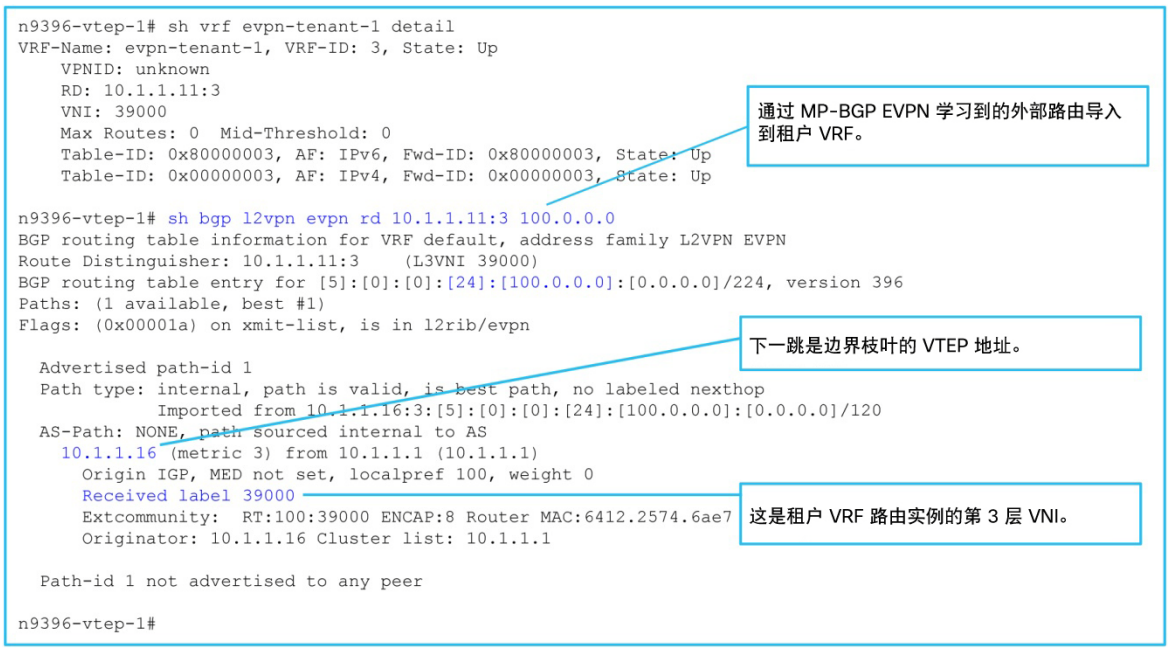
这是通过租户 VRF 中的 OSPF 学习到的外部路由。

外部 OSPF 路由重新分发到 BGP，并通过 MP-BGP L2VPN EVPN 分发到其他 VTEP。

BGP 下一跳是边界枝叶的 VTEP 地址。

MP-BGP EVPN 路由通告给 BGP 对等体。

内部 VTEP 通过 MP-BGP EVPN 学习外部路由：



EVPN VXLAN 边界枝叶节点的可扩展性注意事项

VXLAN 边界枝叶节点是 VXLAN 交换矩阵网络到外部的连接点。这些节点学习外部路由并通过 MP-BGP EVPN 将其重新分发到其他 VTEP。同时，它们向外部通告位于 VXLAN 交换矩阵上的公共子网。

将外部路由分发到 EVPN VXLAN 交换矩阵

边界枝叶可能从外部接收大量的外部路由。由于边界枝叶节点通常是交换矩阵的内部设备的出口网关，因此所有外部路由可能不需要分发到交换矩阵。相反，您可能需要先汇总路由，然后再向 MP-BGP EVPN 通告路由。在某些情况下，按每个租户向交换矩阵通告默认路由可能已足够。减少分发的外部路由数量有助于确保内部 VTEP 设备不会耗尽最长前缀匹配 (LPM) 路由表资源。这种方法还可减轻内部 VTEP 的 MP-BGP EVPN 控制平面负担，从而改善控制平面性能。

向外部发送的 EVPN VXLAN 交换矩阵内部网络通告

需要可从外部访问 EVPN VXLAN 重叠网络中的某些第 3 层子网。边界枝叶节点需要通告这些公共子网的第 3 层可达性信息。MP-BGP EVPN 可以分发内部的 IP 主机路由和外部的子网前缀路由。在边界枝叶与外部路由器之间的路由协议会话中，您可以应用过滤器以避免将内部 IP 主机路由发送到外部。在大多数情况下，外部网络需要使用公共子网的 LPM 前缀路由将流量发送到 VXLAN 交换矩阵。

边界枝叶节点的 EVPN 租户可扩展性

边界枝叶为 VXLAN 重叠网络中的租户提供外部连接。这些租户需要参与它们用作边界枝叶节点的所有租户 VRF 路由实例。在构建大规模多租户设计时，请遵守有关边界枝叶可以支持的 EVPN 第 3 层 VRF 实例最大数量的要求。

边界枝叶节点的 IP 主机路由可扩展性

为了实现对发往内部终端主机的入站流量的最佳转发，边界枝叶需要为租户公共子网中的终端主机执行基于 IP 主机的路由。此要求意味着边界枝叶需要在 IP 主机路由的硬件转发表中学习和设定主机路由。IP 主机表大小决定了可以在租户公共子网中出现的终端主机总数。

MP-BGP EVPN VXLAN 的数据中心互联

虽然重叠传输虚拟化 (OTV) 和虚拟专用局域网服务 (VPLS) 仍是最成熟的第 2 层数据中心互联 (DCI) 解决方案，但是采用 MP-BGP EVPN 控制平面的 VXLAN 可以在某些部署条件下提供一种替代方案。当 VXLAN 部署在数据中心时，将其用于数据中心之间的互联可以简化整体网络设计，并降低操作复杂性，从而为数据中心内和数据中心之间的流量提供一个统一的网络重叠解决方案。

图 21 说明了一个简单的数据中心和采用 MP-BGP EVPN VXLAN 的 DCI 设计。在此设计中，每个数据中心维护其自己的 BGP 自治系统，并部署 EVPN VXLAN 交换矩阵（运行具有路由反射器的 MP-iBGP），以提高简便性和可扩展性。在数据中心之间，DCI 边界枝叶节点互相运行多跳 MP-eBGP EVPN。其结果是两个数据中心连接在一起，形成一个统一的 MP-BGP EVPN 路由域。在控制平面中，通过数据中心之间的 iBGP-eBGP-iBGP 路径分发 EVPN 路由。在数据平面中，当数据中心 A 中的一台终端主机将流量发送到数据中心 B 中的另一台主机时，数据包经过一个 VXLAN 隧道并由数据中心 A 中的入口 VTEP 封装，然后由数据中心 B 中的出口 VTEP 解装。此方法在重叠网络中提供高度有效的 DCI 数据转发。

图 21. 采用统一 MP-BGP EVPN 管理域的 DCI 解决方案

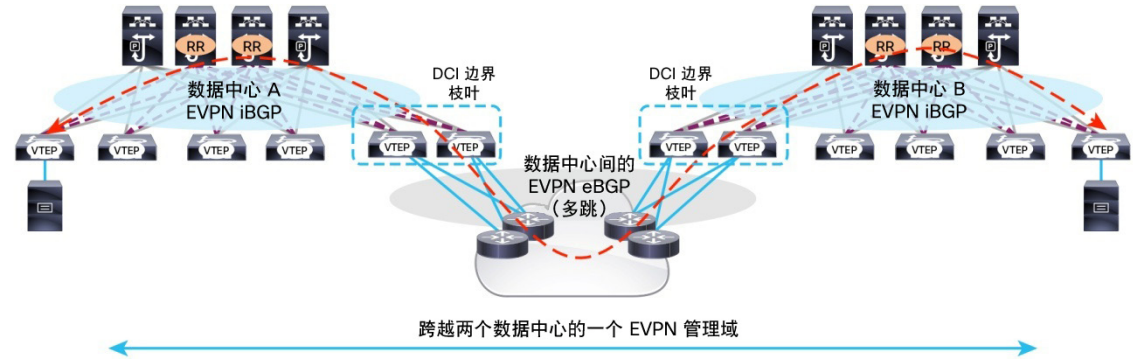
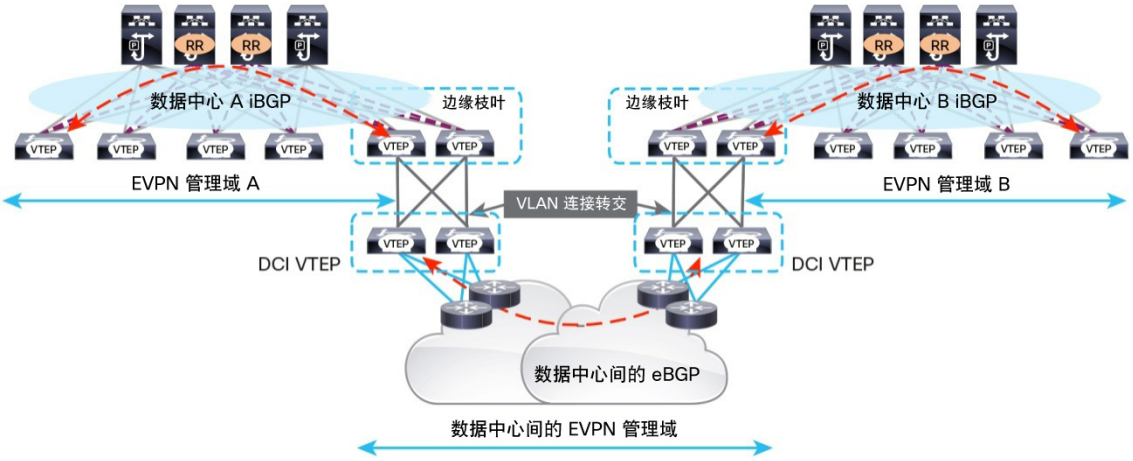


图 22 显示了采用 MP-BGP EVPN 的另一种 DCI 设计。它为每个数据中心提供单独的 MP-iBGP EVPN 域，并通过 DCI VTEP 之间的 MP-eBGP EVPN 跨数据中心域将数据中心连接在一起。如果未直接连接 VTEP，则 DCI VTEP 之间的 MP-eBGP 会话必须是多跳。使用此设计可在每个数据中心灵活部署不同的 EVPN 操作和功能模型。它还使数据中心在数据中心内的 VTEP 对等连接方面具有更大的可扩展性，因为每个数据中心具有其自己的原子 EVPN 域设计。

图 22. 采用单独 MP-BGP EVPN 管理域的 DCI



结论

MP-BGP EVPN 改变了 VXLAN 重叠网络的模式。它引入了控制平面学习，以在任意规模的网络中提供信令统一的转发数据库，而不是依赖泛洪和学习。MP-BGP EVPN 基于一个行业标准草案，以及多个供应商和运营商的协作，他们致力于共同开发一项简单且可互操作的技术。它为重叠网络提供集成的路由和桥接，以优化流量的传送。借助思科 NX-OS 软件的 MP-BGP EVPN 功能和 Cisco Nexus 9000 系列硬件的 VXLAN 路由功能，您可以使用 Cisco Nexus 9000 系列交换机构建具有高可扩展性、强大、高性能的 VXLAN 重叠交换矩阵网络。

相关详细信息

- IETF 草案 - 基于 BGP MPLS 的以太网 VPN：
  https://tools.ietf.org/html/draft-ietf-l2vpn-evpn-11
- IETF 草案 - 采用 EVPN 的网络虚拟化重叠解决方案：
  https://tools.ietf.org/html/draft-ietf-bess-evpn-overlay-00
- IETF 草案 - EVPN 中的集成路由和桥接：
  https://tools.ietf.org/html/draft-ietf-bess-evpn-inter-subnet-forwarding-00
- IETF 草案 - EVPN 中的 IP 前缀通告：
  https://tools.ietf.org/html/draft-rabadan-l2vpn-evpn-prefix-advertisement-02
- RFC 4271 - 边界网关协议 4 (BGP-4)：
  https://tools.ietf.org/html/rfc4271
- RFC 4760 - BGP-4 的多协议扩展：
  https://tools.ietf.org/html/rfc4760
- RFC 4364 - BGP/MPLS IP VPN：
  https://tools.ietf.org/html/rfc4364#page-15
- VXLAN 概述 - Cisco Nexus 9000 系列交换机：
  http://www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/white-paper-c11-729383.html
- Cisco Nexus 9300 平台交换机的 VXLAN 设计：
  http://www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/white-paper-c11-732453.html

美洲总部
Cisco Systems, Inc.
加州圣何西

亚太地区总部
Cisco Systems (USA) Pte.Ltd.
新加坡

欧洲总部
Cisco Systems International BV
荷兰阿姆斯特丹

思科在全球设有 200 多个办事处。 地址、电话号码和传真号码均列在思科网站 **www.cisco.com/go/offices** 中。

美国印刷 C07-734107-01 03/16