OpenStack网络机制实现原理

---

renyl 2015/4/6

目 次

1. 调查背景

本文将针对OpenStack中的Neutron组件所包含的网络类型：FLAT、VLAN、GRE和VXLAN的实现原理进行调查并对其优缺点进行分析。

2. 名词解释

在介绍各网络类型的实现原理之前，需要先了解下网络方面的相关术语及其含义,有如下几个：

1) L2/L3/L4
2) Tap/Tun
3) Bridge
4) Tunnel
5) Promiscuous Mode

2.1.L2/L3/L4

1) L2：是指TCP/IP网络协议中的第二层数据链路层，以Mac地址为基础进行传输。一方面数据链路层接收来自网络层（第三层）的数据帧并为物理层封装这些帧；另一方面数据链路层把来自物理层的原始数据比特封装到网络层的帧中。它起着重要的中介作用。

2) L3：是指TCP/IP网络协议中的第三层网络层，以IP地址为基础进行传输。网络层的主要功能是提供路由，即选择到达目标主机的最佳路径，并沿该路径传送数据包。除此之外，网络层还要能够消除网络拥挤，具有流量控制和拥挤控制的能力。需要注意的是，网络层解决的是网络与网络之间的通信问题，而不是同一网段内部的事。

3) L4:是指TCP/IP网络协议中的第四层传输层，有TCP和UDP两种协议。其中，TCP协议是面向连接的，提供IP环境下的数据可靠传输和差错恢复，其支持的应用层协议主要有Telnet、FTP和SMTP等；UDP协议是面向非连接的，不为IP环境提供可靠性传输，其支持的应用层协议主要有NFS、DNS、TFTP等。

2.2.Tap/Tun

在计算机网络中，Tap/Tun是操作系统内核中的虚拟网络设备，能提供跟硬件实现的网络设备相同的功能。操作系统可以通过Tap/Tun设备向绑定该设备的用户空间的程序发送数据，反之用户空间的程序也可以像操作硬件网络设备那样，通过Tap/Tun设备发送数据。

1) Tap设备等同于一个以太网卡设备，它操作第二层数据包（如以太网数据帧）。Tap设备有完整的物理地址和完整的以太网帧。

2) Tun设备等同于点对点的设备，模拟了网络层设备，它操作第三层数据包(如IP数据包)。Tun设备其实完全不需要有物理地址,它收到和发出的包不需要ARP(Address Resolution Protocol，根据IP地址获取物理地址的一个TCP/IP协议)，也不需要有数据链路层的头。

2.3.Bridge

Bridge工作在数据链路层，可有效地将两个局域网（LAN,Local Area Network）连接起来，通过Mac地址（物理地址）来转发数据帧。

Bridge具有如下特征：
1) Bridge在数据链路层上实现局域网互相通信。
2) Bridge以接收、存储、地址过滤与转发的方式实现互连的网络之间的通信。
3) Bridge能够互连两个采用不同数据链路层协议、不同传输介质与不同传输速度的网络。
4) Bridge需要互连的网络在数据链路层以上采用相同的协议。
5) Bridge可以分隔两个网络之间的流量，有利于改善互连网络的性能与安全性。

2.4.Tunnel

Tunnel是指隧道技术，一种通过使用互联网络的基础设施在网络之间传递数据的方式，主要具有如下特征：
1) 使用隧道传递的数据可以是不同协议的数据帧或包。
2) 隧道协议将其它协议的数据帧或包重新封装然后通过隧道发送。即把下一层（如网络层）的数据包封装到上一层（如应用层）或者同一层（如网络层）的协议中进行传输，从而实现网络之间的穿透。
3) 新的帧头提供路由信息，以便通过互联网传递被封装的负载数据。
4) 隧道的两端（发送端和接收端）必须都要有一个协议能够解析这种封装之后的包，这样双方才可以进行通信。

目前，隧道技术主要有二层隧道和三层隧道：
1) 二层隧道是指：把数据包装入到隧道协议中，数据包依靠数据链路层进行传输。目前二层隧道主要有：PPTP、L2TP、L2F、XVLAN。
2) 三层隧道是指：把数据包装入到隧道协议中，数据包依靠网络层进行传输。目前三层隧道主要有： IPSec、GRE。

2.5.Promiscuous Mode

Promiscuous Mode：指网卡的“混杂模式”，表明一台机器能够接收所有经过它的数据流，而不论其目的地址是否是它。

通常情况下，网卡的默认模式为“非混在模式”，表明一台机器仅接收目的地址为本机的数据流，其它的数据流均被丢弃。

3. FLAT网络

3.1.FLAT实现原理

FLAT网络：是指通过二层交换机只能构建单一的广播域网络，该网络具有如下特征：

1) 在交换局域网内，L2数据帧通过交换机设备进行转发的网络。

2) 交换机在接收到数据帧之后(L2 层叫数据帧，L3 层叫数据包)，先解析出数据帧头中的 Mac 地址，再在转发表（转发表是通过自学习自动建立的）中查找是否有对应 Mac 地址的端口，有的话就从相应端口转发出去；没有的话就“洪泛”（即将数据帧转发到交换机的所有端口）。

注：广播域指的是：广播帧（目标Mac地址全部为1）所能传递到的范围，即能够直接通信的范围。

计算机接收数据帧的策略为：

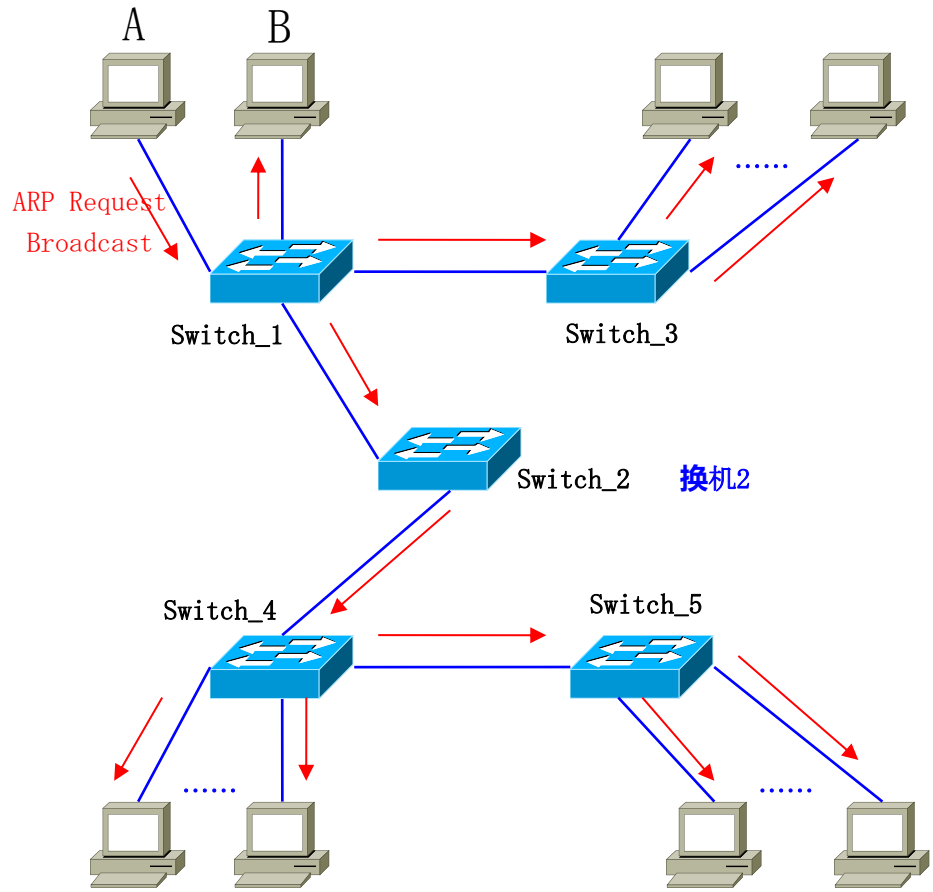
1) 每个端口上的计算机都检查数据帧头中的Mac地址是否与本机网卡的Mac地址一致，一致的话就接收数据帧，不一致就直接丢弃。

2) 如果打开网卡的“混杂模式”，计算机就会接受所有经过该机的数据流。所以在虚拟网桥中，如果希望虚拟机和外部通讯，必须打开桥接到虚拟网桥中物理网卡的混杂模式特性。

FLAT网络下数据流走向：

通过下图的网络结构来解释下，计算机在FLAT网络进行通信时，数据流的走向。

图3-1 FLAT网络下数据流走向



如上图所示：

该交换局域网由5台二层交换机（交换机1～5）连接了大量的计算机构成。假设此时，计算机A需要与计算机B通信，在基于以太网的通信中，必须在数据帧中指定目标Mac地址才能正常通信，因此数据流的走向步骤为：

1) 计算机A必须先广播“ARP请求（ARP Request）信息”，来尝试获取计算机B的Mac地址。
2) 交换机1收到广播帧（ARP请求）后，会将它转发给除接收端口外的其他所有端口，即“洪泛”。
3) 接着，交换机2收到广播帧后也会“洪泛”。
4) 同样交换机3、4、5也还会“洪泛”。
5) 最终ARP请求会被转发到同一网络中的所有计算机上。计算机B响应该ARP请求，其它计算机不响应。

广播帧种类：

实际上广播帧会非常频繁地出现，使用TCP/IP协议栈进行通信时，主要有如下几类：

1) ARP（Address Resolution Protocol，地址解析协议）:根据IP地址获取物理地址的协议。
2) RIP（Routing Information Protocol，路由信息协议）：是内部网关的协议。
3) DHCP（Dynamic Host Configuration Protocol，动态主机配置协议）：动态分配IP地址的协议。

3.2.FLAT优点与缺点

FLAT网络具有如下优点：

1) 网络拓扑结构简单易懂。

2) 数据帧在交换机转发过程中，不会被附加任何Header，使得数据帧在发送/接收时性能较好。

FLAT网络具有如下缺点：

1) 广播信息能够传遍整个网络，导致广播信息不仅消耗了网络整体的带宽，而且收到广播信息的计算机还要消耗一部分CPU时间来对它进行处理，造成了网络带宽和CPU运算能力的大量无谓消耗。

2) 缺少网络隔离，网络安全性较差。

4. VLAN网络

4.1.VLAN实现原理

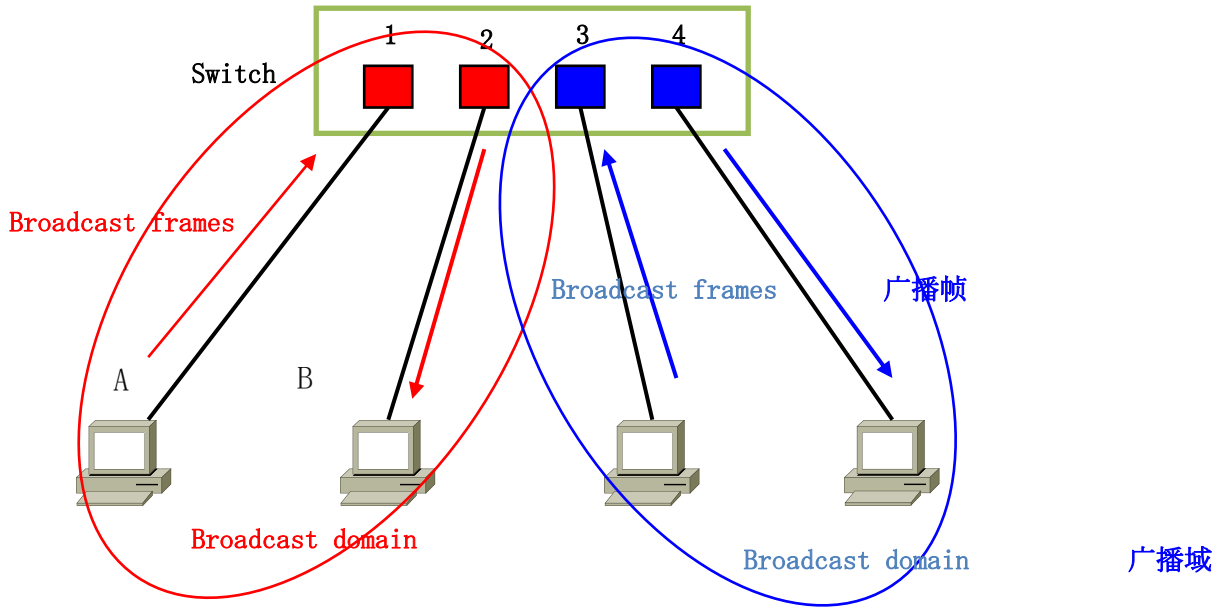
VLAN（Virtual Local Area Network）又称虚拟局域网，是指在交换局域网的基础上，采用网络管理软件构建可跨越不同网段、不同网络的端到端的逻辑网络，该网络具有如下特征：

1） 一个VLAN组成一个逻辑子网（即一个逻辑广播域），它可以覆盖多个网络设备，允许处于不同地理位置的网络用户加入到一个逻辑子网中。

2） VLAN网络使用交换机来分割广播域的，通过一个叫VLAN_ID的标识来区别不同的逻辑子网。

针对交换机通过VLAN来隔离逻辑子网的示意图，如下所示：

图4-1 VLAN隔离的逻辑子网



如上图所示：

交换机上使用不同的VLAN_ID标识来分割逻辑子网，1号端口和2号端口使用一个VLAN_ID，3号端口和4号端口使用另一个VLAN_ID。这样的话，从计算机A发出广播帧的话，收到广播帧的交换机的处理流程如下：

1） 检查接收广播帧的端口属于哪个VLAN_ID，然后把广播帧转发给属于同一VLAN_ID的其它端口。

2） 2号端口连接的计算机B会接收到该广播帧，其它端口的计算机都不会接收到该广播帧。

VLAN网络通过交换机利用VLAN_ID标识的方式有效的隔离了逻辑子网，避免了广播信息传遍整个交换局域网。不过这样的话，同一交换机上的不同逻辑子网想要进行通信（即VLAN间的通信）需要使用路由器或三层交换机提供路由功能。

4.2.VLAN划分方式

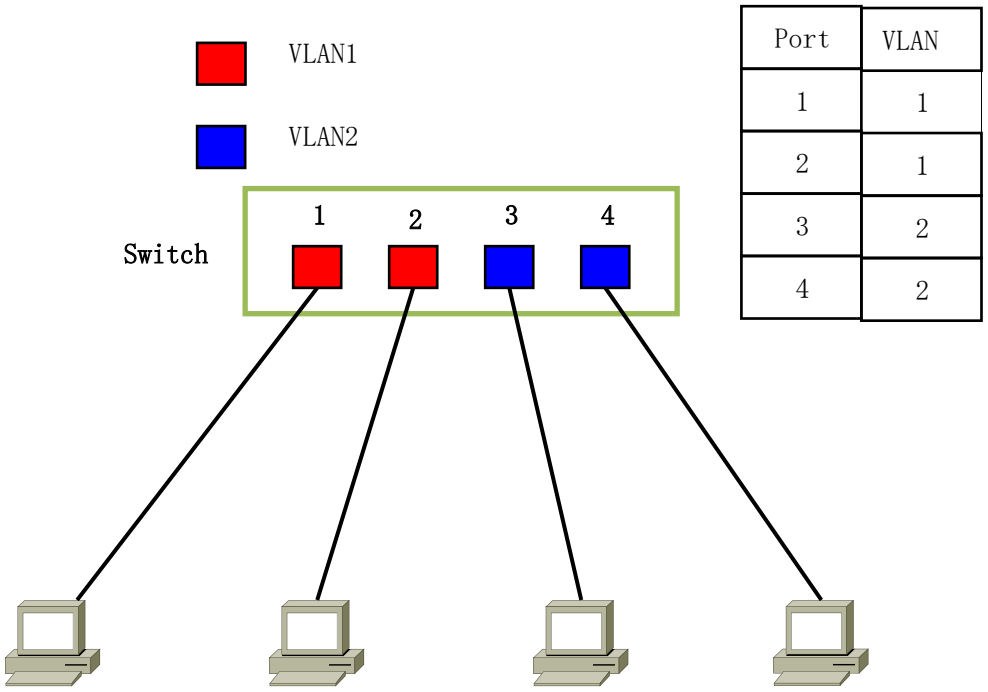
VLAN的划分可以是事先固定的、也可以是根据所连的计算机而动态改变设定。前者被称为“静态VLAN”，后者被称为“动态VLAN”。

4.2.1. 静态划分

静态VLAN又被称为基于端口的VLAN（Port Based VLAN），就是明确指定各端口属于哪个VLAN的设定方法。

关于VLAN网络的静态划分，如下图所示：

图4-2 VLAN网络的静态划分



Port	VLAN
1	1
2	1
3	2
4	2

如上图所示：
1号端口和2号端口被划分为同一个逻辑子网，3号端口和4号端口被划分为同一个逻辑子网。由于需要对每个端口进行指定，当网络中的计算机数目超过一定数字后，设定操作就很费时费力，并且每当计算机变更所连端口时都必须同时更改该端口所属VLAN的设定，这种划分方式明显不够灵活方便。

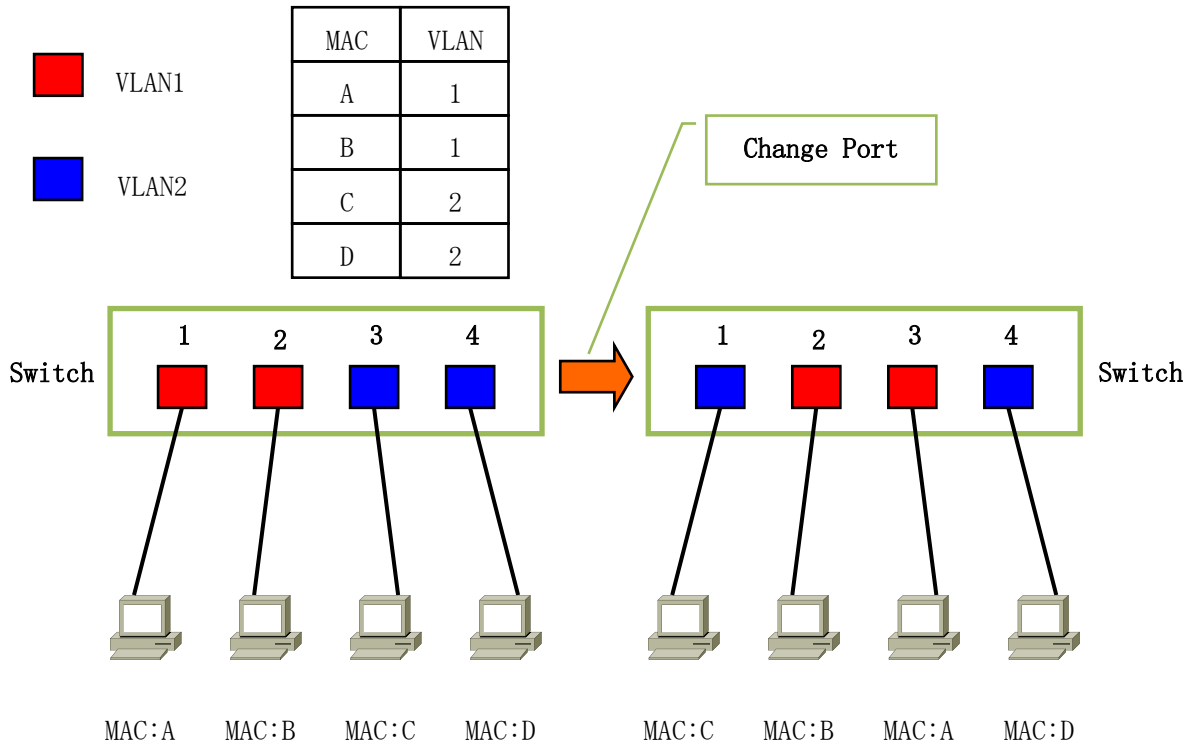
4.2.2. 动态划分

动态VLAN根据每个端口所连的计算机，随时改变端口所属的VLAN，避免了静态VLAN中需要手工更改设定的操作。动态VLAN大致分为如下三类：

1） 基于Mac地址的VLAN（Mac Based VLAN）

基于Mac地址的VLAN划分，就是通过查询并记录端口所连计算机网卡的Mac地址来决定端口的所属VLAN，如下图所示：

图4-3 基于Mac地址的VLAN划分

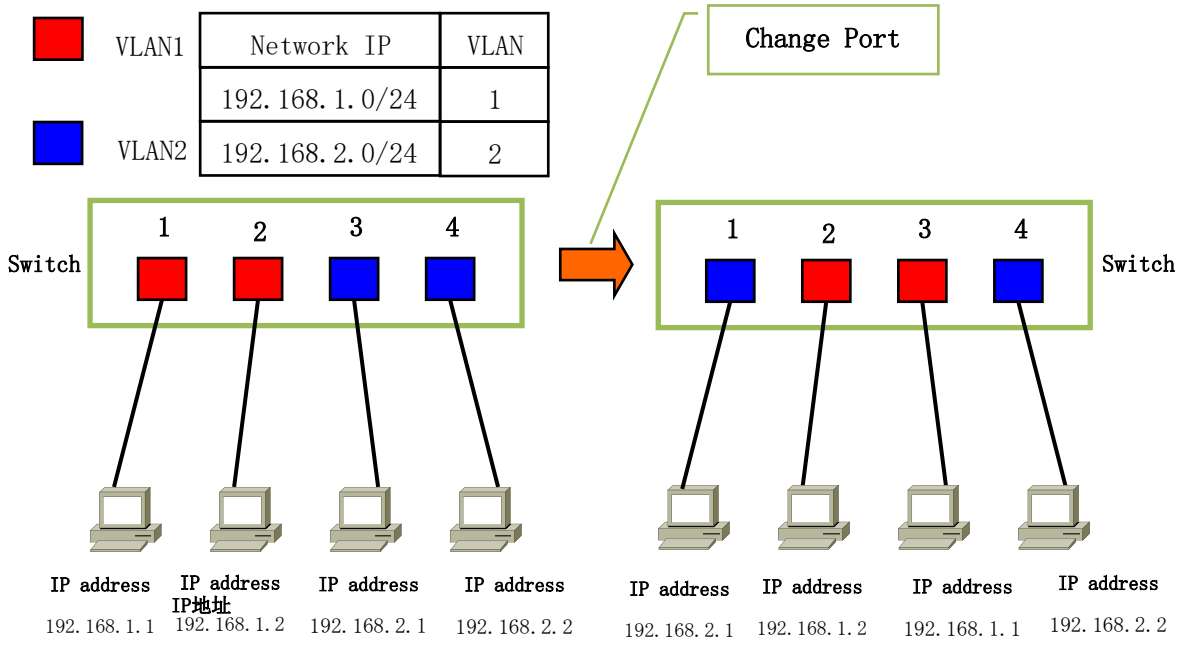


如上图所示：
连接交换机的计算机更改了所连端口号，但是由于计算机的MAC地址没变，因此计算机所属的VLAN仍没有发生变化。不过，这种模式的划分，在设定时仍需统计所连接的所有计算机的MAC地址，并且在计算机更换网卡时还是需要进行更改设定。

2) 基于子网的VLAN（Subnet Based VLAN）

基于子网的VLAN划分，就是通过所连计算机的IP地址来决定端口所属VLAN，如下图所示：

图4-4 基于子网的VLAN划分



如上图所示：
连接交换机的计算机更改了所连端口号，但是由于计算机的IP地址没变，因此计算机所属的VLAN仍没有发生变化。这种模式的划分相比基于MAC地址的划分，能够更为简便地改变网络结构。一般情况下，也都使用基于子网的方法划分VLAN。

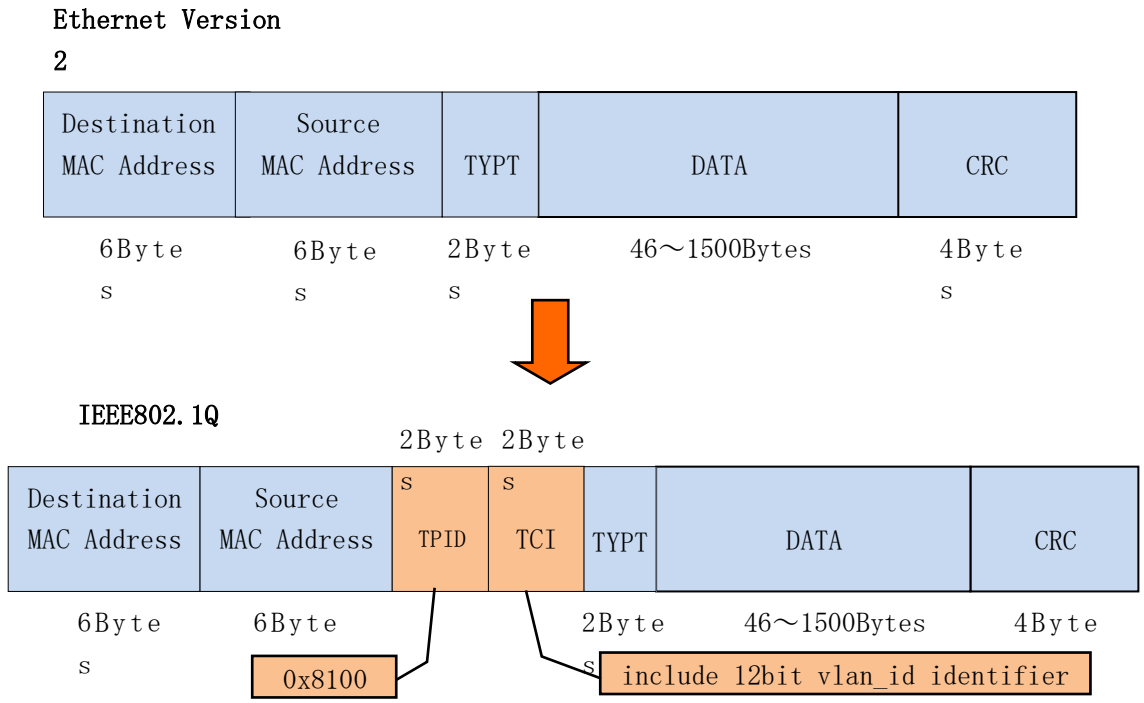
3) 基于用户的VLAN（User Based VLAN）

基于用户的VLAN划分，就是根据交换机各端口所连的计算机上的当前登录用户来决定端口所属VLAN，这种模式的划分，一般用的较少。

4.3.VLAN帧结构

在VLAN网络下，交换机可通过对数据帧添加VLAN识别信息，构建跨越多台交换机的VLAN。VLAN帧的格式(基于IEEE802.1Q标准)如下图所示：

图4-5 VLAN帧结构



如上图所示：

1) 附加的VLAN标识信息位于数据帧中“源MAC地址（Source MAC Address)”和“类别域（Type Field）之间，共占4个字节，2字节的TPID（Tag Protocol Identifier）和2字节的TCI（Tag Control Information)。

2) TCI中的VLAN ID是VLAN网络的识别字段，其占12个bit，因此最多可支持4096（2的12次方）个VLAN网络（其中，0和4095作为预留值，用户不能设置)。

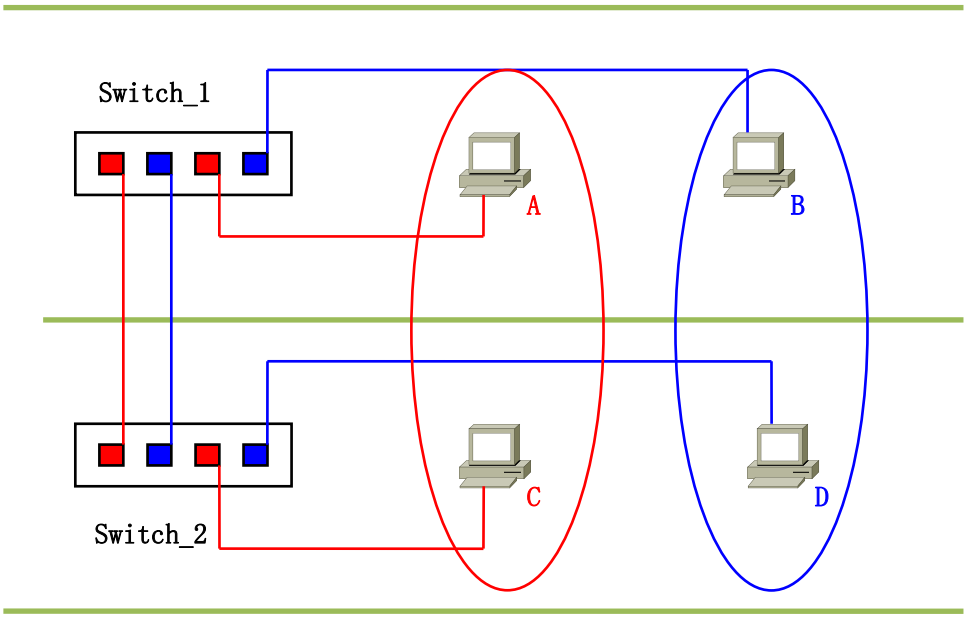
4.4.VLAN模式

VLAN网络的实现是通过交换机来操作的，关于交换机的端口，主要可以分为两种模式：

1) 访问链接（Access Link）：在这种模式下，交换机的端口只属于1个VLAN，一般用于连接计算机的端口。
2) 汇聚链接（Trunk Link）： 在这种模式下，交换机的端口属于多个VLAN，能够转发多个不同VLAN的数据帧，一般用于交换机与交换机（或路由器）之间的连接。

为了对比出汇聚模式下计算机跨交换机进行通信的优点，先来看下，非汇聚模式下跨交换机通信的结构图，如下所示：

图4-6 非汇聚模式下跨交换机通信



如上图所示：

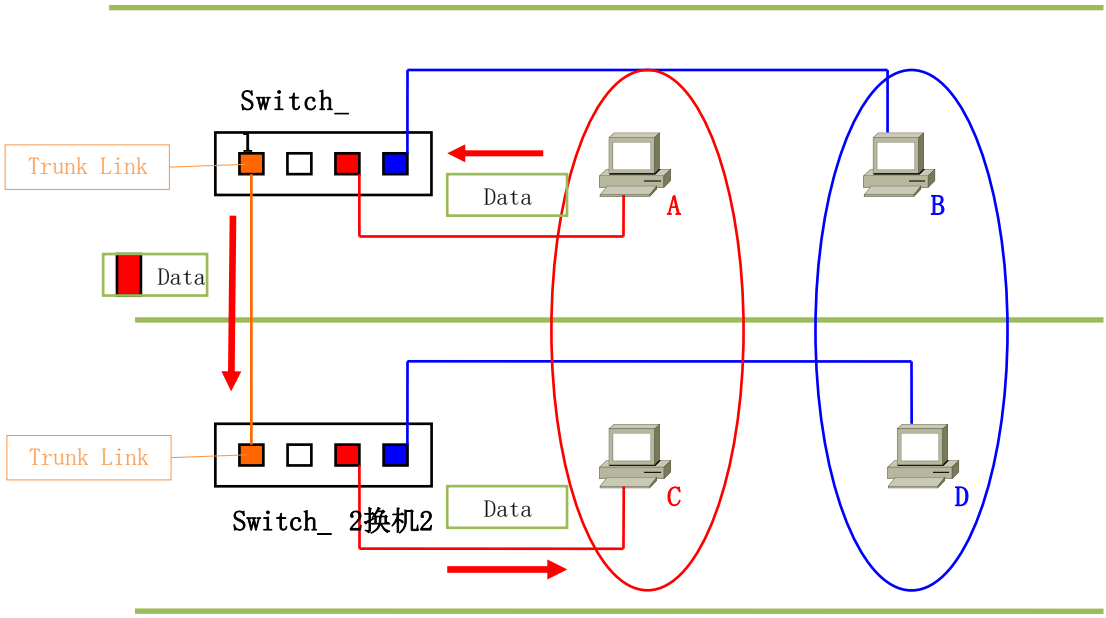
1) 计算机A和计算机C属于同一个VLAN，计算机B和计算机D属于同一个VLAN。
2) 由于计算机A和计算机C所连接的交换机并不相同，它们之间想要进行通信的话，必须要在这两个交换机之间建立一个连接。此时计算机A和计算机C的通信步骤如下：
   a) 计算机A发出数据帧，交换机1收到该数据帧后把其转发到同一VLAN的其它端口。
   b) 交换机2接收到该数据帧，并转发到同一VLAN的其它端口。
   c) 计算机C收到计算机A发送的数据帧。

由上可知：

在上述网络环境下，如果再新建一个VLAN时，为了让这个VLAN能够互通，就需要再交换机之间再新建一个连接，因此，这种访问模式的扩展性和管理效率都不好。

针对汇聚模式下跨交换机通信的结构图，如下所示：

图4-7 汇聚模式下跨交换机通信



如上图所示：

1） 计算机A和计算机C属于同一个VLAN，计算机B和计算机D属于同一个VLAN。

2） 由于计算机A和计算机C所连接的交换机并不相同，它们之间想要进行通信的话，必须要在这两个交换机之间建立一个连接。此时，交换机1和交换机2通过汇聚链接进行连接，计算机A和计算机C的通信步骤如下：

a） 计算机A发送数据帧，交换机1收到该数据帧后，在其上追加VLAN识别信息，然后从汇聚端口转发出去。

b） 交换机2收到该数据帧后，拆开数据帧，根据VLAN识别信息，把该数据帧转发到同一VLAN的端口上去，在转发之前会把VLAN识别信息从该数据帧中删除。（VLAN识别信息的添加和删除都是由交换机处理的，对通信双方的计算机来说完全是透明的）。

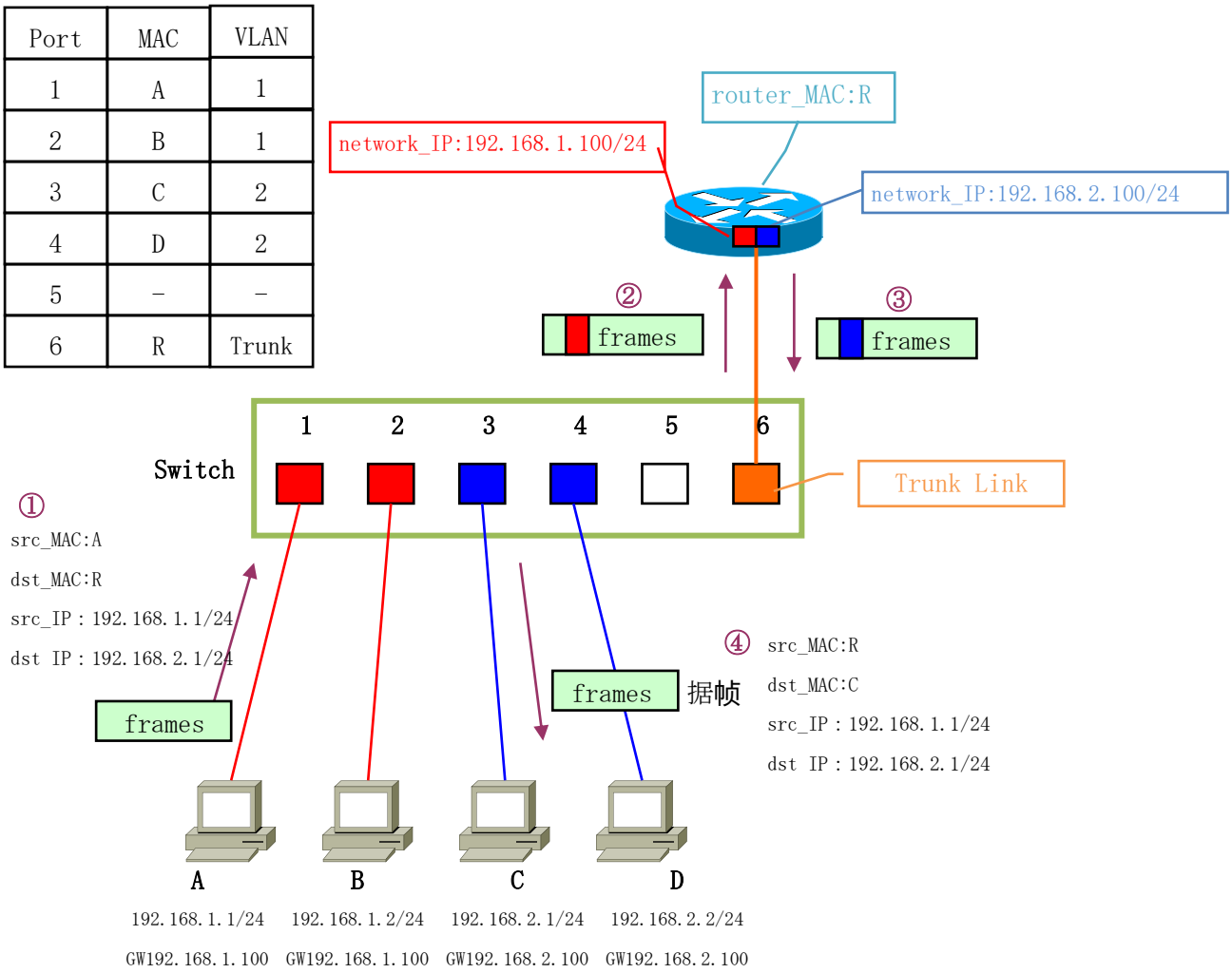
c） 计算机C收到计算机A发送的数据帧。

3） 同样，计算机B和计算机D之间的通信过程跟上述一样。（因为汇聚端口属于多个VLAN，能够转发多个不同VLAN的数据帧）。

4.5.VLAN间通信

不同VLAN间无法通过L2层交换机直接进行通信，可以借助L3层交换机或路由器来完成不同VLAN间的通信，，如下图所示：

图4-7 VLAN间通信



如上图所示：

1） 交换机端口1和端口2为同一个VLAN，端口3和端口4为同一个VLAN，端口号6通过汇聚连接（Trunk Link）方式与路由器相连接。

2） 计算机A和计算机B由于属于同一个VLAN，因此它们之间的通信仅需要交换机处理即可，不需要路由器参与。

3） 计算机A和计算C由于不属于同一VLAN，因此它们之间的通信不仅需要交换机处理，还需要路由器参与处理，具体步骤为：

a） 计算机A从通信目标IP地址（计算机C的IP地址）得知与自己不属于同一个网段，因此会向

设定的默认网卡（Default Gateway，GW）发送数据帧（即目标MAC地址为路由器的MAC地址，而非计算机C的Mac地址）。

b) 交换机在端口1上接收到数据帧后，检索MAC地址列表中与端口1同属一个VLAN的表项，由于汇聚链路被看作属于所有的VLAN，因此可从该端口转发。

c) 由于交换机端口6是汇聚链接，从该端口发送数据帧需要附件VLAN标识信息，于是在原有的数据帧上附加VLAN识别信息后发送数据帧。

d) 路由器收到交换机发来的数据帧后，根据路由表得知目标网络（192.168.2.0/24）属于蓝色VLAN接口，于是交给蓝色VLAN接口去处理。

e) 蓝色VLAN接口根据数据帧的目的IP地址查找对应MAC地址，然后修改该数据帧的目的MAC地址，以及VLAN标识信息，然后发送数据帧。

f) 交换机收到该数据帧后，根据VLAN标识信息从MAC地址列表中进行检索，发现端口3符合要求，于是交换机就将数据帧中的VLAN标识信息删除后然后转发到端口3。

g) 计算机C收到计算机A发送的数据帧。

4.6.VLAN优点与缺点

VLAN网络具有如下优点：

1) 端口隔离：使得一个交换机可以当作多个逻辑交换机来使用。

2) 网络安全：不同VLAN子网之间不能够直接通信，杜绝了广播信息传遍整个网络。

3) 灵活管理：更改用户所属的网络时，不必更换端口和连线，只需更改相关配置即可。

VLAN网络具有如下缺点：

1) VLAN的数量限制：最多4096个VLAN远不能满足大规模云计算数据中心的需求

2) 交换机MAC表耗尽: 虚拟化以及东西向流量导致更多的MAC表项

3) 物理网络基础设施的限制:基于IP子网的区域划分限制了需要二层网络连通性的应用负载的部署。

5. GRE网络

5.1.GRE实现原理

GRE（Generic Routing Encapsulation，通用路由封装）网络：是指对某些网络层协议（如IP和IPX）的数据报文进行封装，从而使得这些被封装的数据报文能够在另一个网络层协议（如IP）中进行传输的网络，该网络具有如下特征：

1) 采用了Tunnel技术，Tunnel是个虚拟的点对点的连接，提供了一条通道使封装的数据报文能够在这个通道上传输。
2) 在一个Tunnel的两端需要分别对数据报进行封装和解封装。
3) 对通信的两端机器来说，数据报的封装和解封装都是两端的路由器进行操作，对两端机器来说这些操作都是透明的。
4) 本质是在遂道的两端的 L4 层建立 UDP 连接传输重新包装的 L3 层包头，在目的地再取出包装后的包头进行解析。

关于GRE隧道互连的网络，其示意图如下所示：

图5-1 X协议网络通过GRE隧道互连

如上图所示：

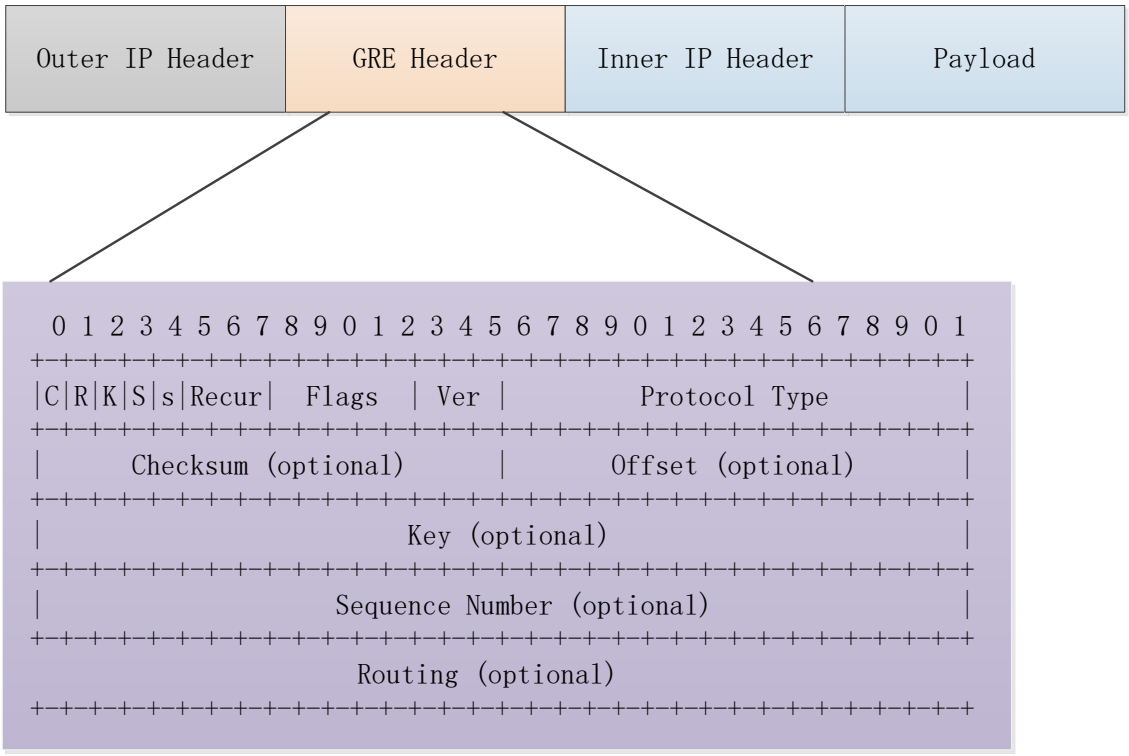
Group1中机器要想和Group2中机器进行通信，需要通过GRE隧道互连，两者之间的通信步骤为：

1) Group1中的机器发送数据包。
2) RouterA连接Group1的接口收到X协议数据包后，首先交给X协议处理。
3) X协议检查报文头中的目的地址来确定如何路由此数据包。
4) 数据包的目的地址要经过Tunnel才能到达，于是RouterA将数据包转给相应的Tunnel接口。
5) Tunnel接口接收到此数据包后进行GRE封装，在封装IP数据包头之后，根据此IP包的目的地址及路由表对数据包进行转发，从相应的网络接口发送出去。
6) RouterB接收到RouterA发送的数据包后，对该数据包进行解析，发现该数据包的目的地址就是自己以及协议字段值为47，从而确定该数据包是GRE封装后的数据包，于是删掉GRE头，并根据内部IP数据包的目的地址及路由表对数据包进行转发，从相应的网络接口发送出去。
7) Group2中的机器接收该数据包。

5.2.GRE 数据包格式

在GRE网络下，数据包在网络传递过程中，路由器会对数据包进行GRE封装/解封装，即对数据包附加/删除GRE Header，GRE 数据包的格式如下图所示（参照RFC1701定义）：

图5-2 GRE 数据包格式



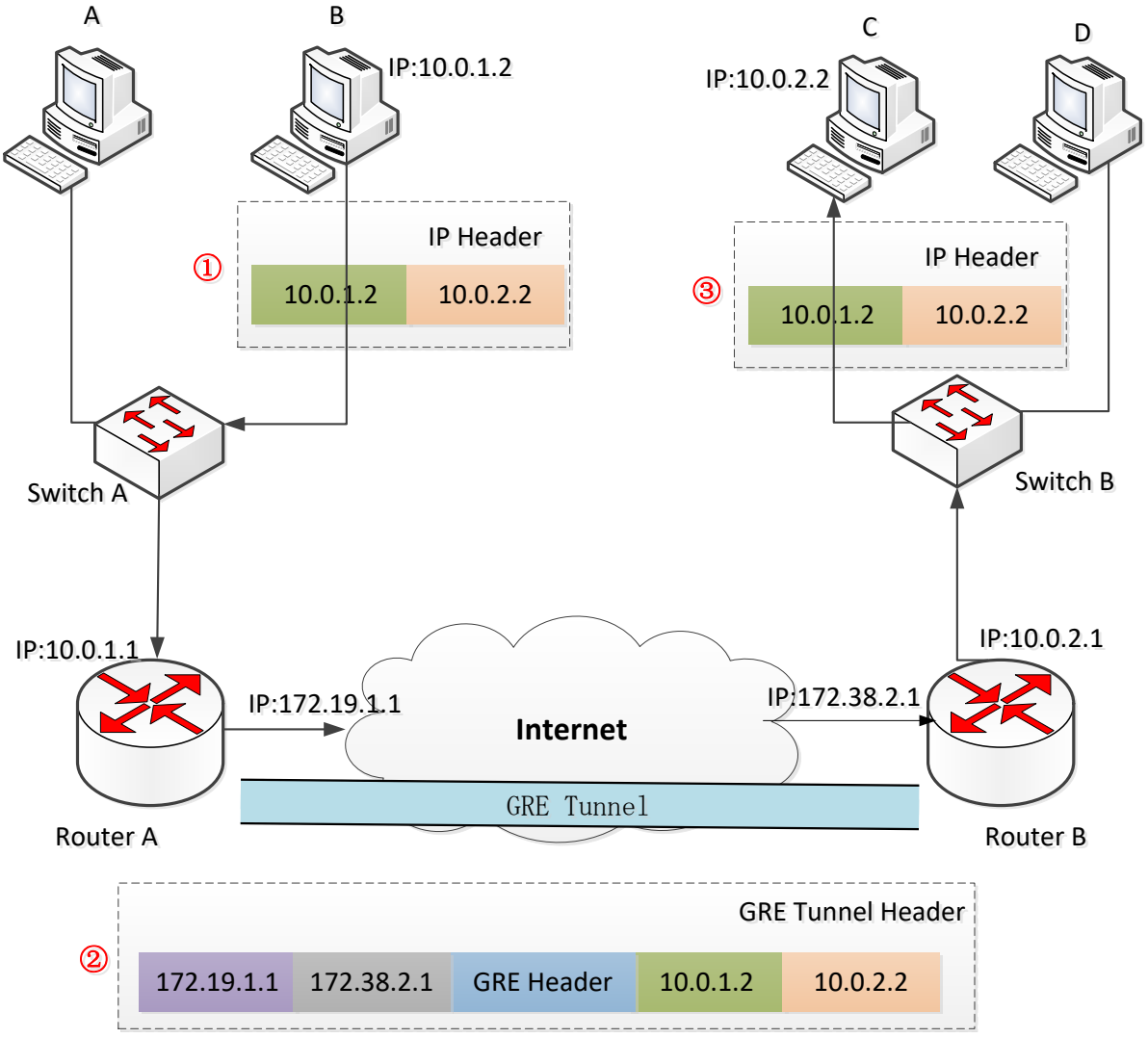
如上图所示：

1) Payload：表示需要封装和传输的数据包。
2) Inner IP Header:表示Payload的IP Header。
3) GRE Header：表示GRE协议的头部，长度为4～20个字节，前4个字节是必选项，第5～20字节是可选项，由第1个字节的相关bit值来决定。详细描述如下：
   a) 0bit（C）：校验和标志位，如果该位值为1，那么“Checksum”和“Offset”部分的4个字节必须出现在GRE Header中。
   b) 1bit（R）:路由标志位，如果该位值为1，那么“Routing”部分的4个字节必须出现在GRE Header中。
   c) 2bit（K）：密钥标志位，如果该位值为1，那么“Key”部分的4个字节必须出现在GRE Header中。这样通信双方将进行通道识别关键字的验证，只有Tunnel两端设置的识别关键字完全一致时才能通过验证，否则将丢弃数据包。
   d) 3bit（S）：序列同步标志位，如果该位值为1，那么“Sequence Number”部分的4个字节必须出现在GRE中。
   e) 4～15bit：一般为0。
   f) 16～31（Protocol Type）：协议类型，如IP协议为0X0800。
4) Outer IP Header:表示外部IP Header,是数据包在网络传输中使用的IP Header。

5.3. GRE通信

不同子网的计算机通过GRE隧道进行通信的示意图，如下所示：

图 5-3 不同子网通过GRE隧道进行通信



如上图所示：

计算机A和计算机B位于同一局域网，计算机C和计算D位于同一局域网，两个局域网通过路由器连接Internet，并通过路由器配置了一个GRE隧道。此时，计算机A想要和计算机C进行通信，通信步骤为：

1） 计算机A发送数据包。

2） SwitchA收到该数据包后，转发给RouterA。

3） RouterA收到该数据包后，发现目的IP地址需要经过Tunnel才能到达，于是RouterA将数据包转给相应的Tunnel接口。

4) Tunnel接口接收到此数据包后进行GRE Header封装，在封装IP数据包头之后，根据此IP包的目的地址及路由表对数据包进行转发，从相应的网络接口发送出去。

5) RouterB接收到RouterA发送的数据包后，对该数据包进行解析，发现该数据包的目的地址就是自己以及协议字段值为47，从而确定该数据包是GRE封装后的数据包，于是删掉GRE头，并根据内部IP数据包的目的地址及路由表对数据包进行转发，从相应的网络接口发送出去。

6) SwitchB收到该数据包，查询MAC表项并从相应的端口转发出去。

7) 计算机C收到计算B发送的数据包。

5.4.GRE优点与缺点

GRE网络具有如下优点：

1) 易于重建：由于是在L3上面包装L3，可以不用变更底层网络架构重建L3通信。

1) 易于迁移：由于是跨不同网络实现二次IP通信，可以方便Guest迁移。

2) 支持网络数量大：由于GRE Header中有32bit用于标识网络，网络数量最多可达到2^32个。

GRE网络具有如下缺点：

1) 传输性能下降：由于隧道两端需要进行封装/解封装处理，以及由于封装造成的数据量增加，会导致使用GRE隧道后设备的数据转发效率有一定程度的下降。

2) 广播风暴严重：由于GRE遂道将虚拟二层打通了，会导致广播风暴更严重。（不过虚拟交换机能够完全知道虚拟机的IP和MAC地址的映射关系，因此不需要通过ARP广播来找MAC地址）

3) 网络扩展性差：GRE隧道只能是点对点的，当有多个节点时，就需要构建多个GRE隧道。

6. VXLAN网络

6.1.VXLAN实现原理

VXLAN（Virtual Extensible LAN，虚拟可扩展局域网）网络：是指对L2层的数据帧进行封装并通过L4层的UDP进行传输的网络，该网络具有如下特征：

1) VXLAN网络是通过一个叫VTEP（VXLAN Tunnel Endpoint）的设备（VTEP可由支持VXLAN的硬件设备或软件来实现）来对数据包封装和解封装，不同于GRE网络是通过路由器来进行数据包封装和解封装。

2) VXLAN网络主要用于封装、转发2层报文，使得多个通过三层连接的网络可以表现的和直接通过一台物理交换机连接处在一个LAN中一样。

3) VXLAN网络提供了将二层网络overlay在三层网络上的能力，VXLAN Header中的VNI（VXLAN Network Identity）有24个bit，数量远远大于4096，并且UDP的封装可以穿越三层网络，比VLAN有更好的扩展性。

4) 由于VXLAN网络中的不同VTEP起初是不“认识的”,于是每个VETP都要关联一个组播地址。当其中一个VTEP发送一个ARP请求时，会发送一个组播IGMP（Internet Group Management Protocol，因特网组管理协议）报文给所有同在这个网络组中的其他VTEP，所有关联这个组播地址的VTEP都会收到这个报文。

5) 对通信的两端机器来说，数据报的封装和解封装都是两端的VTEP设备进行操作，对两端机器来说这些操作都是透明的。

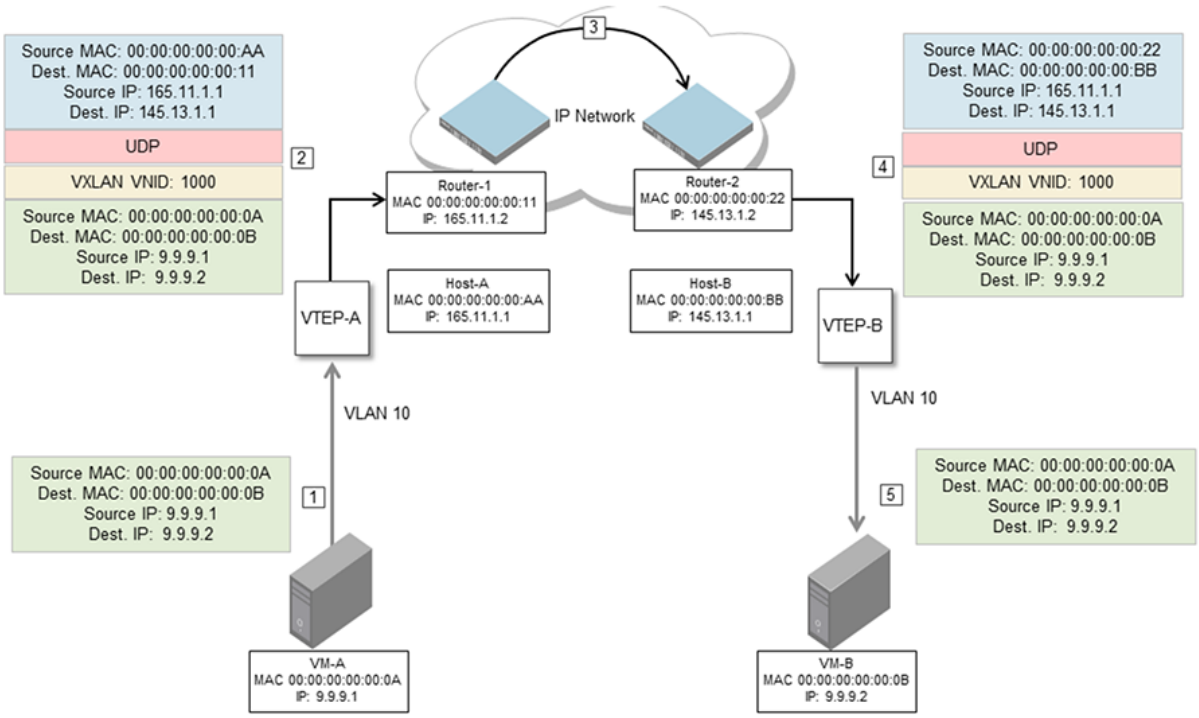
6) VXLAN网络不会在通信两端之间维持一个长连接，所以VXLAN需要一个控制平面来记录对端地址可达情况。控制平面的表为(VNI，内层MAC，外层VTEP_IP)。

7) VXLAN学习地址的时候仍然保存着二层协议的特征，节点之间不会周期性的交换各自的路由表，对于不认识的MAC地址，VXLAN依靠组播来获取路径信息。

8) VXLAN还有自学习的功能，当VTEP收到一个UDP数据报后，会检查自己是否收到过这个虚拟机的数据，如果没有，VTEP就会记录源VNI/源外层IP/源内层MAC对应关系，避免组播学习。

不同子网的虚拟机通过VXLAN隧道进行通信的示意图，如下所示：

图6-1不同子网通过VXLAN隧道进行通信



如上图所示：

VM-A和VM-B位于不同子网下的两个虚拟机，两个子网之间配置了VXLAN隧道，此时虚拟机VM-A想要和虚拟机VM-B进行通信，通信步骤为：

1) 虚拟机VM-A发送数据包。

2) VTEP-A（Host-A）接收到该数据包后，根据其目的IP和Mac地址得知需要通过VXLAN隧道发送，于是对该数据包进行封装，附加VXLAN Header、UDP Header、Outer IP Header和Outer Mac Header，然后发送该数据包。

3) 路由器Router-1接收到该数据包后，对该数据包进行解析，然后根据数据包的目的地址及路由表项修改数据包的Outer IP Header和Outer Mac Header，接着从相应的网络接口发送出去。

4) 路由器Router-2接收到该数据包后，对该数据包进行解析，然后根据数据包的目的地址及路由表项修改数据包的Outer IP Header和Outer Mac Header，接着从相应的网络接口发送出去。

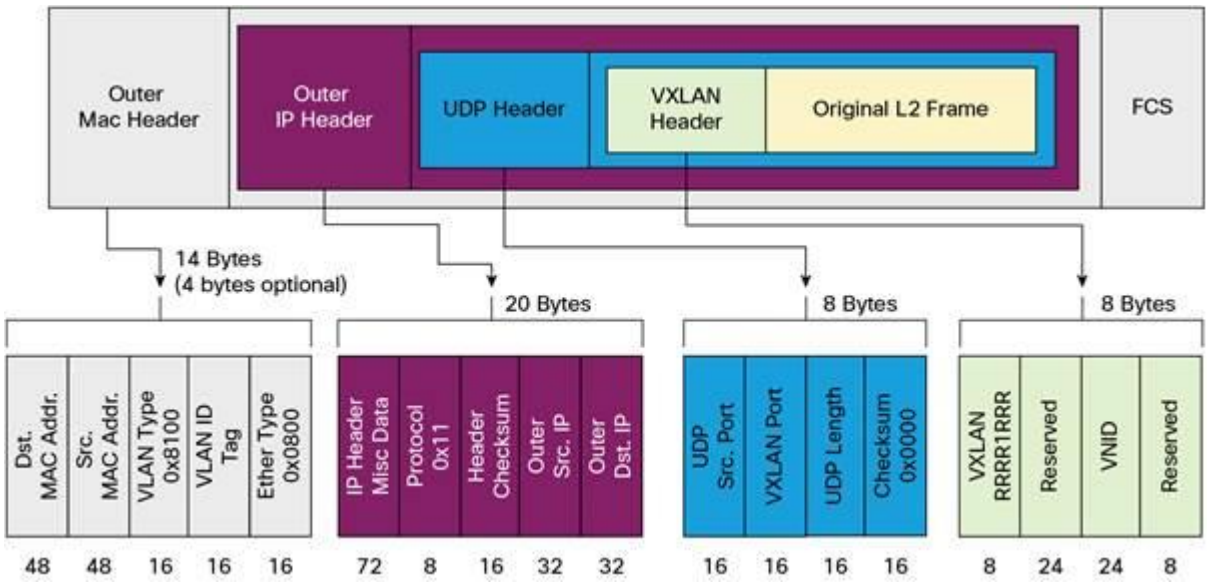
5) VTEP-B(Host-B)接收到该数据包后，删除Outer Mac Header、Outer IP Header、UDP Header和VXLAN Header，然后进行转发。

6) 虚拟机VM-B接收到虚拟机VM-A发送的数据包。

6.2.VXLAN 数据包格式

VXLAN数据包从功能上来说，等同于一个传统的L2层数据帧。每个VXLAN数据包通过一个24bit的VNI（VXLAN Network Identity）来进行标识，关于VXLAN数据包的格式，如下图所示：

图6-2 VXLAN 数据包格式



如上图所示：

1) FCS（Frame Check Sequence）：该部分占4个字节，用于保存CRC（Cyclical Redundancy Check）校验值。

2) Original L2 Frame：原始的L2层数据帧，该数据帧也可以包含一个VLAN Header。

3) VXLAN Header：该部分占8个字节，其中前8个bit为flag标志，中间的24bit为VXLAN数据包的标识符（VNI），VNI是唯一的，只有具有相同的标识符才可以进行通信。

4) Outer UDP Header：该部分占8个字节，其中源端口号是动态被分配的，目的端口号一般都为4789。

5) Outer IP Header：外部IP表头，其中源IP地址为源VTEP_IP,目的IP地址为目的VTEP_IP。

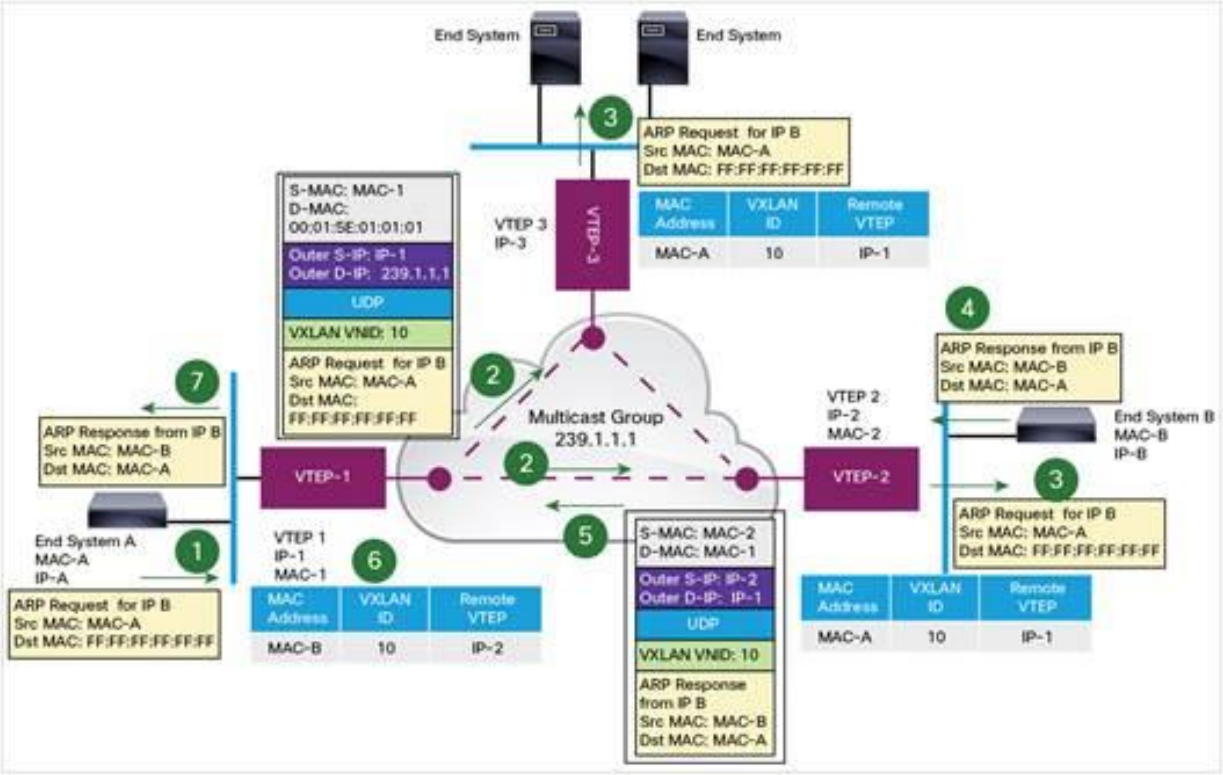
6) Outer Mac Header：外部Mac表头。

6.3.VXLAN通信

6.3.1. 多播通信

由于VXLAN网络中不同VTEP起初是不“认识的”,每个VETP都要关联一个组播地址，当不同VTEP需要进行通信时，会发送一个ARP请求，这个请求会被转化为一个组播IGMP报文发送给同在这个网络中的其它VTEP,所有关联这个组播地址的VTEP会的收到这个报文。关于ARP请求数据包在VXLAN下的走向，如下图所示：

图6-3 VXLAN网络下的ARP请求过程



如上图所示：
终端系统A想要和终端系统B进行通信，但是不知道其Mac地址，于是发送一个ARP请求数据包，通信步骤为：

1) 终端系统A发送一个ARP请求数据包，其目的Mac地址全为1。

2) VTEP-1接收该数据包后，检查其路由表项，发现没有终端系统B的IP映射，于是封装该数据包为一个IP多播数据包（目的IP地址为一个多播地址），然后转发该数据包。

3) VTEP2和VTEP3都收到了该数据包，首先开始解析该数据包，发现其VXLAN ID与自己一致，于是开始解封装（删除UDP Header和VLXAN Header等）该数据包，然后转发到自己的本地网络（Local Network）。同时VTEP-2和VTEP-3也会从Outer IP Header以及从内部Mac Header得知VTEP-1的IP地址和终端系统A的Mac地址，并记录到自己的路由表项中。

4) 终端系统B接收到该数据包，发现其目的IP地址与自己IP地址一致，于是就响应自己的Mac地址进行回复，同时记录下终端系统A的IP地址和Mac地址的映射在本地。

5) VTEP2接收到终端系统B回复的数据包，由于VTEP2现在已经知道终端系统A的Mac和IP的映射，于是使用单播对该数据进行封装，然后转发出去。

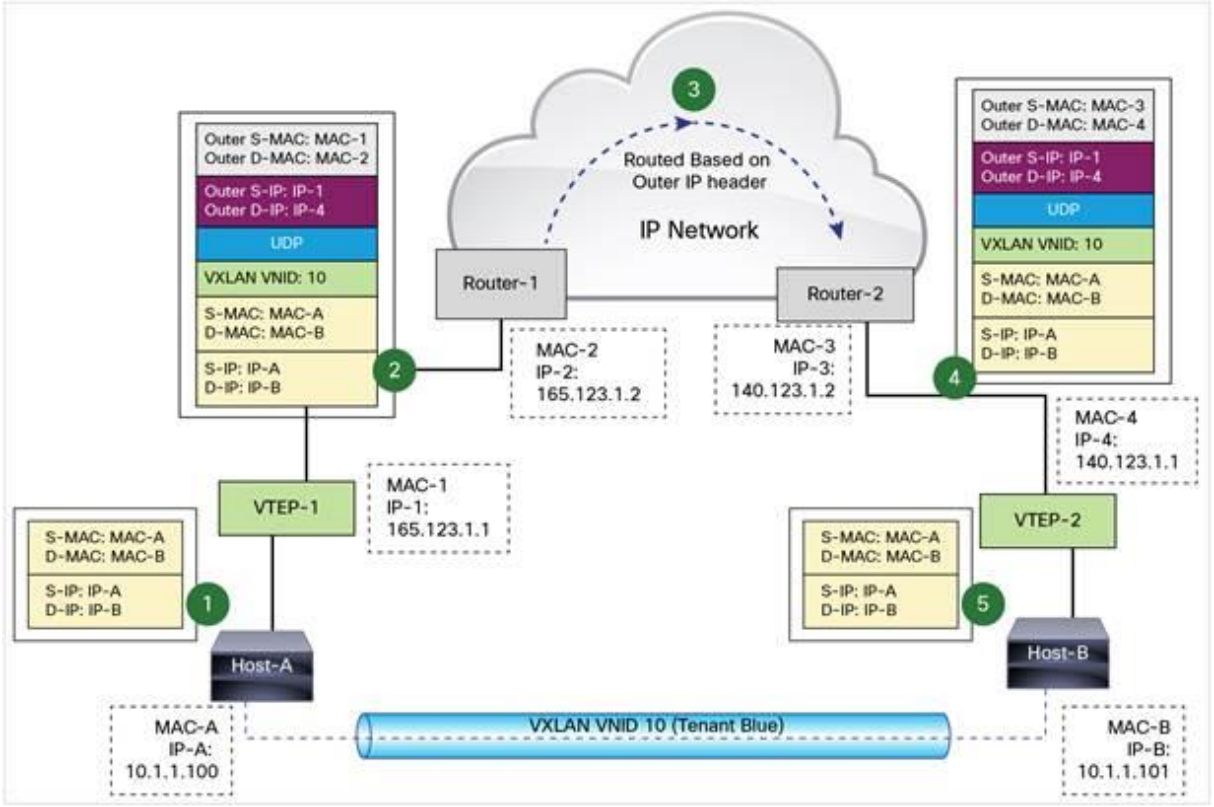
6) VTEP-1接收到VTEP-2发送来的数据包，首先开始解析该数据包，发现其VXLAN ID与自己一致，于是开始解封装（删除UDP Header和VLXAN Header等）该数据包，然后转发到自己的本地网络（Local Network）。同时，也会从Outer IP Header以及从内部Mac Header得知VTEP-2的IP地址和终端系统B的Mac地址，并记录到自己的路由表项中。

7) 终端系统A收到终端系统B的回复数据包，于是解析该数据包，然后记录下终端系统B的IP地址和Mac地址的映射在本地。

6.3.2. 单播通信

当VXLAN网络下的机器通过ARP请求得知想要进行通信机器的IP-Mac映射的关系时，其通信流程如下图所示：

图6-4 VXLAN网路下的单播通信



由上图可知：

1) Host-A发送数据包。

2) VTEP-1接收到该数据包后，根据其目的IP和Mac地址得知需要通过VXLAN隧道发送，于是对该数据包进行封装，附加VXLAN Header、UDP Header、Outer IP Header和Outer Mac Header，然后发送该数据包。

3) 路由器Router-1接收到该数据包后，对该数据包进行解析，然后根据数据包的目的地址及路由表项修改数据包的Outer IP Header和Outer Mac Header，接着从相应的网络接口发送出去。

4) 路由器Router-2接收到该数据包后，对该数据包进行解析，然后根据数据包的目的地址及路由表项修改数据包的Outer IP Header和Outer Mac Header，接着从相应的网络接口发送出去。

5) VTEP-2接收到该数据包后，删除Outer Mac Header、Outer IP Header、UDP Header和VXLAN Header，然后进行转发。

6) Host-B接收到Host-A发送的数据包。

6.4.VXLAN优点与缺点

VLAN网络具有如下优点：

1) 支持网络数量大：由于VXLAN Header中有24bit用于标识网络，网络数量最多可达到2^24个。

2) 扩展性好：VXLAN网络通过L2 over UDP，屏蔽了底层的差异，同时UDP也是 IPv4 的单播和多播，不需要像GRE网络那样针对每个通道（只能是点对点）进行设置，提高了网络的扩展性。

3) 易于迁移：VXLAN 通过重新定义 L2 层帧头，使得两个跨 L3 层的两个子网从 L2 层上能够互通，因此可以让虚拟机跨数据中心进行迁移（以前顶多只能在同一个VLAN里迁移）。

VLAN网络具有如下缺点：

1) 传输性能下降：由于隧道两端需要进行封装/解封装处理，以及由于封装造成的数据量增加，会导致使用VXLAN隧道后设备的数据转发效率有一定程度的下降。

2) 广播风暴严重：当VETP发送一个ARP请求时，都会发送一个组播IGMP报文给所有同在这个网络组中的其他VTEP，所有关联这个组播地址的VTEP会的收到这个报文。

7. 总结

关于FLAT、VLAN、GRE和VXLAN四种网络类型的优缺点，对比如下表所示：

表7-1 不同网络类型的对比

No.	网络类型	支持网络数量	网络性能	扩展性	迁移性	网络适用范围
1	FLAT	1	较好	困难	困难	内部网络
2	VLAN	2^12	较好	困难	困难	内部网络
3	GRE	2^32	较差	困难	容易	外部网络
4	VXLAN	2^24	较差	容易	容易	外部网络

8. 遗留问题

关于OpenStack下网络相关的遗留问题，如下表所示：

表8-1 遗留问题

No.	遗留问题
1	OpenStack环境下Neutron组件的工作原理。
2	Openstack环境下网络数据流的走向。

9. 参考资料

1) https://tools.ietf.org/html/rfc1701
2) https://tools.ietf.org/html/rfc2784
3) http://tools.ietf.org/html/rfc2890
4) http://techbackground.blogspot.jp/2013/06/path-mtu-discovery-and-gre.html
5) http://www.cisco.com/c/en/us/support/docs/ip/generic-routing-encapsulation-gre/64565-gre-tunnel-keepalive.html
6) http://www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/white-paper-c11-729383.html
7) http://www.borgcube.com/blogs/2011/11/vxlan-primer-part-1/
8) http://www.brocade.com/downloads/documents/html_product_manuals/brocade-vcs-gateway-vmware-dp/GUID-5A5F6C36-E03C-4CA6-9833-1907DD928842.html
9) https://tools.ietf.org/html/rfc7348
10) http://www.brocade.com/downloads/documents/html_product_manuals/brocade-vcs-gateway-vmware-dp/GUID-C06F441E-560A-4D83-8DE1-780F602417E5.html