【华为悦读汇】技术发烧友：认识VXLAN-交换机-华为企业互动社区

1 背景

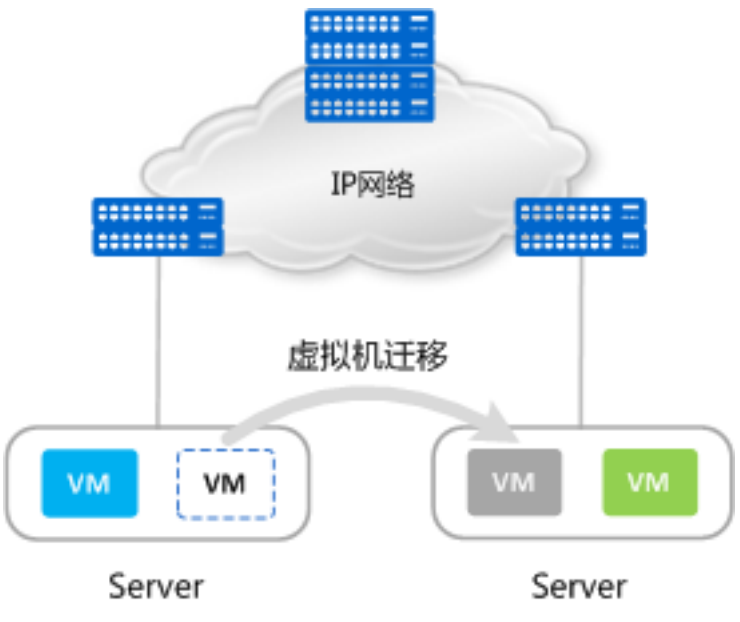
1.1 云计算成为企业IT建设新形态

任何技术的产生，都有其特定的时代背景与实际需求，VXLAN正是为了解决云计算时代虚拟化中的一系列问题而产生的一项技术。

云计算，凭借其在系统利用率高、人力/管理成本低、灵活性/可扩展性强等方面表现出的优势，已经成为目前企业IT建设的新形态；而在云计算中，大量的采用和部署虚拟化是一个基本的技术模式。

服务器虚拟化技术的广泛部署，极大地增加了数据中心的计算密度；同时，为了实现业务的灵活变更，虚拟机VM（Virtual Machine）需要能够在网络中不受限迁移（如图1-1所示）。实际上，对于数据中心而言，虚拟机迁移已经成为了一个常态性业务。

图1-1 虚拟机迁移



1.2 传统数据中心网络面临的挑战

虚拟机数量的快速增长与虚拟机迁移业务的日趋频繁，给传统的“二层+三层”数据中心网络带来了新的挑战：

l 虚拟机规模受网络设备表项规格的限制

对于同网段主机的通信而言，报文通过查询MAC表进行二层转发。服务器虚拟化后，数据中心中VM的数量比原有的物理机发生了数量级的增长，伴随而来的便是虚拟机网卡MAC地址数量的空前增加。此时，处于接入侧的二层设备表示“我要Hold不住了”！

一般而言，接入侧二层设备的规格较小，MAC地址表项规模已经无法满足快速增长的VM数量。

l 传统网络的隔离能力有限

VLAN作为当前主流的网络隔离技术，在标准定义中只有12比特，也就是说可用的VLAN数量只有4000个左右。对于公有云或其它大型虚拟化云计算服务这种动辄上万甚至更多租户的场景而言，VLAN的隔离能力显然已经力不从心。

l 虚拟机迁移范围受限

虚拟机迁移，顾名思义，就是将虚拟机从一个物理机迁移到另一个物理机，但是要求在迁移过程中业务不能中断。要做到这一点，需要保证虚拟机迁移前后，其IP地址、MAC地址等参数维持不变。这就决定了，虚拟机迁移必须发生在一个二层域中。而传统数据中心网络的二层域，将虚拟机迁移限制在了一个较小的局部范围内。

值得一提的是，通过堆叠、SVF、TRILL等技术构建物理上的大二层网络，可以将虚拟机迁移的范围扩大。但是，构建物理上的大二层，难免需要对原来的网络做大的改动，并且大二层网络的范围依然会受到种种条件的限制。

2 VXLAN粉墨登场

传统数据中心网络的种种限制，推动了新技术的产生。于是，
在VMware、Cisco等全球知名厂商的共同推动下，VXLAN粉墨登场。

2.1 VXLAN是什么

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，采用L2 over L4（MAC-in-UDP）的报文封装模式，将二层报文用三层协议进行封装，可实现二层网络在三层范围内进行扩展，同时满足数据中心大二层虚拟迁移和多租户的需求。

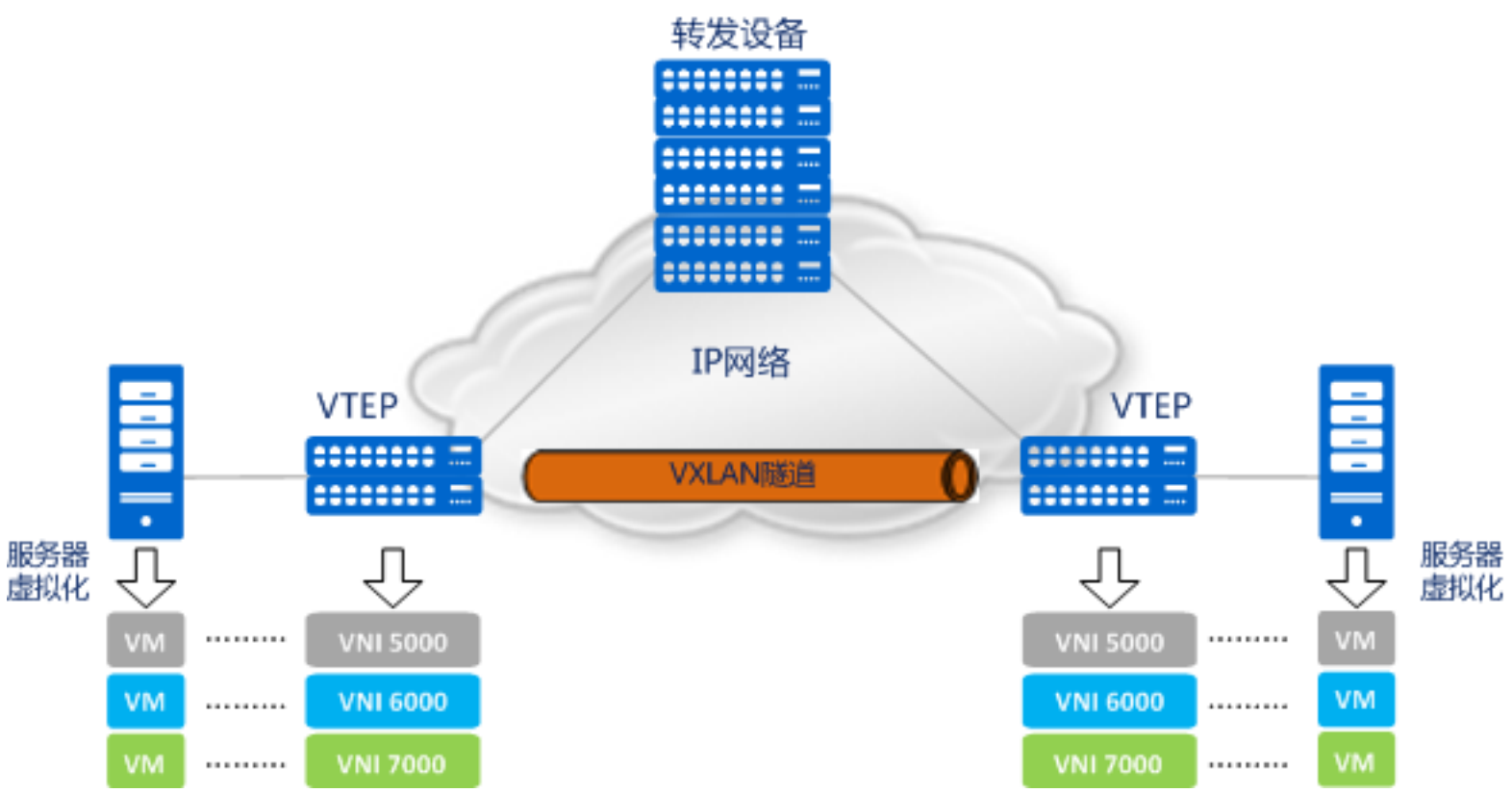
说明

NVO3是基于三层IP overlay网络构建虚拟网络的技术的统称，VXLAN只是NVO3技术之一。除此之外，比较有代表性的还有NVGRE、STT。

在回答VXLAN如何解决前面提到的问题之前，先让我们来了解下VXLAN的网络模型。

2.2 VXLAN网络模型

图2-1 VXLAN网络模型



从上图中可以发现，VXLAN网络中出现了以下传统数据中心网络中没有的新元素：

l VTEP（VXLAN Tunnel Endpoints，VXLAN隧道端点）

VXLAN网络的边缘设备，是VXLAN隧道的起点和终点，VXLAN报文的相关

处理均在这上面进行。总之，它是VXLAN网络中绝对的主角。VTEP既可以是一***立的网络设备（比如华为的CE系列交换机），也可以是虚拟机所在的服务器。那它究竟是如何发挥作用的呢？答案稍候揭晓。

l VNI（VXLAN Network Identifier，VXLAN 网络标识符）

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。VNI是一种类似于VLAN ID的用户标示，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。VXLAN报文封装时，给VNI分配了足够的空间使其可以支持海量租户的隔离。详细的实现，我们将在后文中介绍。

l VXLAN隧道

“隧道”是一个逻辑上的概念，它并不新鲜，比如大家熟悉的GRE。说白了就是将原始报文“变身”下，加以“包装”，好让它可以在承载网络（比如IP网络）上传输。从主机的角度看，就好像原始报文的起点和终点之间，有一条直通的链路一样。而这个看起来直通的链路，就是“隧道”。顾名思义，“VXLAN隧道”便是用来传输经过VXLAN封装的报文的，它是建立在两个VTEP之间的一条虚拟通道。

2.3 见招拆招

看到这里，爱思考的你肯定又要问了，VXLAN网络模型为什么是长这个样子滴？前文已经讲到，VXLAN是为了解决云计算时代虚拟化中的一系列问题而产生的一项技术。下面就让我们来看下，基于图2-1的网络模型，VXLAN是如何见招拆招来解决这一系列问题的。

l 招式一：隐形

对于“虚拟机规模受网络设备表项规格的限制”这个问题，可能有人会想：换成规格大一些的接入交换机（比如跟核心或网关同档次的设备）不就行了。我只能说，如果你是“壕”，确实可以这么做。但是在不提高网络建设成本的前提下，如何能解决问题呢？

既然无法提升设备表项规格，那就只能限制设备上的MAC表项，将大量

VM的MAC地址“隐形”。那么，如何做到隐形呢？这时，就该VTEP出场了。

VTEP会将VM发出的原始报文封装成一个新的UDP报文，并使用物理网络的IP和MAC地址作为外层头，对网络中的其他设备只表现为封装后的参数。也就是说，网络中的其他设备看不到VM发送的原始报文。

如果服务器作为VTEP，那从服务器发送到接入设备的报文便是经过封装后的报文，这样，接入设备就不需要学习VM的MAC地址了，它只需要根据外层封装的报文头负责基本的三层转发就可以了。因此，虚拟机规模就不会受网络设备表项规格的限制了。

当然，如果网络设备作为VTEP，它还是需要学习VM的MAC地址。但是，从对报文进行封装的角度来说，网络设备的性能还是要比服务器强很多。

l 招式二：扩容

对于“传统网络的隔离能力有限”这个问题，VXLAN采用了“扩容”的解决方法，引入了类似VLAN ID的用户标示，也就是前文提到的VNI。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。VTEP在对报文进行VXLAN封装时，给VNI分配了24比特的空间，这就意味着VXLAN网络理论上支持多达16M（即：2^{24}-1）的租户隔离。相比VLAN，VNI的隔离能力得到了巨大的提升，有效得解决了云计算中海量租户隔离的问题。

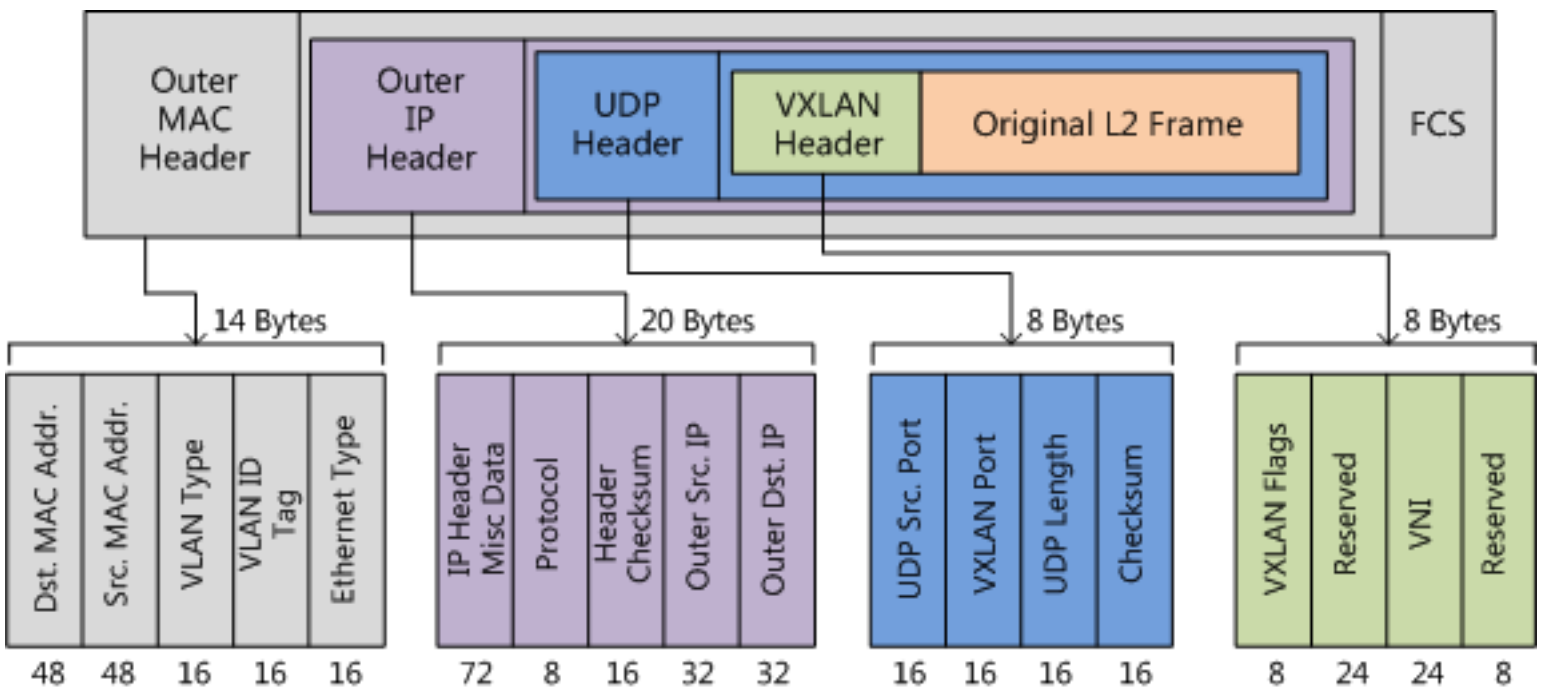
l 招式三：暗度陈仓

前面提到，为了保证业务不中断，VM的迁移就必须发生在同一个二层域内。现在，再回头看下VXLAN网络模型，你是不是惊奇地发现，有了VTEP的封装机制和VXLAN隧道后，所谓的“二层域”就可以轻而易举的突破物理上的界限？也就是说，在IP网络中，“明”里传输的是跨越三层网络的UDP报文，“暗”里却已经悄悄将源VM的原始报文送达目的VM。就好像在三层的网络之上，构建出了一个虚拟的二层网络，而且只要IP网络路由可达，这个虚拟的二层网络想做多大就做多大。现在，你应该明白为什么说VXLAN是一种NVO3技术了吧。

2.4 VXLAN报文长啥样

看过上面的描述，你一定对于封装后的VXLAN报文有了自己的想象。下面就让我们来看下，VXLAN报文到底长啥样。

图2-2 VXLAN报文格式



如你所料，VTEP对VM发送的原始以太帧（Original L2 Frame）进行了以下“包装”：

l 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）。

l 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。

l 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。

l 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的路径上下一跳设备的MAC地址。

2.5 本章小结

本章中，我们介绍了VXLAN的概念、VXLAN网络模型及VXLAN报文的封装格式，了解了VXLAN技术是如何见招拆招解决云计算时代虚拟化中的一系列问题的。看到这里，相信你对于VXLAN已经有了初步的了解。

有了以上的理论基础，想必你一定迫不及待的想进一步了解VXLAN的控制面及转发面的工作原理，比如：

- VTEP如何确定跟谁建立VXLAN隧道?
- VXLAN隧道怎么建立起来的?
- 原始报文如何知道要进入哪条隧道呢?
- VTEP是如何对报文进行封装的呢?

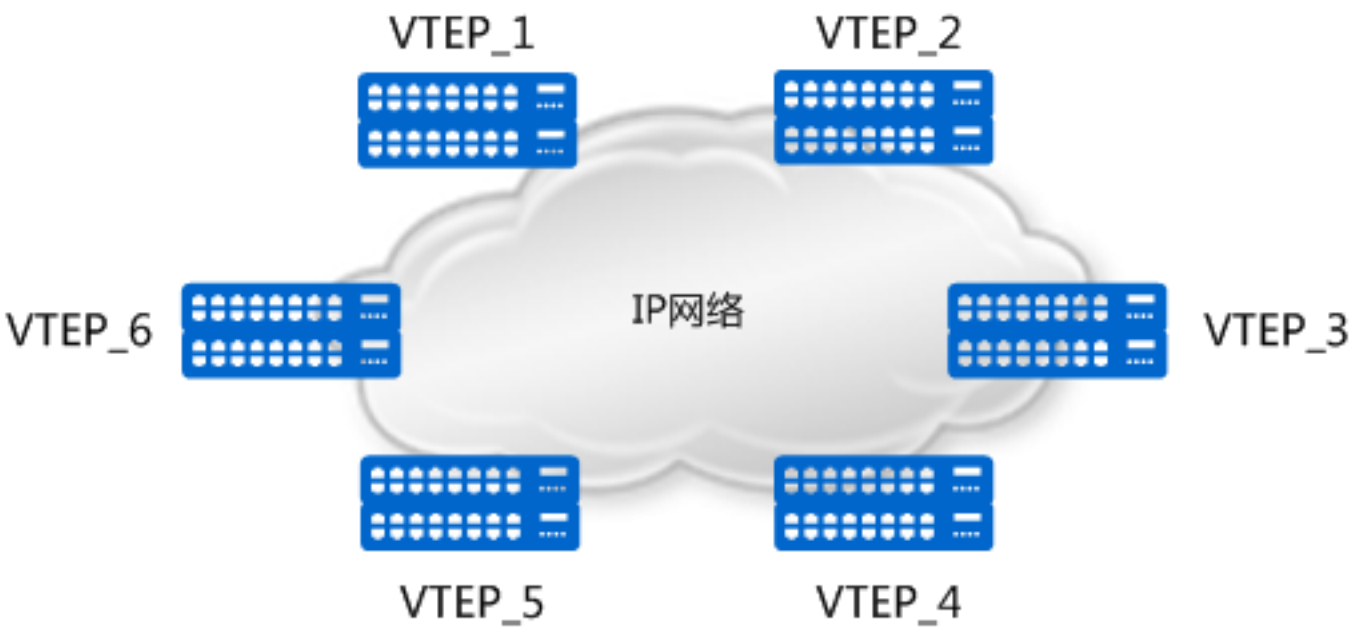
下面我们就以CE系列交换机的实现为例，逐一解答你的疑惑。

3 VXLAN报文转发机制

3.1 建立VXLAN隧道

3.1.1 哪些VTEP间需要建立VXLAN隧道

图3-1 建立VXLAN隧道示意图（1）

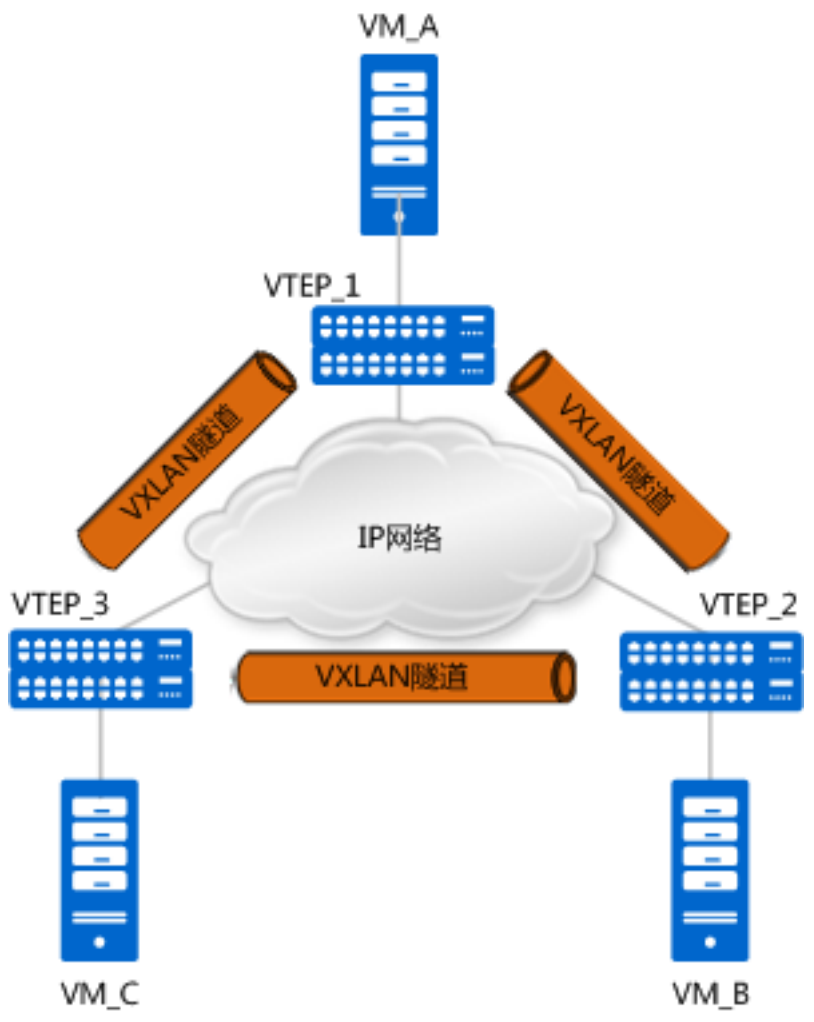


如图3-1所示，网络中存在多个VTEP，那么这其中哪些VTEP间需要建立VXLAN隧道呢?

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3-1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3-2所示。

图3-2 建立VXLAN隧道示意图（2）



3.1.2 什么是“同一大二层域”

前面提到的“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CE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上配置命令行建立起来的。配置如下：

#

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来确定报文封装时该添加哪个VNI。那么，报文根据什么来确定自己属于哪个BD呢?

3.1.3 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”。

3.1.3.1 哪些报文要进入VXLAN隧道

回答这个问题之前，不妨先让我们想下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。我们知道，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，但他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***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CE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***怎样的处理。下面我们就来看下，二层子接口是如何完成这两件事的。

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）。CE系列交换机目前支持三种不同的流封装类型，分别是dot1q、untag和default，它们各自对报文的处理方式如表3-1所示。有了这张表，你就能明白哪些报文要进VXLAN隧道了。

表3-1 不同流封装类型的接口对报文的处理方式

流封装类型	允许进入VXLAN隧道的报文类型	报文进行封装前的处理	收到VXLAN报文并解封装后的处理
dot1q	只允许携带指定VLAN Tag的报文进入VXLAN隧道。 （这里的“指定VLAN Tag”是通过命令进行配置的）	进行VXLAN封装前，先剥掉原始报文的外层VLAN Tag。	进行VXLAN解封装后： 若内层原始报文带有VLAN Tag，则先将该VLAN Tag替换为指定的VLAN Tag，再转发； 若内层原始报文不带VLAN Tag，则先将其添加指定的VLAN Tag，再转发。
untag	只允许不携带VLAN Tag的报文进入VXLAN隧道。	进行VXLAN封装前，不对原始报***处理，即不添加任何VLAN Tag。	进行VXLAN解封装后，不对原始报***处理，即不添加/不替换/不剥掉任何VLAN Tag。

default	允许所有报文进入VXLAN隧道，不论报文是否携带VLAN Tag。	进行VXLAN封装前，不对原始报***处理，即不添加/不替换/不剥掉任何VLAN Tag。	进行VXLAN解封装后，不对原始报***处理，即不添加/不替换/不剥掉任何VLAN Tag。

说明

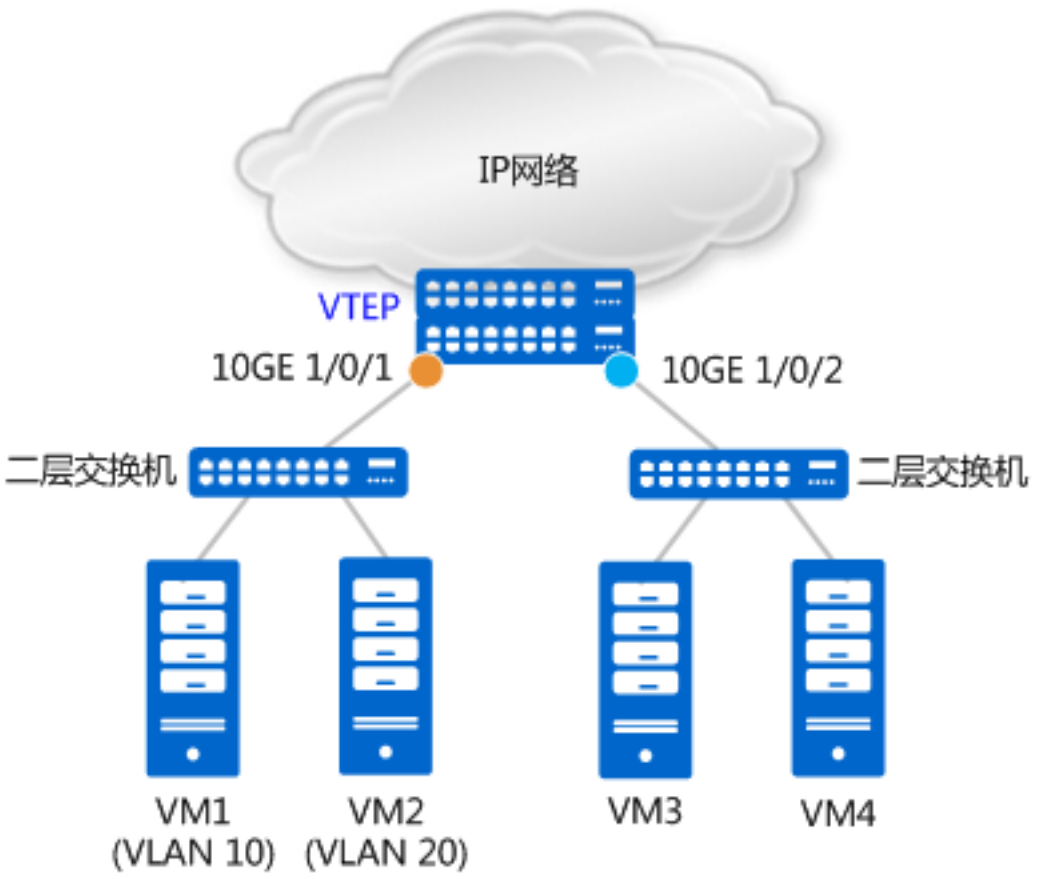
VXLAN隧道两端二层子接口的配置并不一定是完全对等的。正因为这样，才可能实现属于同一网段但是不同VLAN的两个VM通过VXLAN隧道进行通信。

3.1.3.2 二层子接口加入BD

看了上面的描述，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，就可以确定报文属于哪个BD啦！

比如图3-3所示的组网，我们可以分别在VTEP的两个物理接口10GE 1/0/1和10GE 1/0/2上配置不同流封装类型的二层子接口并将其分别加入不同的BD。

图3-3 二层子接口加入BD



基于二层物理接口10GE 1/0/1，分别创建二层子接口10GE 1/0/1.1和10GE 1/0/1.2，且分别配置其流封装类型为dot1q和untag。配置如下：

#

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　//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　//报文进入的是BD 20

#

基于二层物理接口10GE 1/0/2，创建二层子接口10GE 1/0/2.1，且流封装类型为default。配置如下：

#

interface 10GE1/0/2.1 mode l2　//创建二层子接口10GE1/0/2.1

encapsulation **default**　//允许所有报文进入VXLAN隧道

bridge-domain 30　//报文进入的是BD 30

#

此时你可能会有这样的疑问，为什么要在10GE 1/0/1上创建两个不同类型的子接口？是否还可以继续在10GE 1/0/1上创建一个default类型的二层子接口？换句话说，用户应该如何选择配置哪种类型的二层子接口？三种类型的二层子接口之间，是否存在配置约束关系?

3.1.3.3 各类型二层子接口的应用场景

我们先来解答下是否可以在10GE 1/0/1上再创建一个default类型的二层子接口。答案是不可以。其实根据表3-1的描述，这一点很容易理解。因为default类型的二层子接口允许所有报文进入VXLAN隧道，而dot1q和untag类型的二层子接口只允许某一类报文进入VXLAN隧道。这就决定了，default类型的二层子接口跟其他两种类型的二层子接口是不可以在同一物理接口上共存的。否则，报文到了接口之后如何判断要进入哪个二层子接口呢。**所以，default类型的子接口，一般应用在经过此接口的报文均需要走同一条VXLAN隧道的场景，即下挂的VM全部属于同一BD。**例如，图3-3中VM3和VM4均属于BD 30，则10GE 1/0/2上就可以创建default类型的二层子接口。

再来看下为什么可以在10GE 1/0/1上分别创建dot1q和untag类型的二层子接口。如图3-3所示，VM1和VM2分别属于VLAN 10和VLAN 20，且分别属于不同的大二层域BD 10和BD 20，显然他们发出的报文要进入不同的VXLAN隧道。如果VM1和VM2发出的报文在到达VTEP的10GE 1/0/1接口时，一个是携带VLAN 10的Tag的，一个是不携带VLAN Tag的（比如二层交换机上行连接VTEP的接口上配置的接口类型是Trunk，允许通过的VLAN为10和20，PVID为VLAN 20），则为了区分两种报文，就必须要在10GE 1/0/1上分别创建dot1q和untag类型的二层子接口。**所以，当经过同一物理接口的报文既有带VLAN Tag的，又有不带VLAN Tag的，并且他们各自要进入不同的VXLAN隧道，则可以在该物理接口上同时创建dot1q和untag类型的二层子接口。**

当然，现网中可能存在各种不同的组网，小编也不可能一一列举出来。所以在实际应用中，请务必根据组网需求，结合表3-1，合理规划二层子接口的流封装类型。

3.1.4 VXLAN隧道怎么建

现在，我们可以来看下VXLAN隧道是怎么建立起来的了。

一般而言，隧道的建立不外乎手工方式和自动方式两种。

l 手工方式

这种方式需要用户手动指定VXLAN隧道的源和目的IP地址分别为本端和对端VTEP的IP地址，也就是人为的在本端VTEP和对端VTEP之间建立静态VXLAN隧道。

对于CE系列交换机，以上配置是在NVE（Network Virtualization Edge）接口下完成的。配置过程如下：

#

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 : 1.1.1.1

UDP Port : 4789

BUM Mode : head-end

Group Address : -

Peer List : 2.2.2.2 2.2.2.3

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。

l 自动方式

自动方式下VXLAN隧道的建立需要借助于其他的协议，例如BGP。CE系列交换机中，自动方式建立VXLAN隧道主要应用在EVN（Ethernet Virtual Network）和VXLAN的分布式网关场景中。本文不对该方式进行详细讲述，具体实现可参考EVN的相关资料。

3.1.5 如何确定报文要进哪条隧道

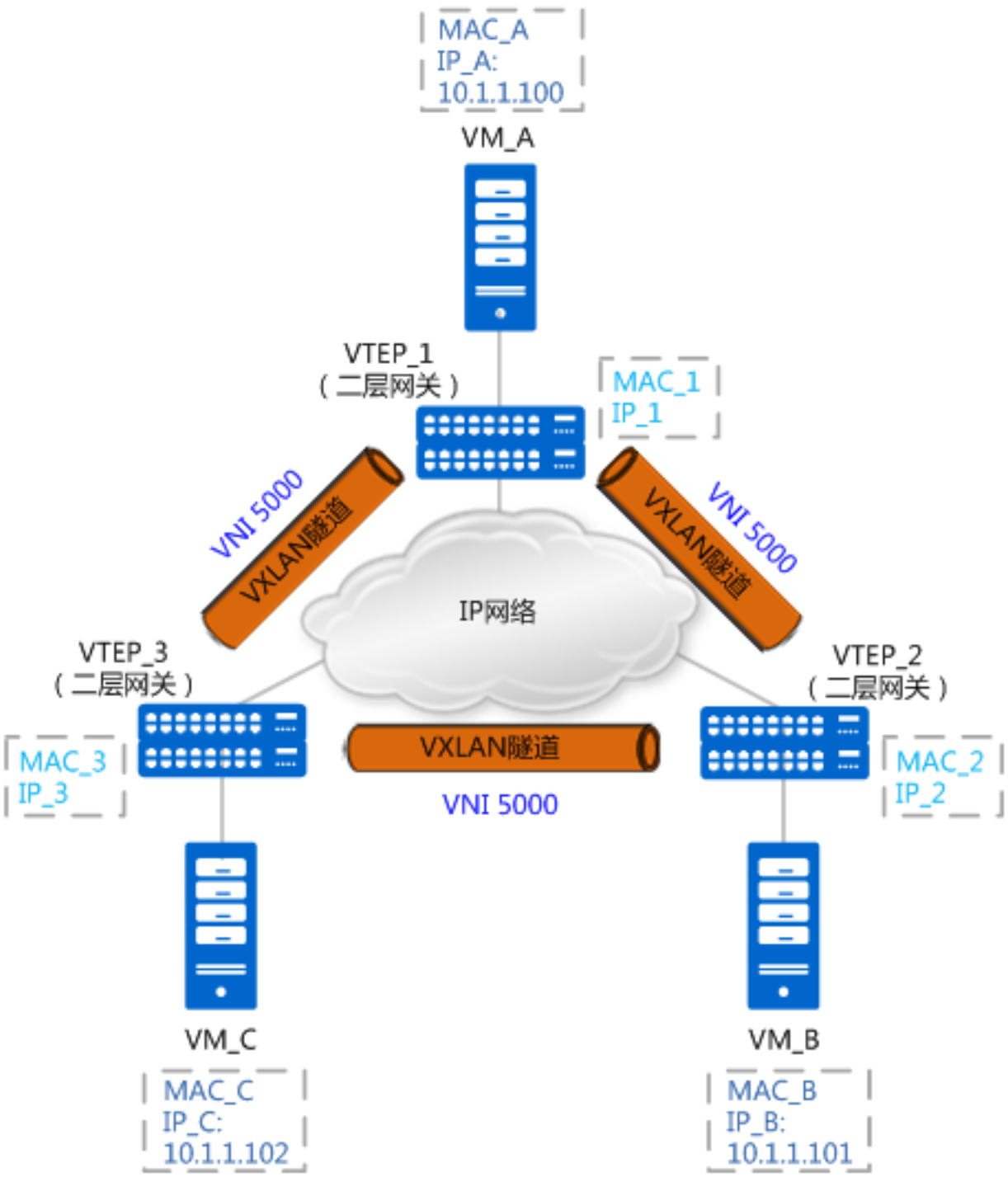
从前面的描述我们知道，属于同一BD的VXLAN隧道可能不止一条，比如前面的头端复制列表中，同一个源端VTEP（1.1.1.1）对应了两个对端VTEP（2.2.2.2和2.2.2.3）。那就带来了另一个问题，报文到底应该走哪一条隧道呢?

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表，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。下面就让我们来看下，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。

3.2 VXLAN网络中报文的转发流程

3.2.1 同子网互通

图3-4 同子网VM互通组网图



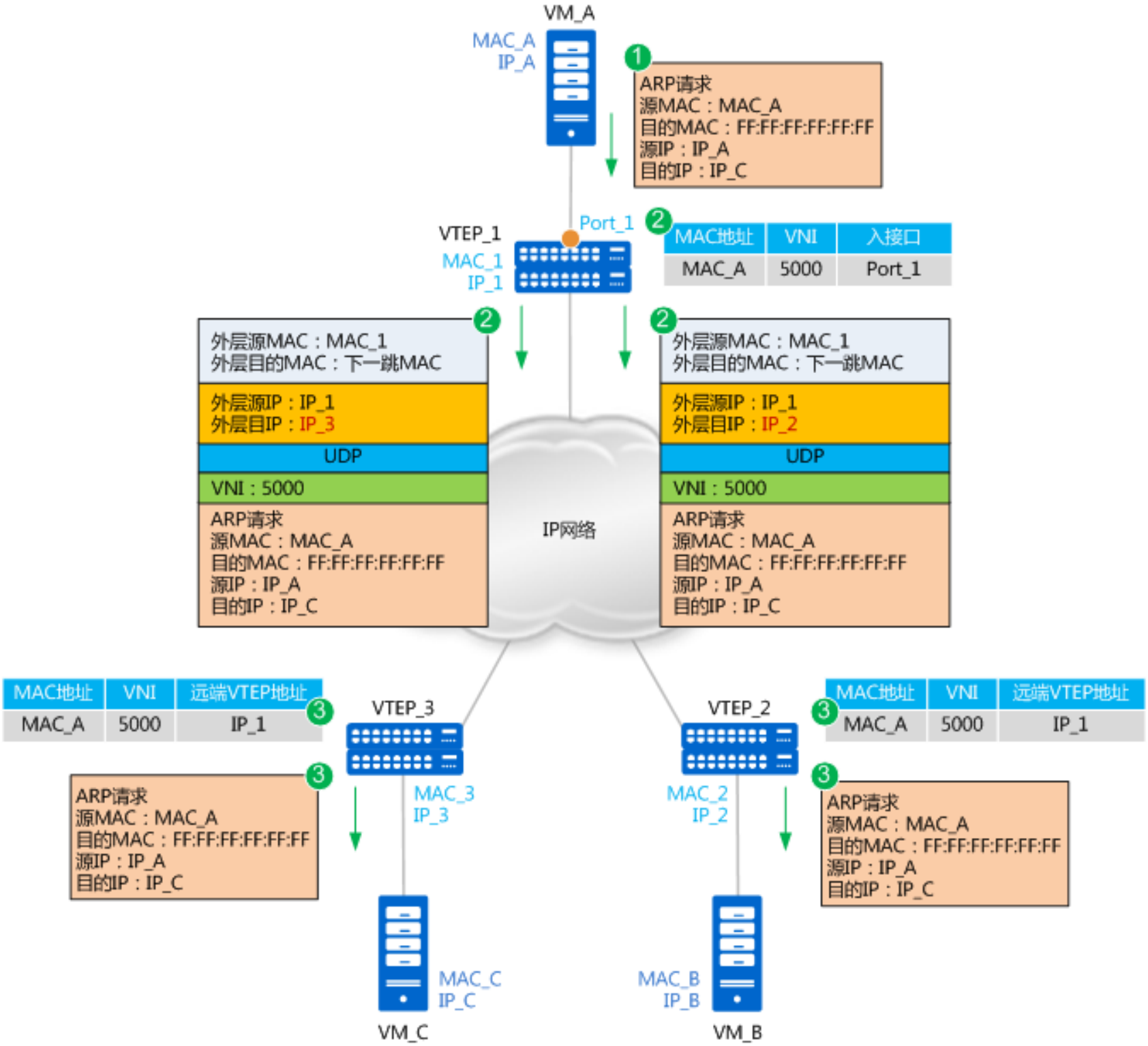
如图3-4所示，VM_A、VM_B和VM_C同属于10.1.1.0/24网段，且同属于VNI 5000。此时，VM_A想与VM_C进行通信。

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。

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。

l ARP请求报文转发流程

图3-5 ARP请求报文转发流程



如图3-5所示，ARP请求报文的转发流程如下：

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。

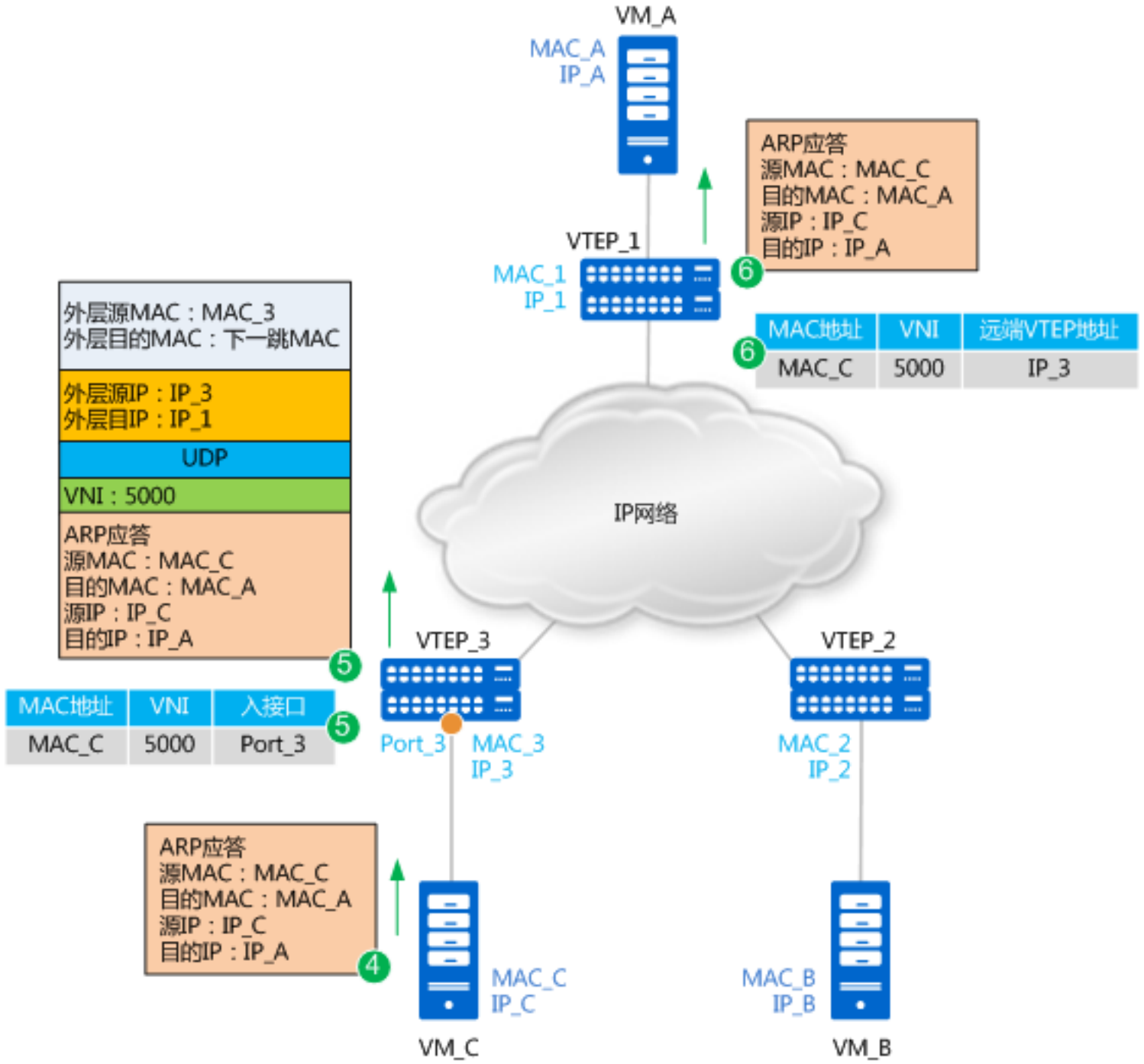
封装后的报文，根据外层MAC和IP信息，在IP网络中进行传输，直至到达对端VTEP。

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。

l ARP应答报文转发流程

图3-6 ARP应答报文转发流程



如图3-6所示，ARP应答报文的转发流程如下：

4 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。

5 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。

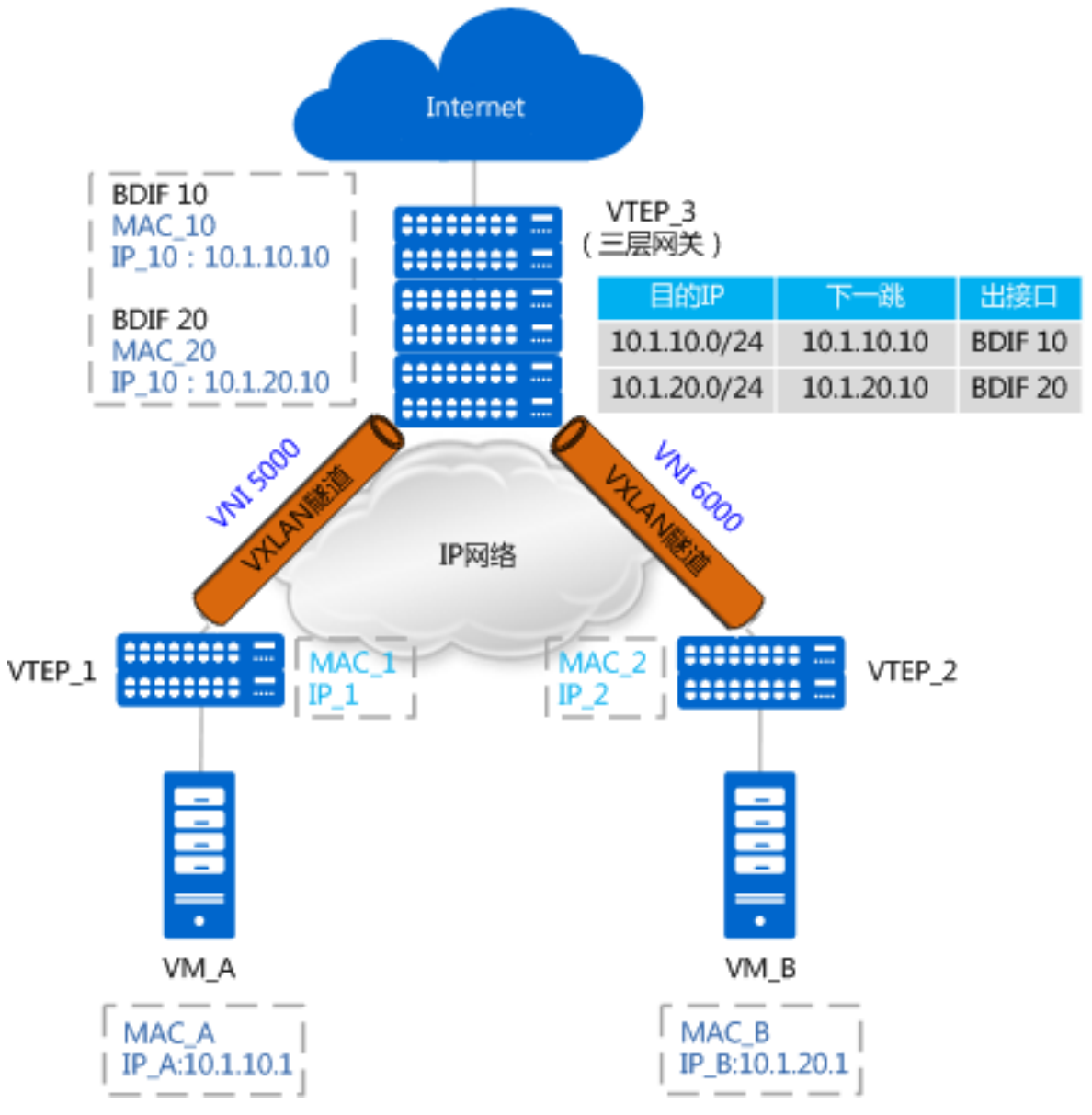
封装后的报文，根据外层MAC和IP信息，在IP网络中进行传输，直至到达对端VTEP。

6 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3-6中所展示的类似，本文就不再赘述啦！

3.2.2 不同子网互通

图3-7 不同子网VM互通组网图



如图3-7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

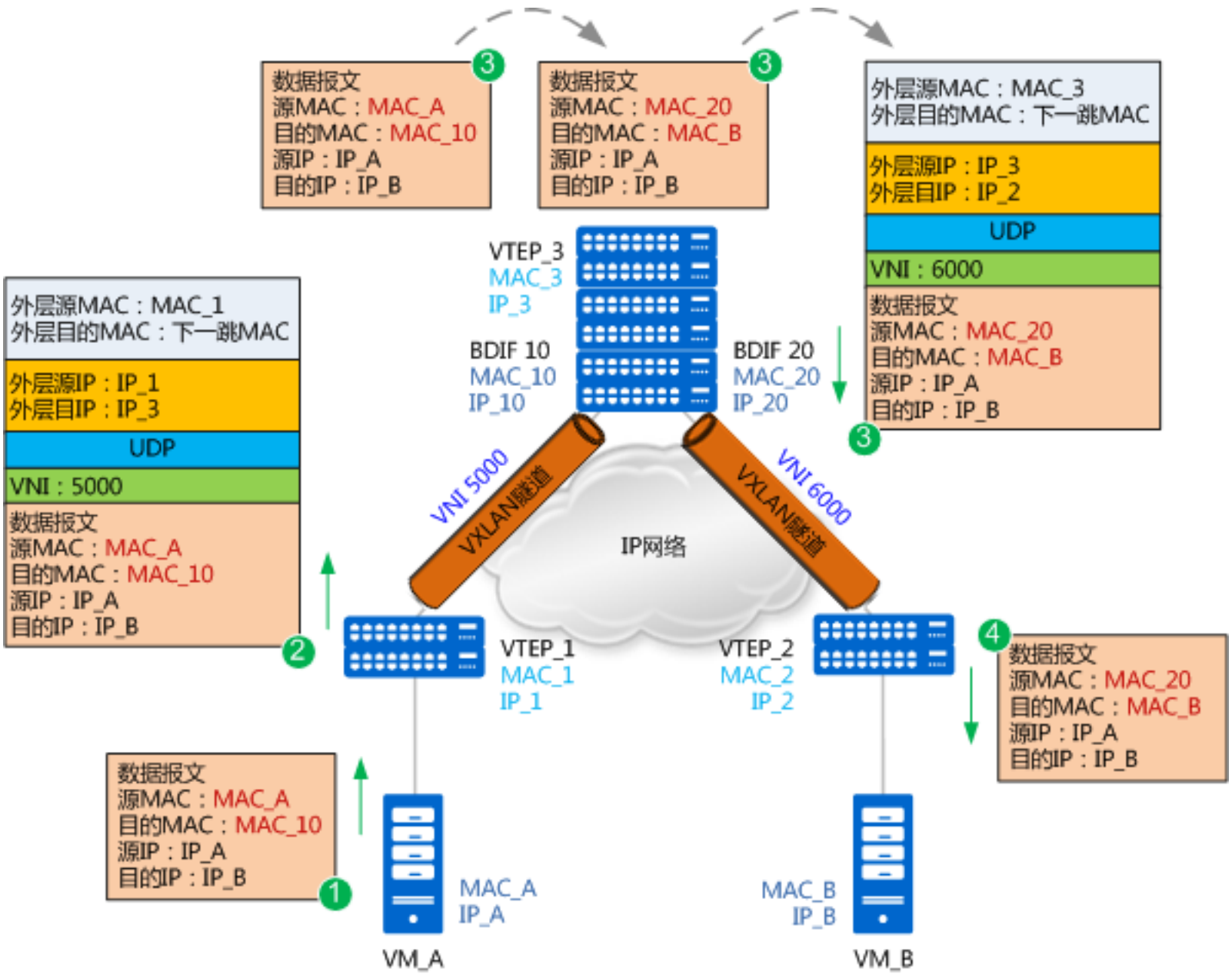
和10.1.20.0/24网段的路由。此时，VM_A想与VM_B进行通信。

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VM之间或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地址学习的过程与同子网互通中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3-8 不同子网VM互通报文转发流程



如图3-8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***如下处理：

l 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。

l 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）。

l 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3-8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

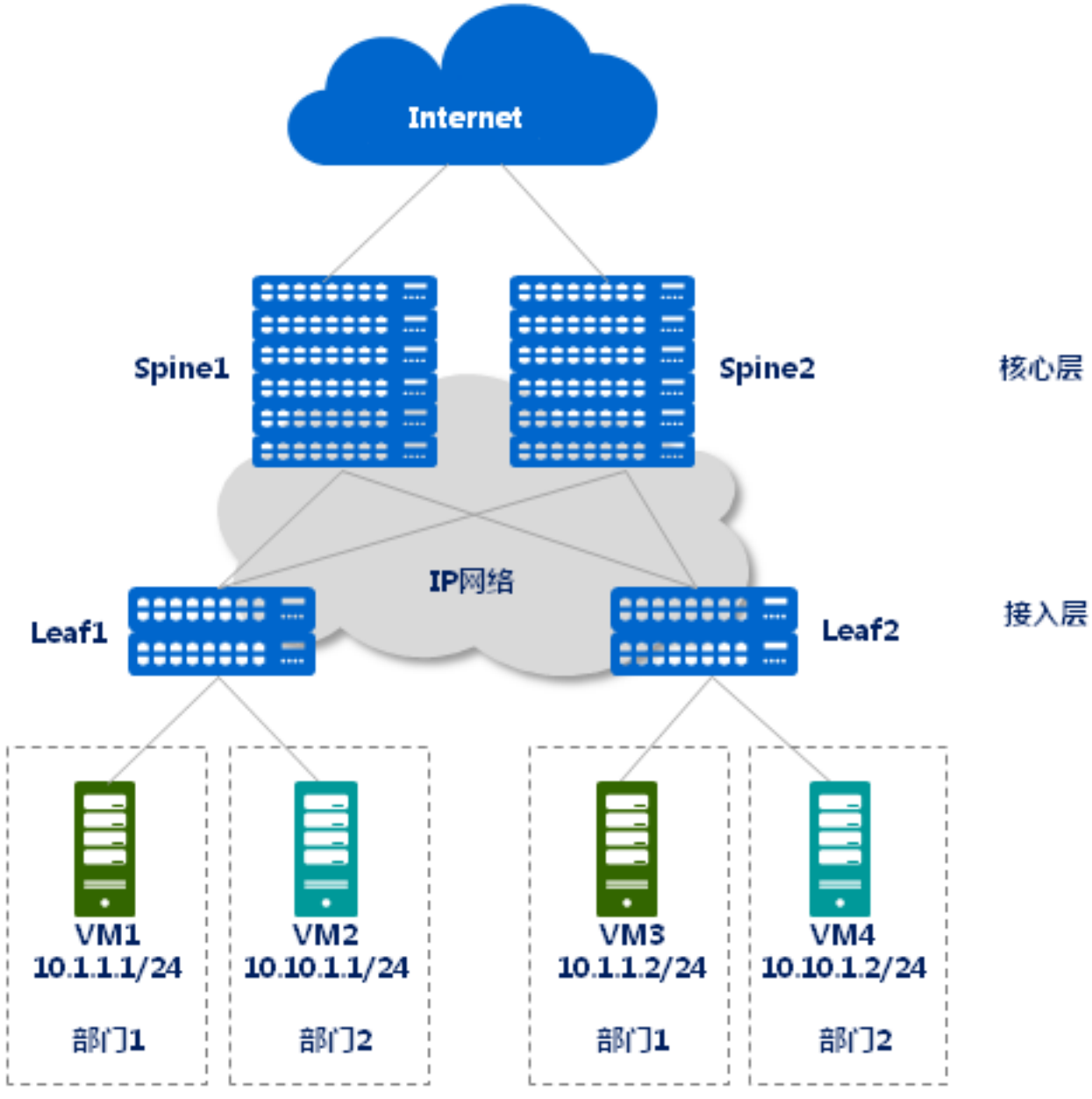
3.3 本章小结

本章主要介绍了VXLAN控制面表项的建立过程及VXLAN网络中报文的转发过程。看到这里，

相信你对于VXLAN已经达到熟悉的阶段了。有了上面的理论基础，接下来，我们可以来了解下VXLAN在现网中是如何部署的了。

4 VXLAN应用部署方式

本篇我们以下图所示的典型的“Spine-Leaf”数据中心组网为例，给大家介绍一下CE系列交换机VXLAN的应用场景和部署方案。



在上图所示的数据中心里，企业用户拥有多个部门（部门1和部门2），每个部门中拥有多个VM（VM1和VM3，VM2和VM4）。同部门的VM属于同一个网段，不同部门的VM属于不同的网段。用户希望同一部门VM之间、不同部门VM之间，VM与Internet之间均可相互访问。

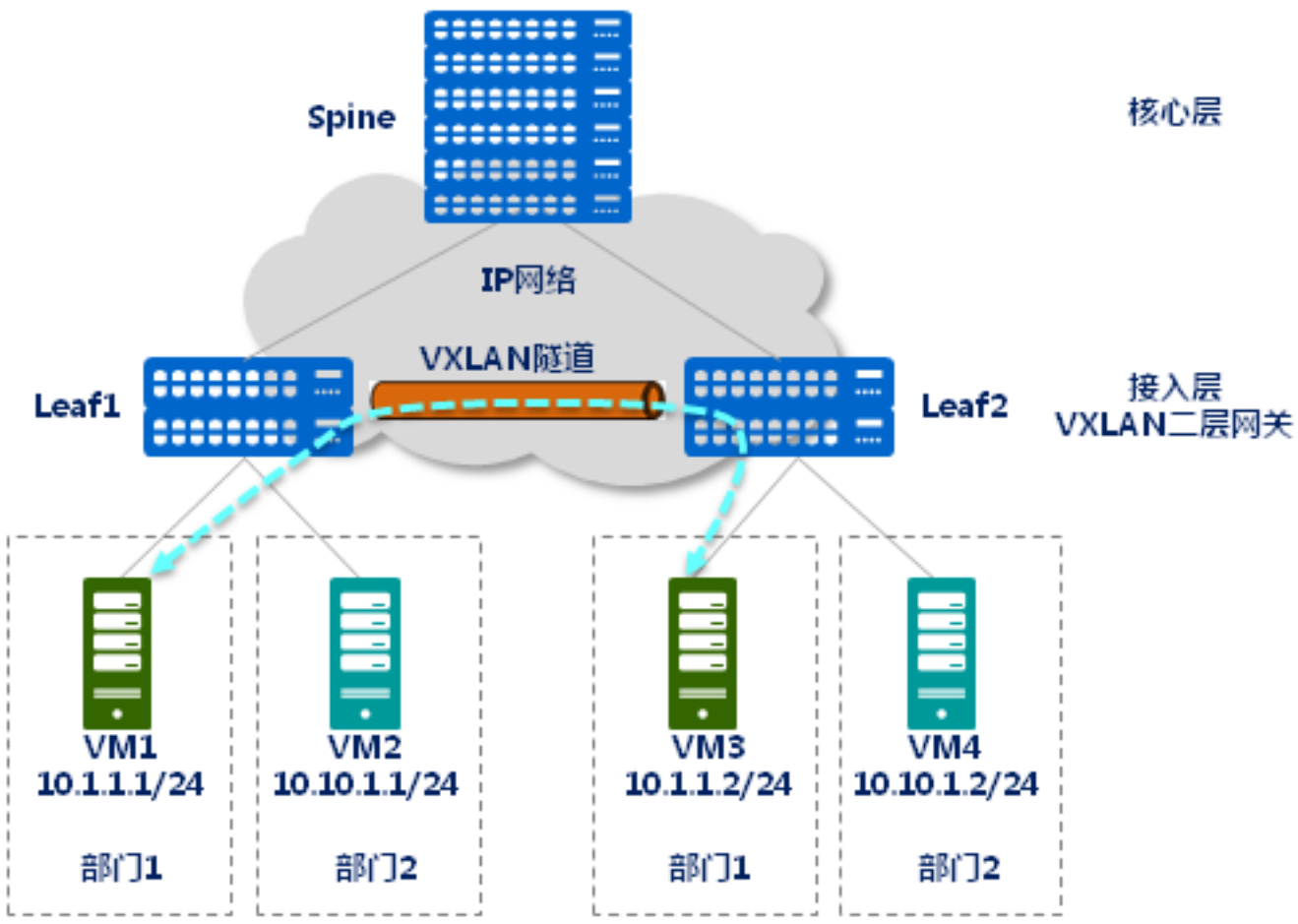
4.1 VXLAN网络的子网互通

4.1.1 相同子网互通

部署方案

如图4-1所示，Leaf1和Leaf2作为VXLAN网络的VTEP，两个Leaf之间搭建VXLAN隧道，并在每个Leaf上部署VXLAN二层网关，即可实现同一部门VM之间的相互通信。此时Spine只作为VXLAN报文的转发节点，不感知VXLAN隧道的存在，可以是任意的三层网络设备。

图4-1 相同子网互通

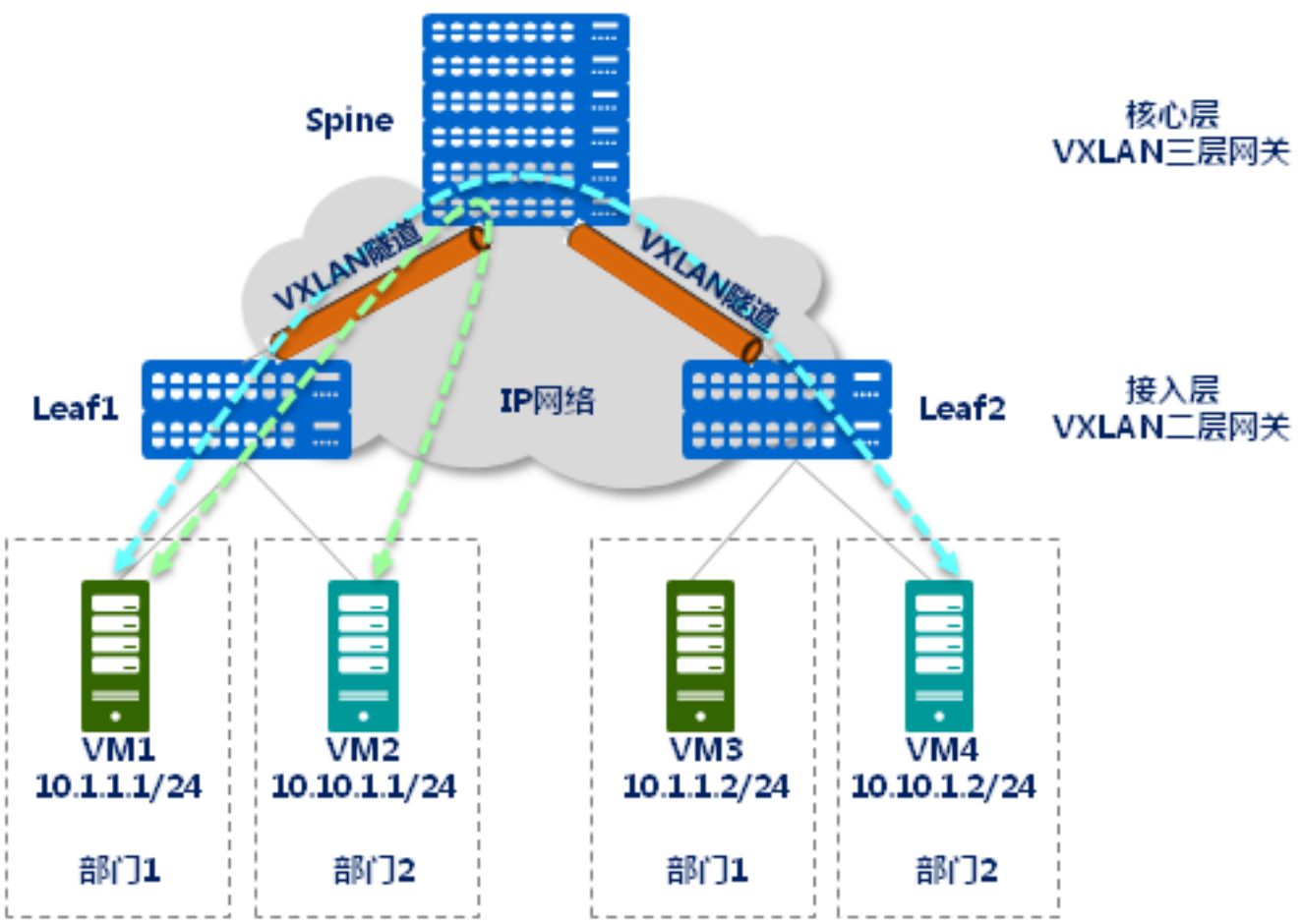


4.1.2 不同子网互通（集中式网关）

部署方案

如图4-2所示，Leaf1、Leaf2和Spine作为VXLAN网络的VTEP，Leaf1和Spine之间、Leaf2和Spine之间分别搭建VXLAN隧道，并在Spine上部署VXLAN三层网关，即可实现不同部门VM之间的相互通信。

图4-2 不同子网互通（集中式网关）



4.1.3 不同子网互通（分布式网关）

出现背景

细心的读者可能已经发现，在不同子网互通（集中式网关）中，同一Leaf（Leaf1）下挂的不同网段VM（VM1和VM2）之间的通信，都需要在Spine上进行绕行，这样就导致Leaf与Spine之间的链路上，存在冗余的报文，额外占用了大量的带宽。同时，Spine作为VXLAN三层网关时，所有通过三层转发的终端租户的表项都需要在该设备上生成。但是，Spine的表项规格有限，当终端租户的数量越来越多时，容易成为网络瓶颈。

分布式网关的出现，很好的解决了这两个问题。

部署方案

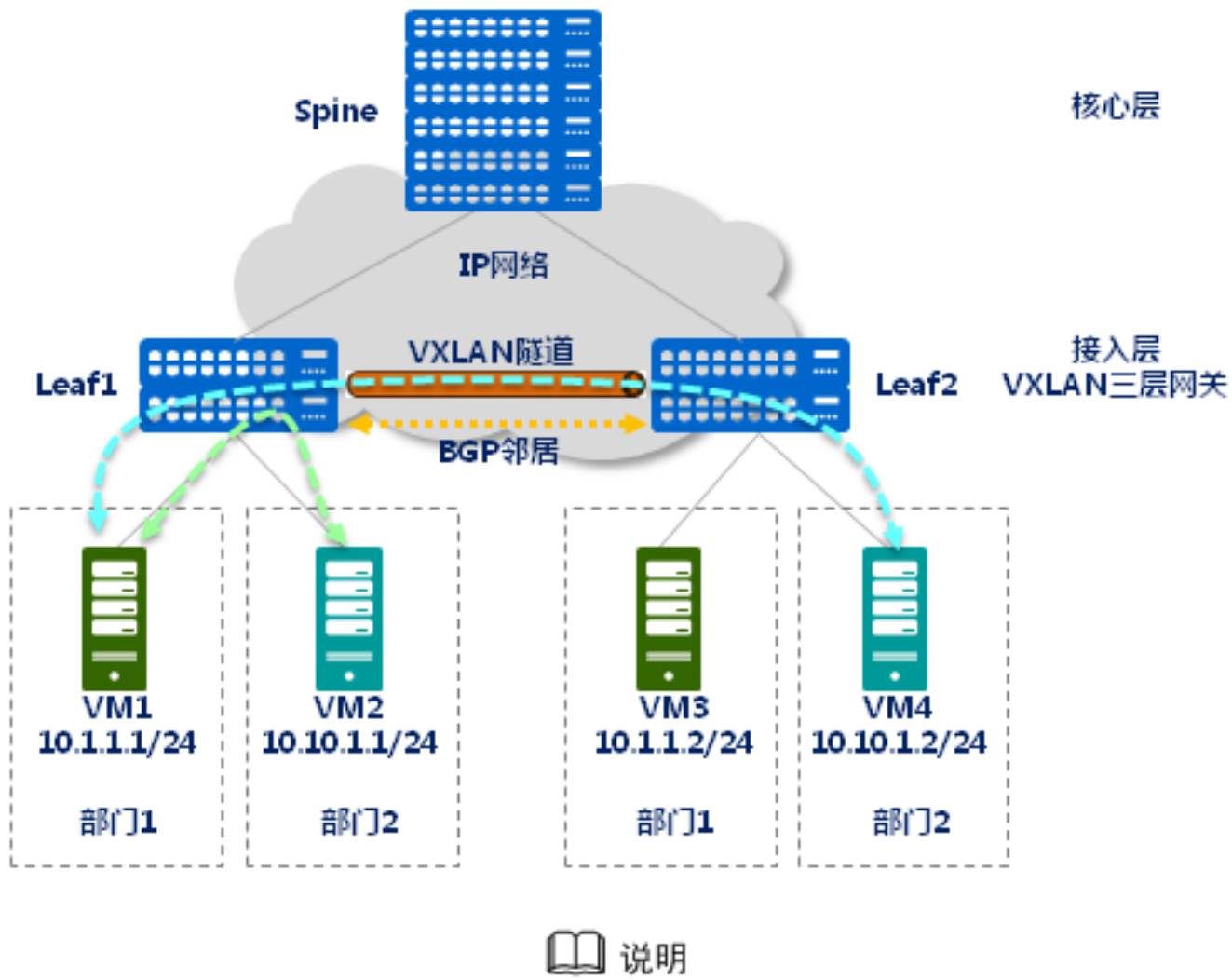
l 同Leaf节点下不同部门VM之间的通信

如图4-3所示，Leaf1作为VXLAN网络的VTEP，在Leaf1上部署VXLAN三层网关，即可实现同Leaf下不同部门VM之间的相互通信。此时，VM1和VM2互访时，流量只需要在Leaf1节点进行转发，不再需要经过Spine节点，从而节约了大量的带宽资源。

l 跨Leaf节点不同部门VM之间的通信

如图4-3所示，Leaf1和Leaf2作为VXLAN网络的VTEP，在Leaf1和Leaf2上部署VXLAN三层网关。两个VXLAN三层网关之间通过BGP动态建立VXLAN隧道，并通过BGP的remote-nexthop属性发布本网关下挂的主机路由信息给其他BGP邻居，从而实现跨Leaf节点不同部门VM之间的相互通信。

图4-3 不同子网互通（分布式网关）



说明

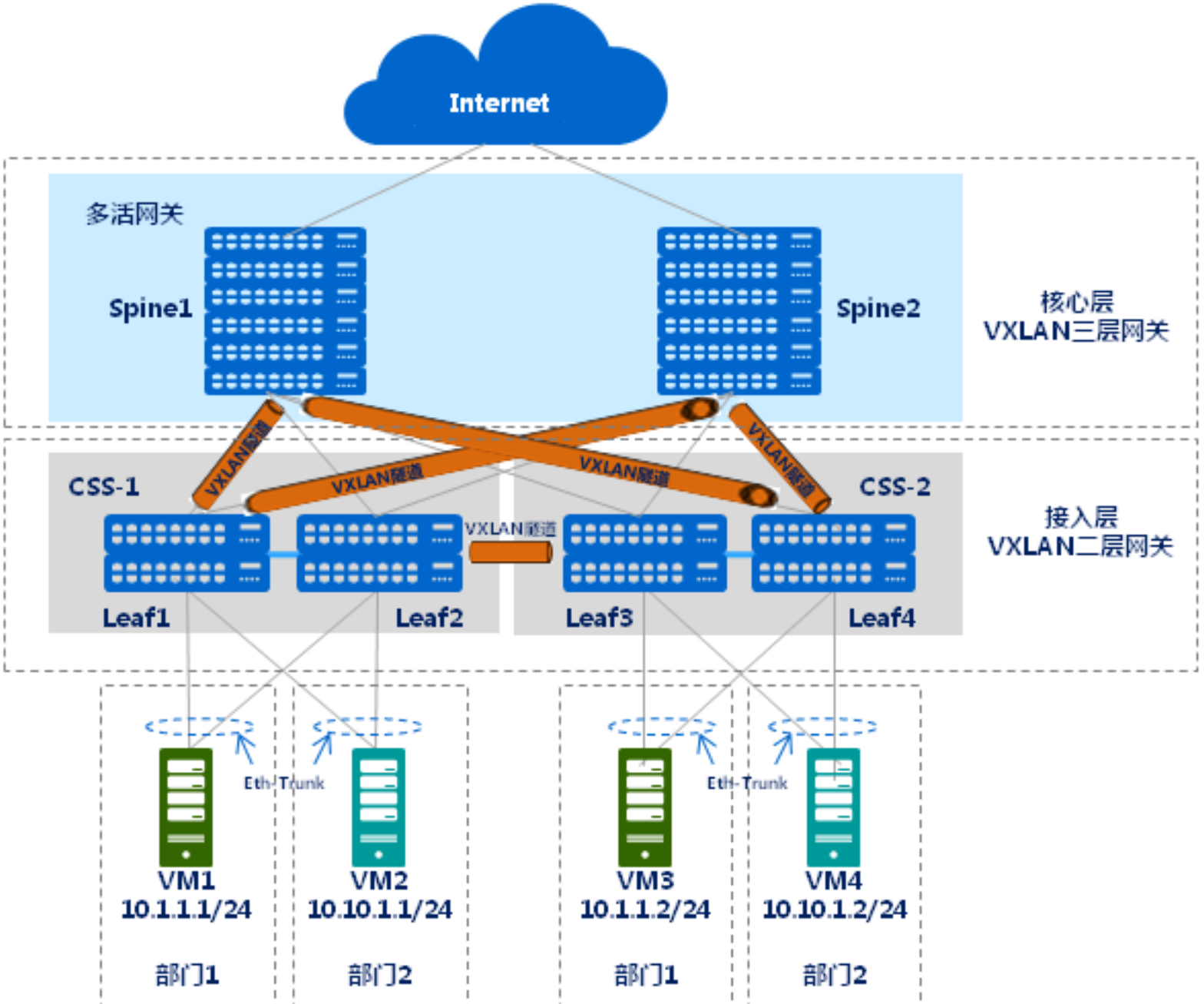
Leaf作为VXLAN三层网关时，只学习其下挂终端租户的表项，而不必像集中式三层网关一样，需要学习网络中所有终端租户的表项，从而解决了集中式三层网关带来表项瓶颈问题。

4.2 VXLAN网络的可靠性

随着网络的快速普及和应用的日益深入，基础网络的可靠性日益成为用户关注的焦点，如何能够保证网络传输不中断对于终端用户而言非常重要。

在VXLAN网络的子网互通中，VM与Leaf之间，Leaf与Spine之间都是通过单归方式接入的。这种组网接入方式，显然已经不能满足用户对VXLAN网络可靠性的需求。

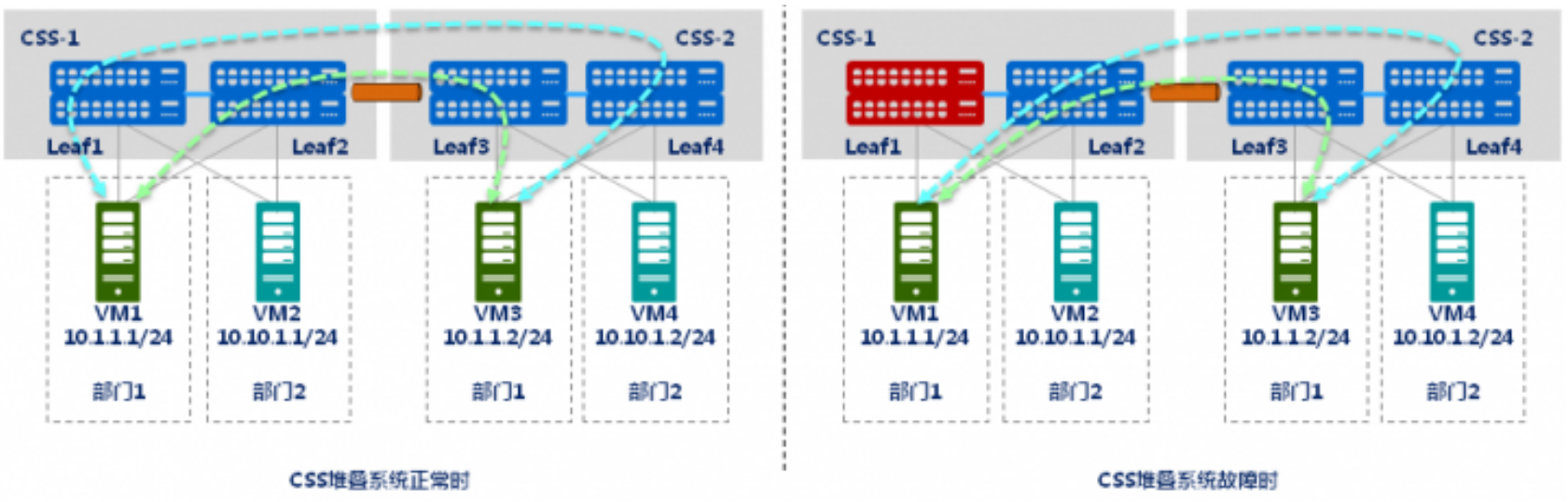
这时，可以按照如下图所示方式，提升VXLAN网络的可靠性。



4.2.1 接入层的可靠性

通常采用堆叠（CSS）方式提升接入层的可靠性。这是因为，接入层的设备数量繁多，堆叠方式可以将多台交换机设备组合在一起，虚拟化成一台交换设备，所有配置均在这一台虚拟交换机上进行，从而简化了接入层设备的运维复杂度。此外，堆叠系统内成员交换机之间在进行冗余备份的同时，能够利用跨设备的Eth-Trunk实现设备间链路的负载分担。

图4-4 接入层的可靠性



如图4-4所示，Leaf1和Leaf2组建为堆叠系统CSS-1，Leaf3和Leaf4组建为堆叠系统CSS-2，VM1~VM4均通过双归的方式接入到各自的CSS系统中。CSS-1和CSS-2作为VXLAN网络的VTEP，两个CSS之间搭建VXLAN隧道，并在每个CSS上部署VXLAN二层网关，从而实现同一部门VM之间的相互通信。

l 当CSS系统正常时，VM1与VM3之间互访的流量，通过CSS-1堆叠系统中的Leaf1和Leaf2进行负载分担转发。

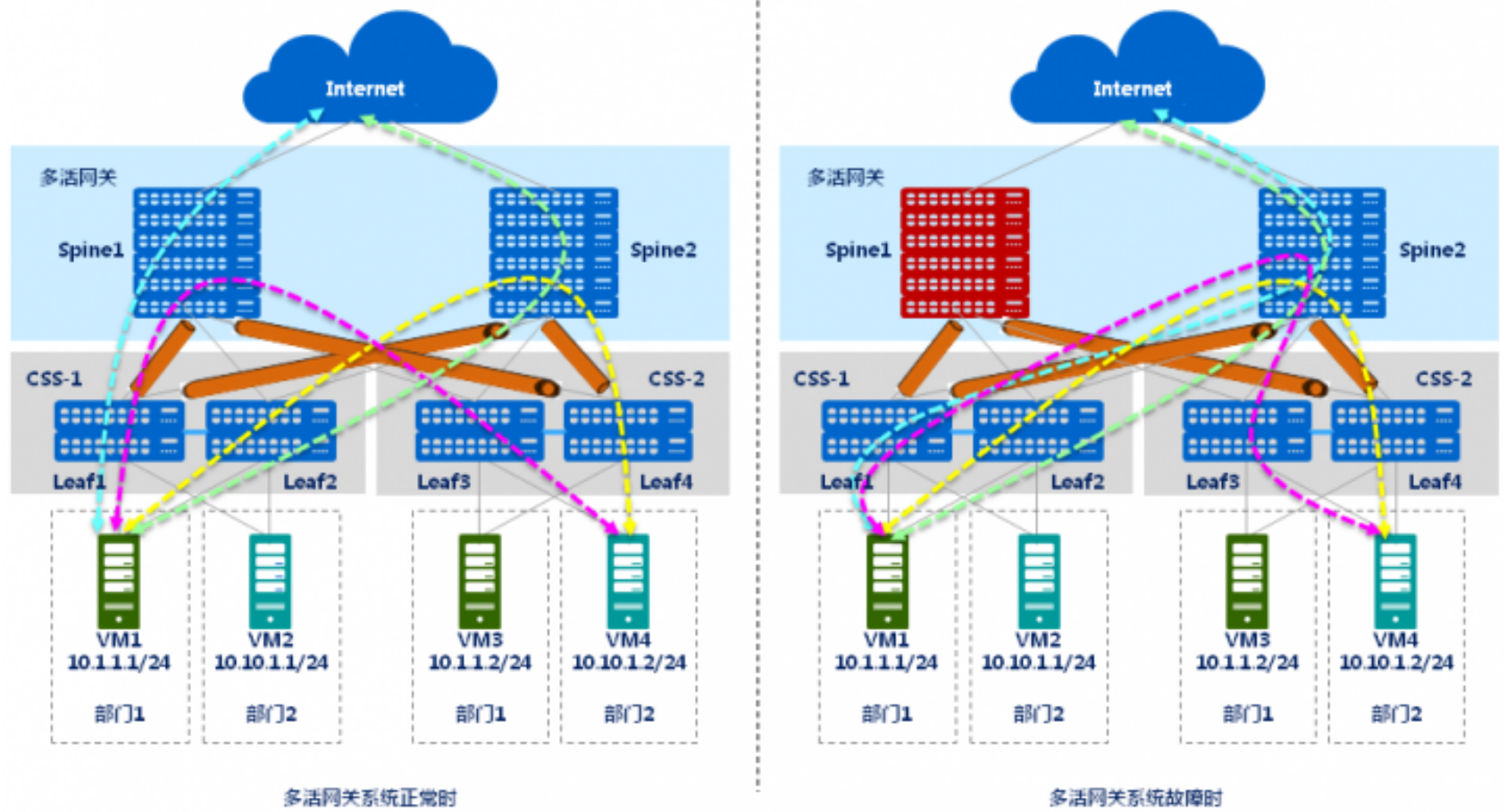
l 当CSS系统故障时（Leaf1故障），VM1与VM3之间互访的流量，全部切换到CSS-1堆叠系统中的Leaf2进行转发，从而实现了流量的不间断，提升了接入层的可靠性。

4.2.2 核心层的可靠性

通常采用多活网关方式提升核心层的可靠性。这是因为，核心层设备物理位置较为分散，传统的设备级备份无法满足要求，必须通过协议级备份来实现。

在多活网关组网中，通过给多台Spine设备部署相同的网关信息，将它们对外模拟成VXLAN网络中的一个虚拟VTEP，然后在所有Spine设备上配置三层网关，使得无论流量发到哪一个Spine，该设备都可以提供服务，将报文正确转发给下一跳设备。此外，多活网关中的多台Spine之间形成负载分担关系，共同进行流量转发。

图4-5 核心层的可靠性



如图4-5所示，Spine1、Spine2分别与接入层的堆叠系统CSS-1和CSS-2之间建立VXLAN隧道，在Spine1、Spine2上配置VXLAN三层网关功能，Spine1、Spine2上部署相同的网关MAC地址、网关IP地址以及源VTEP地址，以便对外模拟成一个虚拟的VTEP，从而实现了不同网段VM之间、VM与外网之间的互通。

l 当多活网关系统正常时，VM1与VM4之间互访的流量、VM1与Internet之间互访的流量，通过Spine1和Spine2进行负载分担转发。

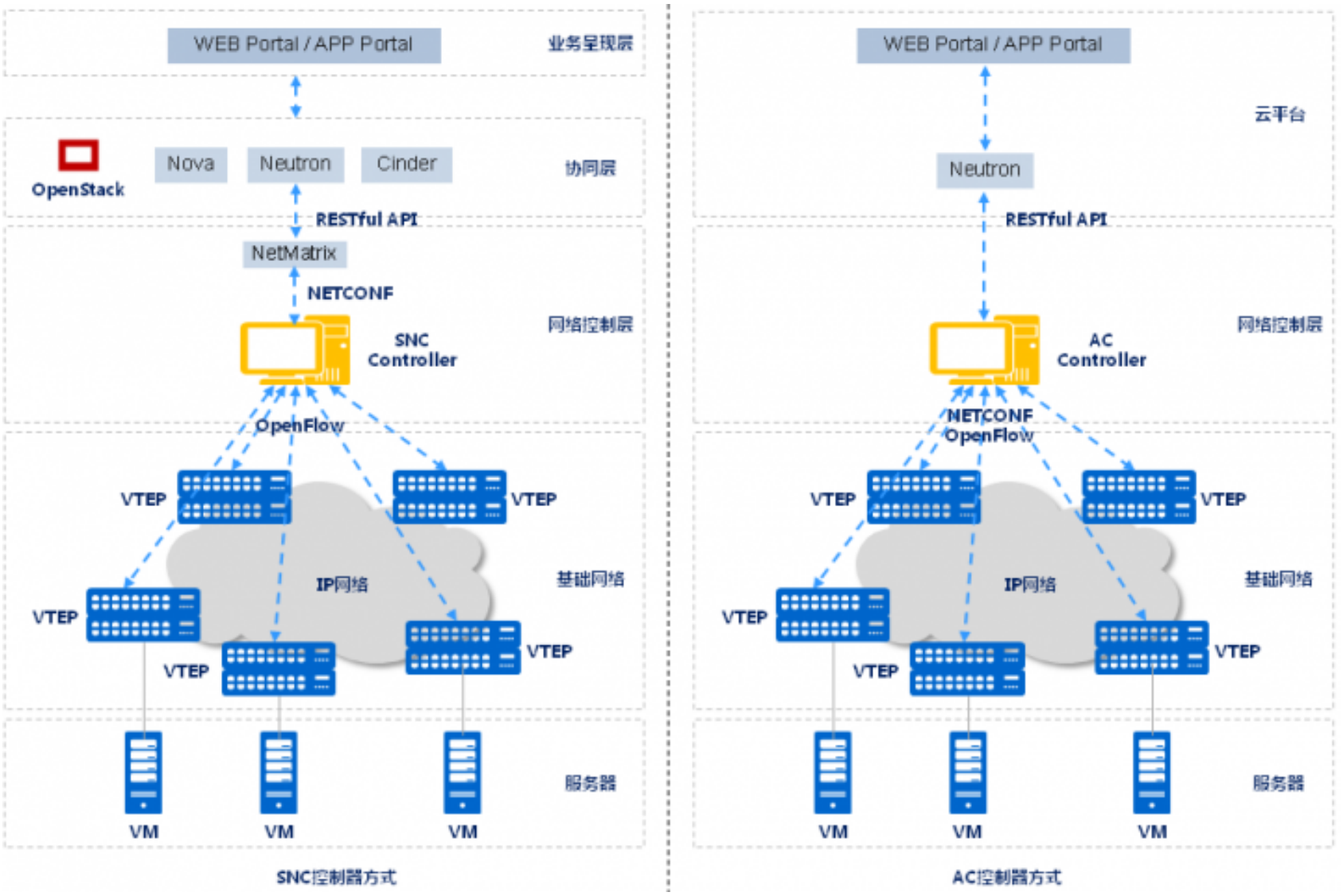
l 当多活网关系统故障时（Spine1故障），VM1与VM4之间互访的流量、VM1与Internet之间互访的流量，全部切换到Spine2进行转发，从而实现了流量的不间断，提升了核心层的可靠性。

4.3 VXLAN网络的部署方案

CE系列交换机支持通过**单机方式**和**控制器方式**来部署VXLAN网络。这两种方式中VXLAN网络的子网互通以及VXLAN网络的可靠性的实现均与前面一致，不同点在于VXLAN的配置下发方式不同：单机方式是通过CLI手动在设备上进行配置，控制器方式是通过控制器向设备下发配置或流表，设备仅作为转发器。

下面小编以图4-6所示组网为例，简单介绍一下当前CE系列交换机支持的VXLAN控制器部署方式：SNC控制器方式和AC控制器方式。

图4-6 控制器部署方案



l SNC控制器方式

SNC控制器方式是指通过SNC控制器动态建立VXLAN隧道。

此方式下，CE系列交换机作为转发器，无需进行VXLAN配置。VXLAN隧道的创建以及指导报文转发的表项，均由SNC控制器通过OpenFlow协议向转发器下发。

l AC控制器方式

AC控制器方式是指通过AC控制器动态建立VXLAN隧道。

此方式下，CE系列交换机作为转发器，需要预先完成部分基础配置（具体配置内容请参考产品配置指南），AC控制器通过NETCONF协议向转发器下发建立VXLAN隧道的配置，通过OpenFlow协议控制报文在隧道中的转发。

5 尾言

本篇内容，我们通过介绍VXLAN出现的时代背景、VXLAN的概念及网络模型、VXLAN报文的封装格式，让你对VXLAN有了初步的了解；通过介绍VXLAN隧道的建立及报文的转发流程，让你熟悉了VXLAN的控制面及转发面的工作机制；通过介绍CE系列交换机VXLAN的应用场景和部署方案，让你进一步了解VXLAN技术在现网中是如何运用的。

总之，VXLAN通过MAC-in-UDP的报文封装，实现了二层报文在三层网络上的透传，在云端上架起了一道道无形的“彩虹”，解决了云计算中虚拟化带来的一系列问题。