什么是 EVPN

文档版本 03
发布日期 2020-11-11

华为技术有限公司

商标声明


和其他华为商标均为华为技术有限公司的商标。
本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受华为公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为公司对本文档内容不做任何明示或默示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

华为技术有限公司

地址： 深圳市龙岗区坂田华为总部办公楼 邮编：518129

网址： https://www.huawei.com

客户服务邮箱： support@huawei.com

客户服务电话： 4008302118

目 录

1 EVPN 简介 1
2 了解 BGP EVPN 中的几种路由类型 3
2.1 EVPN 中定义的五种路由类型概览 3
2.2 EVPN Type2 路由 4
2.3 EVPN Type3 路由 6
2.4 EVPN Type5 路由 7
3 理解 BGP EVPN 作为 VXLAN 控制面的工作过程 9
3.1 同子网 VXLAN 隧道的建立 9
3.2 使用 EVPN 学习 MAC 地址 11
3.3 跨子网 VXLAN 隧道的建立和路由发布 12
4 VXLAN BGP EVPN 网络中流量的转发过程 17
4.1 同子网报文转发 17
4.2 跨子网报文转发 19
5 VXLAN BGP EVPN 网络中的 ARP 广播抑制 21
6 如何配置 VXLAN BGP EVPN 24

1 EVPN 简介

EVPN 基本概念

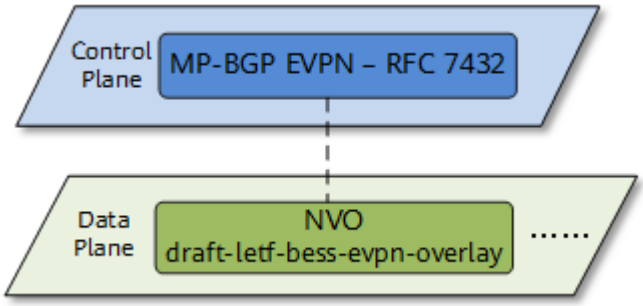
为什么会有EVPN（Ethernet VPN）呢？最初的VXLAN方案（RFC7348）中没有定义控制平面，是手工配置VXLAN隧道，然后通过流量泛洪的方式进行主机地址的学习。这种方式实现上较为简单，但是会导致网络中存在很多泛洪流量、网络扩展起来困难。

为了解决上述问题，人们在VXLAN中引入了EVPN作为VXLAN的控制平面，如图1-1所示（VXLAN是一种NVO协议）。EVPN还能作为一些其他协议的控制面，本文仅描述EVPN作为VXLAN的控制面的相关信息。

说明

VXLAN相关的基本概念，请参见另外一篇文章《什么是VXLAN》。

图 1-1 将 EVPN 作为 VXLAN 的控制平面



EVPN参考了MP-BGP（MultiProtocol BGP）的机制。在深入理解EVPN的工作原理前，我们先对MP-BGP（MultiProtocol BGP）做下简单回顾。

传统的BGP-4使用Update报文在对等体之间交换路由信息。一条Update报文可以通告一类具有相同路径属性的可达路由，这些路由放在NLRI（Network Layer Reachable Information，网络层可达信息）字段中。因为BGP-4只能管理IPv4单播路由信息，为了提供对多种网络层协议的支持（例如IPv6、组播），发展出了MP-BGP。MP-BGP在BGP-4基础上对NLRI作了新扩展。玄机就在于新扩展的NLRI上，扩展之后的NLRI增加了地址族的描述，可以用来区分不同的网络层协议，例如IPv6单播地址族、VPN实例地址族等。

类似的，EVPN也是借用了MP-BGP的机制，在L2VPN地址族下定义了新的子地址族——EVPN地址族，在这个地址族下又新增了一种NLRI，即EVPN NLRI。EVPN NLRI定

义了几种BGP EVPN路由类型，这些路由可以携带主机IP、MAC、VNI、VRF等信息。这样，当一个VTEP学习到下挂的主机的IP、MAC地址信息后，就可以通过MP-BGP路由将这些信息发送给其他的VTEP，从而在控制平面实现主机IP、MAC地址的学习，抑制了数据平面的泛洪。

采用EVPN作为VXLAN的控制平面具有以下优势：

- 可实现VTEP自动发现、VXLAN隧道自动建立，从而降低网络部署、扩展的难度。
- EVPN可以同时发布二层MAC信息和三层路由信息。
- 可以减少网络中的泛洪流量。

欢迎收看视频了解 EVPN

1. 视频**《CloudEngine系列交换机EVPN特性介绍1》**介绍了在相同网段二层互通场景下，通过BGP EVPN实现VXLAN隧道建立及报文转发的过程。
2. 视频**《CloudEngine系列交换机EVPN特性介绍2》**介绍了在不同网段三层互通场景下，通过BGP EVPN实现VXLAN隧道建立及报文转发的过程。

2 了解 BGP EVPN 中的几种路由类型

本节介绍BGP EVPN NLRI中定义的几种路由类型、报文格式及其作用。

2.1 EVPN中定义的五种路由类型概览

2.2 EVPN Type2路由

2.3 EVPN Type3路由

2.4 EVPN Type5路由

2.1 EVPN 中定义的五种路由类型概览

EVPN NLRI定义了如**表2-1**所示的五种EVPN路由类型。其中Type1～Type4是在RFC7432中定义的，Type5是在后来的草案中定义的。

表 2-1 EVPN 路由类型

路由类型	路由描述	RFC/Draft
Type1	Ethernet auto-discovery (AD) route，以太自动发现路由	RFC 7432
Type2	MAC/IP advertisement route，MAC/IP路由	
Type3	Inclusive multicast Ethernet tag route，Inclusive Multicast路由	
Type4	Ethernet segment route，以太网段路由	
Type5	IP prefix route，IP前缀路由	draft-ietf-bess-evpn-prefix-advertisement

其中Type1和Type4是用于EVPN ESI（Ethernet Segment Identifier）多活场景，该场景是一种按照RFC标准定义的方式实现的VXLAN网关多归多活方案，可有效提升

VXLAN接入侧的可靠性，目前仅部分CloudEngine交换机款型支持，详情可参见CloudEngine交换机产品文档中的“**EVPN ESI多活功能**”。

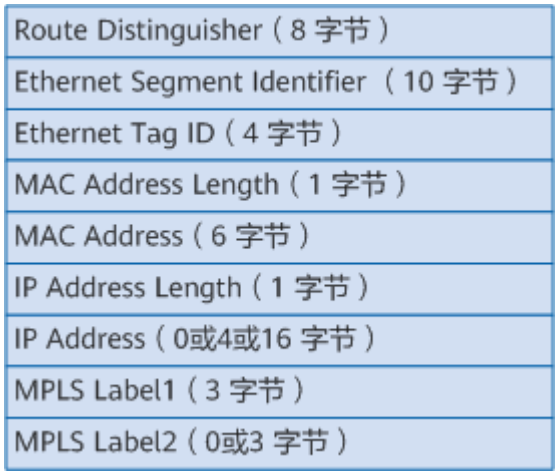
本文主要对常见的EVPN中Type2、Type3、Type5类型的路由进行重点介绍。

2.2 EVPN Type2 路由

格式说明

EVPN Type2路由，也就是MAC/IP路由，主要用于VTEP之间相互通告主机IP、MAC信息。Type2路由的NLRI部分格式如图2-1所示。

图 2-1 Type2 路由的报文格式



各字段的解释如下表所示：

字段	说明
Route Distinguisher	该字段为EVPN实例下设置的RD（Route Distinguisher）值。作用类似于L3VPN的RD值，这里的RD是区分不同的EVPN实例。一个二层广播域BD就对应一个EVPN实例。
Ethernet Segment Identifier	该字段为当前设备与对端连接定义的唯一标识。
Ethernet Tag ID	该字段为当前设备上实际配置的VLAN ID。
MAC Address Length	该字段为此路由携带的主机MAC地址的长度。
MAC Address	该字段为此路由携带的主机MAC地址。
IP Address Length	该字段为此路由携带的主机IP地址的掩码长度。
IP Address	该字段为此路由携带的主机IP地址。

字段	说明
MPLS Label1	该字段为此路由携带的二层VNI，用于标识不同的BD。
MPLS Label2	该字段为此路由携带的三层VNI，用于标识不同的VRF。VXLAN网络中为了实现不同租户之间的隔离，需要通过不同的VRF（L3VPN）来隔离不同租户的路由表，从而将不同租户的路由存放在不同的私网路由表中，而三层VNI就是用来标识这些VRF的。

应用说明

Type2路由在VXLAN网络中的使用场景和作用参见下表。

表 2-2 Type2 路由使用场景说明

场景	Type2路由功能说明
通告主机MAC地址	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之间借助Type2路由，可以相互通告已经获取到的主机MAC。详细的说明请参见本文的**3.2 使用EVPN学习MAC地址**。
通告主机ARP	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，可应用于以下两个子场景： • 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。ARP广播抑制的详情可参见本文中的**5 VXLAN BGP EVPN网络中的ARP广播抑制**。 • 场景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。
通告主机IP路由	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Type2路由，可以相互通告已经获取到的主机IP路由。详细的说明请参见本文的**主机路由发布**。
ND表项扩散	Type2路由可以同时携带主机MAC地址+主机IPv6地址，因此该路由可以用来在VTEP之间传递ND表项，实现ND表项扩散，可用于实现NS组播抑制、防止ND欺骗攻击、分布式网关场景下的IPv6虚拟机迁移。

场景	Type2路由功能说明
通告主机IPv6路由	在分布式网关场景中，要实现跨子网IPv6主机的三层互访，网关设备需要互相学习主机IPv6路由。作为BGP EVPN对等体的VTEP之间通过交换Type2路由，可以相互通告已经获取到的主机IPv6路由。

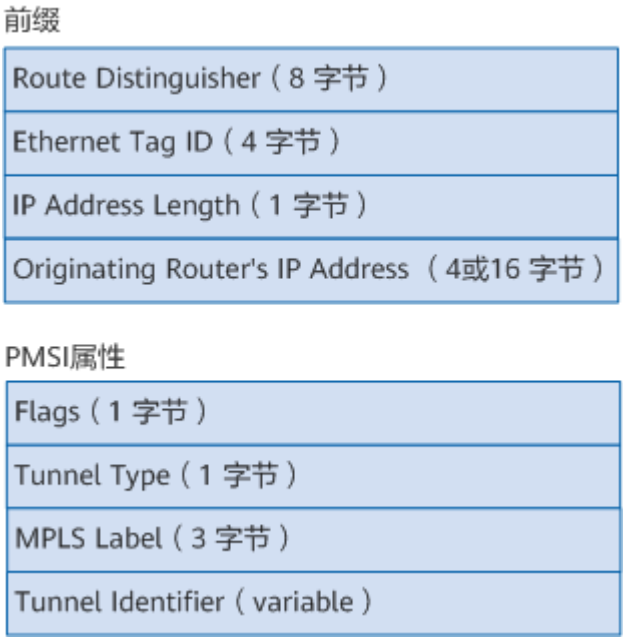
2.3 EVPN Type3 路由

格式说明

EVPN Type3路由主要用于在VTEP之间相互通告二层VNI、VTEP IP信息，以建立头端复制列表，即用于VTEP的自动发现和VXLAN隧道的动态建立：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。

Type3路由的NLRI是由“前缀”和“PMSI”属性组成，报文格式如图2-2所示。其中VTEP IP信息体现在NLRI的**Originating Router's IP Address**字段中，二层VNI信息则体现在PMSI属性的**MPLS Label**中。

图 2-2 Type3 路由的报文格式



各字段的解释如下表所示：

字段	说明
Route Distinguisher	该字段为EVPN实例下设置的RD（Route Distinguisher）值。
Ethernet Tag ID	该字段为当前设备上的VLAN ID。在此路由中为全0。

字段	说明
IP Address Length	该字段为此路由携带的本端VTEP IP地址的掩码长度。
Originating Router's IP Address	该字段为此路由携带的本端VTEP IP地址。
Flags	该字段为标志位，标识当前隧道是否需要叶子节点信息。 在VXLAN场景中，该字段没有实际意义。
Tunnel Type	该字段为此路由携带的隧道类型。目前，在VXLAN场景中，支持的类型只有“6：Ingress Replication”，即头端复制，用于BUM报文转发。
MPLS Label	该字段为此路由携带的二层VNI。
Tunnel Identifier	该字段为此路由携带的隧道信息。目前，在VXLAN场景中，该字段也是本端VTEP IP地址。

应用说明

Type3路由动态建立头端复制列表的过程简介请参见本文的**3.1 同子网VXLAN隧道的建立**。

2.4 EVPN Type5 路由

格式说明

EVPN Type5路由又称IP前缀路由，主要用于传递网段路由。不同于Type2路由只传递32（IPv4）/128（IPv6）位的主机路由，Type5路由可传递0～32/0～128掩码长度的网段路由。

Type5路由的报文格式如**图2-2**所示。

图 2-3 Type5 路由的报文格式

各字段的解释如下表所示：

字段	说明
Route Distinguisher	该字段为EVPN实例下设置的RD（Route Distinguisher）值。
Ethernet Segment Identifier	该字段为当前设备与对端连接定义的唯一标识。
Ethernet Tag ID	该字段为当前设备上实际配置的VLAN ID。
IP Prefix Length	该字段为此路由携带的IP前缀掩码长度。
IP Prefix	该字段为此路由携带的IP前缀。
GW IP Address	该字段为默认网关地址。该字段在VXLAN场景中没有实际意义。
MPLS Label	该字段为此路由携带的三层VNI。

应用说明

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：

- 当携带主机IP地址时，主要用于分布式网关场景中的主机/网段路由通告，请参见本文的**网段路由发布**。
- 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。

3 理解 BGP EVPN 作为 VXLAN 控制面的工作过程

BGP EVPN在VXLAN网络中是如何工作的呢？本节将为您介绍BGP EVPN作为VXLAN控制面的工作过程。

在用BGP EVPN方式部署分布式VXLAN网络的场景中，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。BGP EVPN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通告等，还可以使能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BGP EVPN方式。

本文下面的内容以Underlay网络和Overlay网络均为IPv4为例，介绍EVPN作为VXLAN控制面的工作过程。

3.1 同子网VXLAN隧道的建立

3.2 使用EVPN学习MAC地址

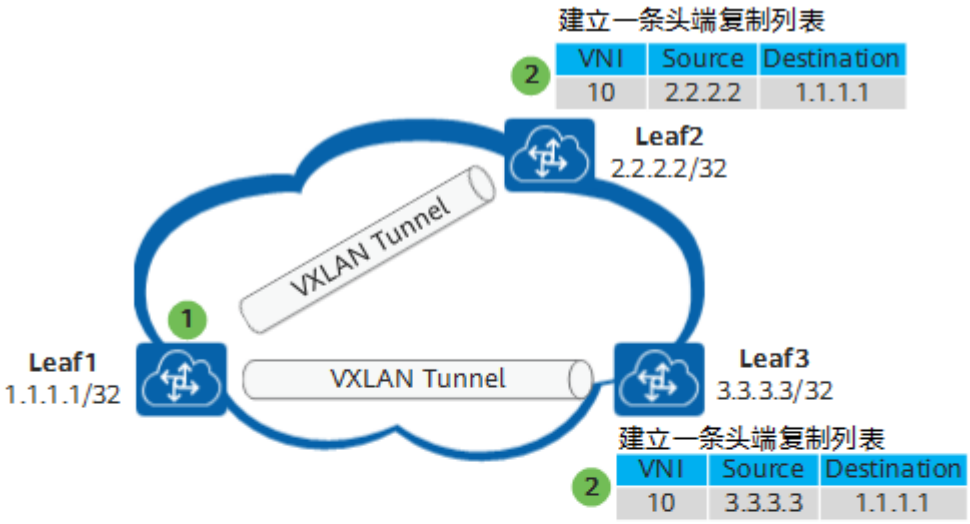
3.3 跨子网VXLAN隧道的建立和路由发布

3.1 同子网 VXLAN 隧道的建立

VXLAN隧道由一对VTEP确定，在同子网互通场景下，因为只需要在同一个二层广播域（BD）内互通，所以只要两端VTEP的IP地址路由可达，VXLAN隧道就可以建立。通过EVPN动态建立VXLAN隧道，就是在两端VTEP之间建立BGP EVPN对等体，然后对等体之间通过交互Type3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。

下面以**图3-1**为列，介绍VTEP之间是如何通过Type3路由建立VXLAN隧道。

图 3-1 同子网 VXLAN 隧道的建立示意图



Leaf1向Leaf2/Leaf3发送的Type3路由

NLRI	Route Type	Inclusive multicast Ethernet tag route (Type 3)
	Route Distinguisher (RD)	EVPN实例RD值（1:10）
	Ethernet Tag ID	0
	IP Address Length	32
	Originating IP Address	Leaf1的VTEP IP（1.1.1.1）
PMSI	Flags	...
	Tunnel Type	6: Ingress Replic
	MPLS Label	二层VNI（10）
	Tunnel Identifier	...
Extended Community		EVPN实例的ERT（0:10）
Other attributes		...

图中Leaf1、Leaf2、Leaf3作为VTEP，以Leaf1向Leaf2、Leaf3发送路由为例。

1. 在Leaf1上完成VTEP IP、二层VNI、EVPN实例等相关配置后（这些配置的样例如下所示），Leaf1会向对等体Leaf2、Leaf3分别发送EVPN Type3路由。路由中会携带二层VNI、本端VTEP IP、EVPN实例的RD、出方向VPN-Target（ERT）等信息。

```
[Leaf1]
bridge-domain 10
 vxlan vni 10    //二层VNI
 evpn
  route-distinguisher 1:10    //EVPN实例的RD
  vpn-target 0:10 export-extcommunity    //EVPN实例的ERT
  vpn-target 100:5000 export-extcommunity
  vpn-target 0:10 import-extcommunity
#
interface Nve1
 source 1.1.1.1    //Leaf1的VTEP IP地址
 vni 10 head-end peer-list protocol bgp
#
```

2. Leaf2、Leaf3收到Leaf1发来的Type3路由后，如果Leaf1的VTEP IP三层路由可达，则建立一条到Leaf1的二层VXLAN隧道；同时，如果本地有相同的VNI，则建立一条头端复制列表，用于后续广播、组播、未知单播报文的转发。

在Leaf2、Leaf3收到Leaf1发送的EVPN路由时，会基于路由携带的RT值（EVPN实例的ERT值）是否与本地EVPN实例的IRT值匹配，来判断是否接纳该路由。

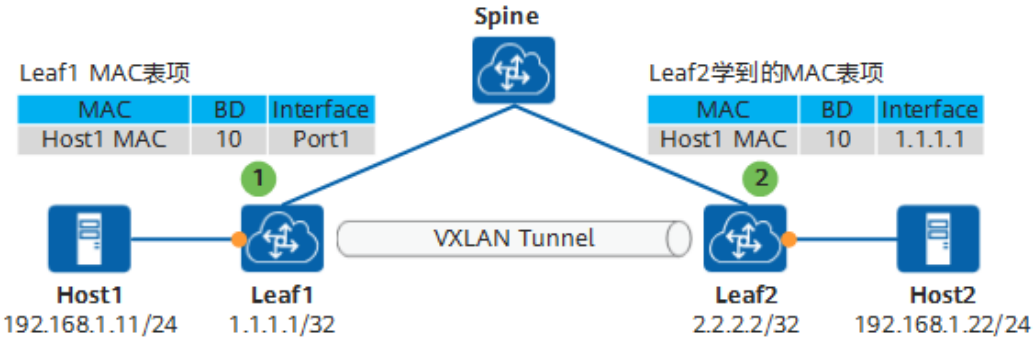
经过以上的流程，Leaf2、Leaf3上就能建立到Leaf1的头端复制列表，指导后续BUM报文的转发。类似的，Leaf1上也会建立到Leaf2、Leaf3的头端复制列表。

3.2 使用 EVPN 学习 MAC 地址

使用EVPN作为VXLAN的控制平面，可以用EVPN来进行MAC学习，以替代数据平面泛洪方式的MAC学习，减少泛洪流量。使用EVPN来进行MAC学习的过程，是通过在VTEP之间传递Type2路由完成的。

下面以图3-2为例，介绍VTEP之间是如何通过EVPN来实现远程主机的MAC学习的。

图 3-2 使用 EVPN 来学习远程主机 MAC 地址的过程示意图



Leaf1向Leaf2发送的Type2路由

NLRI	Route Type	MAC/IP advertisement route (Type 2)
	Route Distinguisher (RD)	EVPN实例RD值（1:10）
	Ethernet Segment Identifier	...
	Ethernet Tag ID	...
	MAC Address Length	MAC地址长度（48）
	MAC Address	Host1的MAC地址（1000-2000-3000）
	IP Address Length	...
	IP Address	...
	MPLS Label1	二层VNI（10）
	MPLS Label2	...
Next-hop		Leaf1的VTEP IP地址（1.1.1.1）
Extended Community		EVPN实例的ERT（0:10）
Other attributes		...

图中Leaf1和Leaf2作为VTEP，分别连接同网段的主机Host1和Host2，以Leaf1向Leaf2发送Type2路由为例。

1. Host1在连接至Leaf1时，通常会触发ARP、DHCP等行为。通过这些流量，Leaf1上就会学习到Host1的MAC信息，记录在本地MAC表中。

   Leaf1学习到本地主机的MAC表项后，会向其对等体Leaf2发送EVPN Type2路由。该路由会携带本端EVPN实例的ERT、VTEP IP地址、二层VNI、Host1的MAC地址等信息。其中本端的EVPN实例的ERT、VTEP IP地址、二层VNI这些信息来源于本端VTEP上的配置，样例如下：

```
[Leaf1]
bridge-domain 10
 vxlan vni 10    //二层VNI
 evpn
  route-distinguisher 10:1
  vpn-target 0:10 export-extcommunity    //EVPN实例的ERT
  vpn-target 100:5000 export-extcommunity
  vpn-target 0:10 import-extcommunity
#
interface Nve1
 source 1.1.1.1   //Leaf1的VTEP IP地址
 vni 10 head-end peer-list protocol bgp
#
```

2. Leaf2收到Leaf1发来的Type2路由后，能够学习到Host1的MAC地址信息，并将其保存在MAC表中，其下一跳为Leaf1的VTEP IP地址。

   需要说明的是，Leaf2收到Leaf1发送的EVPN路由时，能否接纳该路由信息，是需要通过EVPN实例的RT（Route Target）值是否匹配来判断的。RT是一种BGP扩展团体属性，用于控制EVPN路由的发布与接收。也就是说，RT决定了本端的EVPN路由可以被哪些对端所接收，以及本端是否接收对端发来的EVPN路由。

   RT属性分为两类：

   - ERT（Export RT）：本端发送EVPN路由时，携带的RT属性设置为ERT。
   - IRT（Import RT）：本端在收到对端的EVPN路由时，将路由中携带的ERT与本端的IRT进行比较，只有两者相等时才接收该路由，否则丢弃该路由。

   在本例中，Leaf2上接收Leaf1发过来的EVPN路由，则需保证Leaf2上配置的IRT（Import RT）与Leaf1配置的ERT（Export RT）一致，例如Leaf2上EVPN中的IRT配置为**0:10**，与上文中Leaf1上的ERT一致：

```
[Leaf2]
bridge-domain 10
 vxlan vni 10    //二层VNI
 evpn
  route-distinguisher 10:2
  vpn-target 0:10 export-extcommunity
  vpn-target 100:5000 export-extcommunity
  vpn-target 0:10 import-extcommunity    //EVPN实例的IRT
#
```

经过以上的流程，在未发送广播请求的情况下，Leaf2就可以学习到Host1的MAC地址。类似的，Leaf1也可以学习到Host2的MAC地址。

另外需要强调的是，EVPN只是减少了网络中的流量泛洪，并不会完全避免，例如在以下一些场景：

1. 网络中存在“静默”主机的情况，这种情况下主机不会触发ARP、DHCP等行为，导致VTEP学习不到本地主机MAC地址，从而也就无法发送MAC地址信息让其他VTEP学习到。
2. 主机首次通信的过程中，主机会发送ARP广播请求报文，这种也会产生泛洪。这种情况还可以通过ARP广播抑制功能来避免泛洪，可参考本文中的**5 VXLAN BGP EVPN网络中的ARP广播抑制**。

3.3 跨子网 VXLAN 隧道的建立和路由发布

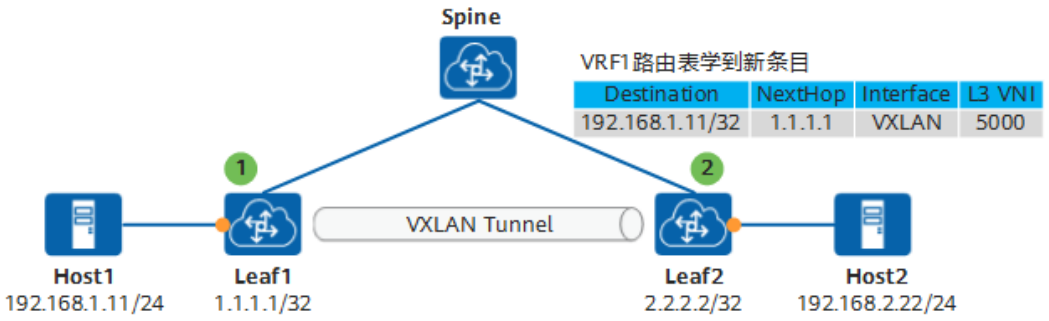
主机路由发布

EVPN Type2路由不仅可以发布主机MAC地址，还可以发布主机路由信息，这是因为Type2路由还可以携带32位掩码的主机IP地址信息。主机路由的发布可以实现分布式网关场景下跨网段主机之间的互通。VTEP之间需要发布下属主机的IP路由，否则对端

VTEP就无法学习到该主机的路由信息，从而没法进行三层转发。简单来说就是“你得告诉我你下面都接了什么网段的路由，否则我怎么知道要发给你呢”。

下面以图3-3为列，介绍VTEP之间是如何使用EVPN来发布主机路由的。

图 3-3 使用 EVPN 发布主机路由的示意图



Leaf1向Leaf2发送的Type2路由

NLRI	Route Type	MAC/IP advertisement route (Type 2)
	Route Distinguisher (RD)	EVPN实例RD值（1:10）
	Ethernet Segment Identifier	...
	Ethernet Tag ID	...
	MAC Address Length	MAC地址长度（48）
	MAC Address	Host1的MAC地址（1000-2000-3000）
	IP Address Length	IP地址掩码长度（32）
	IP Address	Host1的IP地址（192.168.1.11）
	MPLS Label1	二层VNI（10）
	MPLS Label2	三层VNI（5000）
Next-hop		Leaf1的VTEP IP地址（1.1.1.1）
Extended Community		EVPN实例的ERT值（0:10）
Other attributes		...

图中Leaf1和Leaf2作为VTEP，同时作为三层网关，分别连接不同网段的主机Host1和Host2，以Leaf1向Leaf2发送路由为例。

1. Host1在连接至Leaf1时，通常会触发ARP、DHCP等行为。通过这些流量，Leaf1上就会学习到Host1的ARP信息。同时，还可以根据Host1所属的BD域，获取相应的二层VNI、L3VPN实例及L3VPN实例关联的三层VNI信息。

   为什么会有L3VPN和三层VNI呢？因为同一个Leaf下可能接入多个租户的服务器，而为了实现不同租户之间的隔离，所以就在Leaf上通过创建不同的L3VPN来隔离不同租户的路由表，从而将不同租户的路由存放在不同的私网路由表中。而三层VNI就是用来标识这些L3VPN的，当Leaf节点收到对端发送来的数据报文时（报文会携带三层VNI），就根据其三层VNI找到相应的L3VPN，通过查找该L3VPN实例下的路由表来进行转发。

   Leaf1获取的二层VNI、L3VPN实例及L3VPN实例关联的三层VNI信息依赖的关键配置示例如下：

```
[Leaf1]
ip vpn-instance vpn1    //L3VPN实例
 ipv4-family
  route-distinguisher 20:4
  vpn-target 100:5000 export-extcommunity evpn
  vpn-target 100:5000 import-extcommunity evpn
 vxlan vni 5000      //L3VPN实例关联的三层VNI
#
```

```
bridge-domain 10
 vxlan vni 10     //二层VNI
 evpn
  route-distinguisher 10:4
  vpn-target 0:10 export-extcommunity
  vpn-target 100:5000 export-extcommunity
  vpn-target 0:10 import-extcommunity
#
interface Vbdif10    //根据BD信息获取三层Vbdif接口和此接口绑定的L3VPN实例
 ip binding vpn-instance vpn1
 ip address 192.168.1.1 255.255.255.0
 mac-address 0000-5e00-0102
 vxlan anycast-gateway enable
 arp collect host enable
#
```

以上这些总结起来就是Leaf1会获取Host1的：**IP + MAC + Host1所属的二层VNI + VBDIF绑定的L3VPN实例的三层VNI**，然后：

a. Leaf1上的EVPN实例就可以根据这些信息生成EVPN Type2类型的路由（参见上图中的表格），除了获取的Host1的相关信息外，还携带本端EVPN实例的ERT、路由下一跳（本端VTEP IP）、VTEP的MAC等信息，将其发送给对等体Leaf2。

b. Leaf1上的EVPN实例将Host1的IP + MAC + 三层VNI发给本端的L3VPN实例，从而在本端的L3VPN实例中生成本地Host1的路由。

2. Leaf2收到Leaf1发来的Type2路由后，能够学习到Host1的IP地址信息，并将其保存在相应的路由表中，其下一跳为Leaf1的VTEP IP地址，同时记录对应的三层VNI信息，处理过程如下：

a. 检查该路由的ERT与接收端**EVPN实例**的IRT是否相同。如果相同，则接收该路由，同时EVPN实例提取其中包含的主机IP+MAC信息，用于主机ARP通告。

b. 检查该路由的ERT与接收端**L3VPN实例**的IRT是否相同（如下表中的举例所示）。如果相同，则接收该路由，同时L3VPN实例提取其中的主机IP地址+三层VNI信息，在其路由表中生成Host1的路由。该路由的下一跳会被设置为Leaf1的VXLAN隧道接口。

Leaf1（发送端）	Leaf2（接收端）
ip vpn-instance vpn1 ipv4-family route-distinguisher 20:2 vpn-target 100:5000 export-extcommunity evpn vpn-target 100:5000 import-extcommunity evpn vxlan vni 5000 # bridge-domain 10 vxlan vni 10 evpn route-distinguisher 10:2 vpn-target 100:10 export-extcommunity **vpn-target 100:5000 export-extcommunity //发送端EVPN中的ERT** vpn-target 100:10 import-extcommunity #	ip vpn-instance vpn1 ipv4-family route-distinguisher 20:3 vpn-target 100:5000 export-extcommunity evpn **vpn-target 100:5000 import-extcommunity evpn //接收端L3VPN中的IRT（eIRT）** vxlan vni 5000 # bridge-domain 20 vxlan vni 20 evpn route-distinguisher 10:3 vpn-target 100:20 export-extcommunity vpn-target 100:5000 export-extcommunity vpn-target 100:20 import-extcommunity #

c. 接收端EVPN实例或L3VPN实例接收该路由后会通过下一跳获取Leaf1的VTEP IP地址，如果该地址三层路由可达，则建立一条到Leaf1的VXLAN隧道。

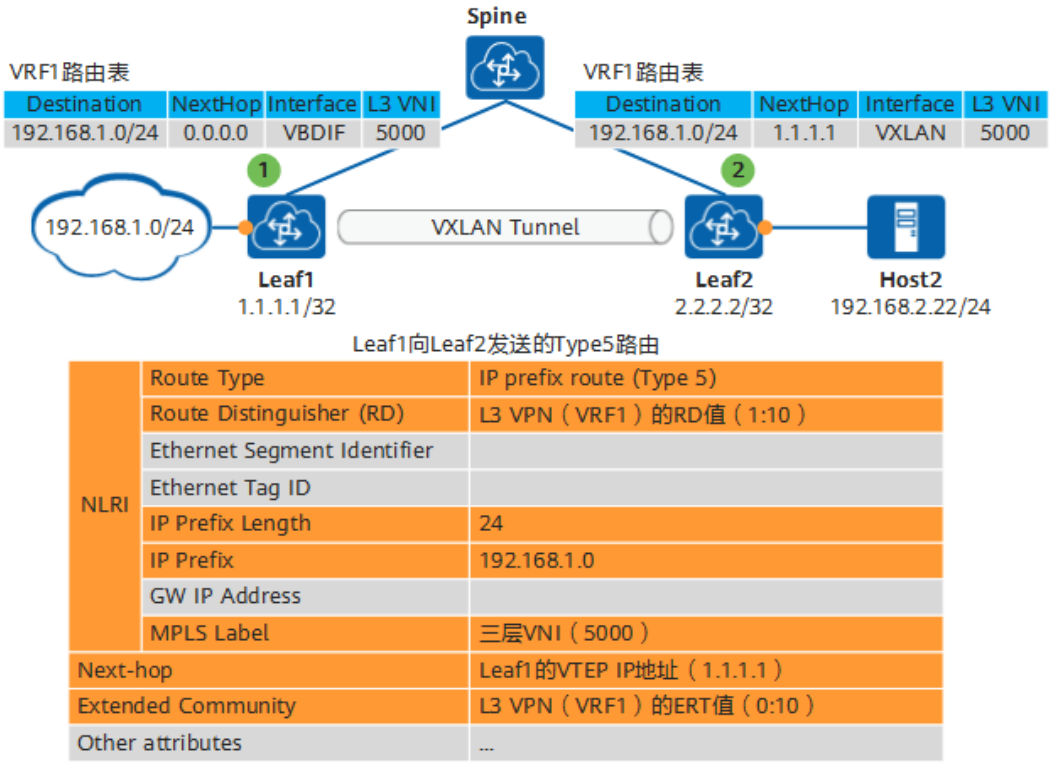
经过以上的流程，Leaf2就可以学习到Host1的IP路由信息，后续转发至Host1的报文时，可以根据查找路由表进行转发。类似的，Leaf1也可以学习到Host2的IP路由信息。

网段路由发布

网段路由的发布流程与主机路由类似，区别在于网段路由是通过Type5路由发布的，Type2路由只能发布32/128位的主机路由。Type5路由也可以发布32/128位的主机路由，在发布32/128位的主机路由时，功能与Type2路由类似。

如果网关设备下连接的网段在整个网络中唯一，则可以配置发布网段路由，否则不能配置发布网段路由。

图 3-4 EVPN 网段路由发布示意图



Leaf1向Leaf2发送的Type5路由

NLRI	Route Type	IP prefix route (Type 5)
	Route Distinguisher (RD)	L3 VPN（VRF1）的RD值（1:10）
	Ethernet Segment Identifier	
	Ethernet Tag ID	
	IP Prefix Length	24
	IP Prefix	192.168.1.0
	GW IP Address	
	MPLS Label	三层VNI（5000）
Next-hop		Leaf1的VTEP IP地址（1.1.1.1）
Extended Community		L3 VPN（VRF1）的ERT值（0:10）
Other attributes		...

下面以图3-4为列，介绍VTEP之间是如何发布网段路由的。图中Leaf1和Leaf2作为VTEP，同时作为三层网关，其中Leaf1连接一个192.168.1.0/24的网段。

1. Leaf1收集到本地IP网段路由，把该IP网段路由通过EVPN Type5路由发送给Leaf2。路由中携带有IP前缀、掩码长度、对应VRF的三层VNI等信息（如上图表格所示）。
2. Leaf2收到Leaf1发来的Type5路由后，能够学习到IP网段路由信息，并将其保存在相应的路由表中，其下一跳为Leaf1的VTEP IP地址，同时记录对应的三层VNI信息。

   Leaf2收到Leaf1发送的EVPN路由时，根据EVPN路由携带的RT值（Type 5路由使用L3VPN实例的ERT值填充）是否与本地L3VPN实例的IRT值匹配，来将网段路由添加到对应VRF的路由表中。如果某VRF的IRT值与EVPN路由携带的RT值相同，则接收该路由，同时提取其中的网段路由+三层VNI信息，在其路由表中生成网段路由。该路由的下一跳会被设置为Leaf1的VTEP IP地址。同时，如果Leaf1的VTEP IP地址三层路由可达，则建立一条到Leaf1的VXLAN隧道。

Leaf1（发送端）	Leaf2（接收端）
ip vpn-instance vpn1 ipv4-family route-distinguisher 20:2 **vpn-target 100:5000 export-extcommunity evpn //Type5路由中发送端的ERT使用L3VPN实例中的ERT（eERT）** vpn-target 100:5000 import-extcommunity evpn vxlan vni 5000 # bridge-domain 10 vxlan vni 10 evpn route-distinguisher 10:2 vpn-target 100:10 export-extcommunity vpn-target 100:5000 export-extcommunity vpn-target 100:10 import-extcommunity #	ip vpn-instance vpn1 ipv4-family route-distinguisher 20:3 vpn-target 100:5000 export-extcommunity evpn **vpn-target 100:5000 import-extcommunity evpn //接收端L3VPN实例中的IRT（eIRT）** vxlan vni 5000 # bridge-domain 20 vxlan vni 20 evpn route-distinguisher 10:3 vpn-target 100:20 export-extcommunity vpn-target 100:5000 export-extcommunity vpn-target 100:20 import-extcommunity #

经过以上的流程，Leaf2就可以学习到Leaf1的网段路由信息，后续转发至该网段的报文时，可以根据查找路由表进行转发。

4 VXLAN BGP EVPN 网络中流量的转发过程

本文下面的内容以Underlay网络和Overlay网络均为IPv4为例，介绍用BGP EVPN部署的分布式VXLAN网络中，报文的转发过程。

4.1 同子网报文转发

4.2 跨子网报文转发

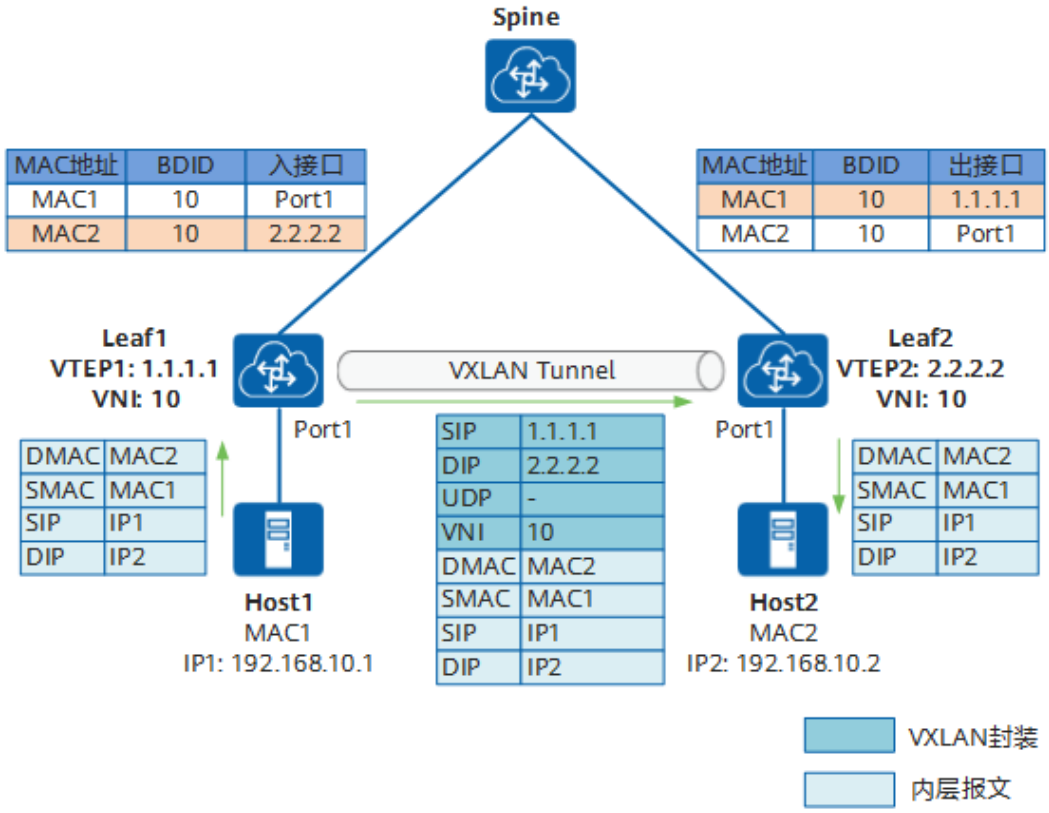
4.1 同子网报文转发

同子网报文转发为二层转发，只在VXLAN二层网关之间进行，三层网关无需感知。

同子网已知单播报文转发

如**图4-1**所示，Host1和Host2同属于一个子网，下面以Host1向Host2发送已知单播报文为例介绍报文在VXLAN网络中的转发流程。

图 4-1 同子网已知单播报文转发示意图



1. Host1发送目的地址为Host2的报文。如果Host1没有Host2的MAC地址，会先发送广播ARP请求来获取Host2的MAC地址，此处该过程不再详述，认为Host1已经获取了Host2的MAC地址。
2. Leaf1收到Host1的报文后，根据报文入端口或VLAN信息判断其所属的BD，并在该BD内查找出接口（通过上一节**3.2 使用EVPN学习MAC地址**可以知道，Leaf1上会学习到Host2的MAC地址，出接口为VTEP 2.2.2.2）。然后Leaf1会对报文进行VXLAN封装后转发。
3. Leaf2接收到VXLAN报文后，根据报文中VNI获取二层广播域，进行VXLAN解封装，获取内层的二层报文。
4. Leaf2根据内层报文的目的MAC地址，从本地MAC表中找到对应的出接口，然后转发给对应的主机Host2。

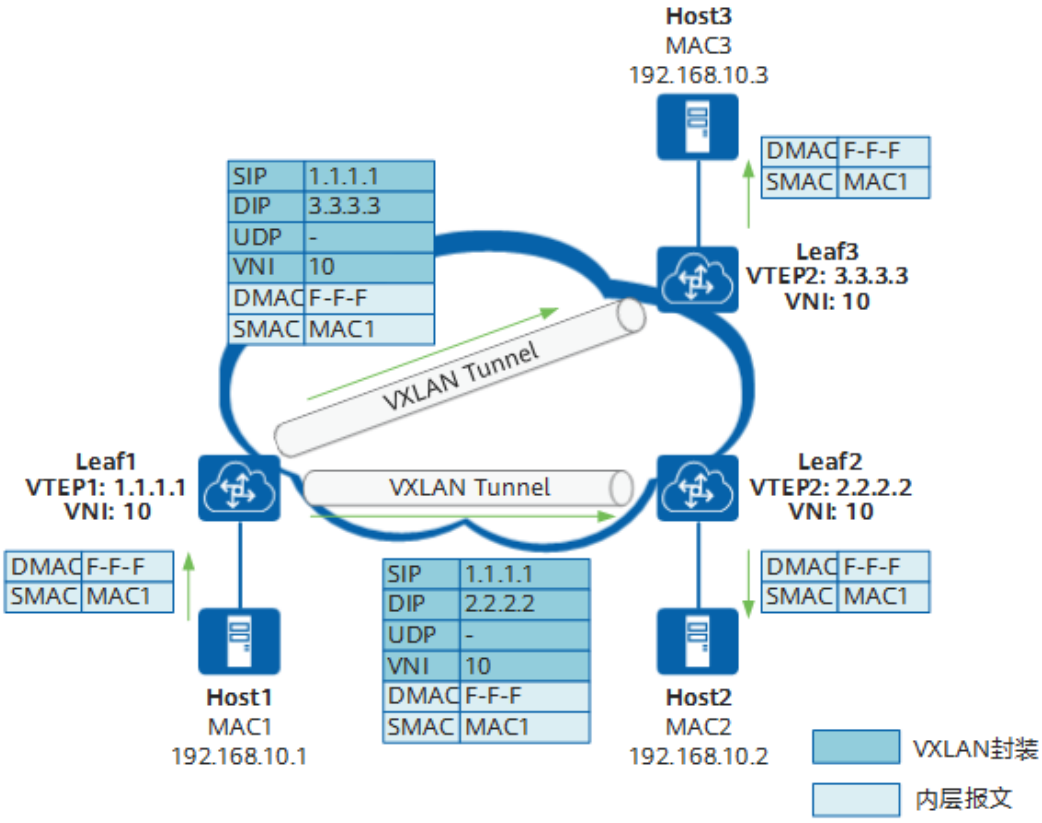
Host2向Host1发送报文的过程与上述过程相同。

同子网 BUM 报文转发

如果是同子网的BUM报文（广播、组播、未知单播），则会向同子网的所有VTEP发送一份报文。

如**图4-2**所示，Host1向外发送广播报文。Leaf1收到Host1的广播报文后，根据报文入端口或VLAN信息判断其所属的BD，并在该BD内查找所有的隧道列表，依据获取的隧道列表进行报文封装后，向所有隧道发送报文，从而将报文转发至同子网的Host2和Host3。

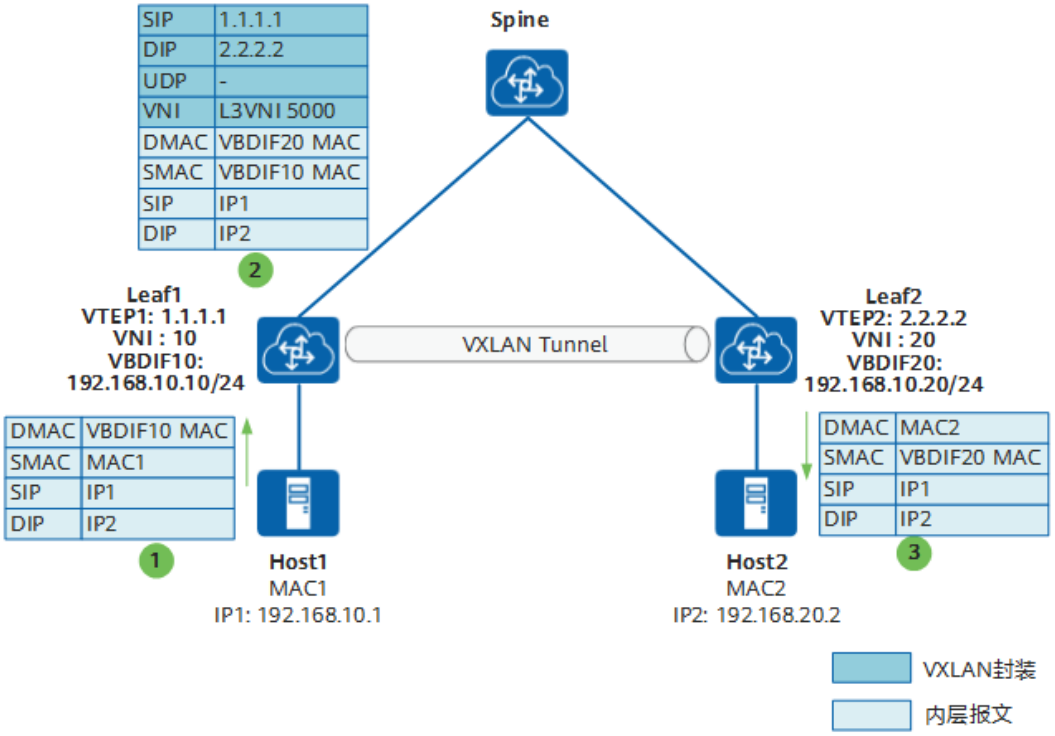
图 4-2 同子网 BUM 报文转发示意图



4.2 跨子网报文转发

如图4-3所示，在分布式网关场景下，Leaf1、Leaf2作为VXLAN的三层网关，进行VXLAN封装及三层转发，Spine仅作为VXLAN报文转发节点，不进行VXLAN报文的处理。

图 4-3 分布式网关场景下跨子网报文转发示意图



以Host1向Host2发送报文为例介绍报文在VXLAN网络中的转发流程：

1. 因为Host1与Host2属于不同网段，所以Host1会先将报文发送给网关（Leaf1），交由网关进行转发。
2. Leaf1接收到来自Host1的报文，根据报文的目的地址判断需要进行三层转发。Leaf1根据报文入端口或VLAN信息判断其所属的BD，找到绑定该BD的L3VPN实例，然后在该L3VPN实例下查找路由表。在前面**3.3 跨子网VXLAN隧道的建立和路由发布**章节已经介绍过，在分布式网关场景下，网关Leaf1会学习到Host2的主机路由。

   Leaf1根据路由获取三层VNI、下一跳等信息，然后进行VXLAN封装，将报文转发至Leaf2。
3. Leaf2收到VXLAN报文后进行解封装，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，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网关的MAC地址，转发给Host2。

Host2向Host1发送报文的过程与上述过程相同。

5 VXLAN BGP EVPN 网络中的 ARP 广播抑制

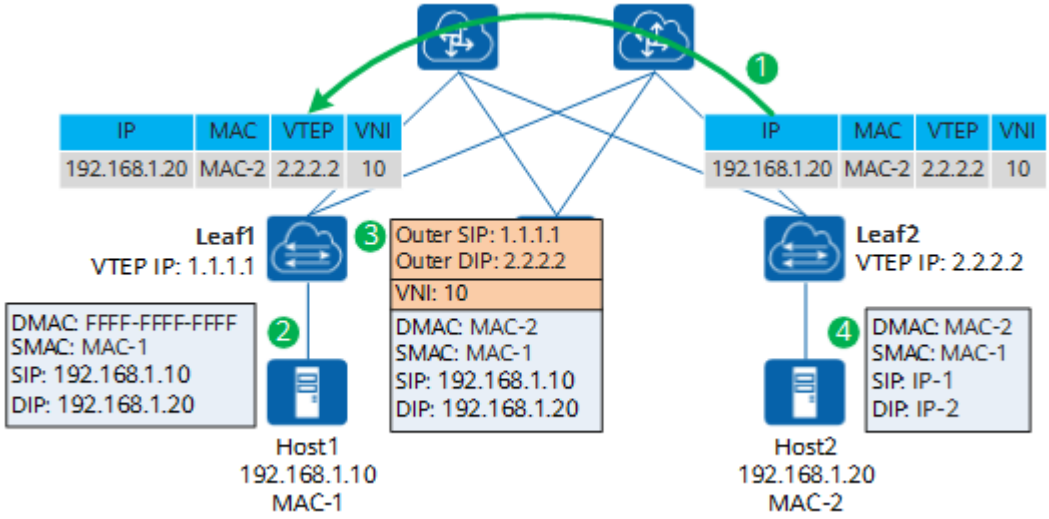
地址解析协议ARP（Address Resolution Protocol）用来将IP地址解析为MAC地址。网络中同网段主机首次通信时，由于没有目标主机的MAC地址信息，因此会发送ARP广播请求来获取目的MAC地址信息。ARP广播请求报文在VXLAN网络中会泛洪转发，大量的ARP报文存在会占用过多的网络资源，导致网络性能下降。

为了抑制ARP广播请求给网络带来的负面影响，可以通过ARP广播抑制功能来尽可能的减少ARP报文在VXLAN网络中的泛洪。ARP广播抑制有两种方式，一种是ARP广播变单播的功能，另一种是ARP二层代答功能。

ARP 广播变单播

ARP广播变单播，顾名思义，就是将ARP广播报文转变成ARP单播，从而以单播形式进行转发。ARP广播变单播的实现思路是在VXLAN三层网关上根据ARP生成ARP广播抑制表（包括主机IP、MAC、VNI、VTEP IP信息），然后通过EVPN将主机信息发送给二层网关；二层网关在收到ARP广播请求后，直接使用学习到的主机MAC替换原来的全F的广播MAC，从而将广播变为单播进行转发。

图 5-1 ARP 广播变单播示意图



以图5-1所示的分布式网关为例，其中Host1和Host2属于同一子网，但是部署在不同的VTEP下。ARP广播变单播过程如下：

1. Leaf2通过Host2发送的ARP报文，可以学习到Host2的ARP表项。然后Leaf2可以根据ARP生成相应的ARP广播抑制表，并通过EVPN向Leaf1发布，这样Leaf1也可以学习到Host2的主机信息。
2. Host1初次访问Host2，发送ARP广播请求来获取Host2的MAC地址。
3. Leaf1收到ARP广播请求后，查询ARP广播抑制表。因为已经有Host2的主机信息，所以Leaf1将ARP请求报文中的全F的广播目的MAC替换为Host2的MAC地址，将ARP广播变为ARP单播，然后再进行VXLAN封装后向Leaf2发送。

   如果Leaf1上没有Host2的ARP广播抑制表，那么依然按照正常的流程进行广播。
4. Leaf2收到VXLAN报文并解封装后，将ARP请求发送给Host2。

可以看出ARP广播变单播功能强依赖于三层网关，需要三层网关学习到主机的ARP信息，如果三层网关学习不到主机ARP，就不能抑制ARP广播。

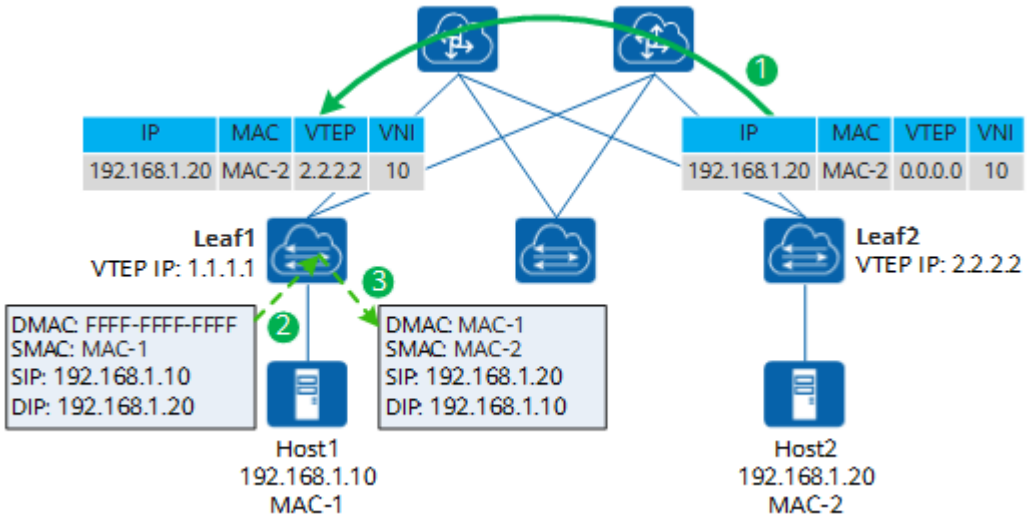
ARP 二层代答

ARP广播变单播的抑制方式需要三层网关的存在，在纯二层网络中，由于不存在三层网关，没有相应的ARP表项，也就无法生成ARP广播抑制表进行ARP抑制。上述二层场景面临的ARP抑制问题，就可以通过ARP二层代答功能来解决。

ARP二层代答的实现思路是在二层网关上收到主机ARP报文，获取ARP报文中的主机信息并生成ARP广播抑制表，然后通过EVPN将主机信息发送给其他二层网关；二层网关在收到ARP广播请求后，根据ARP广播抑制表中的主机信息，直接进行ARP代答。

ARP二层代答是一种有效减少ARP广播报文的机制。使能ARP二层代答之后，二层网关设备在收到ARP请求之后，优先尝试本地代答，只有当本地无法代答时，才会进行广播。

图 5-2 ARP 二层代答示意图



以图5-2为例，其中Host1和Host2属于同一子网，在二层网关设备开启基于BD的ARP二层代答之后，ARP二层代答过程如下：

1. Leaf2上开启ARP二层代答功能后，Leaf2会检测主机发送的ARP报文。当Leaf2接收到Host2的ARP报文后，可以根据ARP生成相应的ARP广播抑制表项，并通过EVPN向Leaf1发布，这样Leaf1也可以学习到Host2的主机信息。

2. Host1初次访问Host2，发送ARP广播请求来获取Host2的MAC地址。
3. Leaf1收到ARP广播请求后，查询ARP广播抑制表。因为已经有Host2的主机信息，所以Leaf1直接对ARP请求进行代答。

   如果Leaf1上没有Host2的ARP广播抑制表，那么依然按照正常的流程进行广播。

6 如何配置 VXLAN BGP EVPN

在华为CloudEngine交换机上配置BGP EVPN的命令步骤、参数说明等，请参见“**配置VXLAN（分布式网关，BGP EVPN方式）**”；相应的配置举例请参见“**VXLAN配置举例**”。

