**EVPN基本原理**

# 一、EVPN基本原理介绍

EVPN（Ethernet Virtual Private Network）是一种用于二层网络互联的VPN技术。EVPN技术采用类似于BGP/MPLS IP VPN的机制，在BGP协议的基础上定义了一种新的NLRI（Network Layer Reachability Information，网络层可达信息）即EVPN NLRI，EVPN NLRI定义了几种新的BGP EVPN路由类型，用于处在二层网络的不同站点之间的MAC地址学习和发布。

原有的VXLAN实现方案没有控制平面，是通过数据平面的流量泛洪进行VTEP发现和主机信息（包括IP地址、MAC地址、VNI、网关VTEP IP地址）学习的，这种方式导致数据中心网络存在很多泛洪流量。为了解决这一问题，VXLAN引入了EVPN作为控制平面，通过在VTEP之间交换BGP EVPN路由实现VTEP的自动发现、主机信息相互通告等特性，从而避免了不必要的数据流量泛洪。

综上所述，EVPN通过扩展BGP协议新定义了几种BGP EVPN路由，这些BGP EVPN路由可以用于传递VTEP地址和主机信息，因此EVPN应用于VXLAN网络中，可以使VTEP发现和主机信息学习从数据平面转移到控制平面。

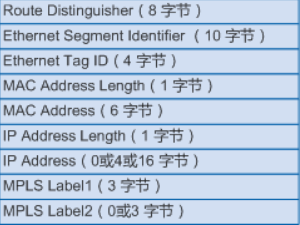
BGP EVPN路由

在EVPN NLRI中定义了如下几种应用于VXLAN控制平面的BGP EVPN路由类型：

## **1. Type2路由：MAC/IP路由**

该类型路由的报文格式如下图所示：

图1 MAC/IP路由的报文格式

****

各字段的解释如下表所示：

****

该类型路由在VXLAN控制平面中的作用包括：

1）主机MAC地址通告

要实现同子网主机的二层互访，两端VTEP需要相互学习主机MAC。作为BGP EVPN对等体的VTEP之间通过交换MAC/IP路由，可以相互通告已经获取到的主机MAC。其中，MAC Address Length和MAC Address字段为主机MAC地址。

2）主机ARP通告

MAC/IP路由可以同时携带主机MAC地址+主机IP地址，因此该路由可以用来在VTEP之间传递主机ARP表项，实现主机ARP通告。其中，MAC Address和MAC Address Length字段为主机MAC地址，IP Address和IP Address Length字段为主机IP地址。此时的MAC/IP路由也称为ARP类型路由。主机ARP通告主要用于以下两种场景：

**1.ARP广播抑制。**当三层网关学习到其子网下的主机ARP时，生成主机信息（包含主机IP地址、主机MAC地址、二层VNI、网关VTEP IP地址），然后通过传递ARP类型路由将主机信息同步到二层网关上。这样当二层网关再收到ARP请求时，先查找是否存在目的IP地址对应的主机信息，如果存在，则直接将ARP请求报文中的广播MAC地址替换为目的单播MAC地址，实现广播变单播，达到ARP广播抑制的目的。(对于硬件二层网关，除了在控制器上使能ARP广播抑制外，还要在设备上预配置BGP-EVPN，控制器上使能BGP-EVPN。BGP-EVPN协议会将学习的ARP表项生成主机信息，**三层网关会将主机信息发布给二层网关**。当二层网关收到ARP广播报文时，会根据学习到的主机信息将广播报文转换为单播报文进行转发，减少BD域内的广播报文，提高网络性能)。

**2.分布式网关场景下的虚拟机迁移。**当一台虚拟机从当前网关迁移到另一个网关下之后，新网关学习到该虚拟机的ARP（一般通过虚拟机发送免费ARP实现），并生成主机信息（包含主机IP地址、主机MAC地址、二层VNI、网关VTEP IP地址），然后通过传递ARP类型路由将主机信息发送给虚拟机的原网关。原网关收到后，感知到虚拟机的位置发生变化，触发ARP探测，当探测不到原位置的虚拟机时，撤销原位置虚拟机的ARP和主机路由。

3）主机IP路由通告

在分布式网关场景中，要实现跨子网主机的三层互访，两端VTEP（作为三层网关）需要互相学习主机IP路由。作为BGP EVPN对等体的VTEP之间通过交换MAC/IP路由，可以相互通告已经获取到的主机IP路由。其中，IP Address Length和IP Address字段为主机IP路由的目的地址，同时MPLS Label2字段必须携带三层VNI。此时的MAC/IP路由也称为IRB（Intergrate综合 Routing and Bridge）类型路由。

说明：

ARP类型路由携带的有效信息有：主机MAC地址+主机IP地址+二层VNI；

IRB类型路由携带的有效信息有：主机MAC地址+主机IP地址+二层VNI+三层VNI。

因此，IRB类型路由包含着ARP类型路由，不仅可以用于主机IP路由通告，也能用于主机ARP通告。

## **2. Type3路由:Inclusive Multicast（集中多播）路由**

该类型路由是由前缀和PMSI属性组成，报文格式如下图所示：

图2 Inclusive Multicast路由的报文格式

****

各字段的解释如下表所示：

****

该类型路由在VXLAN控制平面中主要用于VTEP的自动发现和VXLAN隧道的动态建立。作为BGP EVPN对等体的VTEP，通过Inclusive Multicast路由互相传递二层VNI和VTEP IP地址信息。其中，Originating Router's IP Address字段为本端VTEP IP地址，MPLS Label字段为二层VNI。如果对端VTEP IP地址是三层路由可达的，则建立一条到对端的VXLAN隧道。同时，如果对端VNI与本端相同，则创建一个头端复制表，用于后续BUM报文转发。

（问题：这是VNI相同建立二层vxlan隧道，如果是同vpc不同子网的场景呐？）

## **3. Type5路由:IP前缀路由**

该类型路由的报文格式如下图所示：

图3 IP前缀路由的报文格式

****

各字段的解释如下表所示：

****

该类型路由的IP Prefix Length和IP Prefix字段既可以携带主机IP地址，也可以携带网段地址：

1）当携带主机IP地址时，该类型路由在VXLAN控制平面中的作用与IRB类型路由是一样的，主要用于分布式网关场景中的主机IP路由通告。

2）当携带网段地址时，通过传递该类型路由，可以实现VXLAN网络中的主机访问外部网络。

# 二、BGP EVPN方式部署分布式网关

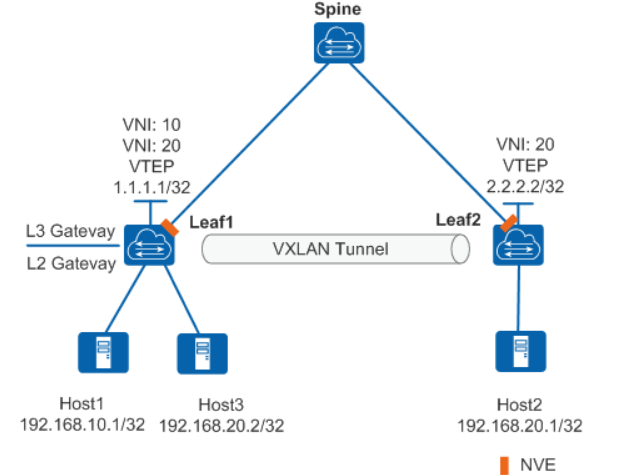
在BGP EVPN方式部署分布式网关的场景中，控制平面的流程包括VXLAN隧道建立、MAC地址动态学习；转发平面的流程包括同子网已知单播报文转发、同子网BUM报文转发、跨子网报文转发。该方式实现的功能全面，支持主机IP路由通告、主机MAC地址通告、主机ARP通告（详情参见EVPN基本原理），可以直接使能ARP广播抑制功能。如果在VXLAN网络中采用分布式网关，推荐使用此方式。

## **1. VXLAN隧道建立**

VXLAN隧道由一对VTEP IP地址确定，创建VXLAN隧道实际上是两端VTEP获取对端VTEP IP地址的过程，只要对端VTEP IP地址是三层路由可达的，VXLAN隧道就可以建立成功。通过BGP EVPN方式动态建立VXLAN隧道，就是在两端VTEP之间建立BGP EVPN对等体，然后对等体之间利用BGP EVPN路由来互相传递VNI和VTEP IP地址信息，从而实现动态建立VXLAN隧道。

在分布式网关的场景中，Leaf同时作为二层网关和三层网关，Spine节点不感知VXLAN隧道，只作为VXLAN报文的转发节点。控制平面只需在Leaf之间建立VXLAN隧道，如图1所示，Leaf1与Leaf2之间建立VXLAN隧道用于Host1和Host2、Host3和Host2之间的通信。对于Host1和Host3之间的通信，由于都属于Leaf1，所以互访的流量只需在Leaf1上处理，无需通过VXLAN隧道转发。

图1 VXLAN隧道示意图

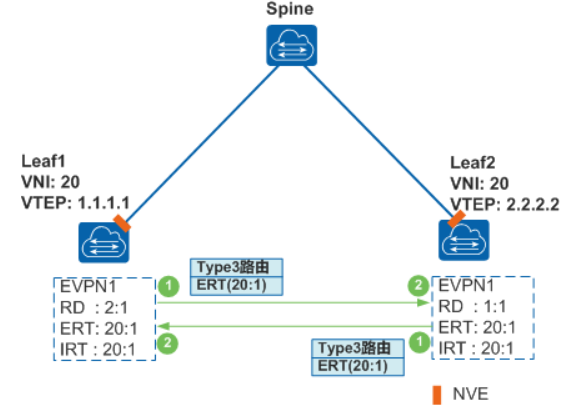
****

在分布式网关场景中，通过BGP EVPN方式动态建立VXLAN隧道有以下两种情况：

### **1.1 同子网互通：**

如图2所示，当处于同一子网的Host3与Host2互通时，只需要进行二层转发。通过BGP EVPN方式建立VXLAN隧道的过程如下：

图2 动态建立VXLAN隧道示意图-1

****

1、首先在Leaf1和Leaf2之间建立BGP EVPN对等体。然后，在Leaf1和Leaf2上分别创建二层广播域，并在二层广播域下配置关联的二层VNI。接下来在二层广播域下创建EVPN实例，配置本端EVPN实例的RD、出方向VPN-Target（ERT）、入方向VPN-Target（IRT）。在配置完本端VTEP IP地址后，Leaf1和Leaf2会生成BGP EVPN路由并发送给对端，该路由携带本端EVPN实例的出方向VPN-Target和BGP EVPN协议新定义的Type3路由即Inclusive Multicast路由。其中，Inclusive Multicast路由如图3所示，由前缀和PMSI属性组成，VTEP IP地址存放在前缀的Originating Router's IP Address字段中，二层VNI存放在PMSI属性的MPLS Label字段中。

图3 Inclusive Multicast路由

****

2、Leaf1和Leaf2在收到对端发来的BGP EVPN路由后，首先检查该路由携带的EVPN实例的出方向VPN-Target，如果与本端EVPN实例的入方向VPN-Target相等，则接收该路由，否则丢弃该路由。在接收该路由后，Leaf1和Leaf2将获取其中携带的对端VTEP IP地址和二层VNI，如果对端VTEP IP地址是三层路由可达的，则建立一条到对端的VXLAN隧道；同时，如果对端二层VNI与本端相同，则创建一个头端复制表，用于后续BUM（Broadcast&Unknown-unicast&Multicast）报文转发。

说明：

VPN-Target是一种32位的BGP扩展团体属性，一个EVPN实例可以配置出方向和入方向两类VPN-Target，两端EVPN实例的VPN-Target要相互匹配（即本端EVPN实例配置的出方向VPN-target值需要与对端EVPN实例配置的入方向VPN-target值相等），才能相互交换EVPN路由，否则VXLAN隧道无法建立成功。如果仅有一端匹配成功可以接收路由，则此端设备上可以建立通往另一端设备的隧道，但是无法传输数据报文，因为另一端设备在收到报文后，将会检查本端是否有通往对端的VXLAN隧道，如果没有的话，将会丢弃报文。

### **1.2 跨子网互通：**

当处于不同子网的Host1与Host2互通时，需要进行三层转发，因此在通过BGP EVPN方式建立VXLAN隧道的过程中，网关Leaf1和Leaf2需要发布下属主机的IP路由。一般情况下，这里发布的是32位主机IP路由，因为在VXLAN网络中，不同的Leaf节点可能连接着相同的网段，所以如果Leaf节点发布的是下属主机IP所在的网段路由，则可能与其他Leaf节点发布的网段路由冲突，进而导致某些Leaf节点的下属主机不可达。只有在如下两种场景中，Leaf节点可发布网段路由：

1）Leaf节点连接的网段在整个VXLAN网络中是唯一的，而且有效的主机明细路由数量较大，此时可发布主机IP所在的网段路由，从而减轻Leaf节点上路由存储的压力。

2）VXLAN网络中的主机需要访问外部网络，此时Leaf节点可在VXLAN网络中发布其连接的外部网段路由，从而使其他Leaf节点学习到去往外部网络的路由。

建立VXLAN隧道之前，需要在Leaf1和Leaf2上进行如下配置准备：

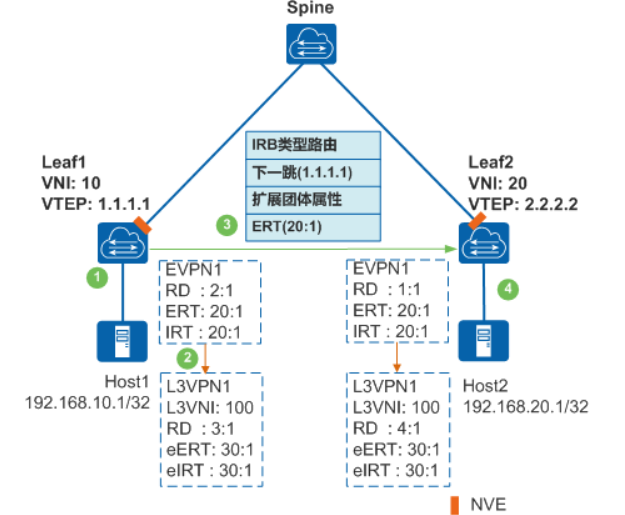
****

****

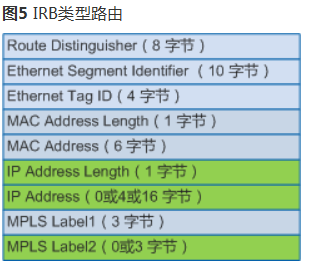
根据主机IP路由发布方式的不同，动态建立VXLAN隧道的过程有以下两种：

#### **1）通过IRB类型路由发布主机IP路由（如图4所示）**

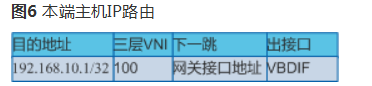
图4 动态建立VXLAN隧道示意图-2

****

1、Host1首次与Leaf1通信时，通过动态ARP报文，Leaf1学习到Host1的ARP表项。同时，Leaf1根据Host1所在的二层广播域找到绑定VBDIF接口的L3VPN实例，获取关联的三层VNI。然后Leaf1上的EVPN实例根据上述信息生成IRB类型路由（ip+mac+L2VNI+L3VNI），如图5所示。其中，主机IP地址存放在IP Address Length和IP Address字段中，三层VNI存放在MPLS Label2字段中。

****

2、Leaf1上的EVPN实例从IRB类型路由中获取Host1的主机IP地址、三层VNI，发给本端L3VPN实例，L3VPN实例在其路由表中保存Host1的主机IP路由（生成本地的L3VPN的主机路由表），如图6所示

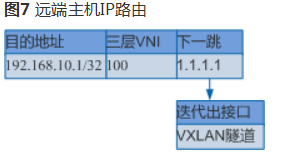
****

3、Leaf1向Leaf2发送BGP EVPN路由，该路由携带本端EVPN实例的ERT、扩展团体属性、路由下一跳属性以及IRB类型路由。其中，扩展团体属性携带的是隧道类型（取值是VXLAN隧道）、本端VTEP的MAC地址；路由下一跳属性携带的是本端的VTEP IP地址。

4、Leaf2收到Leaf1发来的BGP EVPN路由后，同时进行如下处理：

1）检查该路由携带的ERT，如果与本端EVPN实例的IRT相同，则接收该路由，否则丢弃该路由。EVPN实例获取到IRB类型路由后，还能提取到其中包含的ARP类型路由，用于主机ARP通告。

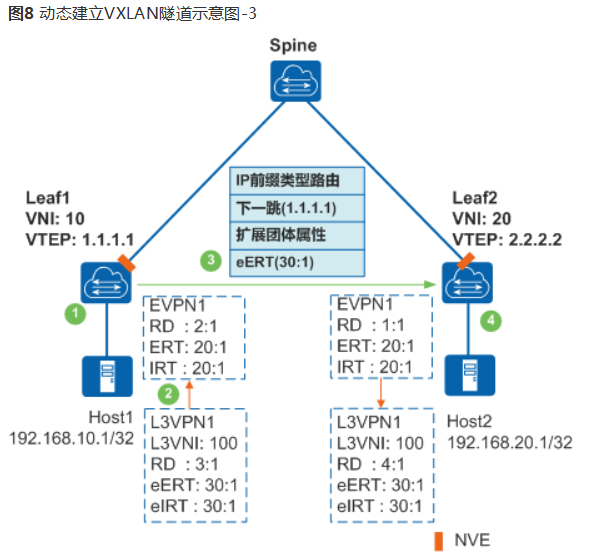
2）检查该路由携带的ERT，如果与本端L3VPN实例的eIRT相同，则接收该路由，否则丢弃该路由。然后，L3VPN实例获取到该路由携带的IRB类型路由，从中提取Host1的主机IP地址、三层VNI，在其路由表中保存Host1的主机IP路由（生成到对端的L3VPN的路由表），并将路由的下一跳迭代出接口设置为VXLAN隧道接口，如图7所示。

****

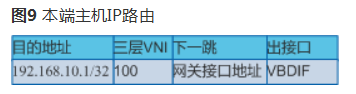
3）在通过EVPN实例或L3VPN实例接收该路由后，Leaf2通过下一跳属性获取Leaf1的VTEP IP地址，如果该VTEP IP地址是三层路由可达的，则建立一条到Leaf1的VXLAN隧道。

Leaf1建立到Leaf2的VXLAN隧道的过程与上述相同，这里不再赘述。

#### **2）通过IP前缀类型路由发布主机IP路由（如图8所示）**

****

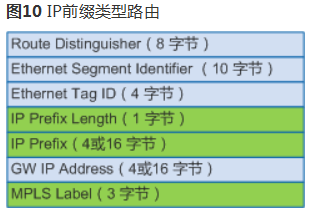
1、首先在Leaf1上将Host1的主机IP地址生成直连路由，然后在Leaf1上配置L3VPN实例引入直连路由，这样Host1的主机IP路由就保存到L3VPN实例的路由表中，并添加L3VPN实例关联的三层VNI，如图9所示。

****

说明：

如果要发布网段路由，这里需要先利用动态路由协议（如OSPF等）发布该网段路由，再配置L3VPN实例引入对应的动态协议路由。

2.在Leaf1上配置L3VPN实例向EVPN实例发布IP路由后，L3VPN实例下Host1的主机IP路由将发布给EVPN实例。然后由EVPN实例生成IP前缀类型路由，如图10所示。其中，主机IP地址存放在IP Prefix Length和IP Prefix字段中，三层VNI存放在MPLS Label字段中。

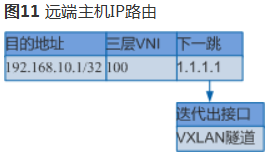
****

3、Leaf1向Leaf2发送BGP EVPN路由，该路由携带本端L3VPN实例的eERT、扩展团体属性、路由下一跳属性以及IP前缀类型路由。其中，扩展团体属性携带的是隧道类型（取值是VXLAN隧道）、本端VTEP的MAC地址；路由下一跳属性携带的是本端的VTEP IP地址。

4、Leaf2收到Leaf1发来的BGP EVPN路由后，同时进行如下处理：

1）检查该路由携带的eERT，如果与本端EVPN实例的IRT相同，则接收该路由，否则丢弃该路由。

2）检查该路由携带的eERT，如果与本端L3VPN实例的eIRT相同，则接收该路由，否则丢弃该路由。然后，L3VPN实例获取到该路由携带的IP前缀类型路由，从中提取Host1的主机IP地址、三层VNI，在其路由表中保存Host1的主机IP路由，并将路由的下一跳迭代出接口设置为VXLAN隧道接口，如图11所示。

****

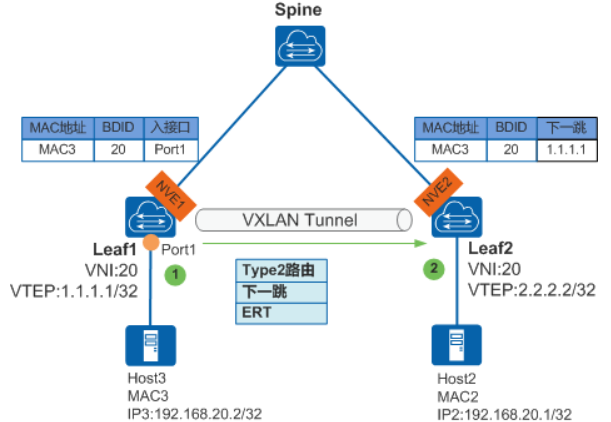
3）在通过EVPN实例或L3VPN实例接收该路由后，Leaf2通过下一跳属性获取Leaf1的VTEP IP地址，如果该VTEP IP地址是三层路由可达的，则建立一条到Leaf1的VXLAN隧道。

Leaf1建立到Leaf2的VXLAN隧道的过程与上述相同，这里不再赘述。

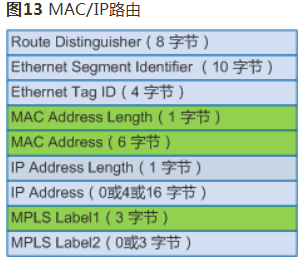
## **2. MAC地址动态学习**

在VXLAN网络中，为了实现终端租户的互通，支持MAC地址动态学习，不需要网络管理员手工维护，大大减少了维护工作量。在分布式网关场景中，跨子网互通需要进行三层转发，MAC地址学习只在本端主机和网关之间通过动态ARP报文实现，这里不再详述。下面结合图12，详细介绍一下同子网主机互通时，MAC地址动态学习的过程：

图12 MAC地址动态学习

****

1、Host3首次与Leaf1通信时，通过动态ARP报文，Leaf1学习到Host3的MAC地址、BDID（二层广播域标识）和报文入接口（即二层子接口对应的物理接口Port1）的对应关系，记录在本地MAC表中。同时，Leaf1根据Host3的ARP表项生成BGP EVPN路由并发送给对等体Leaf2，该路由携带本端EVPN实例的出方向VPN-Target、路由下一跳属性以及BGP EVPN协议新定义的Type2路由即MAC/IP路由。其中，路由下一跳属性携带的是本端VTEM IP地址；MAC/IP路由如图13所示，Host3的MAC地址存放在MAC Address Length和MAC Address字段中，二层VNI存放在MPLS Label1字段中。

****

2、Leaf2收到Leaf1发来的BGP EVPN路由后，首先检查该路由携带的EVPN实例的出方向VPN-Target，如果与本端EVPN实例的入方向VPN-Target相等，则接收该路由，否则丢弃该路由。在接收该路由后，Leaf2获得Host3的MAC地址、BDID和Leaf1上VTEP IP地址（下一跳属性）的对应关系，并保存到本地的MAC表中。

Leaf1学习Host2的主机MAC的过程与上述过程相同，这里不再赘述。

Host3初次与Host2通信时，首先发送目的MAC为全F、目的IP为IP2的ARP请求报文，请求Host2的MAC地址。缺省情况下，Leaf1收到该ARP请求后将在本网段进行广播，为了减少广播报文，此时可以在Leaf1上使能ARP广播抑制功能。这样当Leaf1收到该ARP请求报文时，先根据目的IP检查本地是否有Host2的MAC地址，如果有则将目的MAC替换为Host2的MAC地址，将ARP请求的广播报文变为单播报文，然后通过VXLAN隧道发给Leaf2。Leaf2收到后转发给Host2，Host2收到该ARP请求后学习到Host3的MAC地址，并以单播形式进行ARP应答。Host3收到ARP应答报文后学习到Host2的MAC地址。至此，Host3和Host2互相学习到对方的MAC地址，后续双方将采用单播通信。

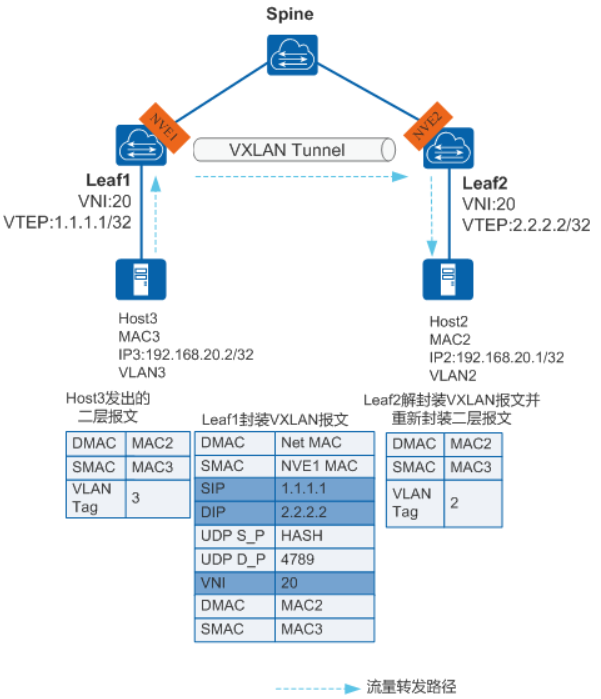
说明：

Leaf节点也可以通过数据转发过程学习主机MAC地址，但是依赖于设备从数据报文中学习MAC地址的能力。在BGP EVPN方式建立VXLAN隧道的场景中，Leaf节点之间可以通过BGP EVPN路由动态学习主机MAC地址，不再依赖数据转发过程。此时，可以关闭Leaf节点从数据报文中学习MAC地址的能力，提升转发效率。

## **3. 同子网已知单播报文转发**

同子网已知单播报文转发只在VXLAN二层网关之间进行，三层网关无需感知。报文转发流程如图14所示。

图14 同子网已知单播报文转发示意图

****

1.Leaf1收到来自Host3的报文，根据报文中接入的端口和VLAN信息获取对应的二层广播域，并在该二层广播域内查找出接口和封装信息。

2.Leaf1上VTEP根据查找到的封装信息对数据报文进行VXLAN封装，然后根据查找到的出接口进行报文转发。

3.Leaf2上VTEP收到VXLAN报文后，根据UDP目的端口号、源/目的IP地址、VNI判断VXLAN报文的合法有效性。然后依据VNI获取对应的二层广播域，进行VXLAN解封装，获取内层的二层报文。

4.Leaf2根据内层二层报文的目的MAC，从本地MAC表找到对应的出接口和封装信息，为报文添加VLAN Tag，转发给对应的主机Host2。

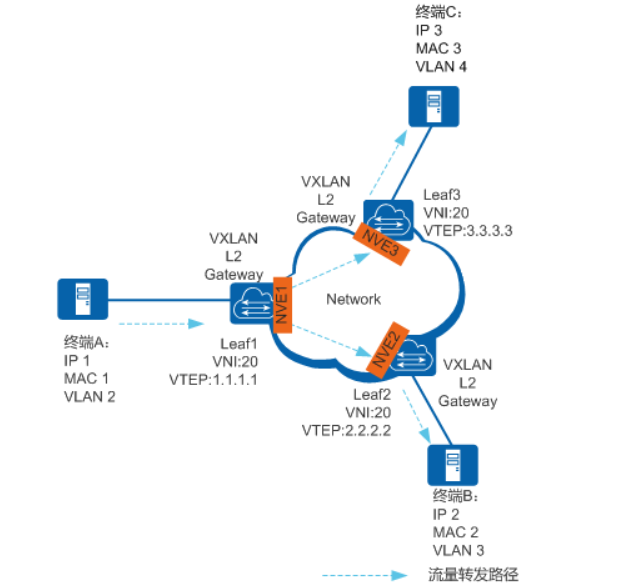
Host2向Host3发送报文的过程类似，这里不再赘述。

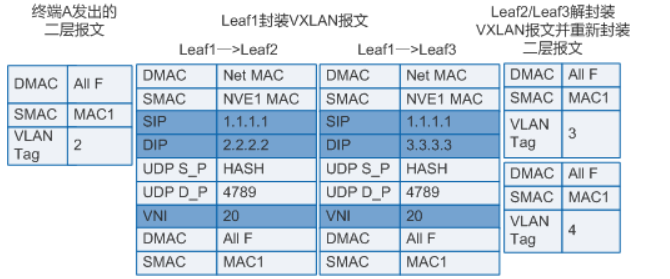
## **4. 同子网BUM报文转发**

同子网BUM报文转发只在VXLAN二层网关之间进行，三层网关无需感知。同子网BUM报文转发可以采用头端复制方式。

头端复制是指，当BUM报文进入VXLAN隧道时，接入端VTEP根据头端复制列表进行报文的VXLAN封装，并将报文发送给头端复制列表中的所有出端口VTEP。BUM报文出VXLAN隧道时，出口端VTEP对报文解封装。BUM报文采用头端复制的转发流程如图15所示。

图15 同子网BUM报文采用头端复制转发示意图

****

****

1.Leaf1收到来自终端A的报文，根据报文中接入的端口和VLAN信息获取对应的二层广播域。

2.Leaf1上VTEP根据对应的二层广播域获取对应VNI的头端复制隧道列表，依据获取的隧道列表进行报文复制，并进行VXLAN封装。然后将封装后的报文从出接口转发出去。

3.Leaf2/Leaf3上VTEP收到VXLAN报文后，根据UDP目的端口号、源/目的IP地址、VNI判断VXLAN报文的合法有效性。然后依据VNI获取对应的二层广播域，进行VXLAN解封装，获取内层二层报文。

4.Leaf2/Leaf3检查内层二层报文的目的MAC，发现是BUM MAC，在对应的二层广播域内的非VXLAN隧道侧进行广播处理，即：Leaf2/Leaf3分别从本地MAC表中找到非VXLAN隧道侧的所有出接口和封装信息，为报文添加VLAN Tag，转发给对应的终端B/C。

说明：

终端B/C向终端A回应报文，参考同子网已知单播报文转发。

头端复制列表详细一点介绍，咱们举个例子：

****

为了解决 VxLan对组播的依赖，所以大家提出给予 HER 的方案，全称是 Head-end Replication，是一种广播/组播流量转成单播的方案。

MAC地址是BUM（广播、未知单播、组播）地址，当BUM报文进入VXLAN隧道，接入端VTEP采用按照头端复制方式进行报文的VXLAN封装（接口收到BUM报文，本地VTEP通过控制平面获取属于同一个VNI的VTEP列表，将收到的BUM报文根据VTEP列表进行复制并发送给属于同一个VNI的所有VTEP）。BUM报文出VXLAN隧道，出口端VETP对报文解封装。（注：通过头端复制方式进行BUM报文的广播，不依赖于组播路由协议）

1、TOR1收到来自终端H1的报文，根据报文接入的端口和VLAN信息获取对应的二层广播域，并判断报文的目的MAC是否为BUM MAC，如果是则进入BUM报文转发流程，否则走单播流程进行转发。

2、在TOR1判断是BUM报文的情况下，TOR1上VTEP根据对应的二层广播域获取对应VNI的头端复制隧道列表，依据获取的隧道列表进行报文复制，并进行VXLAN封装，并从出端口转发。

3、TOR2/TOR3上VTEP收到VXLAN报文后，根据UDP目的端口号、源/目的IP地址、VNI判断VXLAN报文的合法有效性。依据VNI获取对应的二层广播域，然后进行VXLAN解封装，获取内层二层报文。

4、TOR2/TOR3根据获取的内层二层报文进行转发，如果二层报文的目的mac能够命中mac表，则根据mac表中查找到的端口进行报文的封装和转发，否则将报文广播到TOR的用户侧端口。

VxLAN是承载在IP Fabric网络之上的二层网络隧道，和其他的二层网络一样，针对BUM（Broadcast、Unknown-unicast、Multicast，广播、未知单播、组播）报文，行为都是广播。当前，NVE广播的处理行为如下：

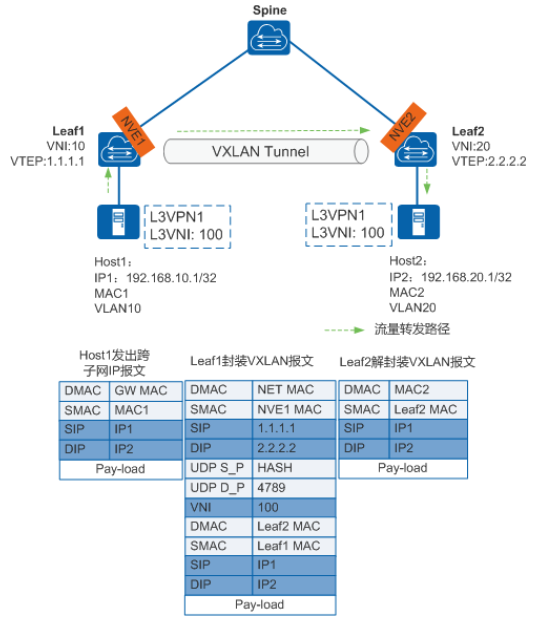
1)从接入侧接收到BUM报文，报文被复制到其他接口侧接口以及所有网络侧接口（即，VXLAN隧道）。

2)从网络侧接收到BUM报文，报文被复制到所有接入侧接口，而不往其他网络侧接口复制（俗称水平分割）。

## **5. 跨子网报文转发**

跨子网报文转发需要通过三层网关实现。在分布式网关场景中，跨子网报文转发的流程如图16所示。

图16 跨子网报文转发示意图

****

1.Leaf1收到来自Host1的报文，检测到报文的目的MAC是网关接口MAC，判断该报文需要进行三层转发。

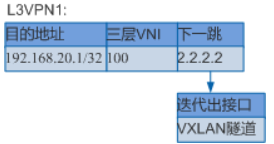
2.Leaf1根据报文的入接口找到对应的二层广播域，然后找到绑定该广播域VBDIF接口的L3VPN实例。根据报文的目的IP地址，查找该L3VPN实例下的路由表（如图17所示），获取该路由对应的三层VNI，以及下一跳地址。再根据下一跳的迭代出接口是VXLAN隧道接口，判断需要进行VXLAN封装：

根据VXLAN隧道的目的IP和源IP地址，获取对应的MAC地址，并将内层目的MAC和源MAC替换。

将三层VNI封装到报文中。

外层封装VXLAN隧道的目的IP和源IP地址，源MAC地址为Leaf1的NVE1接口MAC地址，目的MAC地址为网络下一跳的MAC地址。

图17 L3VPN实例下的主机路由

****

3.封装后的报文根据外层MAC和IP信息在IP网络中传输，送达Leaf2。

4.Leaf2收到VXLAN报文后进行解封装，检测到报文的目的MAC是自己的MAC地址，判断该报文需要进行三层转发。

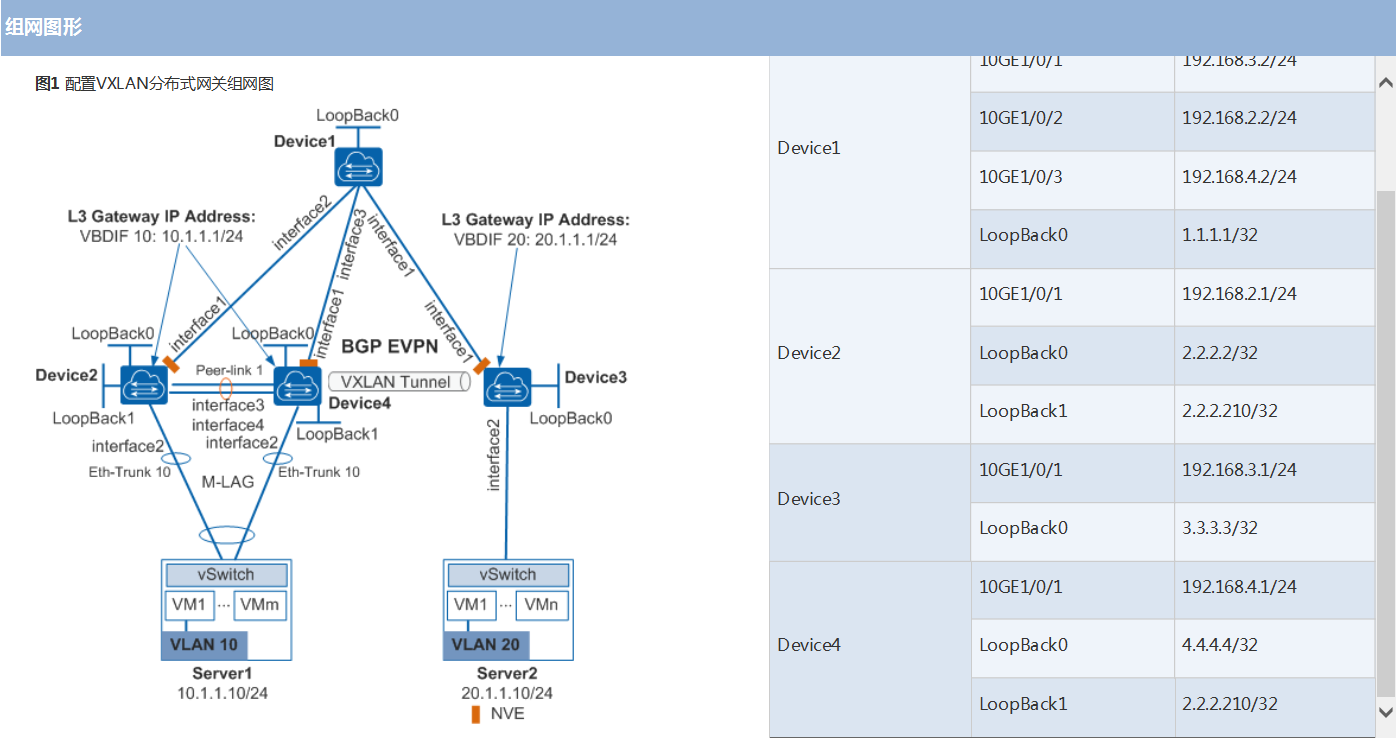
5.Leaf2根据报文携带的三层VNI找到对应的L3VPN实例，通过查找该L3VPN实例下的路由表（如图18所示），获取报文的下一跳是网关接口地址，然后将目的MAC地址替换为Host2的MAC地址，源MAC地址替换为Leaf2的MAC地址，转发给Host2。

图18 L3VPN实例下的主机路由

****

Host2向Host1发送报文的过程类似，这里不再赘述。

# 三、配置分布式网关部署方式的VXLAN示例（BGP EVPN方式）

****

**组网需求**

VXLAN分布式网关可解决VXLAN集中式网关的转发路径不优化、三层网关ARP表项规格瓶颈问题。

如图1所示，某企业在不同的数据中心中都拥有自己的VM，服务器1上的VM1属于VLAN 10，服务器2上的VM1属于VLAN 20，且位于不同网段；服务器1通过Device2和Device4接入VXLAN网络。现需要通过VXLAN分布式网关实现不同数据中心相同VM的互通。其中Device1部署在AS100，Device2、Device4部署在AS200，Device3部署在AS300，Device1、Device2、Device3、Device4使用AS100作为BGP EVPN的进程号。

说明：

本例中interface1、interface2、interface3、interface4分别代表10GE1/0/1、10GE1/0/2、10GE1/0/3、10GE1/0/4。

**配置思路**

采用如下思路配置VXLAN分布式网关：

1.分别在Device1与Device2、Device3、Device4间配置EBGP路由协议。

2.将Device2和Device4配置为根桥，配置相同的桥ID。

3.在Device2与Device4之间配置M-LAG。

4.分别在Device2、Device3和Device4上配置业务接入点实现区分业务流量。

5.使能EVPN作VXLAN控制平面。

6.在Device2、Device3和Device4上建立与Device1之间的EVPN IBGP对等体关系。

7.在Device1上建立与Device2、Device3和Device4之间的BGP EVPN对等体关系，并配置反射器功能。

8.在Device2、Device3和Device4上配置VPN实例和EVPN实例。

9.分别在Device2、Device3、Device4上使能头端复制功能。

10.在Device2、Device3、Device4上配置VXLAN三层网关。

11.在Device1与Device2、Device3、Device4之间配置BGP对邻居发布IRB类型的路由。

**数据准备**

为完成此配置例，需准备如下的数据：

VM所属的VLAN ID分别是VLAN 10和VLAN 20。

网络中设备互连的接口IP地址。

广播域BD ID分别是BD 10和BD 20。

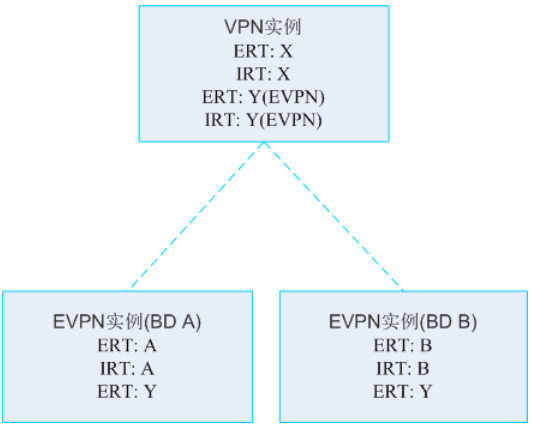
VXLAN网络标识VNI ID分别是VNI 10和VNI 20。

VPN实例下VXLAN网络标识VNI ID是VNI 5010。

EVPN实例的RD值为10:1、20:1、40:1，RT值为10:1、20:1、11:1。

VPN实例的RD值为11:11、44:44、22:22，RT值为1:1、2:2、11:1。

图2 配置RT值

****

RD值表示路由标识，需要全局唯一，因此VPN实例和EVPN实例下均需要配置不同的RD值。

VPN实例和EVPN实例的RT值配置原则如图2所示，其中：

在VPN实例中，除了配置本地VPN实例的ERT X和IRT X外，还需要配置带EVPN参数的ERT Y和IRT Y，用于和EVPN实例交叉，生成主机路由。

在EVPN实例中，除了针对不同的BD配置不同的ERT A、ERT B和IRT A、IRT B外，还必须配置用来和VPN实例交叉的ERT Y。一般情况下不需要配置IRT Y，否则会导致不同BD下的EVPN实例相互扩散MAC地址。

**操作步骤**

1.配置EBGP路由协议

# 配置Device1。Device2、Device3和Device4的配置与Device1类似，这里不再赘述。

<HUAWEI> system-view

[~HUAWEI] sysname Device1

[\*HUAWEI] commit

[~Device1] interface loopback 0

[\*Device1-LoopBack0] ip address 1.1.1.1 32

[\*Device1-LoopBack0] quit

[\*Device1] interface 10ge 1/0/1

[\*Device1-10GE1/0/1] undo portswitch

[\*Device1-10GE1/0/1] ip address 192.168.3.2 24

[\*Device1-10GE1/0/1] quit

[\*Device1] interface 10ge 1/0/2

[\*Device1-10GE1/0/2] undo portswitch

[\*Device1-10GE1/0/2] ip address 192.168.2.2 24

[\*Device1-10GE1/0/2] quit

[\*Device1] interface 10ge 1/0/3

[\*Device1-10GE1/0/3] undo portswitch

[\*Device1-10GE1/0/3] ip address 192.168.4.2 24

[\*Device1-10GE1/0/3] quit

[\*Device1] bgp 100

[\*Device1-bgp] peer 192.168.2.1 as-number 200

[\*Device1-bgp] peer 192.168.3.1 as-number 300

[\*Device1-bgp] peer 192.168.4.1 as-number 200

[\*Device1-bgp] network 1.1.1.1 32

[\*Device1-bgp] quit

[\*Device1] commit

2.配置VXLAN隧道模式并使能VXLAN的ACL扩展功能（仅CE6870EI设备需要配置此步骤）-

# 配置Device2。Device3、Device4的配置与Device2类似，这里不再赘述。

[~Device2] ip tunnel mode vxlan

[\*Device2] assign forward nvo3 acl extend enable

[\*Device2] commit

说明：

配置VXLAN隧道模式、使能VXLAN的ACL扩展功能后，需要保存配置并重启设备才能生效，您可以选择立即重启或完成所有配置后再重启。

3.将Device2和Device4配置为根桥，配置相同的桥ID

说明：

如果组成M-LAG的设备下行接入的是交换设备，必须配置根保护功能。

# 配置Device2。

[~Device2] stp root primary

[\*Device2] stp bridge-address 0039-0039-0039

[\*Device2] interface eth-trunk 10

[\*Device2-Eth-Trunk10] stp edged-port enable

[\*Device2-Eth-Trunk10] commit

[~Device2-Eth-Trunk10] quit# 配置Device4。

[~Device4] stp root primary

[\*Device4] stp bridge-address 0039-0039-0039

[\*Device4] interface eth-trunk 10

[\*Device4-Eth-Trunk10] stp edged-port enable

[\*Device4-Eth-Trunk10] commit

[~Device4-Eth-Trunk10] quit

4.在Device2与Device4之间配置M-LAG

说明：

如果Device2上行接入VXLAN网络的链路出现故障，当用户流量到达Device2时，由于没有可用的上行出接口，Device2会将用户流量全部丢弃。此时，可以配置monitor-link关联Leaf1的上行接口和下行接口，当Device2的上行出接口DOWN时，下行接口状态也会变为DOWN，这样用户侧流量就不会通过Device2进行转发，从而防止流量被丢弃。monitor-link的配置可以参见配置Monitor Link组的上行接口和下行接口。

a.创建Eth-Trunk接口，并将以太物理接口加入Eth-Trunk接口。

服务器上行连接交换机的端口需要绑定在一个聚合链路中且链路聚合模式需要和交换机侧的聚合模式匹配。

# 在Device2上创建Eth-Trunk，配置为LACP模式并加入成员口。Device4的配置与Device2类似，此处不再赘述。

[~Device2] interface eth-trunk 1

[\*Device2-Eth-Trunk1] mode lacp-static

[\*Device2-Eth-Trunk1] trunkport 10ge 1/0/3 to 1/0/4

[\*Device2-Eth-Trunk1] quit

[\*Device2] interface eth-trunk 10

[\*Device2-Eth-Trunk10] mode lacp-dynamic

[\*Device2-Eth-Trunk10] trunkport 10ge 1/0/2

[\*Device2-Eth-Trunk10] quit

[\*Device2] commit

b.分别在Device2和Device4上配置DFS Group。

# 配置Device2。Device4的配置与Device2类似，此处不再赘述。

[~Device2] dfs-group 1

[\*Device2-dfs-group-1] source ip 2.2.2.2

[\*Device2-dfs-group-1] quit

[\*Device2] commit

c.将Device2和Device4之间的链路配置为peer-link。

# 配置Device2。Device4的配置与Device2类似，此处不再赘述。

[~Device2] interface eth-trunk 1

[~Device2-Eth-Trunk1] undo stp enable

[\*Device2-Eth-Trunk1] peer-link 1

[\*Device2-Eth-Trunk1] quit

[\*Device2] commit

d.分别在Device2和Device4上配置绑定DFS和用户侧Eth-Trunk接口。

# 配置Device2。Device4的配置与Device2类似，此处不再赘述。

[~Device2] interface eth-trunk 10

[~Device2-Eth-Trunk10] dfs-group 1 m-lag 1

[\*Device2-Eth-Trunk10] quit

[\*Device2] commit

5.分别在Device2、Device3、Device4上配置业务接入点-

# 配置Device2。Device3、Device4的配置与Device2类似，这里不再赘述。

[~Device2] bridge-domain 10

[\*Device2-bd10] quit

[\*Device2] interface eth-trunk 10.1 mode l2

[\*Device2-Eth-Trunk10.1] encapsulation dot1q vid 10

[\*Device2-Eth-Trunk10.1] bridge-domain 10

[\*Device2-Eth-Trunk10.1] quit

[\*Device2] commit

6.使能EVPN作VXLAN控制平面-

# 配置Device1。Device2、Device3和Device4的配置与Device1类似，这里不再赘述。

[~Device1] evpn-overlay enable

[\*Device1] commit

7.在Device1上建立与Device2、Device3和Device4之间的BGP EVPN对等体关系，并配置反射器功能，指定Device2、Device3和Device4作为反射器的客户机-

# 在Device1上配置BGP EVPN对等体关系

[~Device1] bgp 100 instance evpn1

[\*Device1-bgp-instance-evpn1] peer 2.2.2.2 as-number 100

[\*Device1-bgp-instance-evpn1] peer 2.2.2.2 connect-interface LoopBack0

[\*Device1-bgp-instance-evpn1] peer 3.3.3.3 as-number 100

[\*Device1-bgp-instance-evpn1] peer 3.3.3.3 connect-interface LoopBack0

[\*Device1-bgp-instance-evpn1] peer 4.4.4.4 as-number 100

[\*Device1-bgp-instance-evpn1] peer 4.4.4.4 connect-interface LoopBack0

[\*Device1-bgp-instance-evpn1] l2vpn-family evpn

[\*Device1-bgp-instance-evpn1-af-evpn] peer 2.2.2.2 enable

[\*Device1-bgp-instance-evpn1-af-evpn] peer 2.2.2.2 reflect-client

[\*Device1-bgp-instance-evpn1-af-evpn] peer 3.3.3.3 enable

[\*Device1-bgp-instance-evpn1-af-evpn] peer 3.3.3.3 reflect-client

[\*Device1-bgp-instance-evpn1-af-evpn] peer 4.4.4.4 enable

[\*Device1-bgp-instance-evpn1-af-evpn] peer 4.4.4.4 reflect-client

[\*Device1-bgp-instance-evpn1-af-evpn] undo policy vpn-target

[\*Device1-bgp-instance-evpn1-af-evpn] quit

[\*Device1-bgp-instance-evpn1] quit

[\*Device1] commit

8.在Device2、Device3和Device4上建立与Device1之间的EVPN IBGP对等体关系-

# 在Device2上配置BGP EVPN对等体关系。Device3、Device4的配置与Device2类似，这里不再赘述。

[~Device2] bgp 100 instance evpn1

[\*Device2-bgp-instance-evpn1] peer 1.1.1.1 as-number 100

[\*Device2-bgp-instance-evpn1] peer 1.1.1.1 connect-interface LoopBack0

[\*Device2-bgp-instance-evpn1] l2vpn-family evpn

[\*Device2-bgp-instance-evpn1-af-evpn] peer 1.1.1.1 enable

[\*Device2-bgp-instance-evpn1-af-evpn] quit

[\*Device2-bgp-instance-evpn1] quit

[\*Device2] commit

9.在Device2、Device3和Device4上配置VPN实例和EVPN实例-

# 配置Device2。Device3、Device4的配置与Device2类似，这里不再赘述。

[~Device2] ip vpn-instance vpn1

[\*Device2-vpn-instance-vpn1] vxlan vni 5010

[\*Device2-vpn-instance-vpn1] ipv4-family

[\*Device2-vpn-instance-vpn1-af-ipv4] route-distinguisher 11:11

[\*Device2-vpn-instance-vpn1-af-ipv4] vpn-target 1:1

[\*Device2-vpn-instance-vpn1-af-ipv4] vpn-target 11:1 evpn

[\*Device2-vpn-instance-vpn1-af-ipv4] quit

[\*Device2-vpn-instance-vpn1] quit

[\*Device2] bridge-domain 10

[\*Device2-bd10] vxlan vni 10

[\*Device2-bd10] evpn

[\*Device2-bd10-evpn] route-distinguisher 10:1

[\*Device2-bd10-evpn] vpn-target 10:1

[\*Device2-bd10-evpn] vpn-target 11:1 export-extcommunity

[\*Device2-bd10-evpn] quit

[\*Device2-bd10] quit

[\*Device2] commit

10.在Device2、Device3、Device4上使能头端复制功能

# 配置Device2。Device3、Device4的配置与Device2类似，这里不再赘述。其中Device3的NVE接口不需要配置MAC地址。

[~Device2] interface nve 1

[\*Device2-Nve1] source 2.2.2.210

[\*Device2-Nve1] mac-address 0000-5e00-0101

[\*Device2-Nve1] vni 10 head-end peer-list protocol bgp

[\*Device2-Nve1] quit

[\*Device2] commit

说明：

由于Device2和Device4作为双活接入的网关设备，请确保这两台设备上配置的NVE接口的IP地址和MAC地址相同。

11.在Device2、Device3、Device4上配置VXLAN三层网关

# 在Device2上配置业务环回接口。Device3、Device4的配置与Device2类似，这里不再赘述。（CE6855HI、CE6870EI、CE7855EI不需要配置此步骤）

[~Device2] interface eth-trunk 2

[\*Device2-Eth-Trunk2] service type tunnel

[\*Device2-Eth-Trunk2] quit

[\*Device2] interface 10ge 1/0/5

[\*Device2-10GE1/0/5] eth-trunk 2

[\*Device2-10GE1/0/5] quit

[\*Device2] commit

说明：

成员接口必须是空闲的、没有承载业务的物理接口，对物理接口的状态无要求。

请保证业务环回Eth-Trunk接口的带宽至少是VXLAN三层网关流量所占带宽的两倍。例如：如果用户侧通过VXLAN网络上送到网关的流量为10Gbps，则需要将两个10GE接口加入到业务环回的Eth-Trunk接口中作为其物理成员口。

# 在Device2上配置VXLAN三层网关。Device3、Device4的配置与Device2类似，这里不再赘述，要注意Device2的Vbdif接口的IP地址要与Device3的属于不同网段，Device2与Device4要配置相同的Vbdif接口的IP地址和MAC地址，Device3的Vbdif接口不需要配置MAC地址。

[~Device2] interface Vbdif10

[\*Device2-Vbdif10] ip binding vpn-instance vpn1

[\*Device2-Vbdif10] ip address 10.1.1.1 255.255.255.0

[\*Device2-Vbdif10] mac-address 0000-5e00-0102

[\*Device2-Vbdif10] arp distribute-gateway enable

[\*Device2-Vbdif10] arp collect host enable

[\*Device2-Vbdif10] quit

[\*Device2] commit

说明：

由于Device2和Device4作为双活接入的网关设备，请确保这两台设备上配置的VBDIF接口的IP地址和MAC地址相同。

12.在Device1与Device2、Device3、Device4之间配置BGP对邻居发布IRB类型的路由

# 配置Device1。Device2、Device3、Device4的配置与Device1类似，这里不再赘述。

[~Device1] bgp 100 instance evpn1

[~Device1-bgp-instance-evpn1] l2vpn-family evpn

[~Device1-bgp-instance-evpn1-af-evpn] peer 2.2.2.2 advertise irb

[\*Device1-bgp-instance-evpn1-af-evpn] peer 3.3.3.3 advertise irb

[\*Device1-bgp-instance-evpn1-af-evpn] peer 4.4.4.4 advertise irb

[\*Device1-bgp-instance-evpn1-af-evpn] quit

[\*Device1-bgp-instance-evpn1] quit

[\*Device1] commit

13.检查配置结果

上述配置成功后，在Device2、Device3、Device4上执行display vxlan tunnel命令可查看到VXLAN隧道的信息。以Device2显示为例。

[~Device2] display vxlan tunnel

Number of vxlan tunnel : 1

Tunnel ID Source Destination State Type

--------------------------------------------------------------

4026531841 2.2.2.210 3.3.3.3 up dynamic

配置完成后，不同服务器中的VM1可以相互通信。

**配置文件**

****