Cisco Nexus 9300 平台交换机的 VXLAN 设计

指南

2014 年 10 月

目录

概述

从思科® NX-OS 软件版本 6.1(2)I2(1) 开始，Cisco Nexus® 9300 平台交换机支持虚拟可扩展局域网 (VXLAN) 桥接和网关功能。在其最初的实施中，Cisco Nexus 9300 平台支持基于组播的 VXLAN，也就是说，网络使用底层网络中的组播功能传输重叠 VXLAN 网络的广播、未知单播和组播流量。本文档讨论了 Cisco Nexus 9300 平台上的 VXLAN 功能，以及由作为 VXLAN 隧道终端 (VTEP) 的 Cisco Nexus 9300 平台支持的网络虚拟化设计。VXLAN 技术和基于组播的 VXLAN 不属于本文档的讨论范围。有关这些主题的详细信息，请参阅：
http://www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/white-paper-c11-729383.html

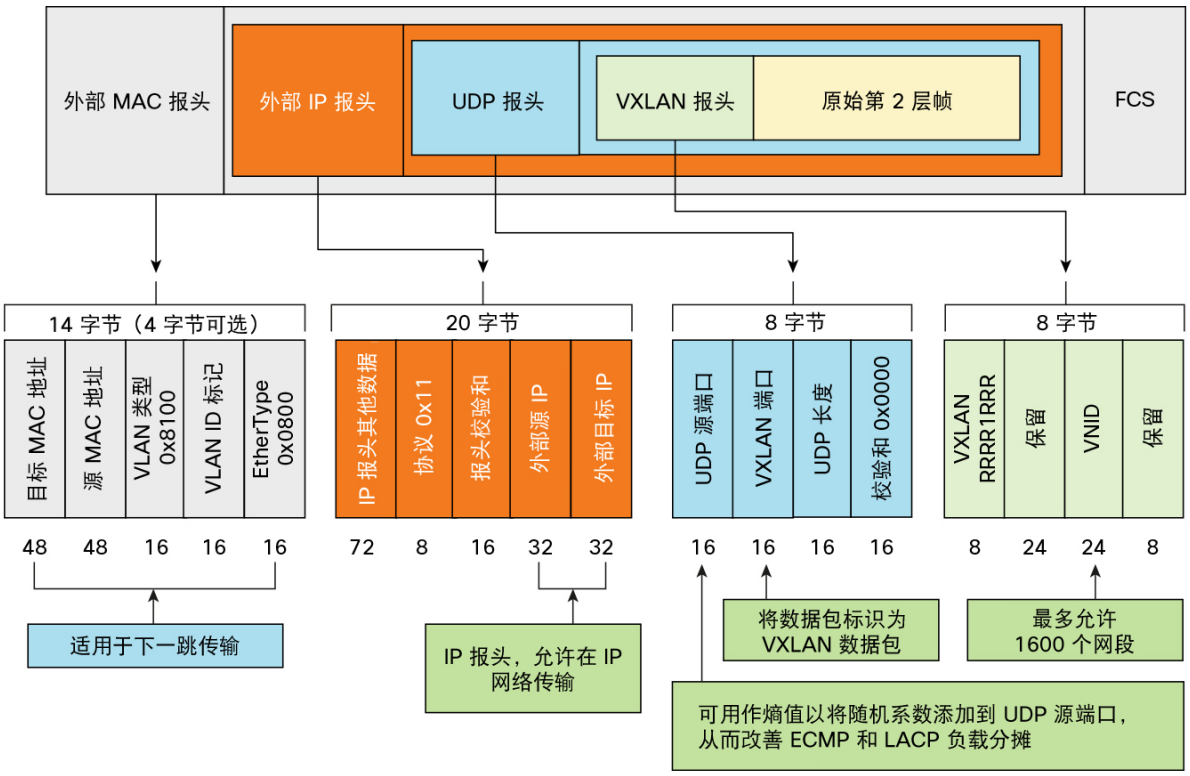
VXLAN 技术概述

数据中心正在面对新的需求，这些需求要求数据中心更高效和优化以降低运营成本，更具可扩展性以支持对数据日益增长的需求，并且更敏捷以支持在这些环境之上运行的各种应用。行业越来越希望虚拟化技术能够提供这些优势，不仅希望计算和存储资源提供这些优势，而且希望网络基础设施也提供这些优势。

VXLAN 是多种可用的网络虚拟化重叠技术之一，提供多个优势。VXLAN 是行业标准协议并使用底层 IP 网络。它将第 2 层网段扩展到第 3 层基础设施，以构建第 2 层重叠逻辑网络。它将以太网帧封装成用户数据报协议 (UDP) 报头，并使用正常的 IP 路由和转发机制将封装的数据包通过底层网络传输到远程 VTEP。

表 1 显示了 VXLAN 数据包格式。该数据包具有 8 字节 VXLAN 报头、UDP 报头、外部 IP 报头和外部 MAC报 头。

图 1. VXLAN 数据包格式 (MAC-in-UDP)



- **VXLAN 报头**：VXLAN 报头中的 24 位 VNID 字段标识了 VXLAN 网段。它为第 2 层网络提供扩展的地址空间。
- **UDP 报头**：UDP 报头中的目的端口指明该数据包是 VXLAN 封装的数据包。VXLAN 最初根据 VXLAN IETF 草案使用与重叠传输虚拟化 (OTV) 相同的 UDP 目的端口 8472，直到 IANA 将端口 4789 分配给 VXLAN。因此，在不同的 VXLAN 实施中可能会看到两个端口。源 UDP 端口是基于原始第 2 层帧头的散列结果，因此源端口号将因流而异。使用可方法可以更好地按流在底层网络中分摊 VXLAN 流量负载。
- **外部 IP 标头**：外部 IP 报头中的源 IP 地址是本地 VTEP 地址。对于广播、未知单播和组播流量，目的 IP 地址是关联的组播组地址的已知单播流量的远程 VTEP 地址。封装的数据包将根据外部报头 IP 地址通过底层传输网络进行路由。
- **外部 MAC 地址或第 2 层报头**：此报头用于将封装的数据包转发到最接近的下一跳设备。

作为网络虚拟化重叠技术，VXLAN 有可能为以下问题提供解决方案：

- **第 2 层网段可扩展性**：VXLAN 具有一个 24 位的虚拟网络标识符 (VNI) 字段，它最多允许同一网络中有 1600 万个唯一的第 2 层网段。尽管当前的网络软件和硬件限制会减少实际部署中的可用 VNI 规模，但是 VXLAN 协议在设计上至少已解除了传统的 IEEE 802.1q VLAN 名称空间中的 4096 VLAN 限制。此改变使组织能够构建具有更多第 2 层网段的数据中心网络，例如在大型多租户网络环境中。
- **第2 层域可扩展性**：许多数据中心应用具有简单的网络视图，并且要求终端主机之间在第 2 层邻接。这些应用的增长使得必须在数据中心内延伸第 2 层域。但是，大型的第 2 层域意味着大型的广播和故障域。为了保持网络稳定性和控制任何网络故障的影响，第 2 层域不能太大。但是此限制与应用的增长发生冲突。VXLAN 可以通过将第 2 层域与网络基础设施分离来解决此困境。基础设施构建为不依赖生成树协议的第 3 层交换矩阵，以预防环路或融合拓扑。第 2 层域位于重叠网络中，有独立的广播和故障域。这种方法使数据中心网络可以发展，而无需冒险创建过大的故障域。
- **第 3 层边界上的第 2 层网段弹性**：数据中心网络通常使用多个第 2 层 Pod 构建，这些 Pod 通过第 3 层汇聚层互联。应用工作负载置于第 2 层连接的各个 pod 中。此方法对数据中心网络内的应用工作负载部署施加了严格限制。采用 MAC-in-IP-UDP 隧道机制，VXLAN 可以在底层第 3 层基础设施中构建第 2 层虚拟网络。应用终端主机可以灵活部署在数据中心网络中，而无需担心底层基础设施的第 3 层边界，并且同时在 VXLAN 重叠网络内保持第 2 层邻接。

术语

以下定义将帮助您了解 VXLAN。

- **虚拟网络标识符 (VNI) 或 VXLAN 网段ID**：系统使用 VNI（也称为 VXLAN 网段 ID）及 VLAN ID 来标识 VXLAN 重叠网络中的 2 层网段。
- **VXLAN 网段**：VXLAN 网段是第 2 层重叠网络，终端设备（包括物理设备和虚拟机）在该网络上通过直接的第 2 层邻接进行通信。
- **VXLAN 隧道终端 (VTEP)**：VTEP 发起和终止 VXLAN 隧道。VTEP 将终端主机第 2 层帧封装在 IP 报头内以通过 IP 传输网络发送它们，并解封从底层 IP 网络收到的 VXLAN 数据包以将其转发到本地终端主机。终端主机不知道 VXLAN。有两种类型的 VTEP：
  - 虚拟 VTEP：基于软件的VTEP；一个示例是虚拟机监控程序主机内支持 VXLAN 的虚拟交换机
  - 物理 VTEP：基于硬件的 VTEP；Cisco Nexus 9300 平台交换机是物理 VTEP

  物理 VTEP 提供基于硬件的高性能，并且能够将 VXLAN 网段与传统的 VLAN 网段桥接，以将第 2 层网段扩展到第 3 层基础设施。

- **VXLAN 网关：**VXLAN 网关连接 VXLAN 和传统的 VLAN 环境。物理 VTEP 设备可以提供基于硬件的 VXLAN 网关功能。图 2 显示了一个示例，在该示例中，一侧的虚拟机监控程序 VTEP 启动 VXLAN 隧道，另一侧的物理 VTEP 设备提供 VXLAN 网关服务，以终止 VXLAN 隧道并将 VXLAN VNI 映射到传统的 VLAN。

图 2. VXLAN 网关

- **VXLAN 桥接：**VXLAN 桥接是 VTEP 设备提供的功能，用于将 VLAN 或 VXLAN 扩展到第 3 层基础设施。图 3 显示了 VLAN 至 VLAN 和 VXLAN 到 VXLAN 的桥接。

图 3. VXLAN 桥接

- **VXLAN 路由：**VXLAN 路由也称为 VXLAN 间路由。它以类似于 VLAN 间路由的方式在重叠网络内的两个 VXLAN VNI 之间提供 IP 路由服务。图 4 显示了 VXLAN 路由的逻辑概念。

图 4. VXLAN 路由

硬件和软件支持

本文档中介绍的解决方案将 Cisco Nexus 9300 平台交换机用作物理 VXLAN VTEP。他们需要下列硬件和软件：

- Cisco Nexus 9300 平台交换机必须用作 VXLAN 拓扑中的 VTEP 设备。
- 建议在 Cisco Nexus 9300 VTEP 交换机上使用思科 NX-OS 版本 6.1(2)I2(2b) 或更高版本。虽然 Cisco Nexus 9300 平台在思科 NX-OS 版本 6.1(2)I2(1) 中开始支持 VXLAN 功能，但是版本 6.1(2)I2(2b) 中已添加许多增强功能。
- VXLAN 功能不需要额外的许可证。但是，底层网络需要内部网关协议 (IGP) 路由和 IP 组播功能的相应许可证。
- 作为底层网络中的主干（或汇聚层）设备的 Cisco Nexus 9500 平台交换机（或其他平台中提供相同或相似的 10 千兆和 40 千兆以太网端口密度和性能的交换机）。

用作 VXLAN VTEP 的 Cisco Nexus 9300 平台

物理 VTEP 设备扮演两个角色：本地 VLAN 中的普通第 2 层交换功能，以及通过 VXLAN 封装将本地 VLAN 扩展到同一第 2 层网段中的远程站点。对于本地 VLAN 中的主机，VTEP 是普通的第 2 层交换机。对于远程 VTEP 设备，该设备用作启动和终止 VXLAN 隧道的 VTEP 对等设备。图 5 显示了 VTEP 设备的逻辑功能。

图 5. 物理 VTEP 功能

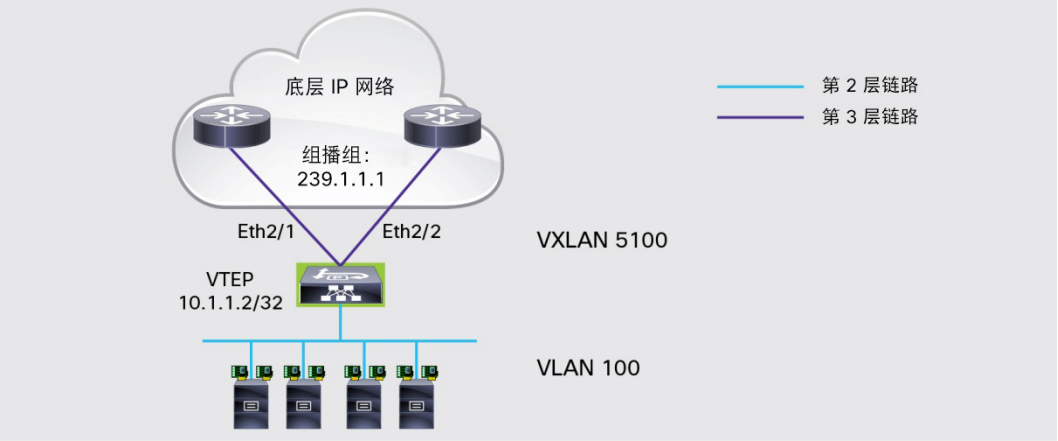
构建和管理 Cisco Nexus 9300 VTEP

本部分演示了如何配置和管理 Cisco Nexus 9300 VTEP。

配置 Cisco Nexus 9300 VTEP

本部分使用图 6 中显示的拓扑示例，演示将 Cisco Nexus 9300 平台交换机配置为 VTEP 的步骤。

图 6. Cisco Nexus 9300 VTEP 示例



要将 Cisco Nexus 9300 平台交换机配置为 VTEP 设备，请执行以下步骤：

第 1 步： 启用 VXLAN 功能。

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

第 2 步： 将 VLAN 映射至 VXLAN VNI。

以下示例将 VLAN 100 映射到 VXLAN VNI 5100：

```
vlan 100
  vn-segment 5100
```

第 3 步： 使用 a/32 IP 地址创建环回接口。

此 IP 地址将用作交换机 VTEP 地址。目前，Cisco Nexus 9300 平台交换机只能有一个 VTEP 地址。需要通过底层 IGP 路由协议通告 VTEP 地址，以便 VTEP 设备可以获得其 VTEP 地址的 IP 可达性。必须在环回接口上启用 IP 组播路由。

以下示例使用 IP 地址 10.1.1.2/32 创建 **interface loopback0**，并在 **ospf 1 area 0** 中通告它。在环回接口下启用协议无关组播 (PIM) 稀疏模式。

```
interface loopback0
  ip address 10.1.1.2/32
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode
```

第 4 步： 创建用作 VXLAN 隧道接口的网络虚拟化终端 (NVE) 接口。

在 NVE 接口下，添加 VXLAN VNI 并将其与底层组播组相关联。

以下示例配置接口 **nve1**，**loopback0** 作为隧道源接口。VNI 5100 添加在 nve1 接口下，并与组播组 239.1.1.1 相关联。这意味着 VXLAN VNI 5100 将在底层网络中使用组播组 239.1.1.1，以传输重叠未知单播、组播和广播流量。

```
interface nve1
  no shutdown
  source-interface loopback0
  overlay-encapsulation VXLAN
  member vni 5100 mcast-group 239.1.1.1
```

除了前面与 VXLAN 相关配置的步骤之外，您还需要在 VTEP 交换机上和底层网络中配置 IGP 路由和组播路由。在本例中，开放最短路径优先 (OSPF) 配置为 IGP，上行链路接口和 **loopback0** 接口上启用了 IP PIM 稀疏模式。自动交汇点 (**auto-rp**) 配置为 PIM 交汇点发现协议。两个上游路由器充当任播交汇点。底层网络配置遵循所选 IGP 和组播路由的最佳实践，因此此处不显示具体配置。

以下是 Cisco Nexus 9300 VTEP 交换机的相关配置：

```
feature pim
feature ospf

interface Ethernet2/1
```

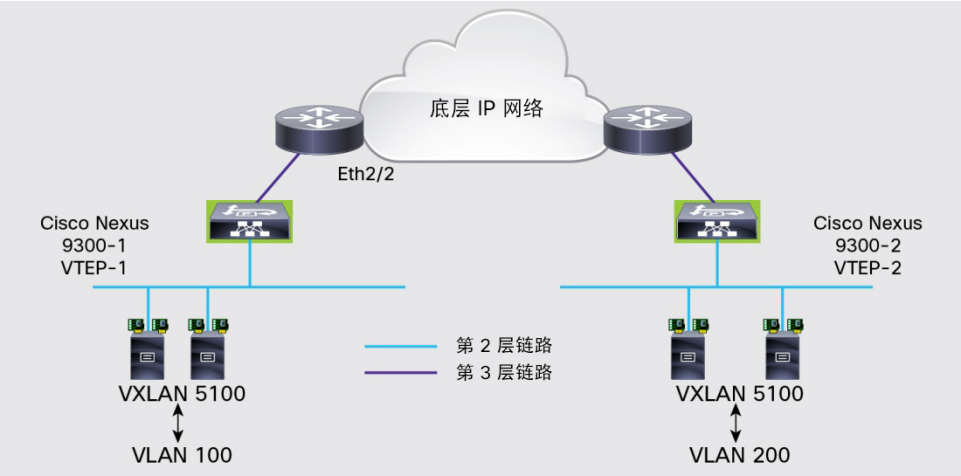
```
  ip address 192.168.1.6/30
  ip ospf network point-to-point
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode
  no shutdown
interface Ethernet2/2
  ip address 192.168.1.10/30
  ip ospf network point-to-point
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode
  no shutdown

interface loopback0
  ip address 10.1.1.2/32
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode

router ospf 1
  router-id 10.1.1.2

ip pim ssm range 232.0.0.0/8
ip pim auto-rp listen
```

注：附录 A 显示了 Cisco Nexus 9300 VTEP 交换机的完整配置示例。

VLAN 间的标记处理和 VLAN 转换

Cisco Nexus 9300 平台交换机遵循 VXLAN 实施的 VXLAN IETF 标准。根据 VXLAN IETF 草案，如果在将数据包封装为 VXLAN 格式以通过底层网络传输数据包之前，存在一个 IEEE 802.1Q VLAN 标记，则入口 VTEP 应删除原始第 2 层数据包中的该标记。远程 VTEP 设备具有关于 VXLAN 的信息，数据包将根据设备自己的 VLAN 至 VXLAN VNI 映射配置放入 VLAN 中。如果利用此机制，同一 VXLAN VNI 的 VTEP 设备可能将 VXLAN VNI 映射到不同的 VLAN。图 7 显示了一个示例，其中 Cisco Nexus 9300 VTEP-1 将 VLAN 100 映射到 VXLAN VNI 5100，而 Cisco Nexus 9300 VTEP-2 将 VXLAN 200 映射到 VXLAN VNI 5100。因此，VTEP-1 背后的 VLAN 100 和 VTEP-2 背后的 VLAN 200 桥接到重叠网络中的一个第 2 层域，这两个 VLAN 内的主机获得直接的第 2 层邻接关系。通过 VXLAN 重叠网络进行的此 VLAN 转换可以为希望将不同的传统 VLAN 连接在一起的组织提供解决方案：例如，在公司收购后或者在内部合并或数据中心迁移期间。

图 7. Cisco Nexus 9300 VTEP 进行的 VLAN 间的标记处理和 VLAN 转换



Cisco Nexus 9300 VTEP 上的主机 MAC 地址管理

Cisco Nexus 9300 VTEP 学习本地主机和远程主机的 MAC 地址，并将其存储在 MAC 地址表中。本地主机 MAC 地址条目使用物理端口信息设定。对于 VTEP 从远程 VTEP 对等体学习的远程主机，具有对等 VTEP 地址的 NVE 接口（VXLAN 隧道接口）被列为端口信息。

本地和远程主机的 MAC 地址条目受同一老化机制约束。默认情况下，MAC 地址表老化计时器为 1800 秒。用户可使用 **mac address-table aging-time** 配置命令配置该计时器。范围为 120 到 91800 秒。

Cisco Nexus 9300 VTEP 交换机中的 MAC 地址表配置示例如下所示：

```
n9396-vtep-1# sh mac address-table
Legend:
        * - primary entry, G - Gateway MAC, (R) - Routed MAC, O - Overlay MAC
        age - seconds since last seen,+ - primary entry using vPC Peer-Link,
        (T) - True, (F) - False
   VLAN     MAC Address      Type      age     Secure NTFY Ports
---------+-----------------+--------+---------+------+----+------------------
*  100     0000.0c07.ac64   dynamic   0         F      F    nve1(10.1.1.100)
*  100     0000.ced0.41cf   dynamic   0         F      F    Eth1/1
*  100     0000.ced1.072f   dynamic   0         F      F    nve1(10.1.1.3)
*  100     6412.2574.6ae7   dynamic   0         F      F    nve1(10.1.1.100)
*  100     6412.2574.9eb7   dynamic   0         F      F    nve1(10.1.1.100)
*  101     0000.0c07.ac65   dynamic   0         F      F    nve1(10.1.1.100)
*  101     0000.ced0.41d1   dynamic   0         F      F    Eth1/1
*  101     0000.ced1.0731   dynamic   0         F      F    nve1(10.1.1.3)
*  101     6412.2574.6ae7   dynamic   0         F      F    nve1(10.1.1.100)
*  101     6412.2574.9eb7   dynamic   0         F      F    nve1(10.1.1.100)
G    -     7c69.f6df.e597   static    -         F      F    sup-eth1(R)
n9396-vtep-1#
```

本地主机

远程主机

监控 Cisco Nexus 9300 VTEP 的 VXLAN 状态

Cisco Nexus 9300 VTEP 维护有关 VXLAN VNI 网段和活动 VTEP 对等设备的信息。以下示例显示了具有分别使用组播组 239.1.1.1 和 239.1.1.2 调配的两个 VXLAN VNI 5100 和 5101 的 VTEP 设备。它还具有两个活动的 远程 VTEP 对等体。

```
n9396-vtep-1# sh nv vni
Interface          VNI        Multicast-group   VNI State
-----------------  --------   ---------------   ---------
nve1               5100       239.1.1.1         up
nve1               5101       239.1.1.2         up


n9396-vtep-1# sh nv peers
Interface          Peer-IP           Peer-State
-----------------  ---------------   -------------
nve1               10.1.1.3           Up
nve1               10.1.1.100         Up


n9396-vtep-1# sh nv peers detail
Peer: 10.1.1.3
  Interface          : nve1
  Peer learnt VNI    : 5100
  Configured VNIs    : 5100-5101
  Provision State    : add-complete
  Route Update       : Yes
  Uptime             : 10:41:46

Peer: 10.1.1.100
  Interface          : nve1
  Peer learnt VNI    : 5100
  Configured VNIs    : 5100-5101
  Provision State    : add-complete
  Route Update       : Yes
  Uptime             : 10:41:35
```

基于组播的 VXLAN 对等体状态基于数据。仅当从远程 VTEP 对等体接收流量时，该对等体才会出现在对等体数据库中。在流量停止后，对等体将在一段时间后老化。对等体老化计时器与主机 MAC 地址条目的老化计时器相关联。当与此对等体相关联的最后一个 MAC 地址条目老化时，对等体也将从对等体数据库中删除。

Cisco Nexus 9300 VTEP 的 VXLAN 统计信息

思科 NX-OS 提供有关 Cisco Nexus 9300 VTEP 的按 VXLAN VNI 网段的统计信息和按 VTEP 对等体的统计信息。命令行界面 (CLI) 监控命令和输出示例如下所示。

按 VXLAN VTEP 对等体的统计信息

```
n9396-vtep-1# sh nv peers 10.1.1.3 interface nve 1 counters
Peer IP: 10.1.1.3
TX
        3954189 unicast packets 4136035670 unicast bytes
        0 multicast packets 0 multicast bytes
RX
        3941862 unicast packets 4123118616 unicast bytes
        0 multicast packets 0 multicast bytes
```

按 VXLAN VNI 的统计信息

```
n9396-vtep-1# sh nv vni 5100 counters
VNI: 5100
TX
        3717075 unicast packets 3886820534 unicast bytes
        2 multicast packets 140 multicast bytes
RX
        3708372 unicast packets 3873416710 unicast bytes
        5 multicast packets 600 multicast bytes
```

Cisco Nexus 9300 VTEP 的组播处理

基于组播的 VXLAN 在底层网络使用 IP 组播功能传输重叠第 2 层网段的广播、组播和未知单播。VXLAN VNI 与传输网络中的组播组相关联。多个 VNI 可以共享一个组播组。每个 VTEP 设备是其加入的组播组中的接收方和来源。按照通常的 IP PIM 稀疏模式实践，如果 VNI 网段包含 n 个 VTEP 设备，则此 VNI 的组播组将具有 n 个 (S, G) 状态和 1 个 (*, G) 状态。因此，随着 VNI 网段数量增加，以及随着与 VXLAN VNI 相关联的组播组数量增加，IP 组播路由表大小可能会变得非常大，带来多播可扩展性方面的挑战并且会增加组播操作和管理的复杂性。

为了帮助降低 VTEP 交换机上的组播路由表增长，使用与 VXLAN VNI 相关联的组播组的 (S, G) 最短路径树 (SPT) 抑制来实施 Cisco Nexus 9300 平台交换机的思科 NX-OS。借助这项增强功能，当 Cisco Nexus 9300 平台交换机用作 VTEP 时，它会使用共享的交汇点树 (*, G) 条目从同一 VNI 中的其他 VTEP 设备接收组播封装的 VXLAN 流量，而不是切换到每个远程 VTEP 源的 SPT。此功能消除了为每个远程 VTEP 维护一个 (S, G) 条目的需要。因此，Cisco Nexus 9300 VTEP 只需为每个与 VNI 相关联的组播组维护两个组播状态：将交汇点作为根的共享树的 (*, G) 条目，以及本地源树的本地 (S, G) 条目。

朝着组播组的交汇点构建 (*, G) 条目。其传出接口为 VXLAN NVE1 接口。此条目用于接收 VXLAN 组播封装的数据包并执行解封。

本地 (S, G) 条目将本地 VTEP 地址（loopback/32 地址）作为源，并将朝着交汇点的上行链路接口作为传出接口。此条目用于执行 VXLAN 组播封装，并朝着组播组的交汇点将封装的组播数据包发送到底层网络。

以下是来自 Cisco Nexus 9300 VTEP 的 **show ip mroute** 输出示例。它显示用于 VXLAN VNI 的组 239.1.1.1 的组播状态。本地 VTEP address 是 10.1.1.2。

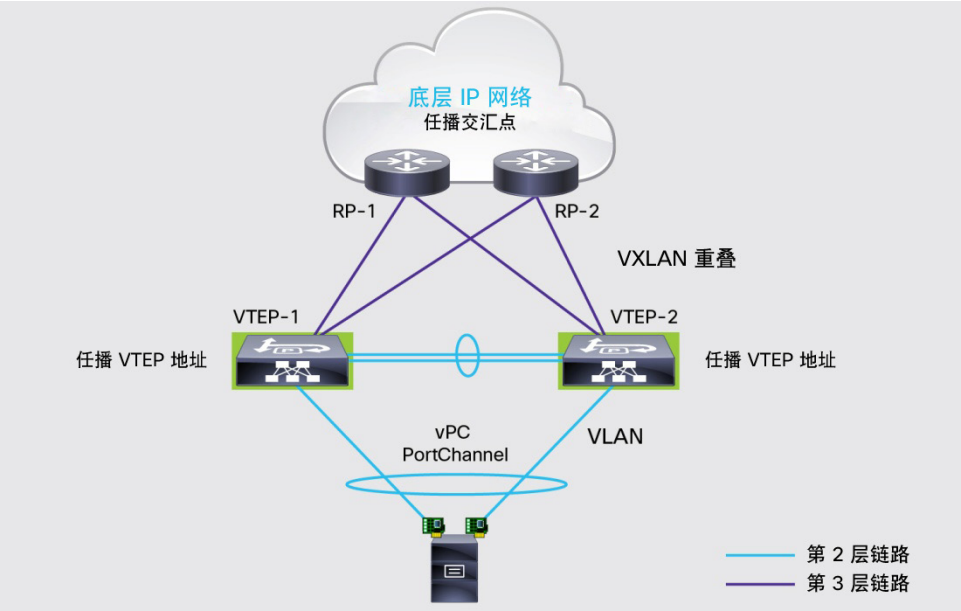
```
n9396-vtep-1# sh ip mroute
IP Multicast Routing Table for VRF "default"

(*, 232.0.0.0/8), uptime: 11:31:35, pim ip
  Incoming interface: Null, RPF nbr: 0.0.0.0, uptime: 11:31:35
  Outgoing interface list: (count: 0)

(*, 239.1.1.1/32), uptime: 11:31:34, nve ip pim
  Incoming interface: Ethernet2/1, RPF nbr: 192.168.1.1, uptime: 11:29:36
  Outgoing interface list: (count: 1)
    nve1, uptime: 11:31:34, nve

(10.1.1.2/32, 239.1.1.1/32), uptime: 11:31:34, nve ip mrib pim
  Incoming interface: loopback0, RPF nbr: 10.1.1.2, uptime: 11:30:38
  Outgoing interface list: (count: 1)
    Ethernet2/1, uptime: 11:30:29, pim
```

请注意，组播封装仅用于传输重叠 VXLAN 网络的广播、未知单播和组播流量。已知的单播流量是单播封装的流量并使用 VTEP 地址传输。

使用 Cisco Nexus 9300 平台交换机构建冗余 vPC VTEP

Cisco Nexus 9300 平台交换机支持 VTEP 冗余，方法是允许一对虚拟 PortChannel (vPC) 交换机用作共享 VTEP 任播地址的逻辑 VTEP 设备（见图 8）。vPC 交换机为冗余主机连接提供 vPC，同时对底层网路中的上游设备逐个运行第 3 层协议。它们加入同一 VXLAN VNI 的组播组，并使用与源相同的任播 VTEP 地址发送 VXLAN 封装的数据包。对于底层网络中的设备（包括组播汇集点和远程 VTEP 设备），两个 vPC VTEP 交换机显示为一个逻辑 VTEP 实体。

图 8. 作为 vPC VTEP 的 Cisco Nexus 9300 平台交换机



要配置 vPC VTEP，请执行以下步骤：

第 1 步： 为主机连接配置 vPC 交换机和 vPC。

使用标准的vPC 配置操作程序。为 vPC 使用思科建议的最佳实践。

第 2 步： 启用 VXLAN 功能。

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

第 3 步： 配置环回接口，将 /32 作为辅助地址。

环回接口上的主要地址很可能将由网络协议（例如 OSPF 和边界网关协议 [BGP]）用作路由器 ID。在这种情况下，两台交换机不能具有相同的主要环回地址。因此，vPC VTEP 将两台交换机之间的环回接口上的相同辅助地址用作 VTEP 任播地址。下面列出了一些示例。

在 vPC 交换机 1 上：

```
interface loopback0
  no ip redirects
  ip address 10.1.1.4/32
  ip address 10.1.1.100/32 secondary  ← 任播 VTEP 地址
  ip ospf network point-to-point
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode
```

在 vPC 交换机 2 上：

```
interface loopback0
  no ip redirects
ip address 10.1.1.5/32
ip address 10.1.1.100/32 secondary  ← 任播 VTEP 地址
ip ospf network point-to-point
ip router ospf 1 area 0.0.0.0
```

第 4 步： 按照正常的 VTEP 配置步骤配置 VXLAN。

```
vlan 100
  vn-segment 5100
interface nve1
  source-interface loopback0
  member vni 5100 mcast-group 239.1.1.1
```

注：附录 B 显示了 Cisco Nexus 9300 vPC VTEP 交换机的完整配置示例。

vPC 对等体必须具有以下相同的配置：

- 将 VLAN 一致映射到虚拟网段 (VN-segment)
- 将 NVE 一致绑定到相同的环回辅助 IP 地址（任播 VTEP 地址）
- 一致的 VNI 至组映射。

Cisco Nexus 9300 vPC VTEP 的基本操作

本部分讨论了 Cisco Nexus 9300 vPC VTEP 的基本操作。

vPC 任播 VTEP 地址

Cisco Nexus 9300 vPC VTEP 交换机将绑定至 VXLAN NVE 隧道的环回接口上的辅助 IP 地址用作任播 VTEP 地址。两台 vPC 交换机需要具有完全相同的辅助环回 IP 地址。它们均在底层网络上通告此任播 VTEP 地址，以便上游设备从两个 vPC VTEP 学习 /32 路由，并且可以在它们之间分摊 VXLAN 单播封装的流量负载。

在 vPC 对等链路出现故障时，正在运行的 vPC 辅助交换机将关闭其绑定至 VXLAN NVE 的环回接口。此关闭会导致辅助 vPC 交换机从其 IGP 通告撤销任播 VTEP 地址，以便底层网络中的上游设备开始讲所有流量仅发送到主要 vPC 交换机。此过程的目的是在对等链路关闭时避免出现 vPC 双活情况。如果利用此机制，当 vPC 对等链路关闭时，连接到辅助 vPC 交换机的孤立设备将不能够接收 VXLAN 流量。

VXLAN 组播和广播、未知单播和组播流量处理

两个 vPC VTEP 交换机独立将 IP PIM 注册数据包发送到 VXLAN VNI 的组播组的交汇点。它们从任播 VTEP 地址获得注册数据包。它们均在其组播路由表中安装相应的 (*, G) 条目（使用输出接口 [OIF] 列表中的 NVE1）。

交汇点设备会看到至任播地址的至少两个等价多路径 (ECMP) 路由：一个路由至每个 VTEP 交换机。它会选择其中一个路径以发送 (S, G) 连接。此处的 S 表示任播 VTEP 地址。从交汇点接收 (S, G) 连接的 VTEP 交换机将在其 (S, G) OIF 列表中朝着交汇点安装上行链路接口。它将成为该组的指定转发器 (DF)。它将使用 (S, G) OIF 列表封装 VXLAN 组播数据包，并通过上行链路接口将其发送到底层网络。

只有作为指定转发器 (DF) 的 VTEP 交换机执行 VXLAN 组播封装和解封。对于来自远程 VTEP 对等体的组播封装流量，交汇点始终将其转发到作为指定转发器的 VTEP 交换机，以便在本地转发后解封。作为指定转发器的 VTEP 交换机将其 (*, G) OIF 列表用于解封。对于需要流动至远程 VTEP 设备的由本地主机生成的广播、未知单播和组播流量，本地主机可以通过 vPC 在两台 VTEP 交换机之间分摊流量负载。在这种情况下，非指定转发器的 VTEP 交换机会通过 vPC 对等链路将流量转发到作为指定转发器的 VTEP 交换机。作为指定转发器的 VTEP 交换机将使用其本地 (S, G) OIF 列表执行 VXLAN 组播封装，并朝着交汇点发送封装的组播数据包。

以下示例显示了 vPC VTEP 交换机上的 VXLAN VNI 组播组的组播条目。

```
vtep-2# sh ip mroute
IP Multicast Routing Table for VRF "default"

(*, 232.0.0.0/8), uptime: 1d05h, pim ip
  Incoming interface: Null, RPF nbr: 0.0.0.0, uptime: 1d05h
  Outgoing interface list: (count: 0)

(*, 239.1.1.1/32), uptime: 1d05h, nve pim ip
  Incoming interface: Ethernet2/1, RPF nbr: 192.168.1.13, uptime: 1d05h
  Outgoing interface list: (count: 1)
    nve1, uptime: 1d05h, nve

(10.1.1.100/32, 239.1.1.1/32), uptime: 1d05h, nve pim mrib ip
  Incoming interface: loopback0, RPF nbr: 10.1.1.100, uptime: 1d05h
  Outgoing interface list: (count: 1)
    Ethernet2/1, uptime: 1d05h, pim

vtep-2#
```

此 VTEP 是组 239.1.1.1 的指定转发器。它为关联的 VXLAN VNI 执行 VXLAN 组播封装和解封。

```
vtep-1# sh ip mroute
IP Multicast Routing Table for VRF "default"

(*, 232.0.0.0/8), uptime: 1d05h, pim ip
  Incoming interface: Null, RPF nbr: 0.0.0.0, uptime: 1d05h
  Outgoing interface list: (count: 0)

(*, 239.1.1.1/32), uptime: 1d05h, nve ip pim
  Incoming interface: Ethernet2/1, RPF nbr: 192.168.1.9, uptime: 1d05h
  Outgoing interface list: (count: 1)
    nve1, uptime: 1d05h, nve

(10.1.1.100/32, 239.1.1.1/32), uptime: 1d05h, nve pim mrib ip
  Incoming interface: loopback0, RPF nbr: 10.1.1.100, uptime: 1d05h
  Outgoing interface list: (count: 0)

vtep-1#
```

此 VTEP 不是组 (239.1.1.1) 的指定转发器。其不会为关联的 VXLAN VNI 执行 VXLAN 多播封装或解封装。

底层网络组播通常设计有冗余交汇点，例如任播交汇点。冗余交汇点设备可以向任一 vPC VTEP 设备独立发送 (S, G) 连接。例如，在图 8 中，RP-1 可以将其连接发送到 VTEP-1，RP-2 可以发送到 VTEP-2。在这种情况下，两台交换机之间的选择过程将选择其中一台交换机作为指定转发器，这是执行 VXLAN 组播封装和解封的唯一一台交换机。

注：此指定转发器选择过程不存在于思科 NX-OS 版本 6.1(2)I2(2a)、6.1(2)I2(3) 和更早的版本中。在这些版本中，对于希望使用 vPC VTEP 部署 VXLAN 的组织，底层网络组播需要具有单个交汇点以避免 VXLAN 组播封装的流量出现潜在的环路。

VXLAN 单播流量处理

对于已知单播流量，两个 vPC VTEP 交换机执行封装和解封。

从本地主机到远程主机的流量将以从接入端口到底层网络上行链路的方向经过 VTEP 设备。VTEP 将封装单播数据包，在外部 IP 报头中，将任播 VTEP 地址作为源，远程 VTEP 地址作为目的地址。

从远程主机流向本地主机的流量将到达 VTEP 设备的上行链路接口。在外部 IP 报头中，封装的单播数据包将远程 VTEP 地址作为源，并将本地任播 VTEP 地址作为目的地址。两个 VTEP 交换机执行解封并将数据包转发到本地目的主机。

vPC VTEP 的 vPC 一致性检查

vPC 一致性检查是配置为 vPC 对的两台交换机用于交换和验证其配置兼容性的一种机制。此检查对于确保 vPC 功能正确运行至关重要。由于 Cisco Nexus 9300 平台交换机支持 vPC VTEP，因此 VXLAN 配置组件添加到 vPC 一致性检查。

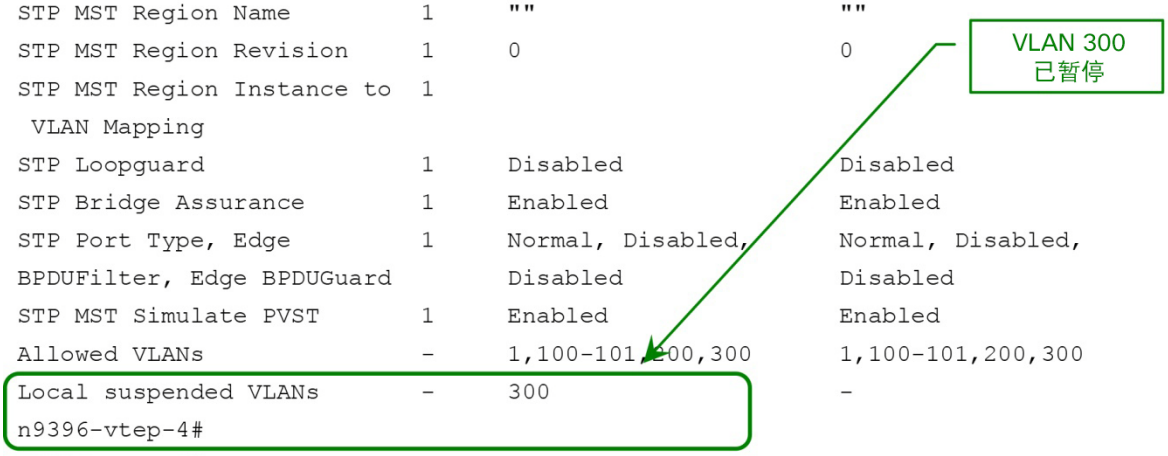
VLAN 至 VXLAN Vn-segment 的映射是一个类型 1 一致性检查参数。两个 VTEP 交换机需要具有相同的映射。具有不匹配的 Vn-segment 映射的 VLAN 将被暂停。如果启用流畅一致性检查，则主要 vPC 交换机将使有问题的 VLAN 保持运行，同时辅助 vPC 交换机暂停它们。如果流畅一致性检查已禁用，两台 vPC 交换机将暂停 VLAN。

以下情况将检测为不一致：

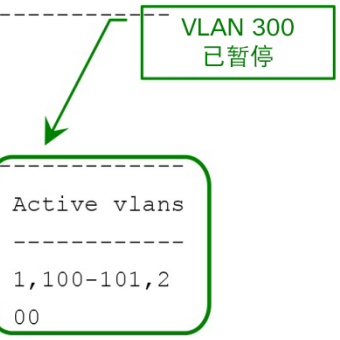
- 一台交换机具有映射到 VN-segment (VXLAN VNI) 的 VLAN，另一台交换机不具有相同 VLAN 的映射。
- 两台交换机具有映射到不同 Vn-segment 的 VLAN。

以下是当两台 vPC 交换机具有映射到不同 Vn-segment 的 VLAN 300 时的 **show** 命令输出示例。

```
n9396-vtep-4# sh vpc consistency-parameters global

    Legend:
        Type 1 : vPC will be suspended in case of mismatch

Name                        Type  Local Value            Peer Value
-------------               ----  ---------------------- -----------------------
Vlan to Vn-segment Map      1     4 Relevant Map(s)      4 Relevant Map(s)
STP Mode                    1     Rapid-PVST             Rapid-PVST
STP Disabled                1     None                   None
STP MST Region Name         1     ""                     ""
STP MST Region Revision     1     0                      0
STP MST Region Instance to  1
 VLAN Mapping
STP Loopguard               1     Disabled               Disabled
STP Bridge Assurance        1     Enabled                Enabled
STP Port Type, Edge         1     Normal, Disabled,      Normal, Disabled,
BPDUFilter, Edge BPDUGuard        Disabled               Disabled
STP MST Simulate PVST       1     Enabled                Enabled
Allowed VLANs               -     1,100-101,200,300      1,100-101,200,300
Local suspended VLANs       -     300                    -
n9396-vtep-4#
n9396-vtep-4#

n9396-vtep-4# sh vpc consistency-parameters vlans

Name                        Type  Reason Code             Pass Vlans
-------------               ----  ---------------------- -----------------------
Vlan to Vn-segment Map      1     vPC type-1             0-299,301-4095
                                  configuration
                                  incompatible -
                                  vn_segment
                                  inconsistent
STP Mode                    1     success                0-4095
STP Disabled                1     success                0-4095
STP MST Region Name         1     success                0-4095
STP MST Region Revision     1     success                0-4095
STP MST Region Instance to  1     success                0-4095
 VLAN Mapping
STP Loopguard               1     success                0-4095
STP Bridge Assurance        1     success                0-4095
STP Port Type, Edge         1     success                0-4095
```



```
BPDUFilter, Edge BPDUGuard
STP MST Simulate PVST       1     success                0-4095
Pass Vlans                  -                            0-299,301-4095
n9396-vtep-4#


n9396-vtep-4# sh vpc br
Legend:
                (*) - local vPC is down, forwarding via vPC peer-link

vPC domain id                     : 100
Peer status                       : peer adjacency formed ok
vPC keep-alive status             : peer is alive
Configuration consistency status  : success
Per-vlan consistency status       : failed
Type-2 inconsistency reason       : Consistency Check Not Performed
vPC role                          : primary
Number of vPCs configured         : 1
Peer Gateway                      : Enabled
Dual-active excluded VLANs        : -
Graceful Consistency Check        : Disabled
Auto-recovery status              : Disabled

vPC Peer-link status
--------------------------------------------------------------------
id   Port   Status Active vlans
--   ----   ------ --------------------------------------------------
1    Po1    up     1,100-101,200

vPC status
----------------------------------------------------------------------
id   Port   Status Consistency Reason                Active vlans
--   ----   ------ ----------- ------                ------------
10   Po10   up     success     success               1,100-101,2
                                                     00

n9396-vtep-4#
```



Cisco Nexus 9300 VTEP 的 VXLAN 设计注意事项

本部分介绍了在为 Cisco Nexus 9300 VTEP 设计 VXLAN 网络时应考虑的要点。

底层网络中的最大传输单位调整

VXLAN 总共增加 50 字节开销，包括：

- 8 字节的 VXLAN 报头
- 8 字节的 UDP 报头
- 20 字节的外部 IP 标头
- 14 字节的外部 MAC 报头

为了避免在通过底层网络发送 VXLAN 封装的数据包时超过最大传输单位 (MTU) 大小，您应该将底层网络中的 MTU 大小增加 50 字节，或者启用巨帧。

底层网络的组播注意事项

组播交汇点配置

Cisco Nexus 9300 平台交换机上的当前 VXLAN 支持在底层网络中使用组播，以在重叠第 2 层网段上传输广播、组播和未知单播流量。通用组播配置的最佳实践适用于 VXLAN 的组播。VXLAN 中的组播的两个注意事项是交汇点位置和所选的协议。通常建议使用冗余交汇点，将交汇点放置在数据路径的中心位置，例如汇聚层或主干层。所选的交汇点协议取决于底层网络内将处于组播转发数据路径的所有设备支持的协议。在 Cisco Nexus 9000 平台上，支持的交汇点协议包括静态 RP、自动 RP 和自举路由器 (BSR)。Cisco Nexus 9000 系列交换机上支持基于 PIM 的任播交汇点和基于组播源发现协议 (MSDP) 的任播交汇点，以实现交汇点冗余。

VXLAN VNI 的组播组共享

Cisco Nexus 9300 平台交换机上的 VXLAN 实施将组播隧道用于广播、未知单播和组播流量转发。理想情况下，VXLAN VNI 与 IP 组播之间的一对一映射提供最佳的组播转发。但是，在使用这种一对一映射时，如果 VXLAN VNI 数量或 VTEP 设备数量增加，会导致底层网络中的所需组播地址空间和组播转发表大小并行增加。有时候，底层网络组播可扩展性和管理可能变得极具挑战性。在这种情况下，将多个 VXLAN 网段映射到一个组播组有助于节省传输网络设备上的组播控制平面资源。此映射提供了一种更容易的方法来实现所需的 VXLAN 可扩展性，而不会给组播扩展的底层网络增加太多负担。

但是请注意，此组播组共享是以组播转发不理想为代价获得的。转发到一个 VNI 的组播组的数据包现在发送到其他共享同一组播组的其他 VNI 的 VTEP。此方法会导致组播数据平面资源的使用效率较低。因此，此解决方案是权衡控制平面可扩展性与数据平面效率的结果。

尽管组播复制和转发不理想，但是具有共享组播组的多个 VNI 不会对 VXLAN VNI 网络之间的第 2 层隔离造成任何影响。在从组播组收到封装的数据包后，VTEP 检查并验证数据包的 VXLAN 报头中的 VNID。如果 VNID 对于 VTEP 言未知，则 VTEP 会丢弃数据包。仅当 VNID 与其中一个 VTEP 的本地 VXLAN VNID 相匹配时，VTEP 才接受并解封数据包，以便进一步查找和转发到其位于此 VXLAN VNI 内的本地主机。其他 VNI 网络将不会接收该数据包。因此，VXLAN VNI 网络之间的分离不受影响。

底层网络中的 ECMP 散列算法

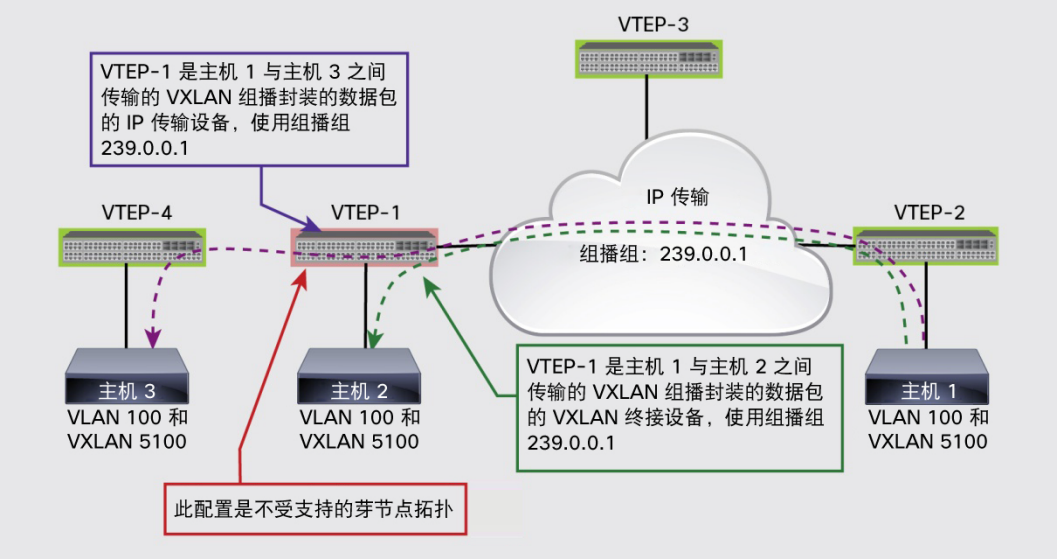
如上一部分所述，Cisco Nexus 9000 系列交换机通过散列原始第 2 层和第 3 层数据包报头，构成 VXLAN 流量的源 UDP 端口。该散列帮助确保每个 VXLAN 封装的流可以通过其外部 IP 报头中的五个元组唯一地标识（源地址和端口、目标地址和端口和协议）。为了为 VXLAN 流量传输实现最佳的负载分摊和分发，底层网络应将基于五元组的散列算法用于 ECMP 和链路汇聚控制协议 (LACP)。

VXLAN 拓扑支持上的限制：不受支持的芽节点拓扑

由于网络转发引擎 (NFE) 的硬件限制，因此 Cisco Nexus 9300 平台交换机当前不支持如图 9 所示的芽节点拓扑。

芽节点是一个 VXLAN VTEP 设备，同时是用于 VXLAN VNI 的同一组播组的 IP 传输设备。在图 9 中，组播组 239.0.0.1 用于 VXLAN VNI。对于从主机 1 到主机 2 的 VXLAN 组播封装流量，VTEP-1 在组 239.0.0.1 中执行组播逆向路径转发 (RPF)。对于使用同一组 239.0.0.1 从主机 1 到主机 3 的 VXLAN 组播封装流量，VTEP-1 是组播数据包的 IP 传输设备。它根据将 239.0.0.1 作为目的地址的外部 IP 报头执行 RPF 检查和 IP 转发。当这两个不同的角色在同一设备上发生冲突时，该设备变成芽节点。由于对 NFE 的硬件限制，Cisco Nexus 9300 平台交换机无法同时执行这两个角色。因此，设计人员需要在将 Cisco Nexus 9300 平台用作 VTEP 的 VXLAN 网络设计中避免使用芽节点拓扑。

图 9. VXLAN 芽节点拓扑



注意：

- NFE 基于被广泛采用的商用芯片。使用同一专用集成电路 (ASIC) 的其他交换机平台需遵守有关芽节点支持的相同硬件限制。在使用这些平台设计 VXLAN 网络时，本文档的读者应了解此限制，并且应避免使用芽节点拓扑。
- 思科很快会为 NFE 上的此硬件限制引入一个软件解决方案，方法是使用也驻留在 Cisco Nexus 9300 平台交换机中的内部开发的应用枝叶引擎 (ALE) ASIC。当这项软件支持变为可用时，Cisco Nexus 9300 VTEP 交换机将支持芽节点拓扑。

作为 VXLAN VTEP 的 Cisco Nexus 9300 平台交换机的设计选项

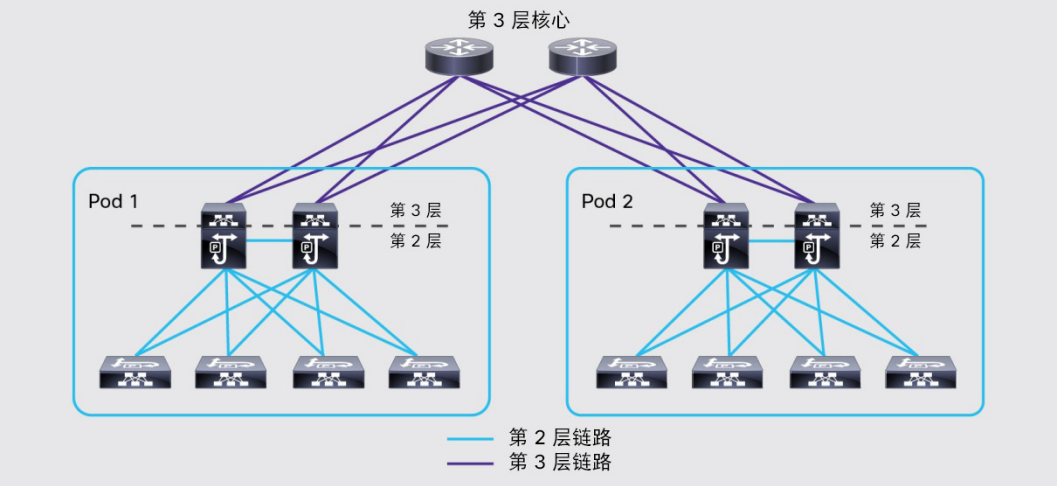
Cisco Nexus 9300 平台在 NX-OS 版本 6.1(2)I2(1) 中开始支持 VXLAN 网关和桥接功能。截至本文档写作时间为止，思科 NX-OS 版本 6.1(2)I2(b) 是适用于 Cisco Nexus 9300 平台的最新版本。在该版本之前，Cisco Nexus 9300 平台交换机不支持 VXLAN 路由。硬件能够支持 VXLAN 路由，但是软件实施将在未来的版本中提供。

本部分中的设计讨论基于 Cisco Nexus 9300 平台上当前可用的 VXLAN 功能，即基于组播的 VXLAN 网关和桥接功能。当 VXLAN 路由功能和以太网 VPN (EVPN) 控制平面在 Cisco Nexus 9300 平台交换机中变为可用时，本部分将修订。

Pod 间的第 2 层扩展设计

传统上，对于要求终端主机之间在第 2 层邻接的应用，数据中心网络中使用典型的第 2 层 POD 设计（图 10）。在第 2 层 Pod 中，一对汇聚交换机充当网络中的第 2 层和第 3 层边界，并且接入交换机运行仅限第 2 层的交换功能。网络服务（例如防火墙和负载均衡器）通常附加到汇聚交换机。对于驻留在 Pod 中的应用 VLAN，汇聚交换机或服务设备用作默认 IP 网关。不同 Pod 之间的流量需要经过第 3 层边界，才可由汇聚交换机或服务设备进行路由。

图 10. 传统的第 2 层数据中心 Pod



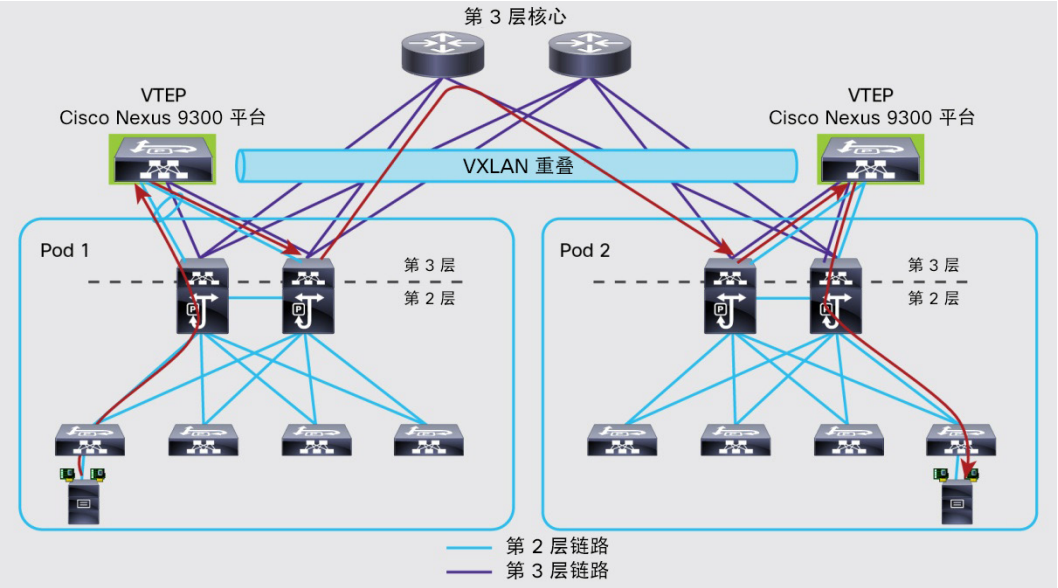
该模式适用于传统的数据中心应用。但是，随着应用工作负载越来越虚拟化，支持工作负载移动性和灵活性成为对数据中心网络的一项新要求。现在，第 2 层网段需要延伸到 Pod 之间的第 3 层边界。

作为第 3 层网络上的第 2 层扩展的重叠技术，VXLAN 是一个可以满足此要求的解决方案。图 11 显示了 Pod 间的第 2 层扩展设计，它将 Cisco Nexus 9300 平台交换机用作 VXLAN VTEP，以在不同 Pod 之间的重叠网络中互联应用 VXLAN。该图显示只有一台 Cisco Nexus 9300 VTEP 交换机连接到每个 Pod 中的汇聚交换机，但是可连接一对 Cisco Nexus 9300 vPC VTEP 以实现冗余。

在此设计中，数据中心 Pod 中的 Cisco Nexus 9300 VTEP 交换机配置为需要通过 VXLAN 扩展的本地 VLAN 的一部分。然后，它们将本地 VLAN 映射到 VXLAN VNI，这些 VNI 由它们之间的 VXLAN 隧道连接在一起。VXLAN 隧道穿越数据中心网络的第 3 层部分，包括 Pod 汇聚层交换机和数据中心核心交换机。图 16 显示了两个 Pod 之间的扩展 VLAN 的流量转发路径。

汇聚交换机继续充当这些 VLAN 的默认网关。因此，路由流量将穿越 Pod，直接通过汇聚交换机，而不经过 VTEP 交换机。

图 11. Pod 间的第 2 层扩展

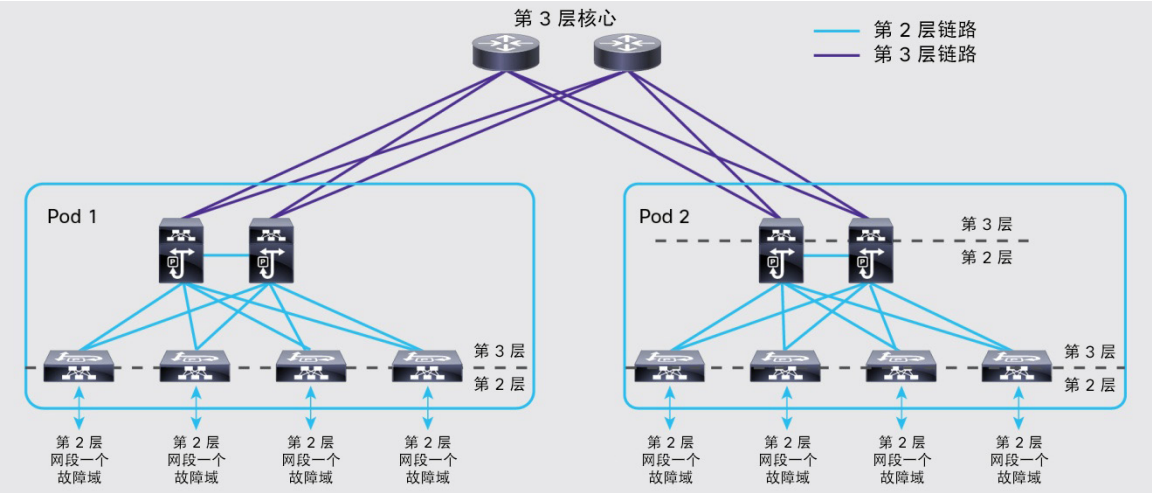


第 3 层数据中心 Pod 设计中的第 2 层扩展

传统的第 2 层 Pod 设计在 Pod 内提供第 2 层邻接，但是它会带来与稳定性和可扩展性相关的多项设计挑战。第 2 层 Pod 是一个第 2 层广播和故障域。通常部署的第 2 层协议（例如生成树协议）在稳定性和可扩展性方面不如第 3 层路由协议。随着第 2 层域增长，其稳定性下降，并且故障域内的故障影响增加。在数据中心网络持续增长时，您应尝试使故障域大小保持在可控范围内，或减少第 2 层域大小。

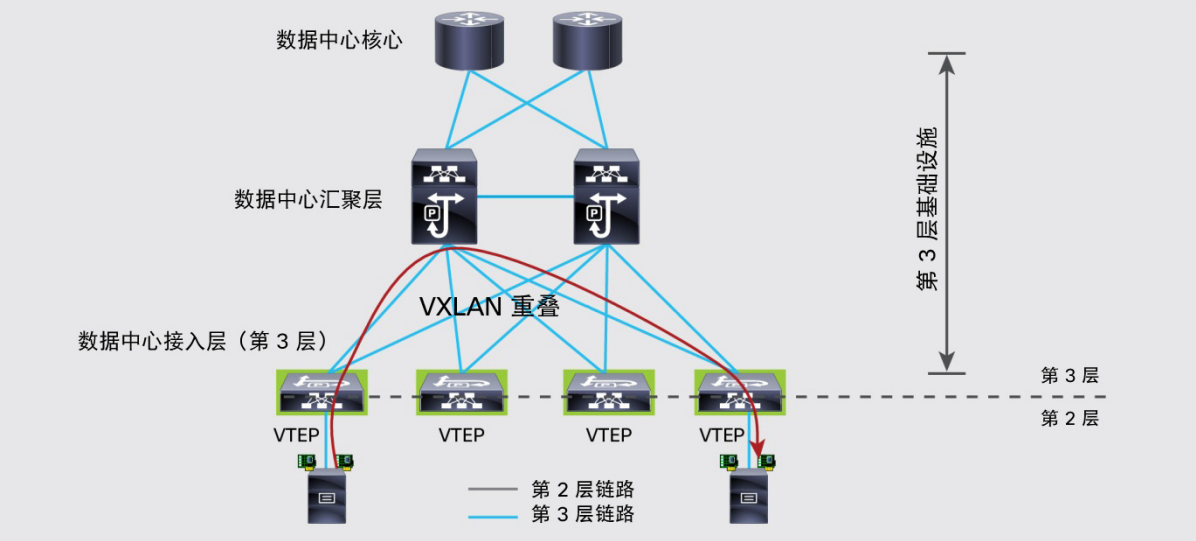
一种越来越普遍使用的方法是将第 3 层路由协议一直运行到接入交换机的第 3 层 Pod 设计。因此，第 2 层域很好地包含在每台接入交换机下，如图 12 所示。如果接入层包括架顶式 (ToR) 接入交换机，则第 2 层域恰好在服务器机架内。此设计可减少网络故障域大小，并显著提高数据中心网络的稳定性。应用的多个实例可以轻松地部署到单独的网络故障域中，这样一个域中发生故障不会危及整个应用的可用性。此方法使 Pod 能够超出第 2 层协议可以稳定保持的大小。

图 12. 第 3 层 Pod 设计



如果第 3 层 Pod 中的某个应用的一部分需要连接到不同接入交换机的主机之间在第 2 层邻接，或者如果第 2 层域需要超出单个服务器机架空间，则需要使用诸如 VXLAN 之类的第 2 层扩展技术，以便在第 3 层 Pod 基础设施之上提供第 2 层重叠。图 13 显示了在 Cisco Nexus 9300 平台交换机上使用 VXLAN 桥接功能的这样一个解决方案。在此设计中，Cisco Nexus 9300 平台交换机部署为第 3 层 ToR 交换机以用于服务器接入连接，并部署为 VXLAN VTEP 设备以在机架之间扩展第 2 层网段。

图 13. 第 3 层 Pod 内的第 2 层扩展



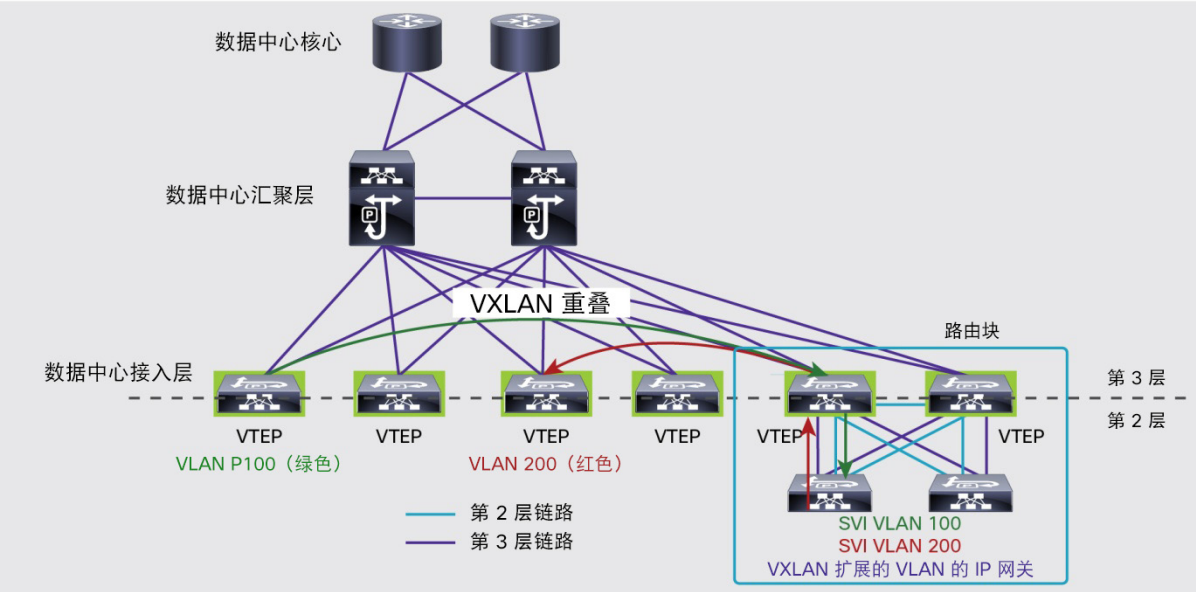
VXLAN 间的路由设计

与传统的 VXLAN 环境一样，在许多情况下要求在 VXLAN 网段之间路由或从 VXLAN 路由到 VLAN 网段。由于当前的思科 NX-OS 版本（版本 6.1(2)I2(3) 及更早版本）不支持 VXLAN 路由，因此需要应用特定的设计才可实现此网络功能。

VXLAN 间的路由设计方案 A：路由块设计

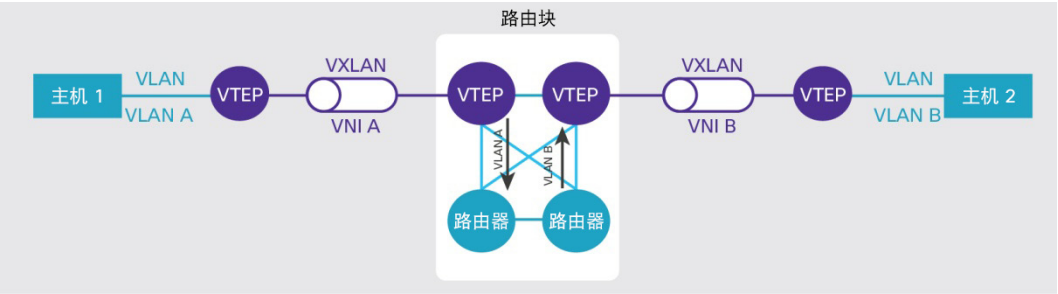
图 14 通过将路由块添加到第 3 层 Pod 网络描绘了一种 VXLAN 路由解决方案。路由块采用单臂路由器设计，包括用于终止 VXLAN 隧道的一个 VTEP 或一对 vPC VTEP，以及充当 VXLAN 扩展的 VLAN 的 IP 网关并为这些 VLAN 执行路由功能的一台或一对路由器。

图 14. VXLAN 路由的路由块设计



对于 VXLAN VNI 内的第 2 层流量，流量会直接在本地 VTEP 与远程 VTEP 之间流动。对于 VXLAN VNI 之间的第 3 层路由流量，流量将首先到达位于路由块中的路由器上的源 VXLAN VLAN IP 子网的 IP 网关，然后将由网关路由器路由到目标 VXLAN VLAN IP 子网。然后，网关路由器会将数据包转回到路由块中的 VTEP，以便在目的 VXLAN 中封装并向目的主机转发。逻辑流量如图 15 所示。

图 15. 路由块 VXLAN 路由设计中的流量



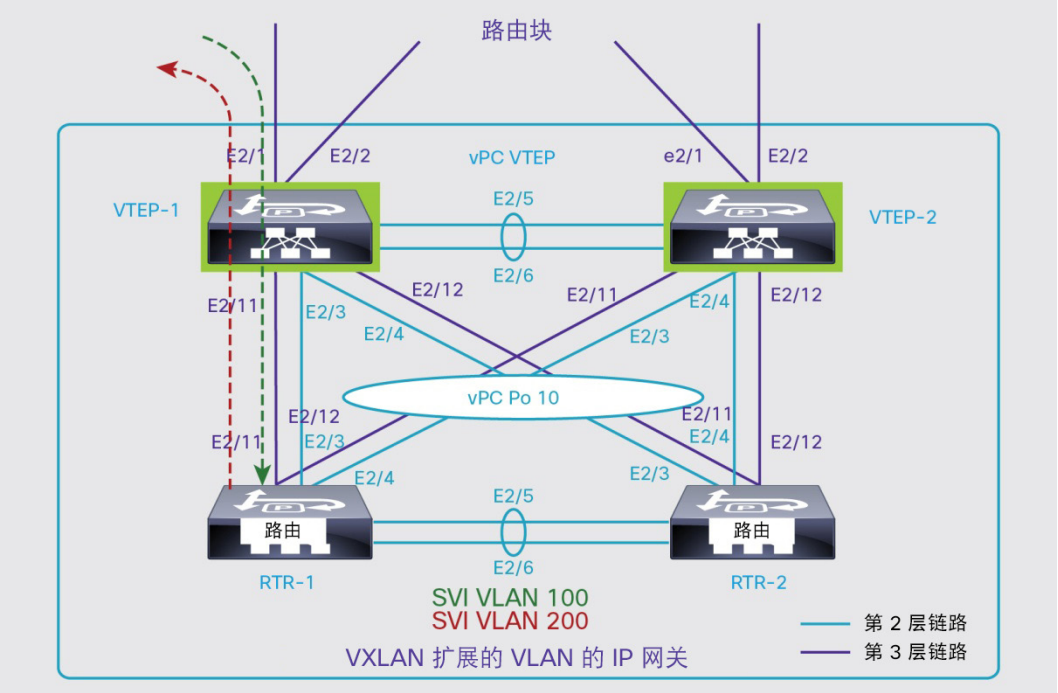
路由块配置

推荐的 VXLAN 路由设计中的路由块包含一对物理 VTEP 或 vPC VTEP（它们将 VXLAN VNI 转换回 VLAN），以及一台或一对路由器（它们用作 VLAN IP 子网之间的 VXLAN IP 子网和路由的 IP 网关）。为实现设备冗余，建议使用冗余的 VTEP 设备（例如作为 vPC VTEP 的一对 Cisco Nexus 9300）和运行第一跳冗余协议（例如热待机路由器协议 [HSRP]）的一对路由器。

图 16 显示了设计有两对 Cisco Nexus 9300 平台交换机的路由块的 VXLAN 路由块示例。一对 Cisco Nexus 9300 平台交换机用作在 VXLAN 与 VLAN之间映射的 vPC VTEP。第二对是 VXLAN 扩展的 VLAN 的IP 网关。在用于第 2 层连接的两对交换机之间存在双倍大小的 vPC。可以安装单独的一组第 3 层链路，以便在 VXLAN VLAN 与非 VXLAN VLAN 或 IP 网络之间路由。附录 A 提供了路由块中的设备的相关配置。

注：在思科 NX-OS 版本 6.1(2)I2(2a) 之前，由于一个已知的软件问题，vPC VTEP 的对等链路和路由块中的路由器的第 2 层链路不能在 Cisco Nexus 9300 平台交换机的 40 千兆以太网链路上。该问题在思科 NX-OS 版本 6.1(2)I2(2a) 中得到解决。

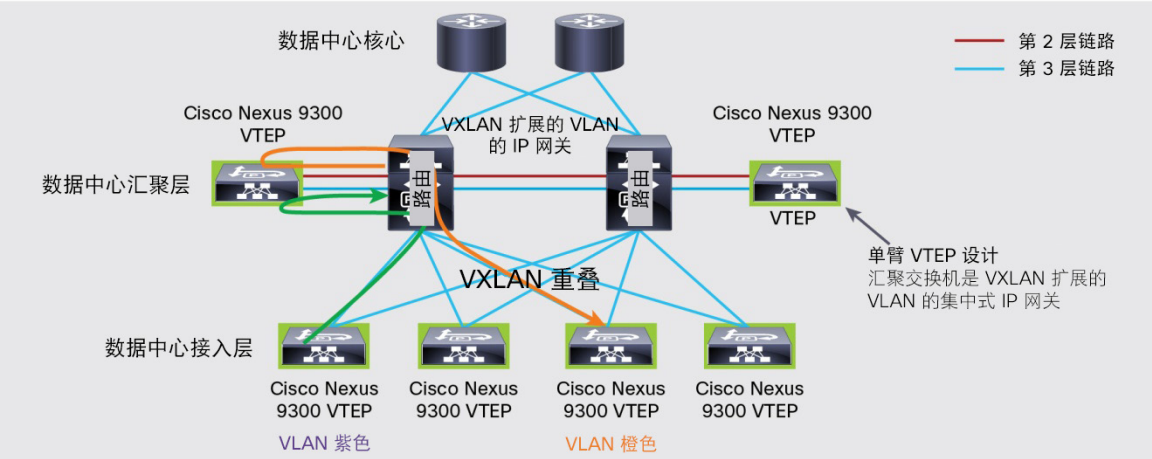
图 16. Cisco Nexus 9300 平台交换机的路由块设计



VXLAN 间的路由设计方案 B：单臂 VTEP 设计

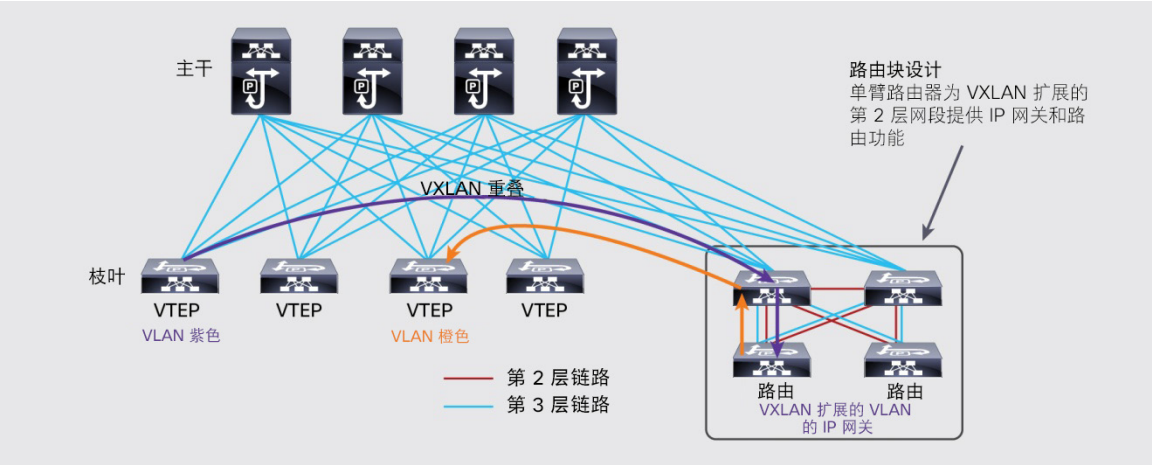
图 17 显示了 VXLAN 间的路由的一种替代设计。它采用单臂 VTEP 设计，其中一个或一对 Cisco Nexus 9300 VTEP 通过第 2 层链路和第 3 层链路连接到汇聚交换机。第 3 层链路用于与机架内的 VTEP 接入交换机建立 VXLAN 隧道，以将主机 VLAN 扩展到第 3 层网络。为汇聚交换机配置了其 IP 子网的主机 VLAN 和交换机虚拟接口 (SVI)。HSRP 和虚拟路由器冗余协议 (VRRP) 可以用于通过在两台汇聚交换机之间部署的第 2 层链路提供第一跳冗余。Cisco Nexus 9300 VTEP 将 VXLAN VNI 映射回 VLAN，并将流量通过第 2 层链路发送到汇聚交换机以进行 VLAN 间的路由。在数据包路由到目标 VLAN IP 子网后，汇聚交换机会通过第 2 层链路将数据包发送回 Cisco Nexus 9300 VTEP，以进行 VXLAN 封装。封装的数据包将通过底层第 3 层网络转发到目标机架。在此设计中，添加的 Cisco Nexus 9300 VTEP 扩展主机 VLAN 网段并将其带到汇聚交换机中。汇聚交换机是 VXLAN 扩展的 VLAN 的集中式 IP 网关。

图 17. VLAN 间路由设计：单臂 VTEP



单臂 VTEP 设计将 VXLAN 扩展的 VLAN 的 IP 网关保持在汇聚交换机上，这样做保留了传统的第 2 层数据中心 Pod 的 IP 网关位置。但是，它可以在未来为将网络迁移到主干-枝叶交换矩阵架构创建块。如图 18 所示，相比之下，路由块设计可以更轻松地将现有的汇聚层和接入层架构转化为真正的主干-枝叶交换矩阵。此架构真正支持跨路由的（第 3 层）交换矩阵在第 2 层邻接。

图 18. 使用 VXLAN 发展为主干-枝叶架构



后续内容

目前 Cisco Nexus 9300 平台交换机仅支持 VXLAN 网关和桥接功能。Cisco NX-OS 的一个已计划的未来版本将把 VXLAN 路由功能引入 Cisco Nexus 9300 平台，这将极大地简化 VXLAN 间路由的网络设计。

此外，思科正在致力于 VXLAN 的 BGP EVPN 控制平面。当前基于组播的 VXLAN 缺乏控制平面，必须依靠泛洪和学习在重叠网络中建立第 2 层转发信息库。底层网络中的组播用于支持重叠泛洪和学习行为。思科 BGP EVPN 控制平面基于标准，不依靠任何交换矩阵控制器。它将提供以下主要优势：

- 消除或减少数据中心的泛洪
- 在重叠网络上实现最优处理多个目标流量（广播、未知单播和组播）
- 为 VXLAN VNI 中的主机提供可靠、快速的地址解析和更新：对于在数据中心内支持工作负载移动性至关重要
- 为 VXLAN 重叠网络提供分布式任播 IP 网关，在第 3 层网络中实现最优的 VXLAN 流量路由

结论

VXLAN 是网络虚拟化技术。它使用 MAC-in-IP-UDP 隧道机制在第 3 层基础设施中构建第 2 层重叠网络。此方法将租户网络视图与共享的通用基础设施分离，使组织可以构建可扩展、可靠的第 3 层数据中心网络，同时保持直接的第 2 层邻接关系。

Cisco Nexus 9300 平台交换机可以是物理 VTEP，提供基于硬件的高性能。Cisco Nexus 9300 平台交换机上的 VXLAN 功能发展迅速，VXLAN 间路由和 EVPN 控制平面功能已计划。这些增强功能问世后，有了 Cisco Nexus 9300 平台交换机的 VXLAN 重叠设计可以被进一步优化和简化。此解决方案在第 3 层交换矩阵为第 2 层重叠提供数据中心网络设计，以帮助提供多租户环境所需的应用工作负载移动性和网络虚拟化。

相关详细信息

- http://www.ietf.org/id/draft-mahalingam-dutt-dcops-VXLAN-09.txt
- https://datatracker.ietf.org/doc/draft-ietf-nvo3-arch/
- http://www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/white-paper-c11-729383.html

附录 A：Cisco Nexus 9300 VTEP 交换机配置示例

```
n9396-vtep-1# sh run

!Command: show running-config
!Time: Thu Jul  3 17:33:40 2014

version 6.1(2)I2(2a)
hostname n9396-vtep-1
vdc n9396-vtep-1 id 1
  allocate interface Ethernet1/1-48
  allocate interface Ethernet2/1-12
  limit-resource vlan minimum 16 maximum 4094
  limit-resource vrf minimum 2 maximum 4096
  limit-resource port-channel minimum 0 maximum 768
  limit-resource u4route-mem minimum 248 maximum 248
  limit-resource u6route-mem minimum 96 maximum 96
  limit-resource m4route-mem minimum 58 maximum 58
  limit-resource m6route-mem minimum 8 maximum 8

feature nxapi
feature bash-shell
feature scp-server
feature ospf
feature pim
feature interface-vlan
feature vn-segment-vlan-based
feature lacp
feature nv overlay

logging level aaa 6
username admin password 5 $1$e8no0GAX$ptdDp5VsZCXG3unIumghO/  role network-admin
no password strength-check
ip domain-lookup
copp profile strict
snmp-server user admin network-admin auth md5 0x66ec86927ebe7a1eac0d1642ba15c553
priv 0x66ec86927ebe7a1eac0d1642ba15c553 localizedkey
rmon event 1 log trap public description FATAL(1) owner PMON@FATAL
rmon event 2 log trap public description CRITICAL(2) owner PMON@CRITICAL
rmon event 3 log trap public description ERROR(3) owner PMON@ERROR
rmon event 4 log trap public description WARNING(4) owner PMON@WARNING
rmon event 5 log trap public description INFORMATION(5) owner PMON@INFO
snmp-server community public group network-admin

ip pim ssm range 232.0.0.0/8
```

```
ip pim auto-rp listen
vlan 1,13,80,90,100-102,110,200,300,999
vlan 100
  vn-segment 5100
vlan 101
  vn-segment 5101
vlan 200
  vn-segment 5200

vrf context management
  ip route 0.0.0.0/0 173.42.127.1

interface Vlan1

interface nve1
  source-interface loopback0
  member vni 5100 mcast-group 239.1.1.1
  member vni 5101 mcast-group 239.1.1.2
  no shutdown

interface Ethernet1/1
  switchport mode trunk
  switchport access vlan 100
  switchport trunk allowed vlan 1,100-101
  storm-control broadcast level 10.00

interface Ethernet1/2

interface Ethernet1/3
  shutdown

interface Ethernet1/4
  shutdown

interface Ethernet1/5

interface Ethernet1/6
  shutdown

interface Ethernet1/7
  shutdown

interface Ethernet1/8
  shutdown
```

```
interface Ethernet1/9
  shutdown

interface Ethernet1/10

interface Ethernet1/11
  shutdown

interface Ethernet1/12
  shutdown

interface Ethernet1/13
  shutdown

interface Ethernet1/14
  shutdown

interface Ethernet1/15
  shutdown

interface Ethernet1/16
  shutdown

interface Ethernet1/17
  shutdown

interface Ethernet1/18
  shutdown

interface Ethernet1/19
  shutdown

interface Ethernet1/20
  shutdown

interface Ethernet1/21
  shutdown

interface Ethernet1/22
  shutdown

interface Ethernet1/23
  shutdown

interface Ethernet1/24
```

```
  shutdown

interface Ethernet1/25
  shutdown

interface Ethernet1/26
  shutdown

interface Ethernet1/27
  shutdown

interface Ethernet1/28
  shutdown

interface Ethernet1/29
  shutdown

interface Ethernet1/30
  shutdown

interface Ethernet1/31
  shutdown

interface Ethernet1/32
  switchport mode trunk
  switchport trunk allowed vlan 1,80,90,100

interface Ethernet1/33

interface Ethernet1/34
  shutdown

interface Ethernet1/35
  no switchport

interface Ethernet1/36
  shutdown

interface Ethernet1/37
  shutdown

interface Ethernet1/38
  shutdown

interface Ethernet1/39
```

```
  shutdown

interface Ethernet1/40
  shutdown

interface Ethernet1/41
  shutdown

interface Ethernet1/42
  shutdown

interface Ethernet1/43
  shutdown

interface Ethernet1/44
  shutdown

interface Ethernet1/45
  shutdown

interface Ethernet1/46
  shutdown

interface Ethernet1/47
  shutdown

interface Ethernet1/48
  shutdown

interface Ethernet2/1
  no switchport
  ip address 192.168.1.6/30
  ip ospf network point-to-point
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode
  no shutdown

interface Ethernet2/2
  no switchport
  ip address 192.168.1.10/30
  ip ospf network point-to-point
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode
  no shutdown
```

```
interface Ethernet2/3
  shutdown

interface Ethernet2/4
  shutdown

interface Ethernet2/5
  shutdown

interface Ethernet2/6
  shutdown

interface Ethernet2/7
  shutdown

interface Ethernet2/8
  shutdown

interface Ethernet2/9
  shutdown

interface Ethernet2/10
  shutdown

interface Ethernet2/11
  shutdown

interface Ethernet2/12
  shutdown

interface mgmt0
  vrf member management
  ip address 173.42.127.15/24

interface loopback0
  ip address 10.1.1.2/32
  ip ospf network point-to-point
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode
line console
line vty
boot nxos bootflash:/n9000-dk9.6.1.2.I2.2a.bin
router ospf 1
  router-id 10.1.1.2
no xml server exec-mode
```

```
logging server 173.42.127.175 7 use-vrf management
logging source-interface mgmt0
```

附录 B：ACI 路由块配置

vPC VTEP 上的 VXLAN 配置

VTEP-1

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

vlan 100
  vn-segment 5100
vlan 101
  vn-segment 5101

interface nve1
  source-interface loopback0
  member vni 5100 mcast-group 239.1.1.1
  member vni 5101 mcast-group 239.1.1.2

interface loopback0
  no ip redirects
  ip address 10.1.1.4/32
  ip address 10.1.1.100/32 secondary
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode

interface Ethernet2/1
  no switchport
  ip address 192.168.1.10/30
  ip ospf network point-to-point
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode
  no shutdown

interface Ethernet2/2
  no switchport
  ip address 192.168.2.10/30
  ip ospf network point-to-point
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode
  no shutdown


feature vpc
```

VTEP-2

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

vlan 100
  vn-segment 5100
vlan 101
  vn-segment 5101

interface nve1
  source-interface loopback0
  member vni 5100 mcast-group 239.1.1.1
  member vni 5101 mcast-group 239.1.1.2

interface loopback0
  no ip redirects
  ip address 10.1.1.5/32
  ip address 10.1.1.100/32 secondary
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode

interface Ethernet2/1
  no switchport
  ip address 192.168.1.14/30
  ip ospf network point-to-point
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode
  no shutdown

interface Ethernet2/2
  no switchport
  ip address 192.168.2.14/30
  ip ospf network point-to-point
  ip router ospf 1 area 0.0.0.0
  ip pim sparse-mode
  no shutdown


feature vpc
```

```
vpc domain 100
  peer-switch
  peer-keepalive destination
172.21.128.79 source 172.21.128.104
  peer-gateway

interface port-channel1
  switchport mode trunk
  spanning-tree port type network
  vpc peer-link

interface port-channel10
  switchport mode trunk
  vpc 10

interface Ethernet2/3
  switchport mode trunk
  channel-group 10 mode active

interface Ethernet2/4
  switchport mode trunk
  channel-group 10 mode active

interface Ethernet2/5
  switchport mode trunk
  channel-group 1 mode active

interface Ethernet2/6
  switchport mode trunk
  channel-group 1 mode active
```

```
vpc domain 100
  peer-switch
  peer-keepalive destination
172.21.128.104 source 172.21.128.79
  peer-gateway

interface port-channel1
  switchport mode trunk
  spanning-tree port type network
  vpc peer-link

interface port-channel10
  switchport mode trunk
  vpc 10

interface Ethernet2/3
  switchport mode trunk
  channel-group 10 mode active

interface Ethernet2/4
  switchport mode trunk
  channel-group 10 mode active

interface Ethernet2/5
  switchport mode trunk
  channel-group 1 mode active

interface Ethernet2/6
  switchport mode trunk
  channel-group 1 mode active
```

路由器配置

路由器 1

```
feature vpc

vpc domain 1
  peer-switch
  peer-keepalive destination 5.5.5.5
source 5.5.5.4 vrf default
  peer-gateway
  ip arp synchronize

interface port-channel1
```

路由器 2

```
feature vpc

vpc domain 1
  peer-switch
  peer-keepalive destination 5.5.5.4
source 5.5.5.5 vrf default
  peer-gateway
  ip arp synchronize

interface port-channel1
```

```
  switchport mode trunk
  spanning-tree port type network
  vpc peer-link

interface port-channel10
  switchport mode trunk
  vpc 10

interface Ethernet2/3
  switchport mode trunk
  channel-group 1 mode active

interface Ethernet2/4
  switchport mode trunk
  channel-group 1 mode active

interface Ethernet2/5
  switchport mode trunk
  channel-group 10 mode active

interface Ethernet2/6
  switchport mode trunk
  channel-group 10 mode active


interface Vlan100
  no shutdown
  ip address 100.0.0.2/24
  hsrp 100
    preempt
    priority 110
    ip 100.0.0.1

interface Vlan101
  no shutdown
  no ip redirects
  ip address 101.0.0.2/24
  no ipv6 redirects
  hsrp 101
    preempt
    priority 110
    ip 101.0.0.1
```

```
  switchport mode trunk
  spanning-tree port type network
  vpc peer-link

interface port-channel10
  switchport mode trunk
  vpc 10

interface Ethernet2/3
  switchport mode trunk
  channel-group 1 mode active

interface Ethernet2/4
  switchport mode trunk
  channel-group 1 mode active

interface Ethernet2/5
  switchport mode trunk
  channel-group 10 mode active

interface Ethernet2/6
  switchport mode trunk
  channel-group 10 mode active


interface Vlan100
  no shutdown
  ip address 100.0.0.3/24
  hsrp 100
    preempt
    ip 100.0.0.1

interface Vlan101
  no shutdown
  no ip redirects
  ip address 101.0.0.3/24
  no ipv6 redirects
  hsrp 101
    preempt
    ip 101.0.0.1
```

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 C11-732453-00 10/14