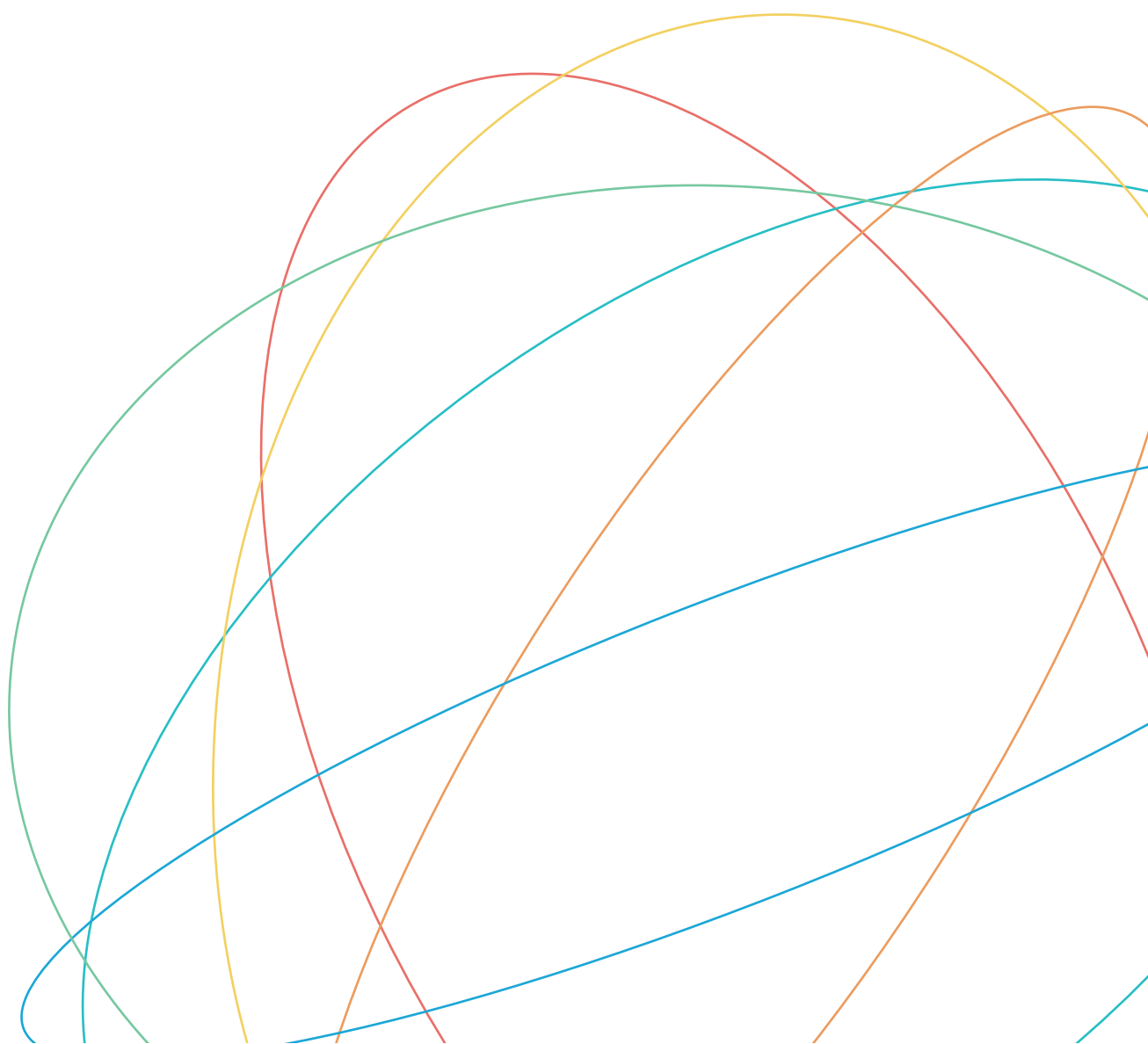


华为园区交换机VXLAN 技术白皮书



1 术语表

英文缩写	英文全称	中文全称
VXLAN	Virtual Extensible Local Area Network	虚拟扩展局域网
NVo3	Network Virtualization over L3	三层网络虚拟化
BUM	Broadcast, Unknown unicast and Multicast	广播、多播、未知单播
VNI	Virtual Network Instance	虚拟网络实例
VM	Virtual Machine	虚拟机
VTEP	VXLAN Tunnel End Point	VxLAN隧道端点
SDN	Software Defined Networking	软件定义网络

2 摘要

VxLAN (Virtual eXtensible Local Area Network) 是应用于数据中心虚拟化的一种 NVo3(Network Virtualization over Layer 3) 技术，主要解决多租户接入扩展和 VM 迁移等场景应用。VxLAN 是一个在 L3 层网络叠加 L2 层的方案，采用 MAC-in-UDP 封装来延伸 L2 层网络。每个被叠加的 L2 层被称为一个 VxLAN 段，通过一个被称为 VxLAN 网络标识符 (VNI) 的 24 位段 ID 来标识，从而可以提供多达 16M 的 VxLAN 段。

3 产生背景

服务器虚拟化后，数据中心内部 VM 的数量比原有的物理机发生了数量级的变化，与之对应的虚拟机虚拟网卡的 MAC 地址数量也相应的增加，这对原有的 LSW 交换机的地址容量能力产生了很大冲击。其次，大规模数据中心以及公有云的场景下，原有的数据中心用来划分虚拟网络的 VLAN 技术不再能够满足需求，因为现有 VLAN 只有 4094 个虚拟网络标识可用，远不能满足数据中心的租户隔离的需求。另一方面，云数据中心中虚拟机可以进行一定范围的迁移，虚拟机迁移一个基本要求是需要保持 IP 和 MAC 地址不变，因此虚拟机只能在二层网络下迁移，如果虚拟机迁移的物理服务器和虚拟机不在同一个网络，则需要同步修改迁移目的

物理网络的 IP 地址，广播域等配置，增加了网络配置的复杂性。

VxLAN(Virtual eXtensible Local Area Network，全称虚拟扩展局域网)这一标准就是在这种背景下产生的，VxLAN 是一种进行大二层虚拟网络扩展的隧道封装技术，很好地解决了上述问题。目前这个技术已标准化（RFC7348），并已经成为业界主流的虚拟网络技术之一。

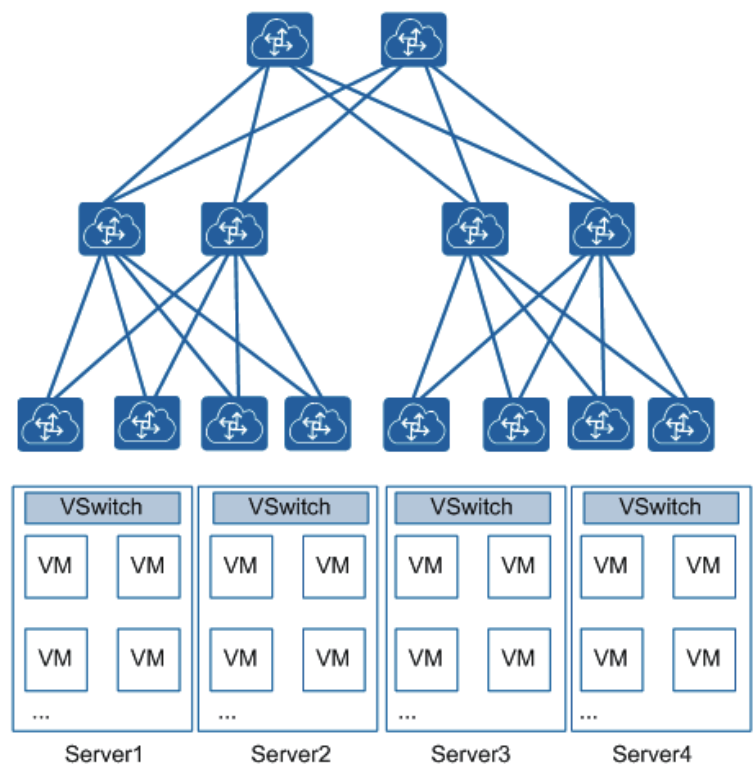


图 1 服务器虚拟化示意图

一台服务器可虚拟多台虚拟机，而一台虚拟机相当于一台主机。

4 技术简介

VxLAN 是一种二层网络叠加三层网络的 NVo3 技术，主要的技术原理就是引入一种 MAC in UDP 格式的封装报文，UDP 报文作为隧道承载层，而原有的二层数据报文内容作为隧道净荷来传输。由于外层采用了 UDP 作为传输手段，就可以让净荷数据轻而易举的在三层网络中传送。达到在三层网络上延伸二层网络的目的。

4.1VxLAN 报文的格式

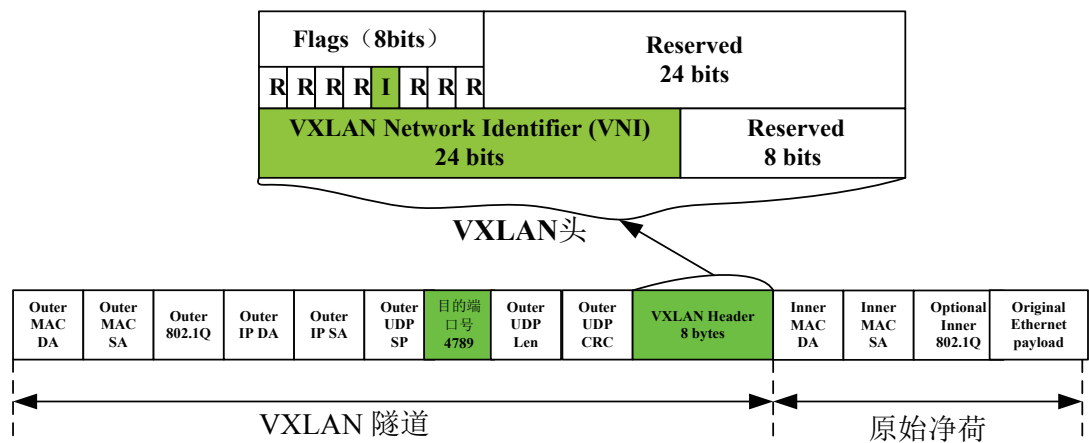


图 2 VxLAN 报文格式

如上，包括 VxLAN 外层封装和内层的原始净荷，其中：

- Flags (8 bits) 其中 I 必须被设置为 1，才是有效的。R 需设为 0。
- VxLAN Segment ID/VxLAN Network Identifier (VNI) – 为 24bit，是虚拟网络的标识。
- Reserved fields (24 bits and 8 bits) – 必须被设置为 0。
- VxLAN 外层隧道的目的端口号为 4789，是专为 VxLAN 分配的端口号。

说明：VxLAN 引入了一个 VxLAN ID (VNI - VxLAN Network Identifier) 来作为虚拟网络的标识，为 24bits，与原有的 VLAN 12bits 长度相比，VxLAN ID 长度增加了一倍，理论可以支持 16M 个虚拟网络。

4.2 VTEP

在 VxLAN 标准中，引入了一个 VTEP(VxLAN Tunnel End Point) 的逻辑体，来完成 VxLAN 报文的封装和解封装。VTEP 与物理网络相连，分配有物理网络的 IP 地址，该地址与虚拟网络无关，VTEP 可以软件或硬件实现，即位于 VM 所在的 Server 中，也可以由独立的硬件设备实现。

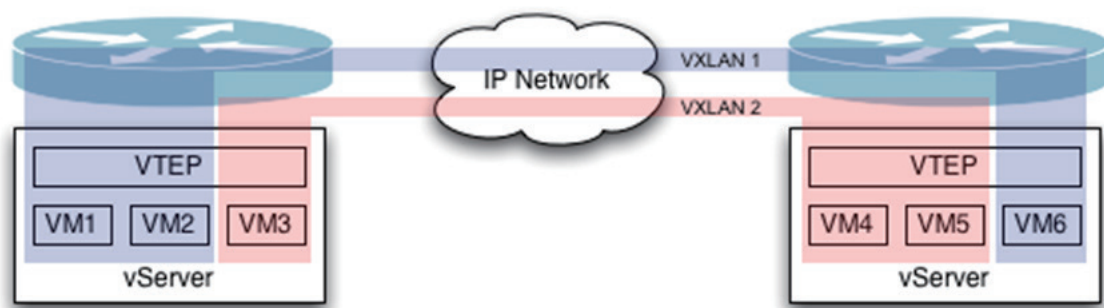


图 3 VxLAN 虚拟网络的示意图

4.3 VxLAN 已知单播流量转发

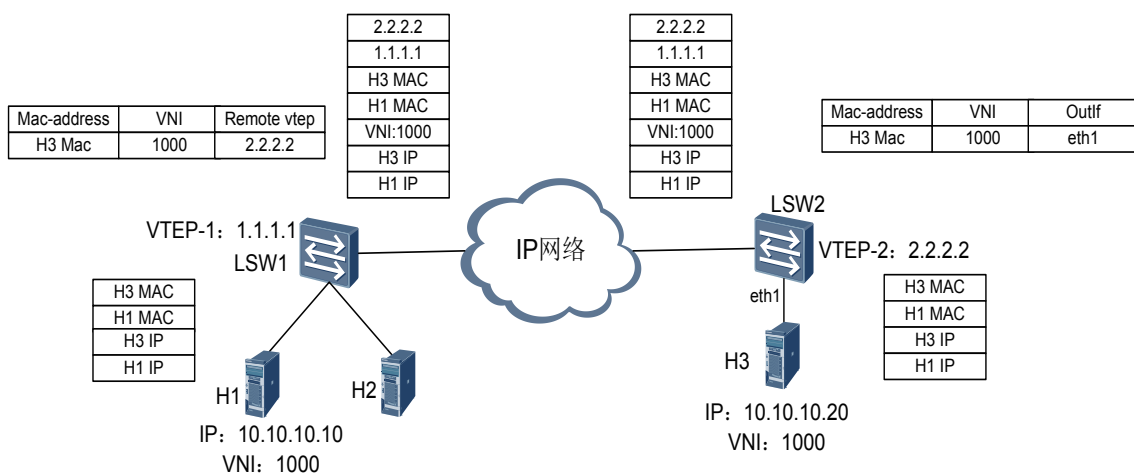


图 4 VxLAN 已知单播流量转发

- 1、LSW1 收到来自终端 H1 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，并判断报文的目的 MAC 是否为已知单播 MAC，如果是已知单播报文，则走已知单播流程，否则走 BUM 报文转发流程。
- 2、LSW1 判断是已知单播的情况下，LSW1 上根据查找到的 VNI 和远端 VTEP 地址进行 VxLAN 封装和报文转发。封装后的 VxLAN 报文通过 LSW1 和 LSW2 之间的 IP 网络进行转发。
- 3、LSW2 上收到 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、LSW2 根据内层二层报文的目的 MAC，查 MAC 表找到的出接口和封装信息，为报文添加 VLAN Tag，转发给对应的终端 H3。

4.4 VxLAN BUM 流量转发

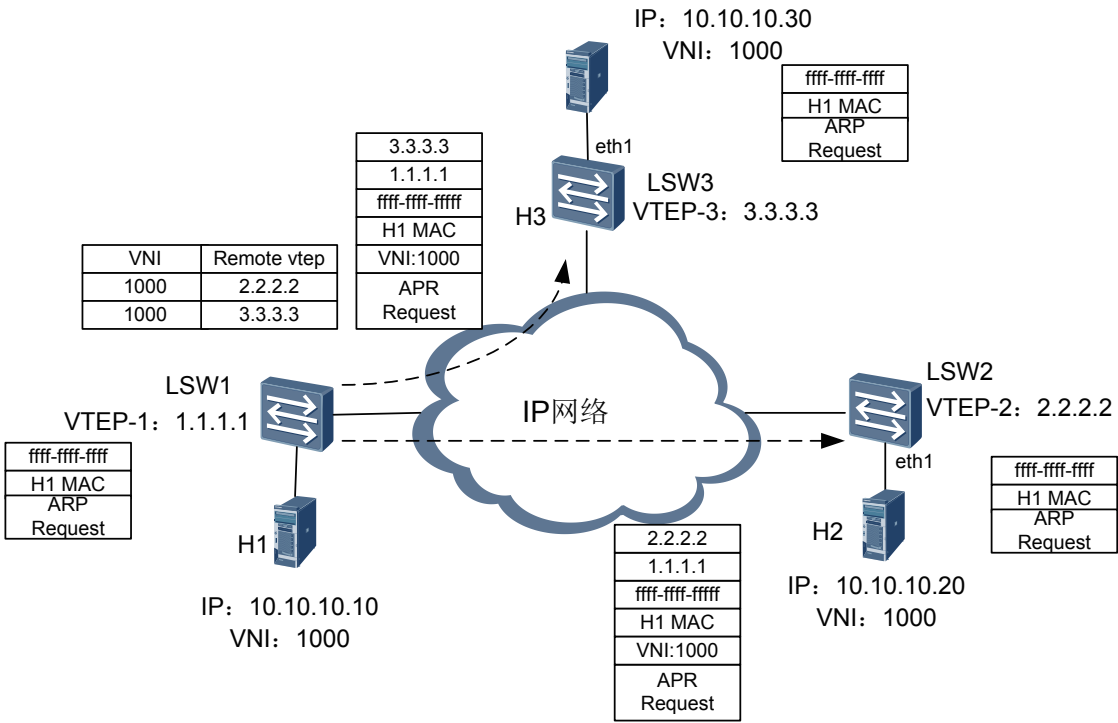


图 5 VxLANBUM 流量转发

- 1、LSW1 收到来自终端 H1 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 LSW1 判断是 BUM 报文的情况下，LSW1 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、LSW2/LSW3 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、LSW2/LSW3 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封装和转发，否则将报文广播到 LSW 的用户侧端口。

注：VxLAN 侧的 BUM 报文只能向用户侧广播，不能广播到 VxLAN 侧，否则会导致环路。

4.5 VxLAN 技术的价值

随着数据中心在物理网络基础设施上实施服务器虚拟化的快速发展，VxLAN 技术具有以下价值：

- 1、通过 24 比特的 VNI 可以支持多达 16M 的 VxLAN 段的网络隔离，对用户进行隔离和标识不再受到限制，可满足海量租户。
- 2、除 VxLAN 网络边缘设备，网络中的其他设备不需要识别虚拟机的 MAC 地址，减轻了设备的 MAC 地址学习压力，提升了设备性能。
- 3、通过采用 MAC in UDP 封装来延伸二层网络，实现了物理网络和虚拟网络解耦，租户可以规划自己的虚拟网络，不需要考虑物理网络 IP 地址和广播域的限制，让云数据中心组网更加灵活，扩容更加方便，不再有二层网络的限制大大降低了网络管理的难度。

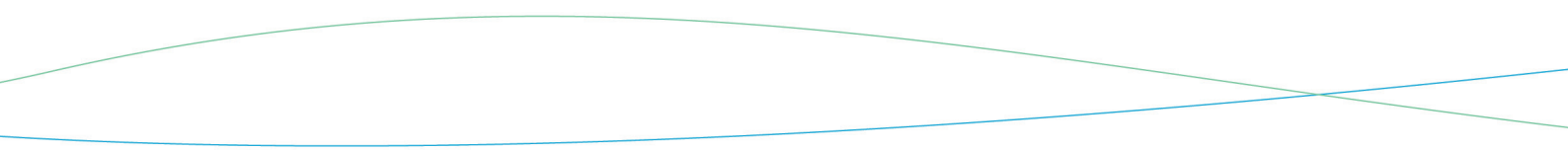
5 解决方案

5.1 组网模型

目前，交换机支持通过单机方式和控制器方式来部署 VxLAN 网络。

- 单机方式：传统网络部署方式，按照网络规划登录到每台设备上配置。云计算数据中心中，此方式无法协同云平台实现网络的自动化部署。
- 控制器方式：在大二层网络中，为了方便控制与部署引入了控制器。控制器是统一的网络控制平台，实现网络资源统一协调及管理，协同云平台实现业务和网络的自动化部署。

SDN 方案中，通常需要采用控制器方式部署 VxLAN 网络。



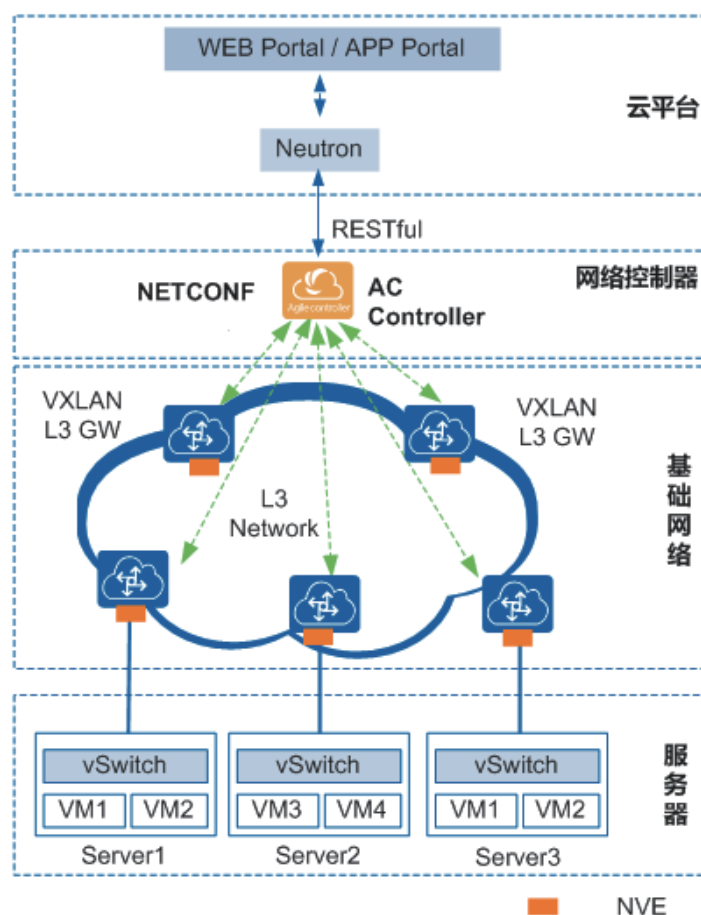


图 6 基于 AC 控制器 +VxLAN 解决方案的网络框架示意图

云平台：实现对网络资源、计算资源和存储资源的按需调度，提供业务管理、运维界面。Neutron 为云平台的组件，用于提供网络服务。

网络控制器：完成网络建模和网络实例化。AC 控制器通过 RESTful 接口，接收来自云平台网络模型配置信息，并转化为相应的配置命令，并通过 NETCONF 协议与网络设备对接建立的通道，将命令下发到基础网络层的网络设备。

基础网络：物理网络和虚拟网络统一规划，支持基于硬件的 VxLAN 网关提高业务性能，同时支持对传统 VLAN 网络的兼容。

5.2 VxLAN 二层网关

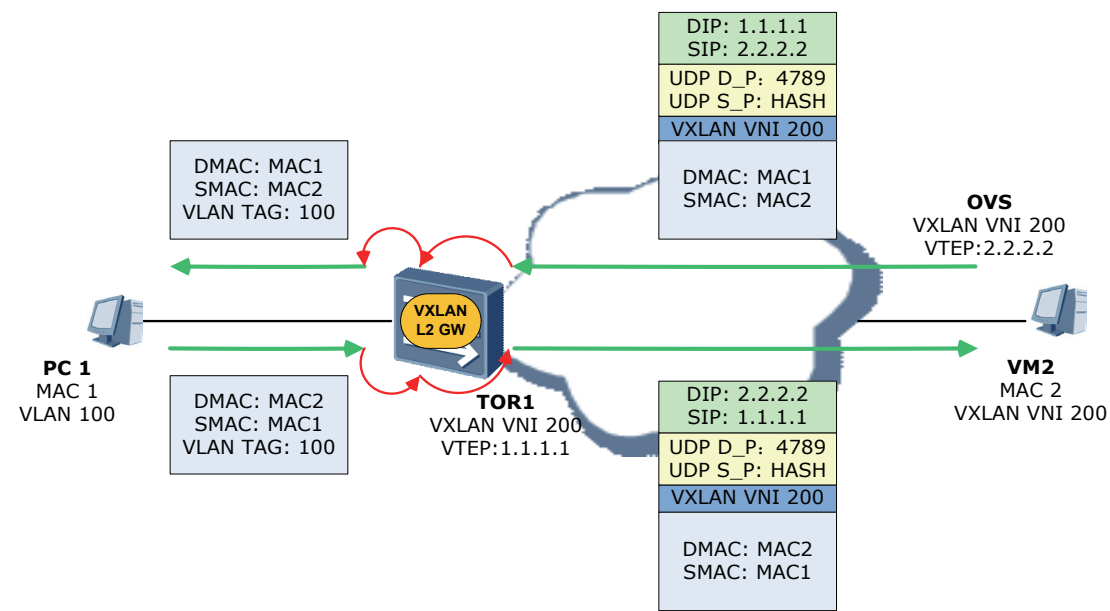


图 7 VxLAN 二层网关

VxLAN 二层网关用于传统以太网络和虚拟化 VxLAN 网络的互通, 如上图所示, VxLAN 二层网关工作流程如下:

- 1、VxLAN 二层网关从传统网络收到以太网报文, 根据报文接入的 Port 和 VLAN 信息获取对应的二层广播域; 查 MAC 表找到出接口和封装信息, 进行封装与转发; 缺省需要剥掉原二层头的 VLAN TAG; 然后新加 VxLAN 封装, 包含 VxLAN 头 VNI、VxLAN UDP 头、VxLAN 隧道头 (DIP\SIP) 以及外层二层头; 完成 VxLAN 报文封装后, 根据 VxLAN 隧道头目的 IP 查路由表发送报文。
- 2、VxLAN 二层网关从 VxLAN 网络收到 VxLAN 报文, 首先根据 VxLAN 隧道表检查 VxLAN UDP D_PORT、VxLAN 隧道 DIP\SIP、VxLAN VNI 的合法有效性; 然后根据 VxLAN VNI 获取对应的二层广播域; 然后解 VxLAN 封装, 获取内层二层报文, 根据二层报文目的 MAC 地址查 MAC 表发送报文。

5.3 VxLAN 集中式三层网关

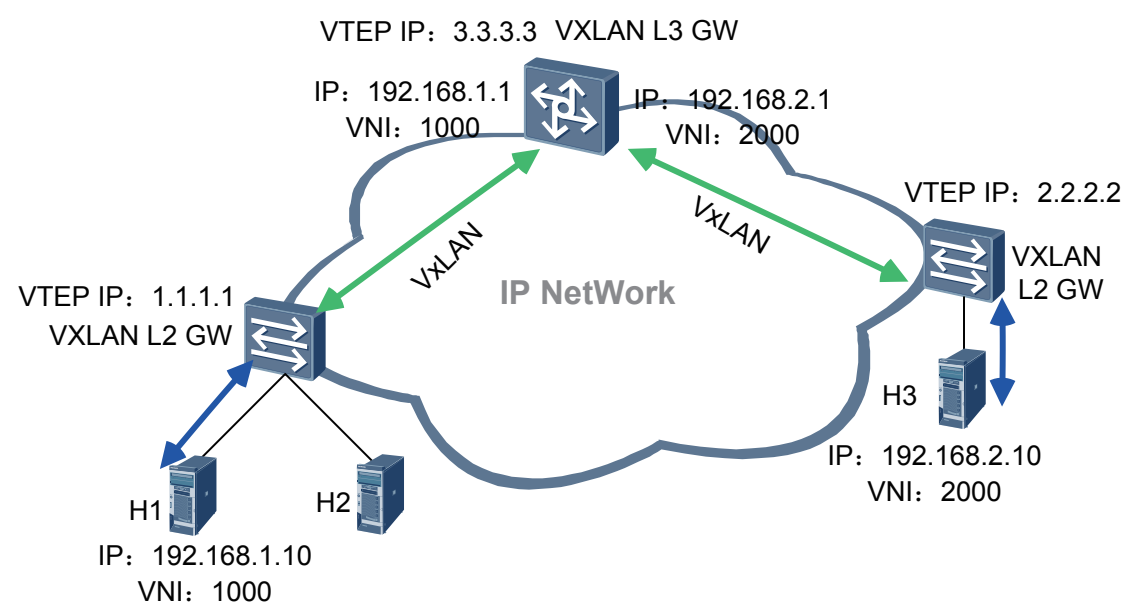


图 8 VxLAN 三层网关

不同网段的 VxLAN 网络之间通信，以及不同网段 VxLAN 网络和非 VxLAN 网络的通信，需要通过 VxLAN 三层网关实现，如上图所示，H1 和 H3 属于两个不同的 VxLAN 网络，通过 VxLAN 三层网关互通，VxLAN 三层网关工作流程如下：

- 1、VxLAN 三层网关收到 VxLAN 报文后解封装获取内层二层报文，判断二层报文目的 MAC 为本机 MAC，入三层转发流程。根据内层报文的目的 IP 地址查路由表，进行三层转发。
- 2、根据路由表获取下一跳 IP 地址，根据 IP 地址查 ARP 表，如果 ARP 表出接口为 VxLAN 隧道，则按照 ARP 表中的 VxLAN 隧道封装信息进行 VxLAN 封装，然后将报文发送出去。

6 典型应用

6.1 同网段终端用户通信的应用

数据中心用来在 Internet 网络基础设施上加速信息的传递，企业、运营商都在大力建设数据中心。目前，规模化、虚拟化、云计算已成为数据中心的发展方向，同时，数据中心为适应更大的业务量并降低维护成本，逐渐向大二层技术及虚拟化迁移。

随着数据中心在物理网络基础设施上实施服务器虚拟化的快速发展，作为 NVo3 技术之一的 VXLAN 技术具有很强的适应性，为数据中心提供了良好的解决方案。

如图示，某企业在不同的数据中心中都拥有 VM，且位于同一网段。现需要实现不同数据中心相同 ID 的 VM 的互通。

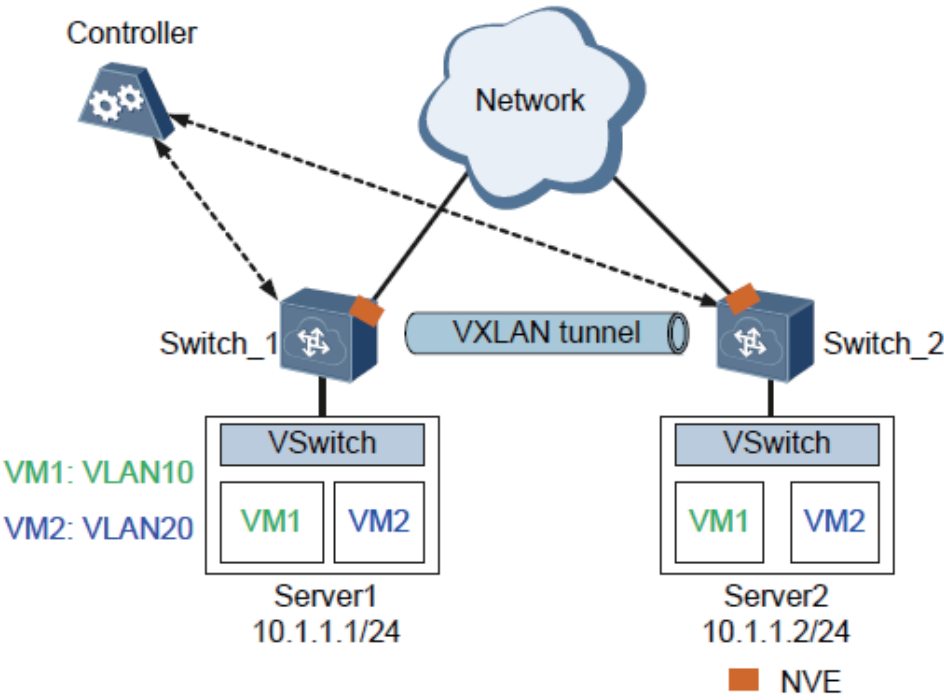


图 9 同网段终端用户通信的应用组网图

特性部署

将交换机设备作为 VXLAN 二层网关，交换机设备之间建立 VXLAN 隧道，通过 VXLAN 二层网关实现同一网段终端用户互通。

- 1、控制器和交换机设备之间通过 NETCONF 协议建立控制通道。
- 2、通过 NETCONF 协议，网管人员在控制器上配置 VXLAN。
- 3、交换机根据收到的信息建立 VXLAN 二层网关，转发器间建立 VXLAN 隧道。
- 4、通过 VXLAN 隧道，依据 MAC 地址表项，实现同一网段终端用户 (VM) 互通。

6.2 不同网段终端用户通信的应用

如图示，某企业在不同的数据中心中都拥有 VM，且位于不同网段。现需要实现不同数据中心同一企业的 VM 互通

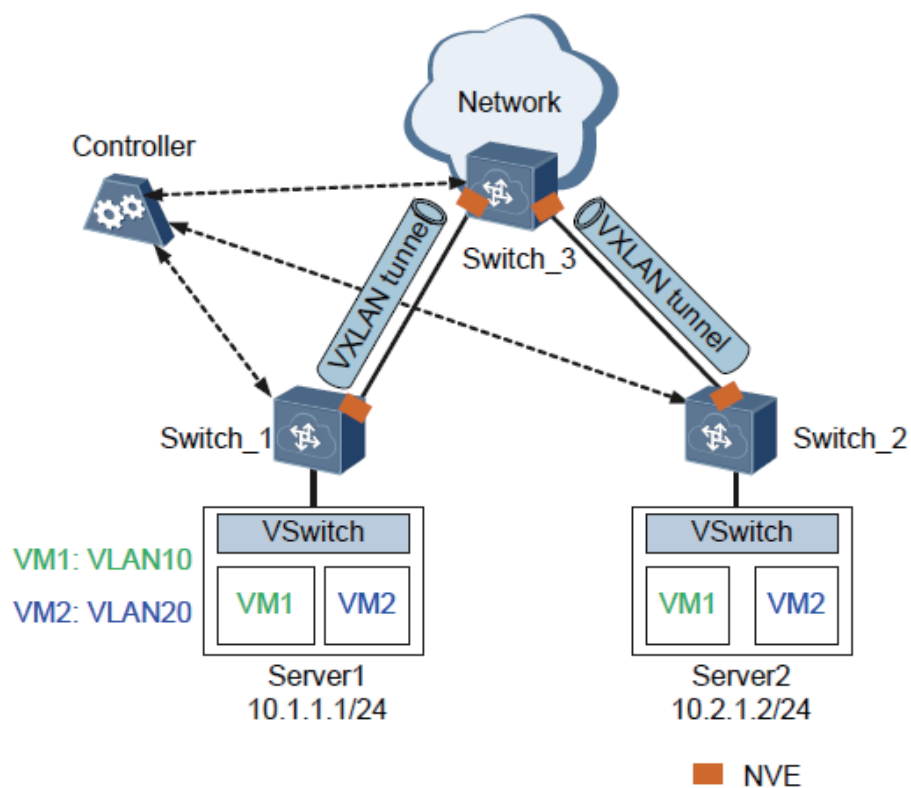


图 10 同网段终端用户通信的应用组网图

特性部署：

将 Switch_3 设备作为 VXLAN 三层网关，其他交换机设备作为 VXLAN 二层网关，交换机设备之间建立 VXLAN 隧道，通过 VXLAN 三层网关实现不同网段终端用户互通。

- 1、控制器和交换机设备之间通过 NETCONF 协议建立控制通道。
- 2、通过 NETCONF 协议，网管人员在控制器上配置 VXLAN。
- 3、交换机 Switch_1、Switch_2 根据收到的信息建立 VXLAN 二层网关，Switch 间建立 VXLAN 隧道。交换机 Switch_3 根据收到的信息建立 VXLAN 三层网关。
- 4、在控制器上，为 VXLAN 三层网关创建三层逻辑接口 VBDIF 接口配置 IP 地址。

6.3 在虚拟机迁移场景中的应用

当前数据中心网络中企业通过部署服务器虚拟化来达到整合 IT 资源、提升资源利用效率、降低开支的目的。随着虚拟化水平的不断提高，物理服务器上虚拟机的数量在不断增加，虚拟化环境下运行的应用数量也在不断上升，为虚拟网络带来了很大的挑战。

如图所示，某企业在数据中心中有两个群集 Cluster，其中工程部门和财务部门都在 Cluster1 上，营销部门在 Cluster2 上。Cluster1 上显示计算空间不足，而 Cluster2 未充分利用。网络管理员需要将工程部门迁移到 Cluster2 上，而不影响业务。

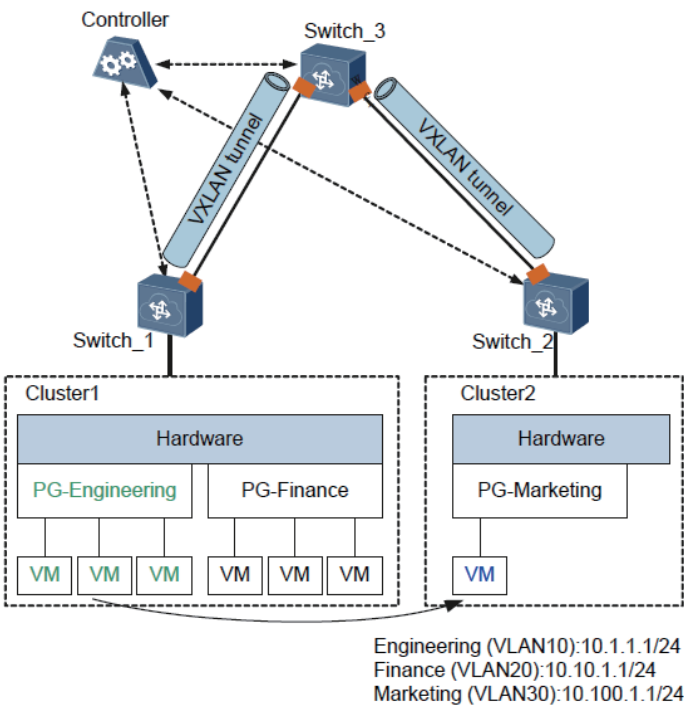


图 11 企业分布组网图

特性部署：

为了保证工程部门迁移过程中业务不中断，则需要保证工程部门的 IP 地址、MAC 地址等参数保持不变，这就要求两个 Cluster 属于一个二层网络。如果使用传统方法解决此问题，这可能需要网络管理员购买新的物理设备以分离流量，并可能导致诸如 VLAN 散乱、网络成环以及系统和管理开销等问题。

为了成功将工程部门迁移到 Cluster2，可通过 VXLAN 实现。VXLAN 是 MAC in UDP 的网络虚拟化技术，只要物理网络支持 IP 转发，所有 IP 路由可达的终端用户即可建立一个大范围二层网络。

- 1、控制器和交换机设备之间通过 NETCONF 协议建立控制通道。
- 2、通过 NETCONF 协议，网管人员在控制器上配置 VXLAN。
- 3、交换机根据收到的信息建立 VXLAN 隧道和 VXLAN 三层网关。
- 4、通过 VXLAN 隧道和 VXLAN 三层网关，工程部门在迁移过程中可保证网络无感知。

工程部门的迁移流程如下：

- 1、工程部门从 Cluster1 迁移到 Cluster2。
- 2、工程部门对应的 VM 发送免费 ARP 报文或 RARP（Reverse Address Resolution Protocol）报文告知 Switch_2 和其他设备。
- 3、Switch_2 和 Switch_1 收到 VM 的免费 ARP 报文或 RARP 报文后，更新 MAC 表，更新为迁移后的 VM 对应的 MAC 地址表。
- 4、Switch_3 收到 VM 的免费 ARP 报文或 RARP 报文后，更新 ARP 表和 MAC 表，更新为迁移后的 VM 对应的 MAC 地址表和 ARP 表。

通过 VXLAN 隧道，工程部门在迁移过程中可保证网络无感知。工程部门从 Cluster1 迁移到 Cluster2 后，终端租户会发送免费 ARP 或 RARP 报文，所有网关设备上保存的原 VM 对应的 MAC 地址表和 ARP 表都将会被刷新，更新为迁移后的 VM 对应的 MAC 地址表和 ARP 表。

版权所有 © 华为技术有限公司 2018。保留一切权利。

非经华为技术有限公司书面同意，任何单位和个人不得擅自摘抄、复制本手册内容的部分或全部，并不得以任何形式传播。

商标声明

、HUAWEI、华为、是华为技术有限公司的商标或者注册商标。

在本手册中以及本手册描述的产品中，出现的其他商标、产品名称、服务名称以及公司名称，由其各自的所有人拥有。

免责声明

本文档可能含有预测信息，包括但不限于有关未来的财务、运营、产品系列、新技术等信息。由于实践中存在很多不确定因素，可能导致实际结果与预测信息有很大的差别。因此，本文档信息仅供参考，不构成任何要约或承诺。华为可能不经通知修改上述信息，恕不另行通知。

华为技术有限公司
深圳龙岗区坂田华为基地
邮编：518129
电话：+86 755 28780808

www.huawei.com