什么是 VXLAN

文档版本 02
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://e.huawei.com

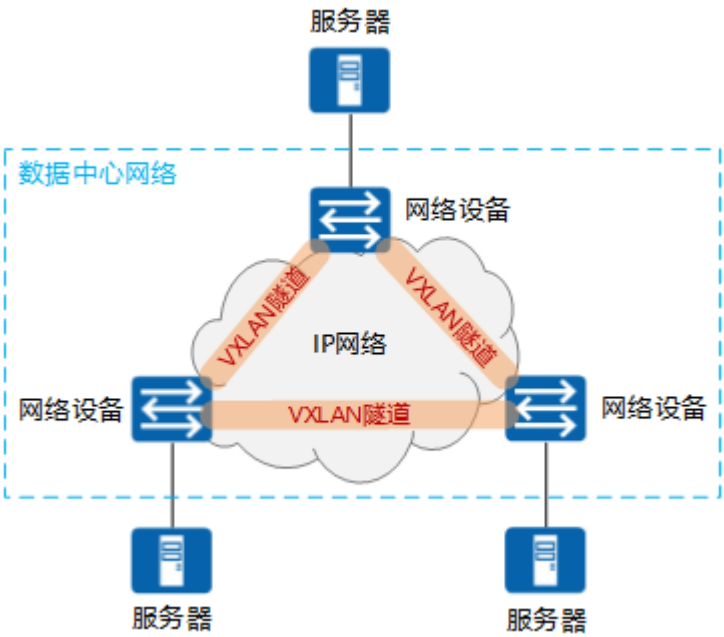
目 录

1 什么是 VXLAN

VXLAN（Virtual eXtensible Local Area Network，虚拟扩展局域网），是由IETF定义的NVO3（Network Virtualization over Layer 3）标准技术之一，是对传统VLAN协议的一种扩展。VXLAN的特点是将L2的以太帧封装到UDP报文（即L2 over L4）中，并在L3网络中传输。

如图1-1所示，VXLAN本质上是一种隧道技术，在源网络设备与目的网络设备之间的IP网络上，建立一条逻辑隧道，将用户侧报文经过特定的封装后通过这条隧道转发。从用户的角度来看，接入网络的服务器就像是连接到了一个虚拟的二层交换机的不同端口上（可把蓝色虚框表示的数据中心VXLAN网络看成一个二层虚拟交换机），可以方便地通信。

图 1-1 VXLAN 是一种隧道技术



VXLAN已经成为当前构建数据中心的主流技术，是因为它能很好地满足数据中心里虚拟机动态迁移和多租户等需求。

2 为什么需要 VXLAN

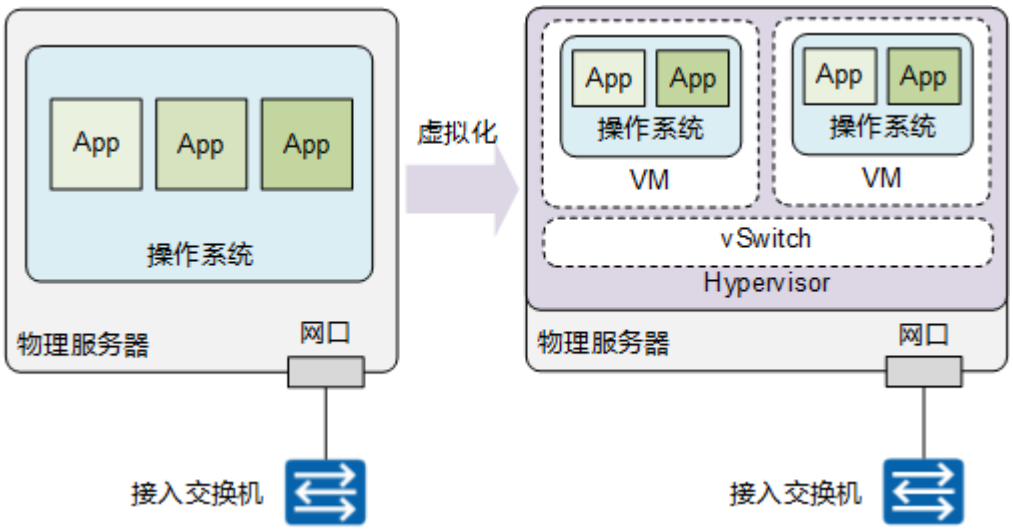
为什么需要VXLAN呢？这和数据中心服务器侧的虚拟化趋势紧密相关，一方面服务器虚拟化后出现了虚拟机动态迁移，要求提供一个无障碍接入的网络；另一方面，数据中心规模越发庞大，租户数量激增，需要网络提供隔离海量租户的能力。采用VXLAN可以满足上述两个关键需求。

虚拟机动态迁移，要求提供一个无障碍接入的网络

什么是服务器虚拟化技术?

传统的数据中心物理服务器利用率太低，平均只有10%～15%，浪费了大量的电力能源和机房资源，所以出现了服务器虚拟化技术。如图2-1所示，服务器虚拟化技术是把一台物理服务器虚拟化成多台逻辑服务器，这种逻辑服务器被称为虚拟机（VM）。每个VM都可以独立运行，有自己的操作系统、APP，当然也有自己独立的MAC地址和IP地址，它们通过服务器内部的虚拟交换机（vSwitch）与外部实体网络连接。

图 2-1 服务器虚拟化示意



通过服务器虚拟化，可以有效地提高服务器的利用率，降低能源消耗，降低数据中心的运营成本，所以虚拟化技术目前得到了广泛的应用。

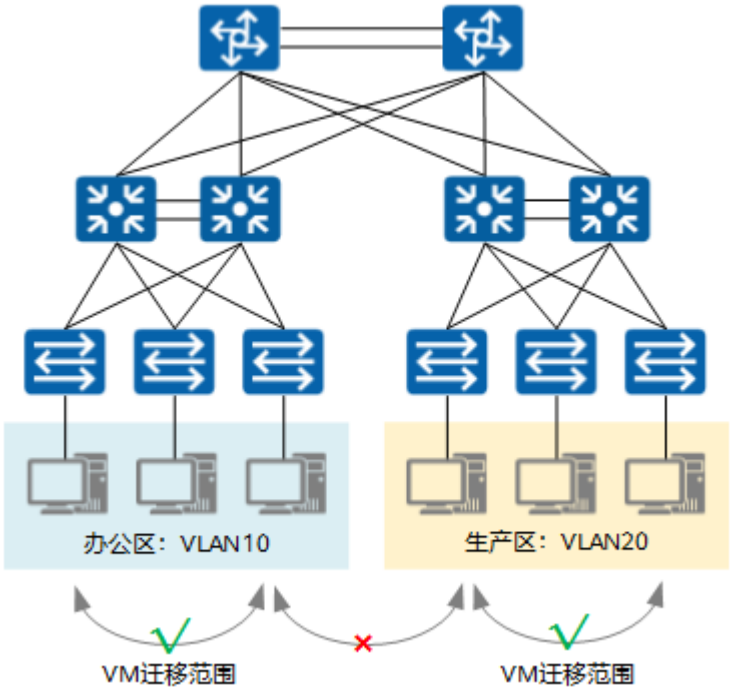
什么是虚拟机动态迁移?

所谓虚拟机动态迁移，就是在保证虚拟机上服务正常运行的同时，将一个虚拟机系统从一个物理服务器移动到另一个物理服务器的过程。该过程对于最终用户来说是无感知的，从而使得管理员能够在不影响用户正常使用的情况下，灵活调配服务器资源，或者对物理服务器进行维修和升级。

在服务器虚拟化后，虚拟机动态迁移变得常态化，为了保证迁移时业务不中断，就要求在虚拟机迁移时，不仅虚拟机的IP地址不变，而且虚拟机的运行状态也必须保持原状（例如TCP会话状态），所以虚拟机的动态迁移只能在同一个二层域中进行，而不能跨二层域迁移。

如图2-2所示，传统的二三层网络架构限制了虚拟机的动态迁移范围，迁移只能在一个较小的局部范围内进行，应用受到了极大的限制。

图 2-2 传统的二三层网络架构限制了虚拟机的动态迁移范围



为了打破这种限制，实现虚拟机的大范围甚至跨地域的动态迁移，就要求把VM迁移可能涉及的所有服务器都纳入同一个二层网络域，这样才能实现VM的大范围无障碍迁移。

VXLAN如何满足虚拟机动态迁移时对网络的要求?

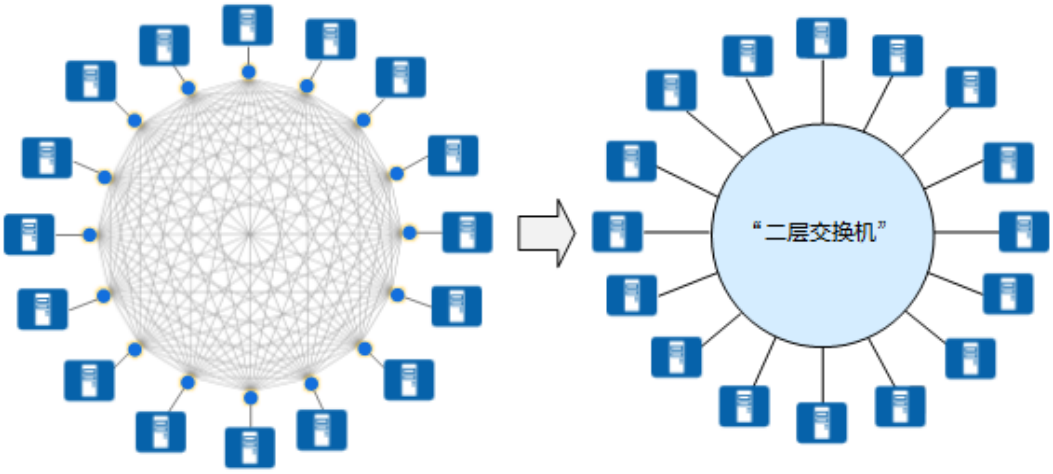
众所周知，同一台二层交换机可以实现下挂服务器之间的二层通信，而且服务器从该二层交换机的一个端口迁移到另一个端口时，IP地址是可以保持不变的。这样就可以满足虚拟机动态迁移的需求了。VXLAN的设计理念和目标正是由此而来的。

从上一个小节我们可以知道，VXLAN本质上是一种隧道技术，当源和目的之间有通信需求时，便在数据中心IP网络之上创建一条虚拟的隧道，透明转发用户数据。而数据中心内相互通信的需求众多，这种隧道的建立方式几乎是全互联形态才能满足通信需求。

VXLAN可以提供一套方法论，在数据中心IP网络基础上，构建一张全互联的二层隧道虚拟网络，保证任意两点之间都能通过VXLAN隧道来通信，并忽略底层网络的结构和细节。从服务器的角度看，VXLAN为它们将整个数据中心基础网络虚拟成了一台巨大

的“二层交换机”，所有服务器都连接在这台虚拟二层交换机上。而基础网络之内如何转发都是这台“巨大交换机”内部的事情，服务器完全无需关心。

图 2-3 VXLAN 将整个数据中心基础网络虚拟成了一台巨大的“二层交换机”



基于这种“二层交换机”的模型，就很容易理解为什么VXLAN可以实现VM动态迁移了：将虚拟机从“二层交换机”的一个端口换到另一个端口，完全无需变更IP地址。

说明

使用这种理念的技术协议，除了VXLAN外，还有NVGRE、STT等，本文仅对VXLAN进行说明。

数据中心租户数量激增，要求提供一个可隔离海量租户的网络

众所周知，在传统的VLAN网络中，标准定义所支持的可用VLAN数量只有4000个左右。服务器虚拟化后，一台物理服务器中承载了多台虚拟机，每个虚拟机都有独立的IP地址和MAC地址，相当于接入数据中心的服务器成倍扩大了。另外，公有云或其它大型虚拟化云数据中心动辄需容纳上万甚至更多租户，VLAN的能力显然已经力不从心。

VXLAN如何来解决上述问题呢？VXLAN在VXLAN帧头中引入了类似VLAN ID的网络标识，称为VXLAN网络标识VNI（VXLAN Network ID），由24比特组成，理论上可支持多达16M的VXLAN段，从而满足了大规模不同网络之间的标识、隔离需求。下文我们会介绍VNI的详细作用。

3 VXLAN 与 VLAN 之间有何不同

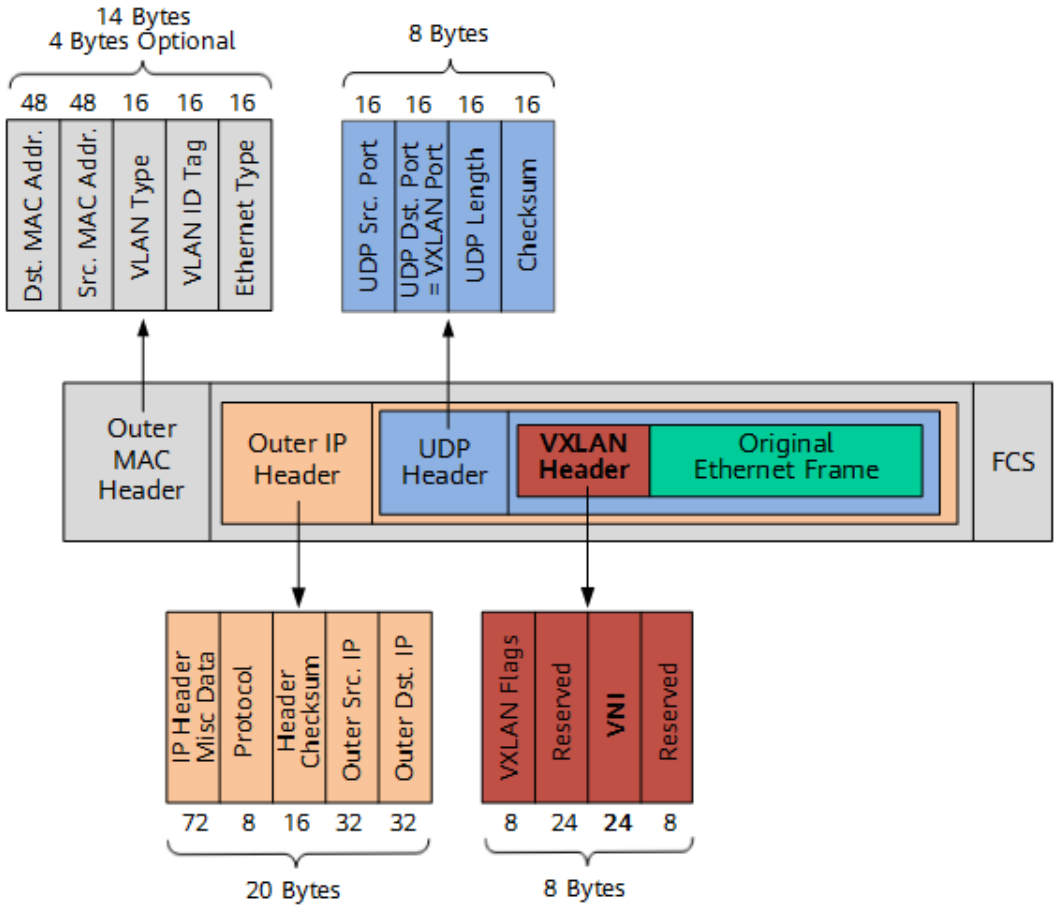
VLAN作为传统的网络隔离技术，在标准定义中VLAN的数量只有4000个左右，无法满足大型数据中心的租户间隔离需求。另外，VLAN的二层范围一般较小且固定，无法支持虚拟机大范围的动态迁移。

VXLAN完美地弥补了VLAN的上述不足，一方面通过VXLAN中的24比特VNI字段（如图3-1所示），提供多达16M租户的标识能力，远大于VLAN的4000；另一方面，VXLAN本质上在两台交换机之间构建了一条穿越数据中心基础IP网络的虚拟隧道，将数据中心网络虚拟成一个巨型“二层交换机”，满足虚拟机大范围动态迁移的需求。

虽然从名字上看，VXLAN是VLAN的一种扩展协议，但VXLAN构建虚拟隧道的本领已经与VLAN迥然不同了。

下面就让我们来看下，VXLAN报文到底长啥样。

图 3-1 VXLAN 报文格式（以外层 IP 头为 IPv4 格式为例）



如上图所示，VTEP对VM发送的原始以太帧（Original L2 Frame）进行了以下“包装”：

- VXLAN Header

  增加VXLAN头（8字节），其中包含24比特的**VNI**字段，用来定义VXLAN网络中不同的租户。此外，还包含**VXLAN Flags**（8比特，取值为00001000）和两个保留字段（分别为24比特和8比特）。

- UDP Header

  VXLAN头和原始以太帧一起作为UDP的数据。UDP头中，目的端口号（VXLAN Port）固定为4789，源端口号（UDP Src. Port）是原始以太帧通过哈希算法计算后的值。

- Outer IP Header

  封装外层IP头。其中，源IP地址（Outer Src. IP）为源VM所属VTEP的IP地址，目的IP地址（Outer Dst. IP）为目的VM所属VTEP的IP地址。

- Outer MAC Header

  封装外层以太头。其中，源MAC地址（Src. MAC Addr.）为源VM所属VTEP的MAC地址，目的MAC地址（Dst. MAC Addr.）为到达目的VTEP的路径中下一跳设备的MAC地址。

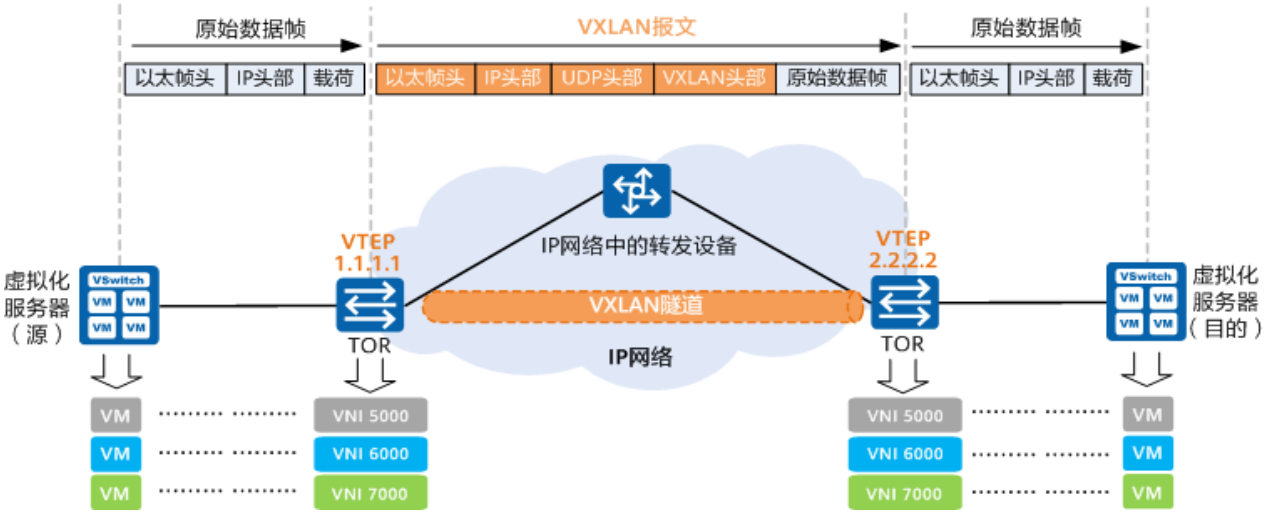
4 VXLAN 隧道是如何建立的

本节将为您介绍VXLAN隧道的建立过程，并在这个过程中更好地理解VXLAN的工作原理。

4.1 什么是 VXLAN 中的 VTEP 和 VNI

下面让我们来进一步了解VXLAN的网络模型以及一些常见的概念。如图4-1所示，两台服务器之间通过VXLAN网络进行通信。

图 4-1 VXLAN 网络模型示意



从上图中可以发现，VXLAN在两台TOR交换机之间建立了一条隧道，将服务器发出的原始数据帧加以“包装”，好让原始报文可以在承载网络（比如IP网络）上传输。当到达目的服务器所连接的TOR交换机后，离开VXLAN隧道，并将原始数据帧恢复出来，继续转发给目的服务器。

另外，VXLAN网络中出现了一些传统数据中心网络中没有的新元素，如VTEP、VNI等，它们的作用是什么呢？下面将向您介绍这几个新元素。

什么是 VXLAN VTEP

如**图4-1**所示，VTEP（VXLAN Tunnel Endpoints，VXLAN隧道端点）是VXLAN网络的边缘设备，是VXLAN隧道的起点和终点，VXLAN对用户原始数据帧的封装和解封装均在VTEP上进行。

VTEP是VXLAN网络中绝对的主角，VTEP既可以是一台独立的网络设备（比如华为的CloudEngine系列交换机），也可以是在服务器中的虚拟交换机。源服务器发出的原始数据帧，在VTEP上被封装成VXLAN格式的报文，并在IP网络中传递到另外一个VTEP上，并经过解封转还原出原始的数据帧，最后转发给目的服务器。

VTEP在VXLAN隧道的建立和转发过程中的详细工作，可以参见下文的**4 VXLAN隧道是如何建立的**

什么是 VXLAN VNI

前文提到，以太网数据帧中VLAN只占了12比特的空间，这使得VLAN的隔离能力在数据中心网络中力不从心。而VNI的出现，就是专门解决这个问题的。

如**图4-1**所示，VNI（VXLAN Network Identifier，VXLAN 网络标识符），VNI是一种类似于VLAN ID的用户标识，一个VNI代表了一个租户，属于不同VNI的虚拟机之间不能直接进行二层通信。如**图3-1**所示，VXLAN报文封装时，给VNI分配了24比特的长度空间，使其可以支持海量租户的隔离。

VNI在VXLAN隧道的建立和转发过程中的详细工作，可以参见下文的**4 VXLAN隧道是如何建立的**

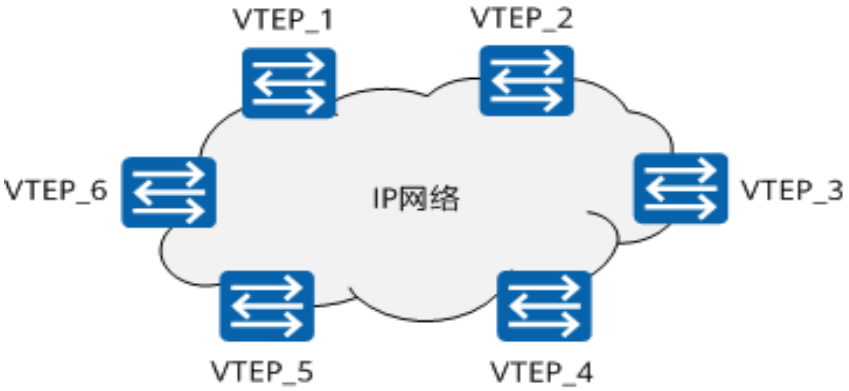
另外，在**分布式网关**部署场景下，VNI还可分为二层VNI和三层VNI，它们的作用不同：

- 二层VNI是普通的VNI，以1：1方式映射到广播域BD，实现VXLAN报文同子网的转发（详情可参见下文的**什么是“同一大二层域”**）。
- 三层VNI和VPN实例进行关联，用于VXLAN报文跨子网的转发（三层VNI的工作详情将在另外一篇EVPN相关的文档中展开描述）。

4.2 哪些 VTEP 之间需要建立 VXLAN 隧道

一条VXLAN隧道是由两个VTEP来确定建立的。数据中心网络中存在很多个VTEP，如**图4-2**所示，那么哪些VTEP间需要建立VXLAN隧道呢？

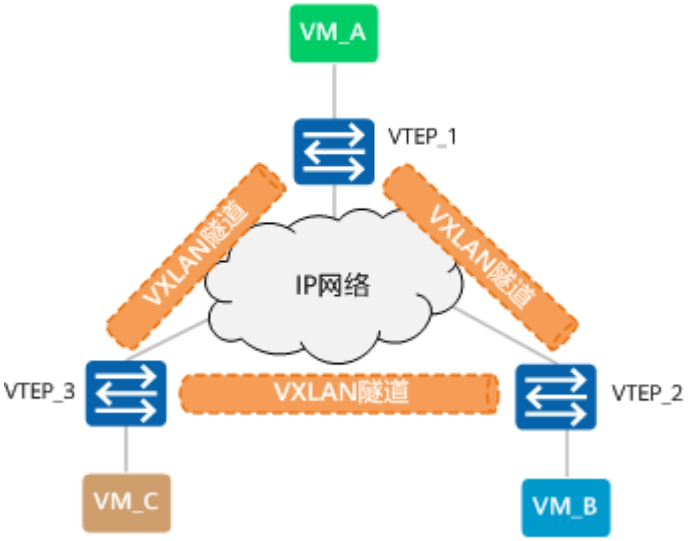
图 4-2 建立 VXLAN 隧道示意图（1）



如前所述，通过VXLAN隧道，“二层域”可以突破物理上的界限，实现大二层网络中VM之间的通信。所以，连接在不同VTEP上的VM之间如果有“大二层”互通的需求，这两个VTEP之间就需要建立VXLAN隧道。换言之，同一大二层域内的VTEP之间都需要建立VXLAN隧道。

例如，假设图4-2中VTEP_1连接的VM、VTEP_2连接的VM以及VTEP_3连接的VM之间需要“大二层”互通，那VTEP_1、VTEP_2和VTEP_3之间就需要两两建立VXLAN隧道，如图4-3所示。

图 4-3 建立 VXLAN 隧道示意图（2）



什么是“同一大二层域”

上文提到的“同一大二层域”，就类似于传统网络中VLAN（虚拟局域网）的概念，只不过在VXLAN网络中，它有另外一个名字，叫做Bridge-Domain，简称BD。

我们知道，不同的VLAN是通过VLAN ID来进行区分的，那不同的BD是如何进行区分的呢？其实前面已经提到了，就是通过VNI来区分的。对于CloudEngine系列交换机而言，BD与VNI是1：1的映射关系，这种映射关系是通过在VTEP设备上配置命令行建立起来的，配置样例如下：


```
bridge-domain 10   //表示创建一个“大二层广播域”BD，其编号为10
 vxlan vni 5000  //表示在BD 10下，指定与之关联的VNI为5000
#
```

VTEP设备会根据以上配置生成BD与VNI的映射关系表，该映射表可以通过命令行查看，如下所示：

```
<HUAWEI> display vxlan vni
Number of vxlan vni : 1
VNI          BD-ID          State
------------------------------------------
5000         10             up
```

有了映射表后，进入VTEP的报文就可以根据自己所属的BD来确定报文在进行VXLAN封装时，该添加哪个VNI标识。那么，报文根据什么来确定自己属于哪个BD呢?

如何确定报文属于哪个 BD

这里要先澄清下，VTEP只是交换机承担的一个角色而已，只是交换机功能的一部分。也就是说，并非所有进入到交换机的报文都会走VXLAN隧道（也可能报文就是走普通的二三层转发流程）。所以，我们在回答“如何确定报文属于哪个BD”之前，必须先要回答“哪些报文要进入VXLAN隧道”。

哪些报文要进入VXLAN隧道?

回答这个问题之前，不妨先让我们回想一下VLAN技术中，交换机对于接收和发送的报文是如何进行处理的。报文要进入交换机进行下一步处理，首先得先过接口这一关，可以说接口掌控着对报文的“生杀大权”。传统网络中定义了三种不同类型的接口：Access、Trunk、Hybrid。这三种类型的接口虽然应用场景不同，但它们的最终目的是一样的：一是根据配置来检查哪些报文是允许通过的；二是判断对检查通过的报文做怎样的处理。

其实在VXLAN网络中，VTEP上的接口也承担着类似的任务，只不过在CloudEngine系列交换机中，这里的接口不是物理接口，而是一个叫做“二层子接口”的逻辑接口。类似的，二层子接口主要做两件事：一是根据配置来检查哪些报文需要进入VXLAN隧道；二是判断对检查通过的报文做怎样的处理。在二层子接口上，可以根据需要定义不同的流封装类型（类似于传统网络中不同的接口类型）。CloudEngine系列交换机目前支持的流封装类型有dot1q、untag、qinq和default四种类型：

- dot1q：对于带有一层VLAN Tag的报文，该类型接口只接收与指定VLAN Tag匹配的报文；对于带有两层VLAN Tag的报文，该类型接口只接收外层VLAN Tag与指定VLAN Tag匹配的报文。
- untag：该类型接口只接收不带VLAN Tag的报文。
- qinq：该类型接口只接收带有指定两层VLAN Tag的报文。
- default：允许接口接收所有报文，不区分报文中是否带VLAN Tag。不论是对原始报文进行VXLAN封装，还是解封装VXLAN报文，该类型接口都不会对原始报文进行任何VLAN Tag处理，包括添加、替换或剥离。

说明

VXLAN隧道两端二层子接口的配置并不一定是完全对等的。正因为这样，才可能实现属于同一网段但是不同VLAN的两个VM通过VXLAN隧道进行通信。

除二层子接口外，还可以将VLAN作为业务接入点。将VLAN绑定到广播域BD后，加入该VLAN的接口即为VXLAN业务接入点，进入接口的报文由VXLAN隧道处理。

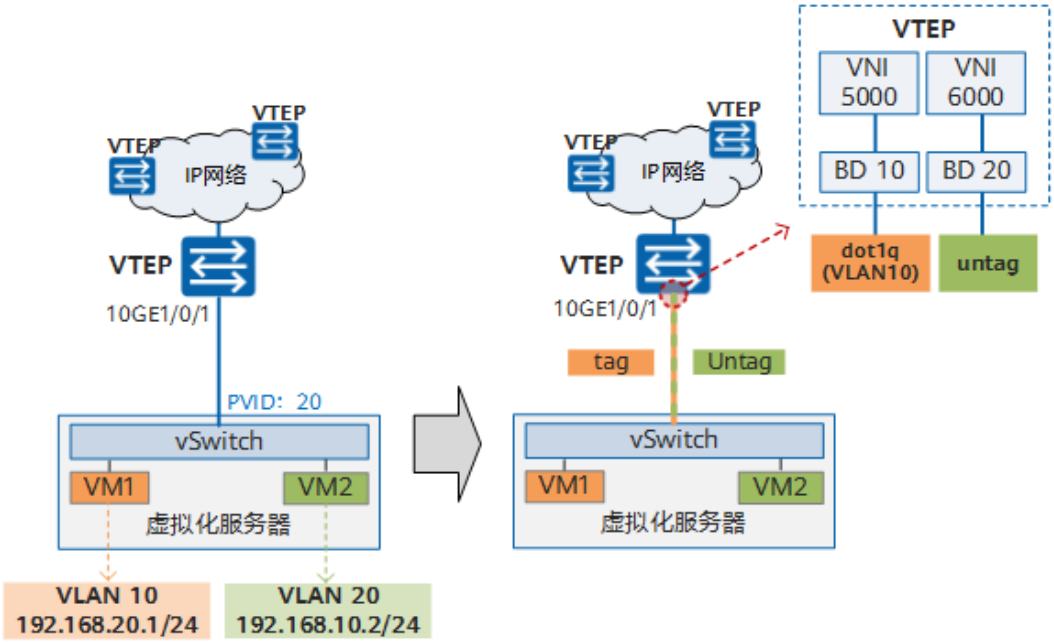
将二层子接口加入BD

现在我们再来回答“如何确定报文属于哪个BD”就非常简单了。其实，只要将二层子接口加入指定的BD，然后根据二层子接口上的配置，设备就可以确定报文属于哪个BD啦！

比如图4-4所示的组网，一台虚拟化服务器中有两个不同VLAN的虚拟机VM1（VLAN 10）和VM2（VLAN 20），它们与其他虚拟机通信时需要接入VXLAN网络。此时我们可以分别在VTEP的物理接口10GE 1/0/1上，分别针对VM1和VM2封装不同的二层子接口，并将其分别加入不同的BD。这样后续VM1和VM2的流量将会进入不同的VXLAN隧道继续转发。

在这个举例中，vSwitch的上行口配置成Trunk模式，且PVID为20。这样vSwitch发给VTEP的报文中，既有带tag的VM1流量，又有untag的VM2流量，此时在VTEP的接入口上创建两个二层子接口，分别配置为dot1q和untag的封装类型。

图 4-4 将二层子接口加入 BD



下面就基于上图，结合CloudEngine交换机上的配置举例进行说明。

在CloudEngine交换机的接入物理接口10GE 1/0/1上，分别创建二层子接口10GE 1/0/1.1和10GE 1/0/1.2，并分别配置其流封装类型为dot1q和untag。

```
interface 10GE1/0/1.1 mode l2   //创建二层子接口10GE1/0/1.1
 encapsulation dot1q vid 10   //只允许携带VLAN Tag 10的报文进入VXLAN隧道
 bridge-domain 10   //指定报文进入的是BD 10
#
interface 10GE1/0/1.2 mode l2   //创建二层子接口10GE1/0/1.2
 encapsulation untag   //只允许不携带VLAN Tag的报文进入VXLAN隧道
 bridge-domain 20   //指定报文进入的是BD 20
#
```

4.3 VXLAN 隧道是怎么建立的

现在，我们可以来看下VXLAN隧道是怎么建立起来的。一般而言，隧道的建立不外乎手工方式和自动方式两种。

手工方式建立 VXLAN 隧道

这种方式需要用户手动指定VXLAN隧道的源IP为本端VTEP的IP、目的IP为对端VTEP的IP，也就是人为地在本端VTEP和对端VTEP之间建立静态VXLAN隧道。

对于CloudEngine系列交换机，以上配置是在NVE（Network Virtualization Edge）接口下完成的，配置举例如下：

```
interface Nve1   //创建逻辑接口NVE 1
 source 1.1.1.1   //配置源VTEP的IP地址（推荐使用Loopback接口的IP地址）
 vni 5000 head-end peer-list 2.2.2.2
 vni 5000 head-end peer-list 2.2.2.3
#
```

其中，**vni 5000 head-end peer-list 2.2.2.2**和**vni 5000 head-end peer-list 2.2.2.3**的配置，表示属于VNI 5000的对端VTEP有两个，IP地址分别为2.2.2.2和2.2.2.3。根据这两条配置，VTEP上会生成如下所示的一张表：

```
<HUAWEI> display vxlan vni 5000 verbose
  BD ID              : 10
  State              : up
  NVE                : 288
  Source Address     : 1.1.1.1
  Source IPv6 Address    : -
  UDP Port           : 4789
  BUM Mode           : head-end
  Group Address      : -
  Peer List          : 2.2.2.2 2.2.2.3
  IPv6 Peer List     : -
```

根据上表中的**Peer List**，本端VTEP就可以知道属于同一BD（或同一VNI）的对端VTEP都有哪些，这也就决定了同一大二层广播域的范围。当VTEP收到BUM（Broadcast&Unknown-unicast&Multicast，广播&未知单播&组播）报文时，会将报文复制并发送给Peer List中所列的所有对端VTEP（这就好比广播报文在VLAN内广播）。因此，这张表也被称为“头端复制列表”。当VTEP收到已知单播报文时，会根据VTEP上的MAC表来确定报文要从哪条VXLAN隧道走。而此时Peer List中所列的对端，则充当了MAC表中“出接口”的角色。

在后面的报文转发流程中，你将会看到头端复制列表是如何在VXLAN网络中指导报文进行转发的。

自动方式建立 VXLAN 隧道

自动方式下VXLAN隧道的建立需要借助于EVPN（Ethernet VPN）协议，这部分内容请参见**《什么是EVPN》**。

如何确定报文要进哪条隧道?

属于同一BD的VXLAN隧道可能不止一条，比如上文的头端复制列表中，同一个源端VTEP（1.1.1.1）对应了两个对端VTEP（2.2.2.2和2.2.2.3）。那就带来了另一个问题，报文到底应该走哪一条隧道呢?

我们知道，基本的二三层转发中，二层转发依赖的是MAC表，如果没有对应的MAC条目，则主机发送ARP广播报文请求对端的MAC地址；三层转发依赖的是FIB表。在VXLAN中，其实也是同样的道理。在下一小节中，将介绍VXLAN网络中报文的转发流程，相信看完下面的内容，关于“如何确定报文要进哪条隧道”的疑惑也就迎刃而解了。

5 VXLAN 网关有哪些种类

VXLAN 二层网关与三层网关

和VLAN类似，不同VNI之间的主机，以及VXLAN网络和非VXLAN网络中的主机不能直接相互通信。为了满足这些通信需求，VXLAN引入了VXLAN网关的概念。VXLAN网关分为二层网关和三层网关：

- VXLAN二层网关：用于终端接入VXLAN网络，也可用于同一VXLAN网络的子网通信。
- VXLAN三层网关：用于VXLAN网络中跨子网通信以及访问外部网络。

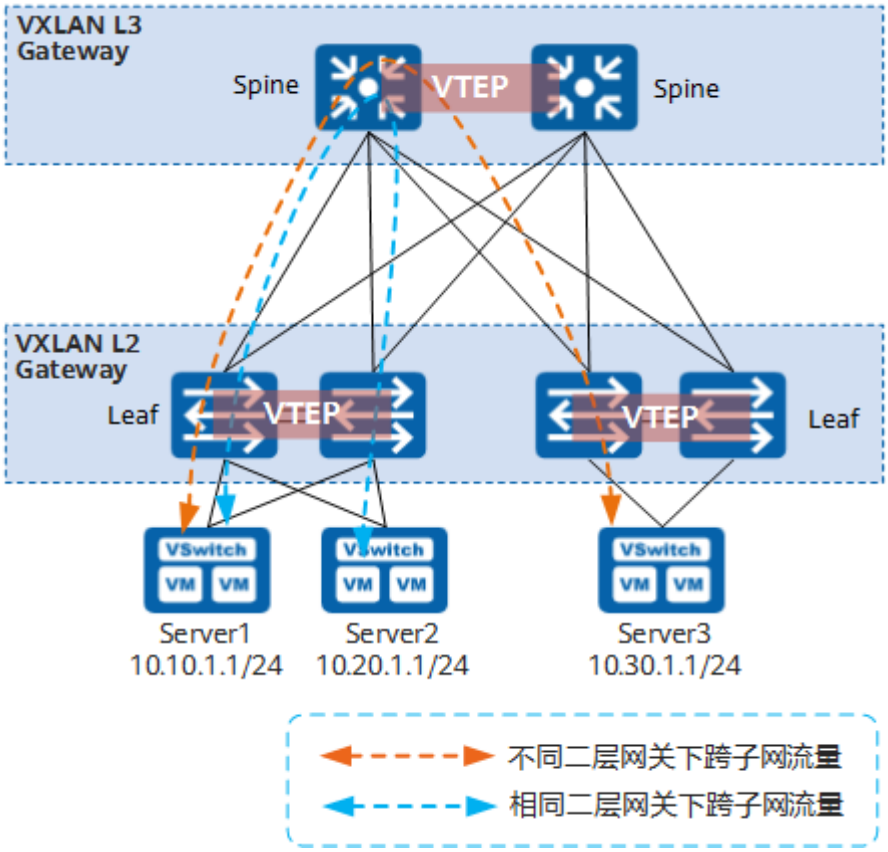
VXLAN 集中式网关与分布式网关

根据三层网关部署方式的不同，VXLAN三层网关又可以分为集中式网关和分布式网关。

VXLAN集中式网关

集中式网关是指将三层网关集中部署在一台设备上，如下图所示，所有跨子网的流量都经过这个三层网关转发，实现流量的集中管理。

图 5-1 VXLAN 集中式网关组网图



部署集中式网关的优点和缺点如下：

- 优点：对跨子网流量进行集中管理，网关的部署和管理比较简单。
- 缺点：
  - 转发路径不是最优：同一二层网关下跨子网的数据中心三层流量都需要经过集中三层网关绕行转发（如图中蓝色虚线所示）。
  - ARP表项规格瓶颈：由于采用集中三层网关，通过三层网关转发的终端的ARP表项都需要在三层网关上生成，而三层网关上的ARP表项规格有限，这不利于数据中心网络的扩展。

VXLAN分布式网关

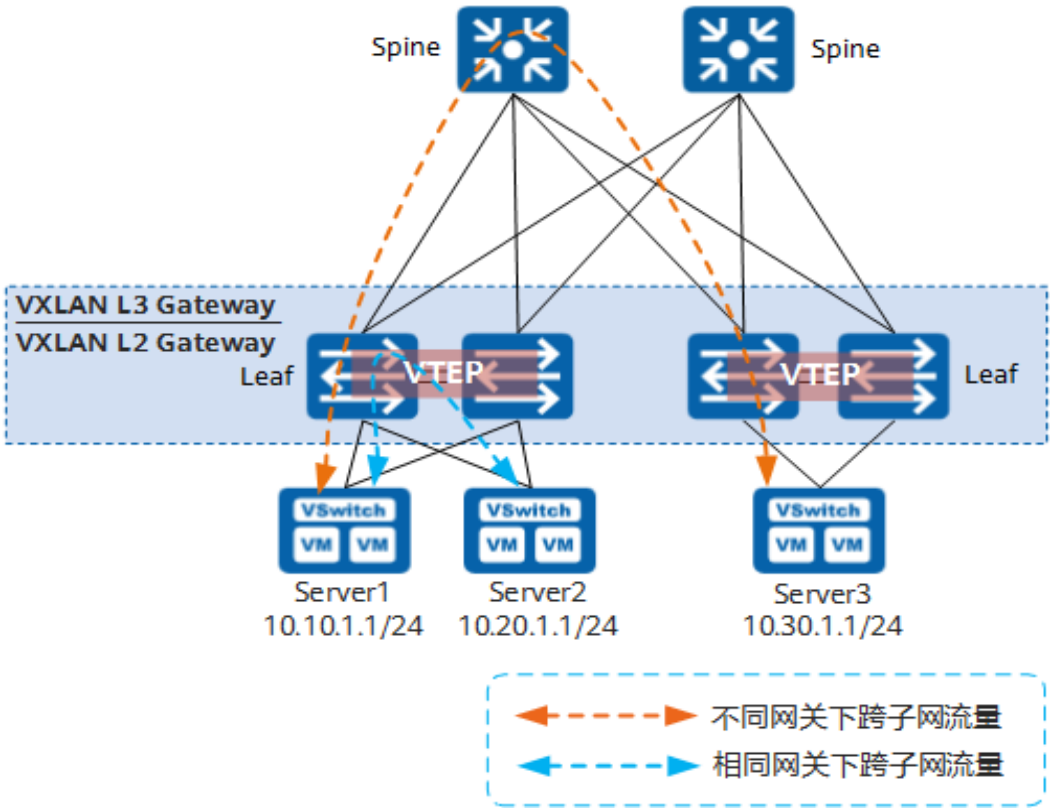
通过部署分布式网关可以解决集中式网关部署的缺点。VXLAN分布式网关是指在典型的“Spine-Leaf”组网结构下，将Leaf节点作为VXLAN隧道端点VTEP，每个Leaf节点都可作为VXLAN三层网关（同时也是VXLAN二层网关），Spine节点不感知VXLAN隧道，只作为VXLAN报文的转发节点。如下图所示，Server1和Server2不在同一个网段，但是都连接到同一个Leaf节点。Server1和Server2通信时，流量只需要在该Leaf节点上转发，不再需要经过Spine节点。

部署分布式网关时：

- Spine节点：关注于高速IP转发，强调的是设备的高速转发能力。
- Leaf节点：
  - 作为VXLAN网络中的二层网关设备，与物理服务器或VM对接，用于解决终端租户接入VXLAN虚拟网络的问题。

- 作为VXLAN网络中的三层网关设备，进行VXLAN报文封装/解封装，实现跨子网的终端租户通信，以及外部网络的访问。

图 5-2 VXLAN 分布式网关示意图



VXLAN分布式网关具有如下特点：

- 同一个Leaf节点既可以做VXLAN二层网关，也可以做VXLAN三层网关，部署灵活。
- Leaf节点只需要学习自身连接服务器的ARP表项，而不必像集中三层网关一样，需要学习所有服务器的ARP表项，解决了集中式三层网关带来的ARP表项瓶颈问题，网络规模扩展能力强。

说明

分布式VXLAN网络中，推荐使用BGP EVPN作为VXLAN网络的控制面，关于BGP EVPN的相关介绍，请参见**《什么是EVPN》**。

6 VXLAN 网络中报文是如何转发的

本节以集中式VXLAN网络（手工方式建立VXLAN隧道）为例，分别介绍相同子网内、不同子网间是如何进行通信的，帮助您理解上文所介绍到的概念。

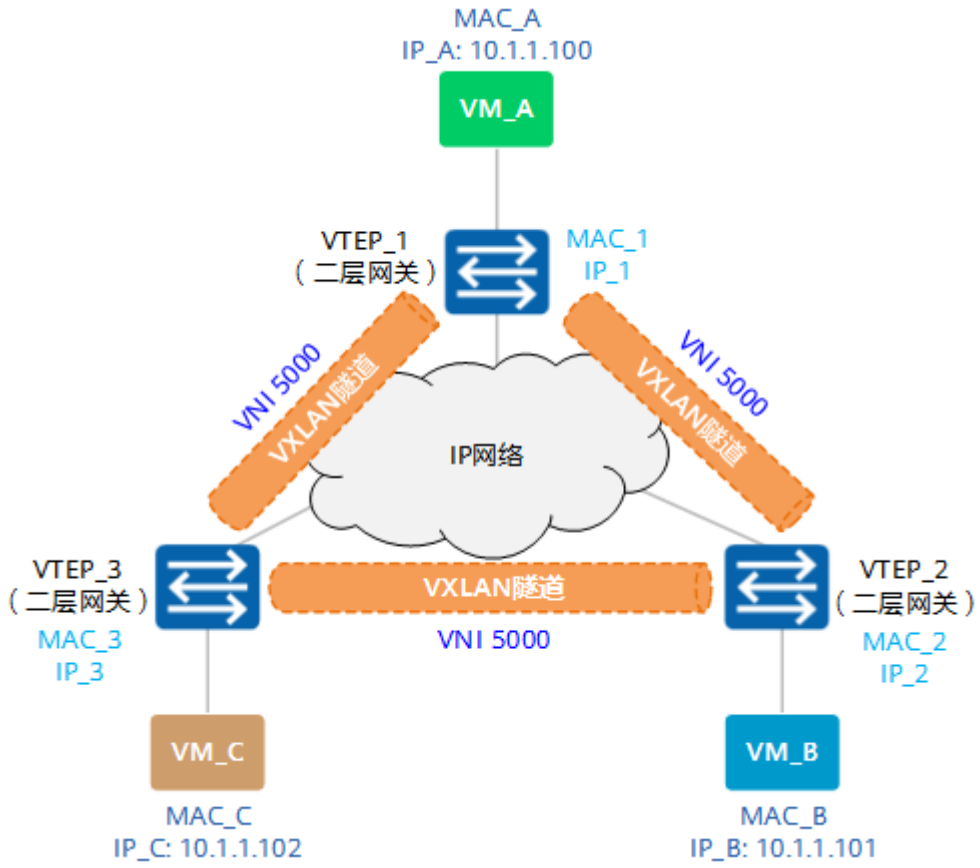
部署了BGP EVPN的分布式VXLAN网络中报文的转发过程，请参见**《什么是EVPN》**。

6.1 集中式 VXLAN 中同子网互通流程

如**图6-1**所示，VM_A、VM_B和VM_C同属于10.1.1.0/24网段，且同属于VNI 5000。此时，VM_A想与VM_C进行通信。

由于是首次进行通信，VM_A上没有VM_C的MAC地址，所以会发送ARP广播报文请求VM_C的MAC地址。

图 6-1 同子网 VM 互通组网图

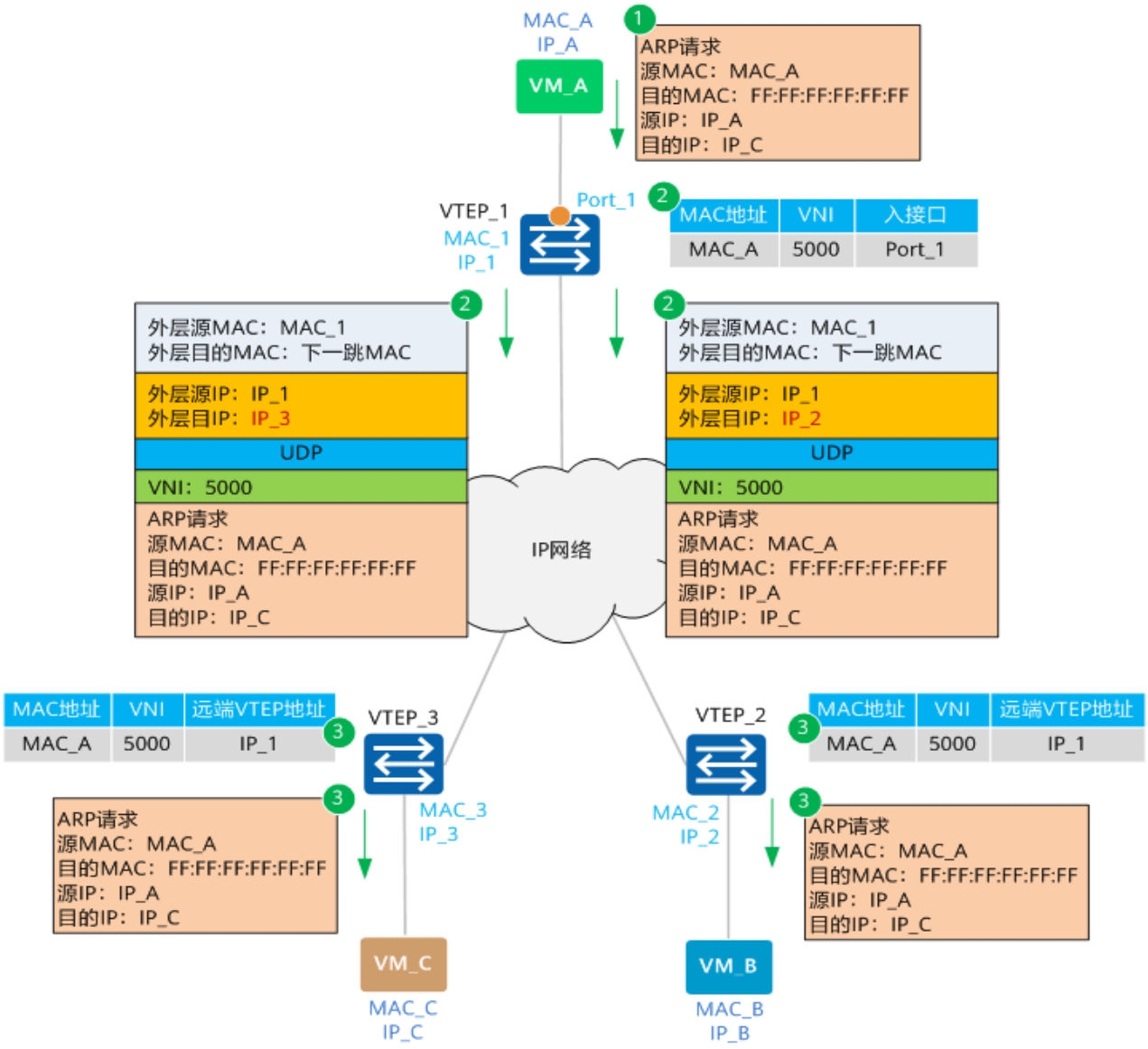


下面就让我们根据ARP请求报文及ARP应答报文的转发流程，来看下MAC地址是如何进行学习的。

ARP 请求报文转发流程

结合图6-2，我们来一起了解一下ARP请求报文的转发流程。

图 6-2 ARP 请求报文转发流程示意



1. VM_A发送源MAC为MAC_A、目的MAC为全F、源IP为IP_A、目的IP为IP_C的ARP广播报文，请求VM_C的MAC地址。
2. VTEP_1收到ARP请求后，根据二层子接口上的配置判断报文需要进入VXLAN隧道。确定了报文所属BD后，也就确定了报文所属的VNI。同时，VTEP_1学习MAC_A、VNI和报文入接口（Port_1，即二层子接口对应的物理接口）的对应关系，并记录在本地MAC表中。之后，VTEP_1会根据头端复制列表对报文进行复制，并分别进行封装。

   可以看到，这里封装的外层源IP地址为本地VTEP（VTEP_1）的IP地址，外层目的IP地址为对端VTEP（VTEP_2和VTEP_3）的IP地址；外层源MAC地址为本地VTEP的MAC地址，而外层目的MAC地址为去往目的IP的网络中下一跳设备的MAC地址。封装后的报文，根据外层MAC和IP信息，在IP网络中进行传输，直至到达对端VTEP。
3. 报文到达VTEP_2和VTEP_3后，VTEP对报文进行解封装，得到VM_A发送的原始报文。同时，VTEP_2和VTEP_3学习VM_A的MAC地址、VNI和远端VTEP的IP地址

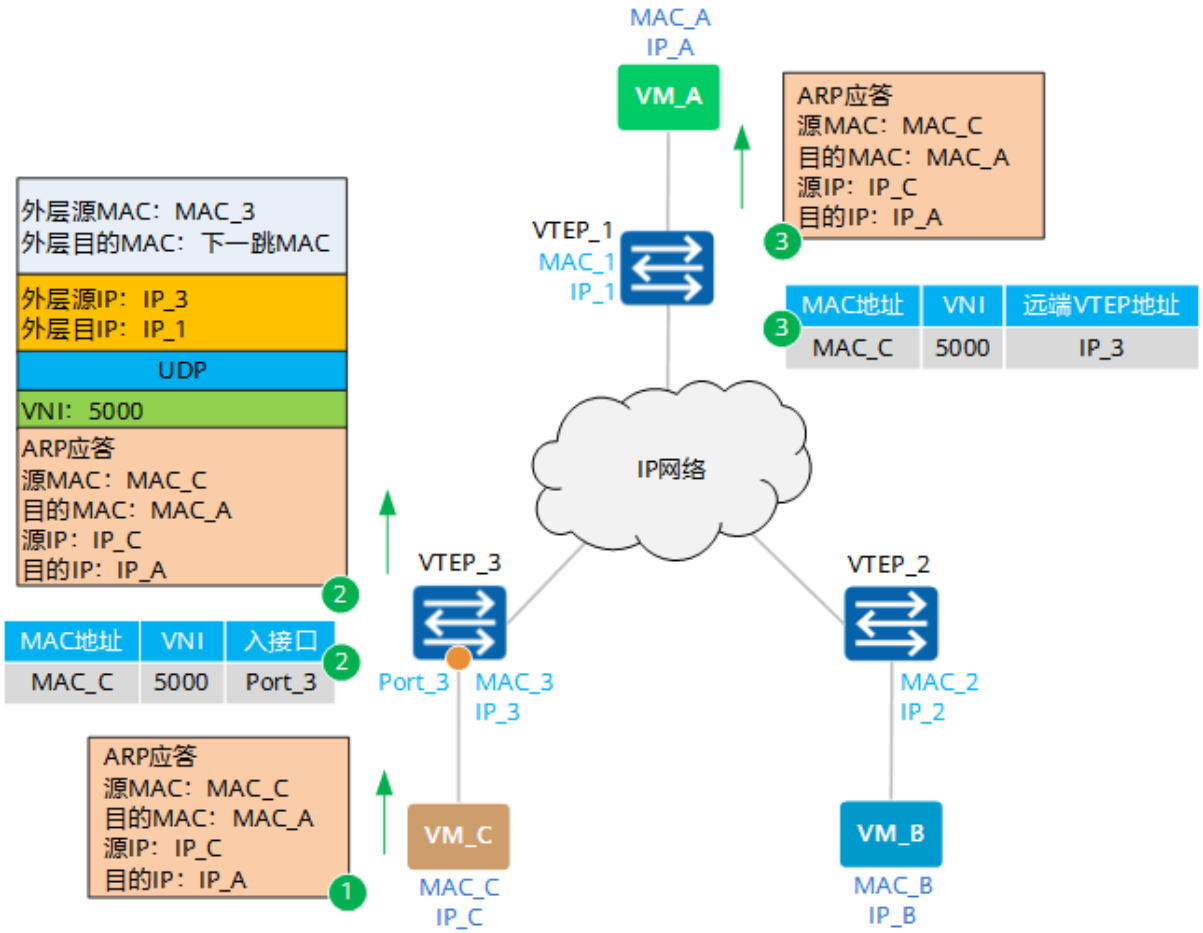
（IP_1）的对应关系，并记录在本地MAC表中。之后，VTEP_2和VTEP_3根据二层子接口上的配置对报文进行相应的处理并在对应的二层域内广播。

VM_B和VM_C接收到ARP请求后，比较报文中的目的IP地址是否为本机的IP地址。VM_B发现目的IP不是本机IP，故将报文丢弃；VM_C发现目的IP是本机IP，则对ARP请求做出应答。下面，让我们看下ARP应答报文是如何进行转发的。

ARP 应答报文转发流程

结合图6-3，我们来一起了解一下ARP应答报文的转发流程。

图 6-3 ARP 应答报文转发流程示意



1. 由于此时VM_C上已经学习到了VM_A的MAC地址，所以ARP应答报文为单播报文。报文源MAC为MAC_C，目的MAC为MAC_A，源IP为IP_C、目的IP为IP_A。
2. VTEP_3接收到VM_C发送的ARP应答报文后，识别报文所属的VNI（识别过程与步骤②类似）。同时，VTEP_3学习MAC_C、VNI和报文入接口（Port_3）的对应关系，并记录在本地MAC表中。之后，VTEP_3对报文进行封装。

   可以看到，这里封装的外层源IP地址为本地VTEP（VTEP_3）的IP地址，外层目的IP地址为对端VTEP（VTEP_1）的IP地址；外层源MAC地址为本地VTEP的MAC地址，而外层目的MAC地址为去往目的IP的网络中下一跳设备的MAC地址。

   封装后的报文，根据外层MAC和IP信息，在IP网络中进行传输，直至到达对端VTEP。

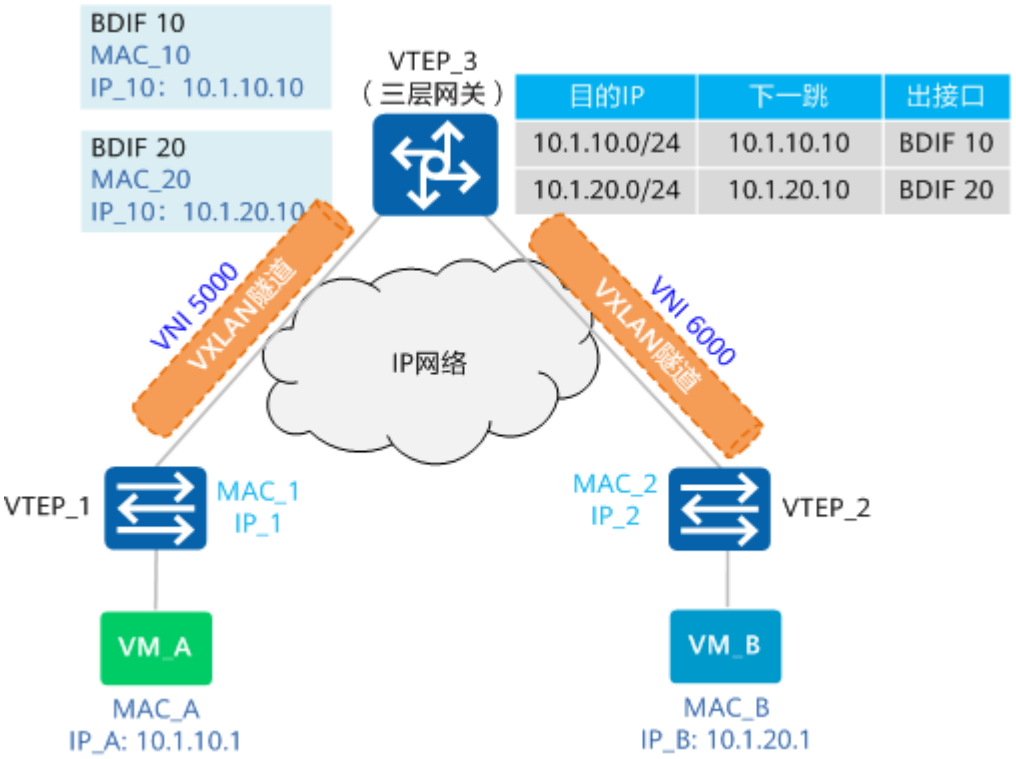
3. 报文到达VTEP_1后，VTEP_1对报文进行解封装，得到VM_C发送的原始报文。同时，VTEP_1学习VM_C的MAC地址、VNI和远端VTEP的IP地址（IP_3）的对应关系，并记录在本地MAC表中。之后，VTEP_1将解封装后的报文发送给VM_A。

至此，VM_A和VM_C均已学习到了对方的MAC地址。之后，VM_A和VM_C将采用单播方式进行通信。单播报文的封装与解封装过程，与图6-3中所展示的类似，本文就不再赘述啦！

6.2 集中式 VXLAN 中不同子网互通流程

如图6-5所示，VM_A和VM_B分别属于10.1.10.0/24网段和10.1.20.0/24网段，且分别属于VNI 5000和VNI 6000。VM_A和VM_B对应的三层网关分别是VTEP_3上BDIF 10和BDIF 20的IP地址。VTEP_3上存在到10.1.10.0/24网段和10.1.20.0/24网段的路由。此时，VM_A想与VM_B进行通信。

图 6-4 不同子网 VM 互通流程示意

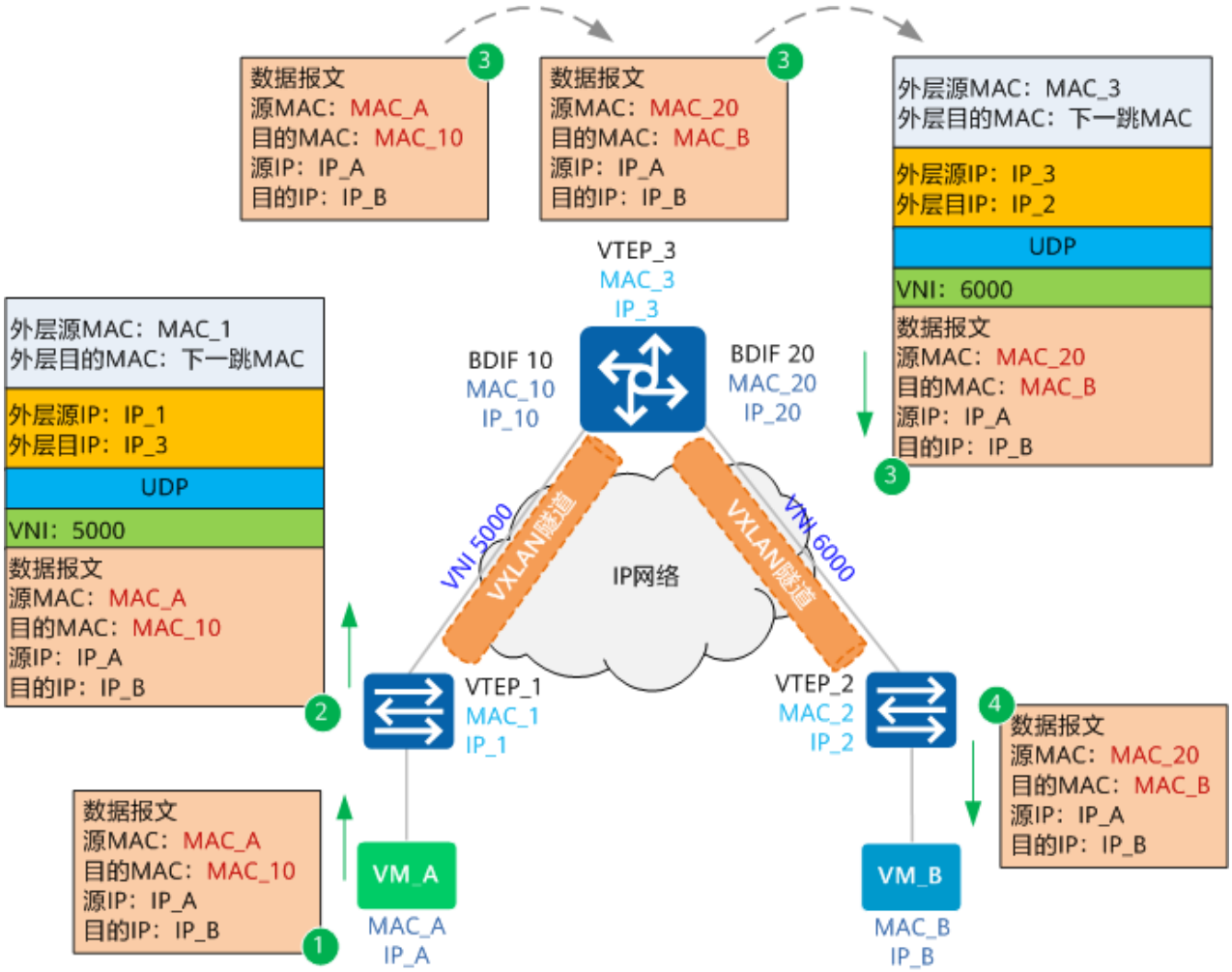


说明

BDIF接口的功能与VLANIF接口类似，是基于BD创建的三层逻辑接口，用以实现不同子网之间的通信，或VXLAN网络与非VXLAN网络之间的通信。

由于是首次进行通信，且VM_A和VM_B处于不同网段，VM_A需要先发送ARP广播报文请求网关（BDIF 10）的MAC。获得网关的MAC后，VM_A先将数据报文发送给网关；之后网关也将发送ARP广播报文请求VM_B的MAC，获得VM_B的MAC后，网关再将数据报文发送给VM_B。以上MAC地址学习的过程与**6.1 集中式VXLAN中同子网互通流程**中MAC地址学习的流程一致，不再赘述。现在假设VM_A和VM_B均已学到网关的MAC、网关也已经学到VM_A和VM_B的MAC，下面就让我们看下数据报文是如何从VM_A发送到VM_B的。

图 6-5 不同子网 VM 互通报文转发流程



如上图所示，数据报文从VM_A发送到VM_B的流程如下：

1. VM_A先将数据报文发送给网关。报文的源MAC为MAC_A，目的MAC为网关BDIF 10的MAC_10，源IP地址为IP_A，目的IP为IP_B。
2. VTEP_1收到数据报文后，识别此报文所属的VNI（VNI 5000），并根据MAC表项对报文进行封装。可以看到，这里封装的外层源IP地址为本地VTEP的IP地址（IP_1），外层目的IP地址为对端VTEP的IP地址（IP_3）；外层源MAC地址为本地VTEP的MAC地址（MAC_1），而外层目的MAC地址为去往目的IP的网络中下一跳设备的MAC地址。

   封装后的报文，根据外层MAC和IP信息，在IP网络中进行传输，直至到达对端VTEP。
3. 报文进入VTEP_3，VTEP_3对报文进行解封装，得到VM_A发送的原始报文。然后，VTEP_3会对报文做如下处理：
   - VTEP_3发现该报文的目的MAC为本机BDIF 10接口的MAC，而目的IP地址为IP_B（10.1.20.1），所以会根据路由表查找到IP_B的下一跳。
   - 发现下一跳为10.1.20.10，出接口为BDIF 20。此时VTEP_3查询ARP表项，并将原始报文的源MAC修改为BDIF 20接口的MAC（MAC_20），将目的MAC修改为VM_B的MAC（MAC_B）。
   - 报文到BDIF 20接口时，识别到需要进入VXLAN隧道（VNI 6000），所以根据MAC表对报文进行封装。这里封装的外层源IP地址为本地VTEP的IP地址

（IP_3），外层目的IP地址为对端VTEP的IP地址（IP_2）；外层源MAC地址为本地VTEP的MAC地址（MAC_3），而外层目的MAC地址为去往目的IP的网络中下一跳设备的MAC地址。

封装后的报文，根据外层MAC和IP信息，在IP网络中进行传输，直至到达对端VTEP。

4. 报文到达VTEP_2后，VTEP_2对报文进行解封装，得到内层的数据报文，并将其发送给VM_B。

VM_B回应VM_A的流程与上述过程类似，本文就不再赘述啦！

说明

VXLAN网络与非VXLAN网络之间的互通，也需要借助于三层网关，其实现与**图6-5**的不同点在于：报文在VXLAN网络侧会进行封装，而在非VXLAN网络侧不需要进行封装。报文从VXLAN侧进入网关并解封装后，就按照普通的单播报文发送方式进行转发。

7 如何配置 VXLAN

在华为CloudEngine交换机上配置VXLAN的命令步骤、参数说明、注意事项以及配置举例，请**进入华为技术支持网站**，选择相应的CloudEngine交换机款型后，打开华为CloudEngine交换机的产品文档，进入“**配置 > VXLAN配置指南**”节点继续阅读。

8 什么是 BGP EVPN

在VXLAN网络中，我们还常常遇到BGP EVPN协议，因为BGP EVPN协议可以用来作为VXLAN网络的控制面协议，可以帮助VTEP之间实现主机IP、MAC地址的学习，抑制数据平面的泛洪等，在实际应用中十分广泛。关于BGP EVPN在VXLAN网络中的工作原理和关键配置介绍，请参见**《什么是EVPN》**。

