# H3C SDN Overlay技术白皮书

# 目　录

# 1 概述

## 1.1 产生背景

随着企业业务的快速扩展，IT 作为基础设施，其快速部署和高利用率成为主要需求。云计算可以为之提供可用的、便捷的、按需的资源，成为当前企业 IT 建设的常规形态，而在云计算中大量采用和部署的虚拟化几乎成为一个基本的技术模式。部署虚拟机需要在网络中无限制地迁移到目的物理位置，虚拟机增长的快速性以及虚拟机迁移成为一个常态性业务。传统的网络已经不能很好满足企业的这种需求，面临着如下挑战：

- 虚拟机迁移范围受到网络架构限制

虚拟机迁移的网络属性要求，当其从一个物理机上迁移到另一个物理机上，虚拟机需要不间断业务，因而需要其 IP 地址、MAC 地址等参数维持不变，如此则要求业务网络是一个二层网络，且要求网络本身具备多路径多链路的冗余和可靠性。传统的网络生成树（STP，Spaning Tree Protocol）技术不仅部署繁琐，且协议复杂，网络规模不宜过大，限制了虚拟化的网络扩展性。基于各厂家私有的 IRF/vPC 等设备级的（网络 N:1）虚拟化技术，虽然可以简化拓扑、具备高可靠性，但是对于网络有强制的拓扑形状要求，在网络的规模和灵活性上有所欠缺，只适合小规模网络构建，且一般适用于数据中心内部网络。

- 虚拟机规模受网络规格限制

在大二层网络环境下，数据流均需要通过明确的网络寻址以保证准确到达目的地，因此网络设备的二层地址表项大小（即 MAC 地址表），成为决定了云计算环境下虚拟机的规模上限，并且因为表项并非百分之百的有效性，使得可用的虚拟机数量进一步降低。特别是对于低成本的接入设备而言，因其表项一般规格较小，限制了整个云计算数据中心的虚拟机数量，但如果其地址表项设计为与核心或网关设备在同一档次，则会提升网络建设成本。虽然核心或网关设备的 MAC 与 ARP 规格会随着虚拟机增长也面临挑战，但对于此层次设备能力而言，大规格是不可避免的业务支撑要求。减小接入设备规格压力的做法可以是分离网关能力，如采用多个网关来分担虚拟机的终结和承载，但如此也会带来成本的巨幅上升。

- 网络隔离/分离能力限制

当前的主流网络隔离技术为 VLAN（或 VPN），在大规模虚拟化环境部署会有两大限制：一是 VLAN 数量在标准定义中只有 12 个比特单位，即可用的数量为 4K，这样的数量级对于公有云或大型虚拟化云计算应用而言微不足道，其网络隔离与分离要求轻而易举会突破 4K；二是 VLAN 技术当前为静态配置型技术，这样使得整个数据中心的网络几乎为所有 VLAN 被允许通过（核心设备更是如此），导致任何一个 VLAN 的未知目的广播数据会在整网泛滥，无节制消耗网络交换能力与带宽。

上述的三大挑战，完全依赖于物理网络设备本身的技术改良，目前看来并不能完全解决大规模云计算环境下的问题，一定程度上还需要更大范围的技术革新来消除这些限制，以满足云计算虚拟化的网络能力需求。在此驱动力基础上，逐步演化出 Overlay 网络技术。

## 1.2 技术优点

Overlay 是一种叠加虚拟化技术，主要具有以下优点：

- 基于 IP 网络构建 Fabric。无特殊拓扑限制，IP 可达即可；承载网络和业务网络分离；对现有网络改动较小，保护用户现有投资。
- 16M 多租户共享，极大扩展了隔离数量。
- 网络简化、安全。虚拟网络支持 L2、L3 等，无需运行 LAN 协议，骨干网络无需大量 VLAN Trunk。
- 支持多样化的组网部署方式，支持跨域互访。
- 支持虚拟机灵活迁移，安全策略动态跟随。
- 转发优化和表项容量增大。消除了 MAC 表项学习泛滥，ARP 等泛洪流量可达范围可控，且东西向流量无需经过网关。

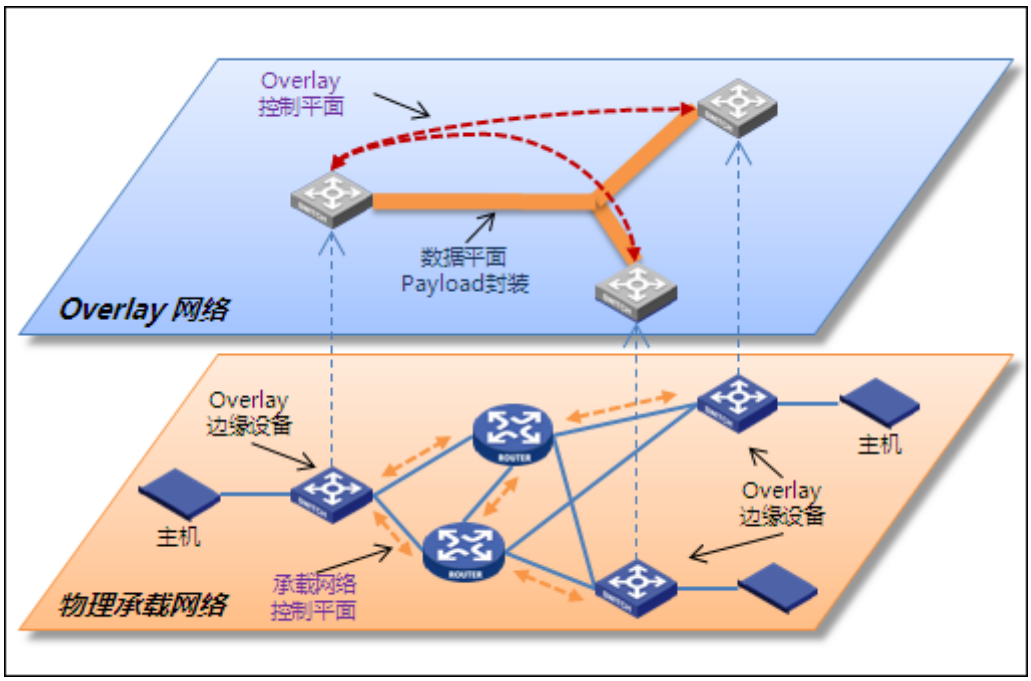
# 2 Overlay技术介绍

## 2.1 Overlay的概念介绍

在网络技术领域，Overlay是一种网络架构上叠加的虚拟化技术模式，其大体框架是对基础网络不进行大规模修改的条件下，实现应用在网络上的承载，并能与其它网络业务分离，并且以基于IP的基础网络技术为主（如 图 1 所示）。

- Overlay 网络是指建立在已有网络上的虚拟网，由逻辑节点和逻辑链路构成。
- Overlay 网络具有独立的控制和转发平面，对于连接在 overlay 边缘设备之外的终端系统来说，物理网络是透明的。
- Overlay 网络是物理网络向云和虚拟化的深度延伸，使云资源池化能力可以摆脱物理网络的重重限制，是实现云网融合的关键。

图1 Overlay 网络概念图




## 2.2 Overlay的解决方法

针对前文提到的三大挑战，Overlay 给出了完美的解决方法。

- 针对虚拟机迁移范围受到网络架构限制的解决方式

Overlay 把二层报文封装在 IP 报文之上，因此，只要网络支持 IP 路由可达就可以部署 Overlay 网络，而 IP 路由网络本身已经非常成熟，且在网络结构上没有特殊要求。而且路由网络本身具备良好的扩展能力，很强的故障自愈能力和负载均衡能力。采用 Overlay 技术后，企业不用改变现有网络架构即可用于支撑新的云计算业务，极方便用户部署。

- 针对虚拟机规模受网络规格限制的解决方式

虚拟机数据封装在 IP 数据包中后，对网络只表现为封装后的网络参数，即隧道端点的地址，因此，对于承载网络（特别是接入交换机），MAC 地址规格需求极大降低，最低规格也就是几十个（每个端口一台物理服务器的隧道端点 MAC）。当然，对于核心/网关处的设备表项（MAC/ARP）要求依然极高，当前的解决方案仍然是采用分散方式，通过多个核心/网关设备来分散表项的处理压力。

- 针对网络隔离/分离能力限制的解决方式

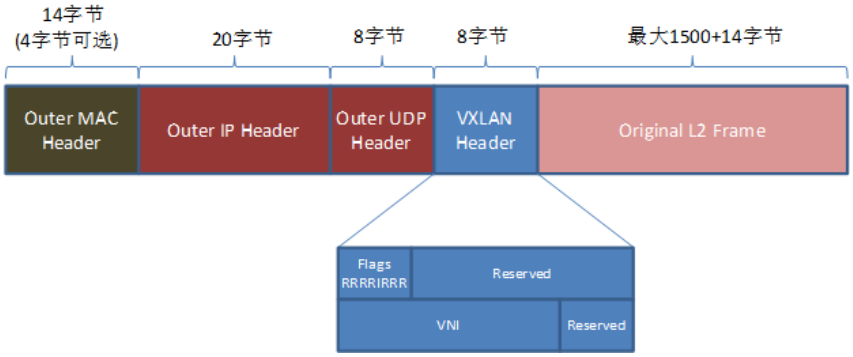
针对 VLAN 只能支持数量 4K 以内的限制，在 Overlay 技术中扩展了隔离标识的位数，可以支持高达 16M 的用户，极大扩展了隔离数量。

# 3 Overlay技术实现

## 3.1 Overlay网络基础架构

VXLAN（Virtual eXtensible LAN，可扩展虚拟局域网络）是基于IP网络、采用“MAC in UDP”封装形式的二层VPN技术，具体封装的报文格式如 图 2 所示。VXLAN可以基于已有的服务提供商或企业IP网络，为分散的物理站点提供二层互联功能，主要应用于数据中心网络。

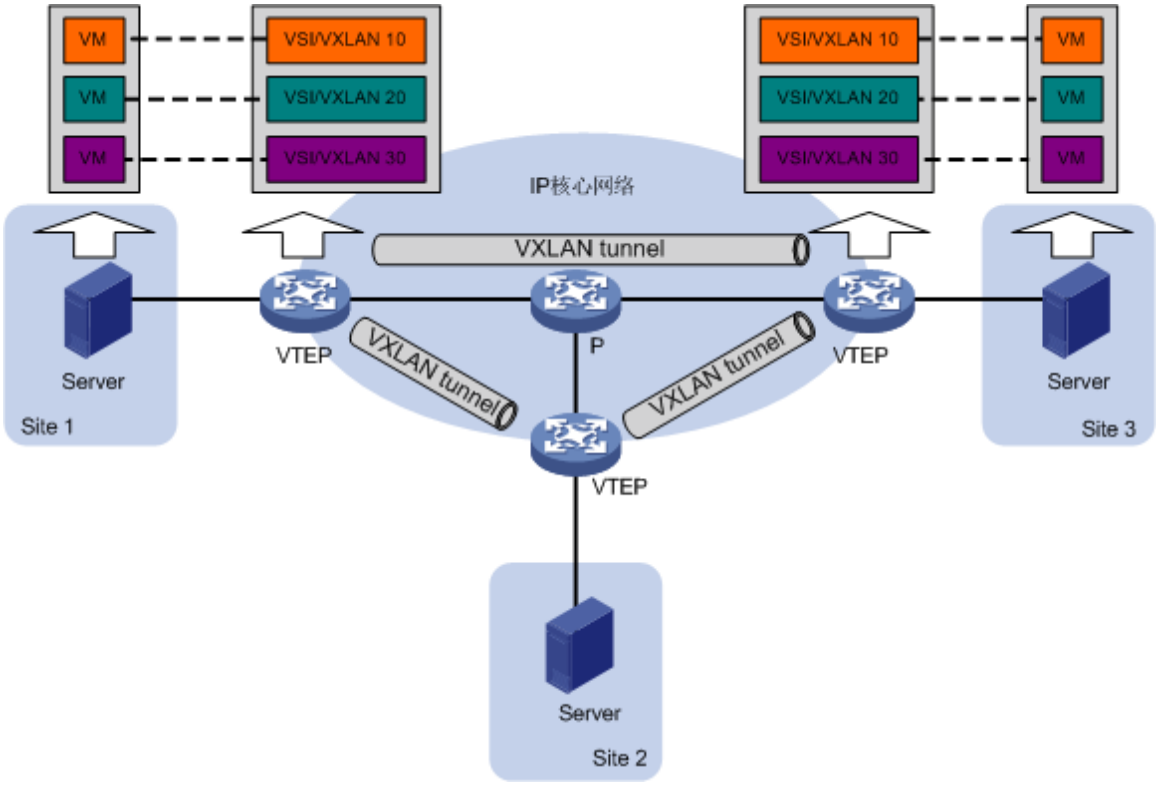
图2 VXLAN 报文格式




VXLAN 技术已经成为目前 Overlay 技术事实上的标准，得到了非常广泛的应用。

以VXLAN技术为基础的Overlay网络架构模型如 图 3 所示：

图3 Overlay 网络的基础架构




- VM（Virtual Machine，虚拟机）

在一台服务器上可以创建多台虚拟机，不同的虚拟机可以属于不同的 VXLAN。属于相同 VXLAN 的虚拟机处于同一个逻辑二层网络，彼此之间二层互通。

两个 VXLAN 可以具有相同的 MAC 地址，但在一个 VXLAN 范围段内不能有一个重复的 MAC 地址。

- VTEP（VXLAN Tunnel End Point，VXLAN 隧道端点）

VXLAN 的边缘设备，进行 VXLAN 业务处理：识别以太网数据帧所属的 VXLAN、基于 VXLAN 对数据帧进行二层转发、封装/解封装 VXLAN 报文等。

VXLAN 通过在物理网络的边缘设置智能实体 VTEP，实现了虚拟网络和物理网络的隔离。VTEP 之间建立隧道，在物理网络上传输虚拟网络的数据帧，物理网络不感知虚拟网络。VTEP 将从虚拟机发出/接受的帧封装/解封装，而虚拟机并不区分 VNI 和 VXLAN 隧道。

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

VXLAN 采用 24 比特标识二层网络分段，使用 VNI 来标识二层网络分段，每个 VNI 标识一个 VXLAN，类似于 VLAN ID 作用。VNI 占用 24 比特，这就提供了近 16M 可以使用的 VXLAN。VNI 将内部的帧封装（帧起源在虚拟机）。使用 VNI 封装有助于 VXLAN 建立隧道，该隧道在第三层网络之上覆盖第二层网络。

- VXLAN 隧道

在两个 VTEP 之间完成 VXLAN 封装报文传输的逻辑隧道。业务报文在入隧道时进行 VXLAN 头、UDP 头、IP 头封装后，通过三层转发透明地将封装后的报文转发给远端 VTEP，远端 VTEP 对其进行出隧道解封装处理。

- VSI（Virtual Switching Instance，虚拟交换实例）

VTEP 上为一个 VXLAN 提供二层交换服务的虚拟交换实例。

## 3.2 Overlay网络部署需求

### 3.2.1 VXLAN网络和传统网络互通的需求

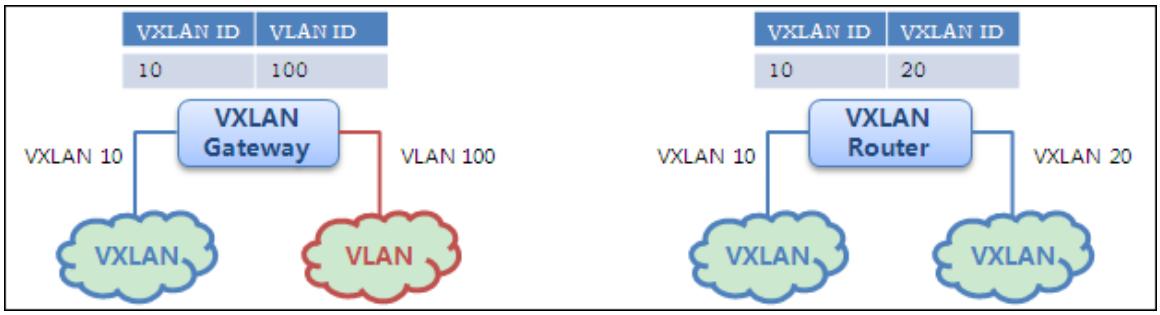
为了实现 VLAN 和 VXLAN 之间互通，VXLAN 定义了 VXLAN 网关。VXLAN 网关上同时存在 VXLAN 端口和普通端口两种类型端口，它可以把 VXLAN 网络和外部网络进行桥接、完成 VXLAN ID 和 VLAN ID 之间的映射和路由。和 VLAN 一样，VXLAN 网络之间的通信也需要三层设备的支持，即 VXLAN 路由的支持。同样 VXLAN 网关可由硬件设备和软件设备来实现。

当收到从 VXLAN 网络到普通网络的数据时，VXLAN 网关去掉外层包头，根据内层的原始帧头转发到普通端口上；当有数据从普通网络进入到 VXLAN 网络时，VXLAN 网关负责打上外层包头，并根据原始 VLAN ID 对应到一个 VNI，同时去掉内层包头的 VLAN ID 信息。相应的如果 VXLAN 网关发现一个 VXLAN 包的内层帧头上还带有原始的二层 VLAN ID，会直接将这个包丢弃。

如 图 4 左侧所示，VXLAN网关最简单的实现应该是一个Bridge设备，仅仅完成VXLAN到VLAN的转换，包含VXLAN到VLAN的 1：1、N：1 转换，复杂的实现可以包含VXLAN Mapping功能实现跨VXLAN转发，实体形态可以是vSwitch、物理交换机。

如 图 4 右侧所示，VXLAN路由器（也称为VXLAN IP GW）最简单的实现可以是一个Switch设备，支持类似VLAN Mapping的功能，实现VXLAN ID之间的Mapping，复杂的实现可以是一个Router设备，支持跨VXLAN转发，实体形态可以是NFV形态的路由器、物理交换机、物理路由器。

图4 VXLAN 网关和 VXLAN 路由简单实现




### 3.2.2 VXLAN网络安全需求

同传统网络一样，VXLAN 网络同样需要进行安全防护。

VXLAN 网络的安全资源部署需要考虑两个需求：

- VXLAN 和 VLAN 之间互通的安全控制

传统网络和 Overlay 网络中存在流量互通，需要对进出互通的网络流量进行安全控制，防止网络间的安全问题。针对这种情况，可以在网络互通的位置部署 VXLAN 防火墙等安全资源，VXLAN 防火墙可以兼具 VXLAN 网关和 VXLAN 路由器的功能，该功能可以称之为南北向流量安全。

- VXLAN ID 对应的不同 VXLAN 域之间互通的安全控制

VM 之间的横向流量安全是在虚拟化环境下产生的特有问题，在这种情况下，同一个服务器的不同 VM 之间的流量可能直接在服务器内部实现交换，导致外部安全资源失效。针对这种情况，可以考虑使用重定向的引流方法进行防护，又或者直接基于虚拟机进行防护，这个功能可以称之为东西向流量安全。

网络部署中的安全资源可以是硬件安全资源，也可以是软件安全资源，还可以是虚拟化的安全资源。

### 3.2.3 Overlay网络虚拟机位置无关性

通过使用 MAC-in-UDP 封装技术，VXLAN 为虚拟机提供了位置无关的二层抽象，Underlay 网络和 Overlay 网络解耦合。终端能看到的只是虚拟的二层连接关系，完全意识不到物理网络限制。

更重要的是，这种技术支持跨传统网络边界的虚拟化，由此支持虚拟机可以自由迁移，甚至可以跨越不同地理位置数据中心进行迁移。如此以来，可以支持虚拟机随时随地接入，不受实际所在物理位置的限制。

所以 VXLAN 的位置无关性，不仅使得业务可在任意位置灵活部署，缓解了服务器虚拟化后相关的网络扩展问题，而且使得虚拟机可以随时随地接入、迁移，是网络资源池化的最佳解决方式，可以有力地支持云业务、大数据、虚拟化的迅猛发展。

### 3.2.4 Overlay与SDN的结合

Overlay 技术与 SDN 可以说天生就是适合互相结合的技术组合。前面谈到的 Overlay 网络虚拟机物理位置无关特性就需要有一种强有力的集中控制技术进行虚拟机的管理和控制。而 SDN 技术恰好可以完美的做到这一点。接下来就让我们继续分析 Overlay 技术和 SDN 技术相结合带来的应用场景。

# 4 H3C SDN Overlay模型设计

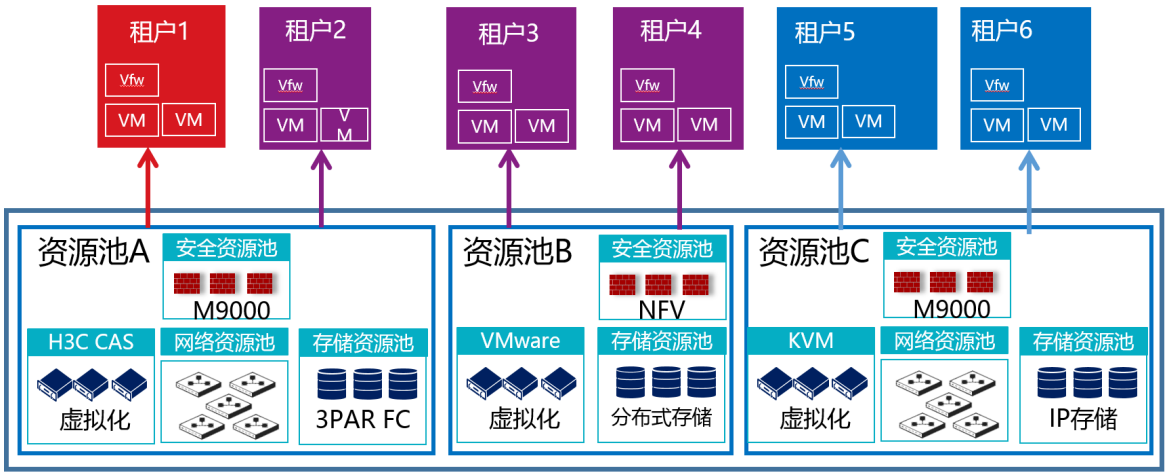
## 4.1 H3C SDN Overlay模型设计

在数据中心虚拟化多租户环境中部署和配置网络设施是一项复杂的工作，不同租户的网络需求存在差异，且网络租户是虚拟化存在，和物理计算资源位置无固定对应关系。通过传统手段部署物理网络设备为虚拟租户提供网络服务，一方面可能限制租户虚拟计算资源的灵活部署，另一方面需要网络管理员执行远超传统网络复杂度的网络规划和繁重的网络管理操作。在这种情况下，VPC（Virtual Private Cloud，虚拟私有云）技术就应运而生了。VPC 对于网络层面，就是对物理网络进行逻辑抽象，构架弹性可扩展的多租户虚拟私有网络，对于私有云、公有云和混合云同样适用。

H3C 的 SDN 控制器称为 VCF 控制器。H3C 通过 VCF 控制器控制 Overlay 网络从而将虚拟网络承载在数据中心传统物理网络之上，并向用户提供虚拟网络的按需分配，允许用户像定义传统 L2/L3 网络那样定义自己的虚拟网络，一旦虚拟网络完成定义，VCF 控制器会将此逻辑虚拟网络通过 Overlay 技术映射到物理网络并自动分配网络资源。VCF 的虚拟网络抽象不但隐藏了底层物理网络部署的复杂性，而且能够更好的管理网络资源，最大程度减少了网络部署耗时和配置错误。

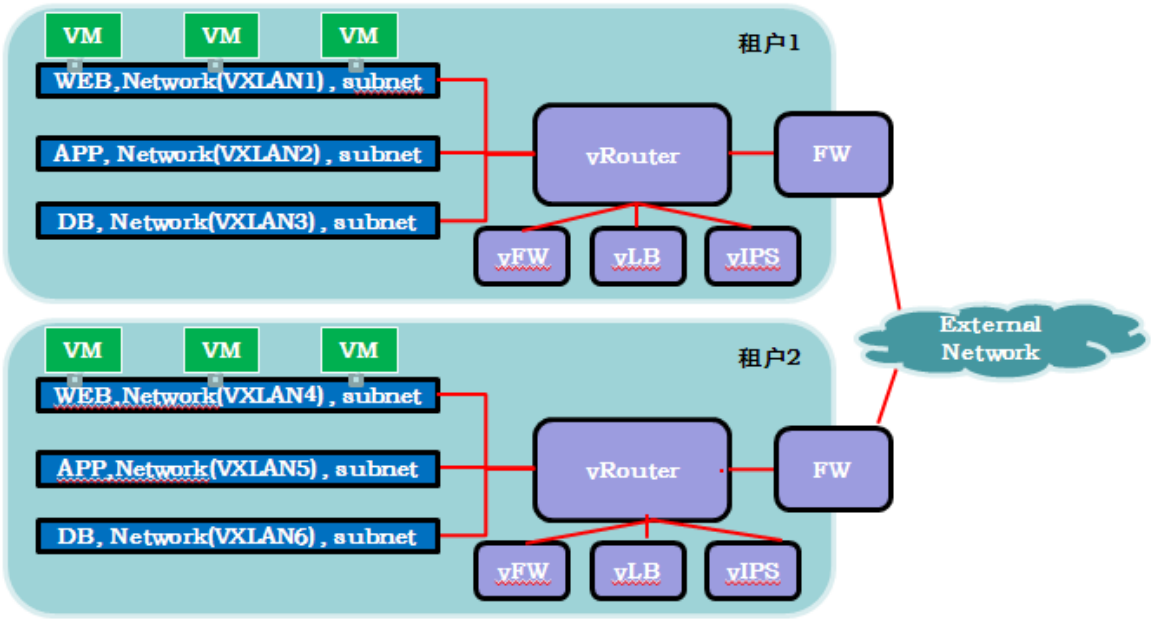
VCF 控制器将虚拟网络元素组织为“资源池”，VCF 控制器控制了“网络资源池”的按需分配，进而实现虚拟网络和物理网络的 Overlay 映射。

图5 VPC 多租户资源池场景




VCF控制器的虚拟网络元素的抽象方式与OpenStack网络模型兼容，如 图 6 所示：

图6 VPC 多租户资源池场景

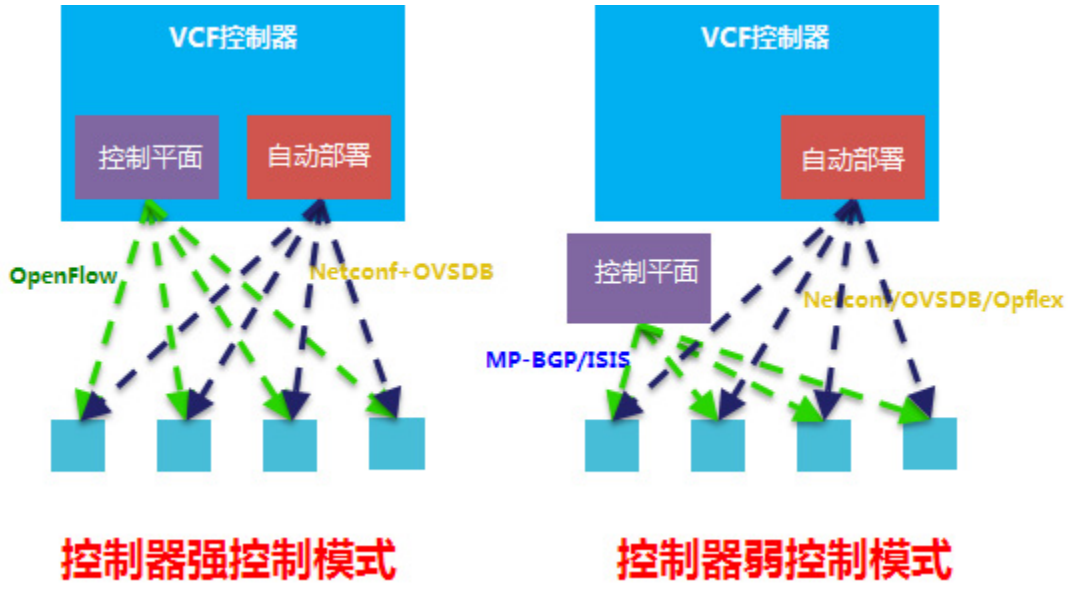



虚拟网络的各个要素如下表：

| 元素名称 | 描述 |
| --- | --- |
| Tenant | 租户 |
| Network | 一个虚拟的二层隔离网络，可以看作是一个虚拟或逻辑的交换机 |
| Subnet | 一个IPv4或IPv6地址块，对应于三层子网 |
| Port | 一个虚拟的或逻辑的交换机端口 |
| vRouter | 代表逻辑三层网关/网络，分散在各个虚拟设备上 |
| vFW、vLB、vIPS | 网络服务功能，为每个租户提供独立的FW、LB及IPS服务 |
| Security Group | vSwitch上的安全组功能 |

## 4.2 SDN控制器模型介绍

图7 SDN 控制器模型




从控制器是否参与转发设备的转发控制来看，当前主要有两种控制器类型：

- 控制器弱控制模式

弱控制模式下，控制平面基于网络设备自学习，控制器不在转发平面，仅负责配置下发，实现自动部署。主要解决网络虚拟化，提供适应应用的虚拟网络。

弱控制模式的优点是转发控制面下移，减轻和减少对控制器的依赖。

- 控制器强控制模式

在强控制模式下，控制器负责整个网络的集中控制，体现 SDN 集中管理的优势。

基于 OpenFlow 的强控制使得网络具备更多的灵活性和可编程性。除了能够给用户提供适合应用需要的网络，还可以集成 FW 等提供安全方案；可以支持混合 Overlay 模型，通过控制器同步主机和拓扑信息，将各种异构的转发模型同一处理；可以提供基于 OpenFlow 的服务链功能对安全服务进行编排，可以提供更为灵活的网络诊断手段，如虚拟机仿真和雷达探测等。

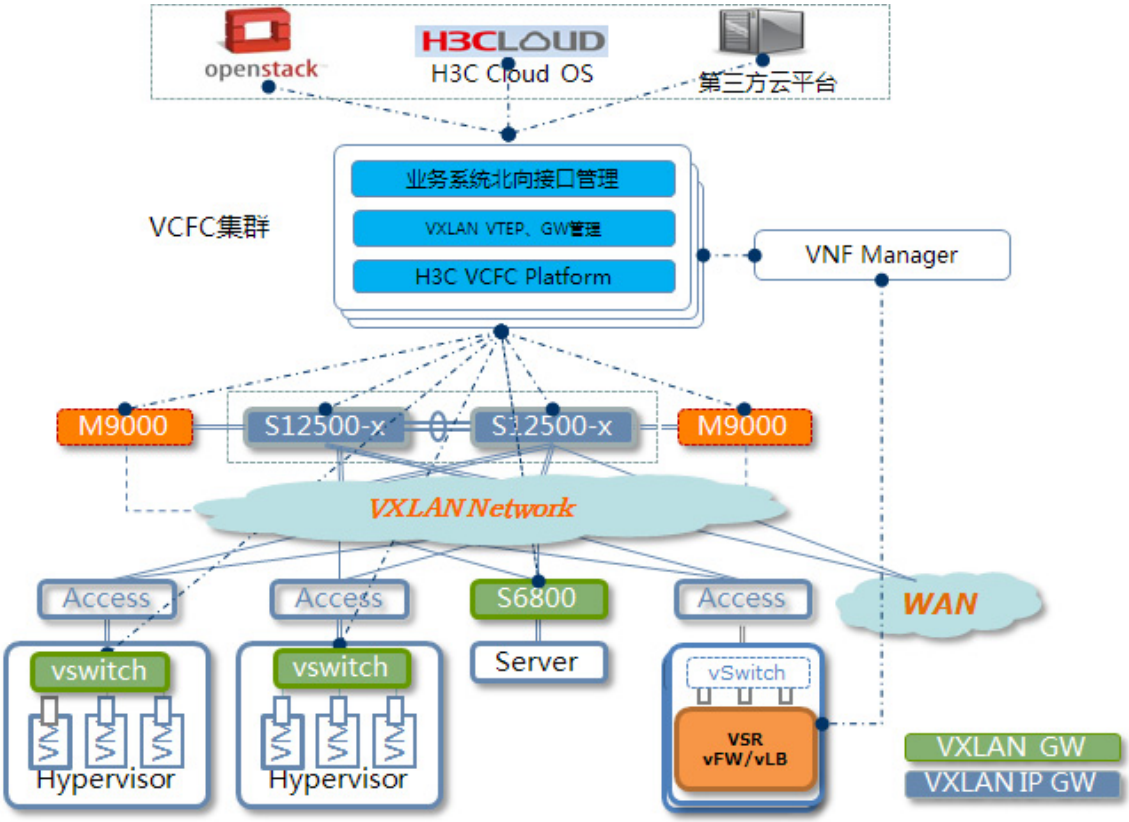
用户可能会担心强控制模式下控制器全部故障对网络转发功能的影响，这个影响因素可以通过下述两点来降低和消除：

- 通过控制器集群增加控制器可靠性，避免单点故障
- 逃生机制：设备与所有控制器失联后，切换为自转模式，业务不受影响。

考虑到强控制模式可以支持混合 Overlay 模型，可以额外支持安全、服务链等灵活、可编程的功能，并且可靠性又可以通过上述方式加强，我们建议使用强控制模式来实现 SDN Overlay。

## 4.3 H3C SDN Overlay组件介绍

图8 H3C SDN Overlay 组件




如 图 8 所示，H3C SDN Overlay主要包含如下组件：

- 云管理系统

可选，负责计算，存储管理的云平台系统，目前主要包括 OpenStack、VMware vCenter 和 H3Cloud OS。

- VCF Controller 集群

必选，VCF Controller 实现对于 VPC 网络的总体控制。

- VNF Manager

VNF Manager 实现对 NFV 设备（如 vFW、vLB）的生命周期管理。

- VXLAN GW

必选，VXLAN GW 包括 vSwitch、S6800、VSR 等，实现虚拟机、服务器等各种终端接入到 VXLAN 网络中。

- VXLAN IP GW

必选，VXLAN IP GW 包括 S12500-X、S9800、VSR 等，实现 VXLAN 网络和经典网络之间的互通。

- 虚拟化平台

可选，vSwitch 和 VM 运行的 Hypervisor 平台，目前主要包括 CAS、VMware、KVM 等。

- Service 安全设备

可选，包括 VSR、vFW、vLB 和 M9000、安全插卡等设备，实现东西向和南北向服务链服务节点的功能。

## 4.4 SDN Overlay网络与云对接

公有云或私有云（VPC）对网络的核心需求：

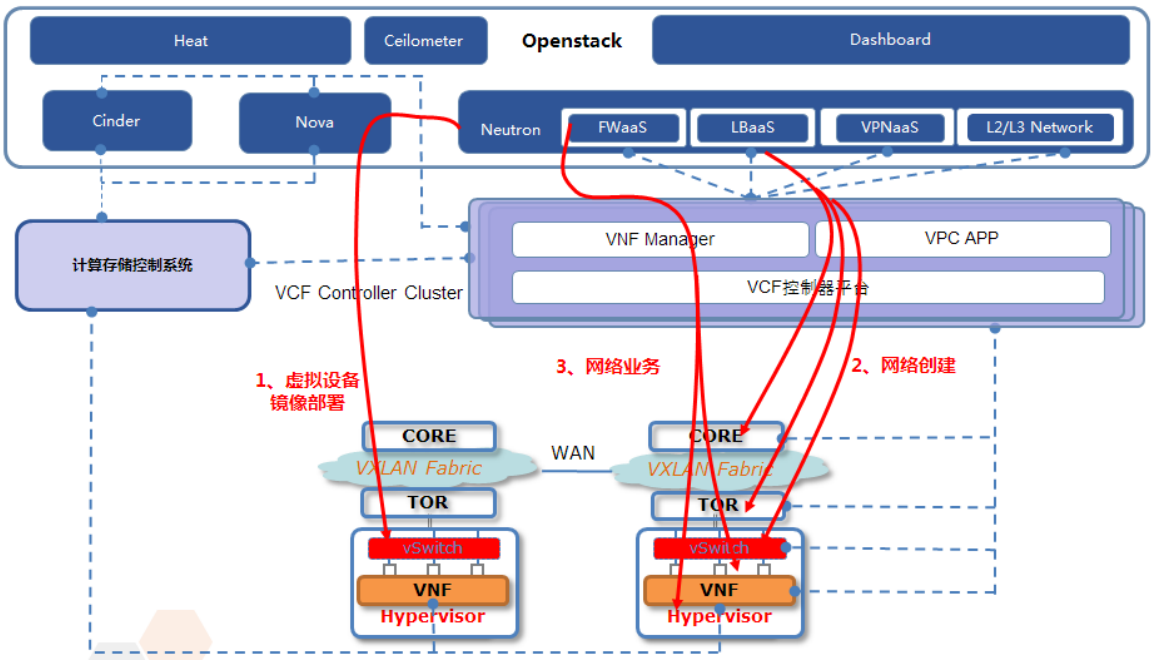
- 租户隔离
- 网络自定义
- 资源大范围灵活调度
- 应用与网络位置无关
- 网络资源池化与按需分配
- 业务自动化

H3C 提出的解决方案：

- 利用 VXLAN Overlay 提供一个“大二层”网络环境，满足资源灵活调度的需求；
- 由 SDN 控制器 VCFC 实现对整个 Overlay 网络的管理和控制；
- 由 VXLAN GW 实现服务器到 VXLAN 网络的接入；
- 由 VXLAN IP GW 实现 VXLAN 网络与传统网络的对接；
- NFV 设备（VSR/vFW/vLB）实现东西向和南北向服务链服务节点的功能；
- SDN 控制器与云管理平台对接，可实现业务的自动化部署。

### 4.4.1 SDN Overlay与OpenStack对接

图9 SDN Overlay 与 OpenStack 对接




如 图 9 所示，与标准的OpenStack对接：采用在Neutron Server中安装VCFC插件的方式，接管OpenStack网络控制。OpenStack定义的插件如 表 1 所示：

表1 OpenStack 定义的插件

| 可对接 Neutron 插件举例 | 可对接对象举例 |
|---|---|
| ml2 | network |
| | subnet |
| | port |
| l3 | router |
| | floatingip |
| vpnaas | vpnservice/ikepolicy |
| fwaas | firewall/firewall_policy |
| lbaas | memberpool |

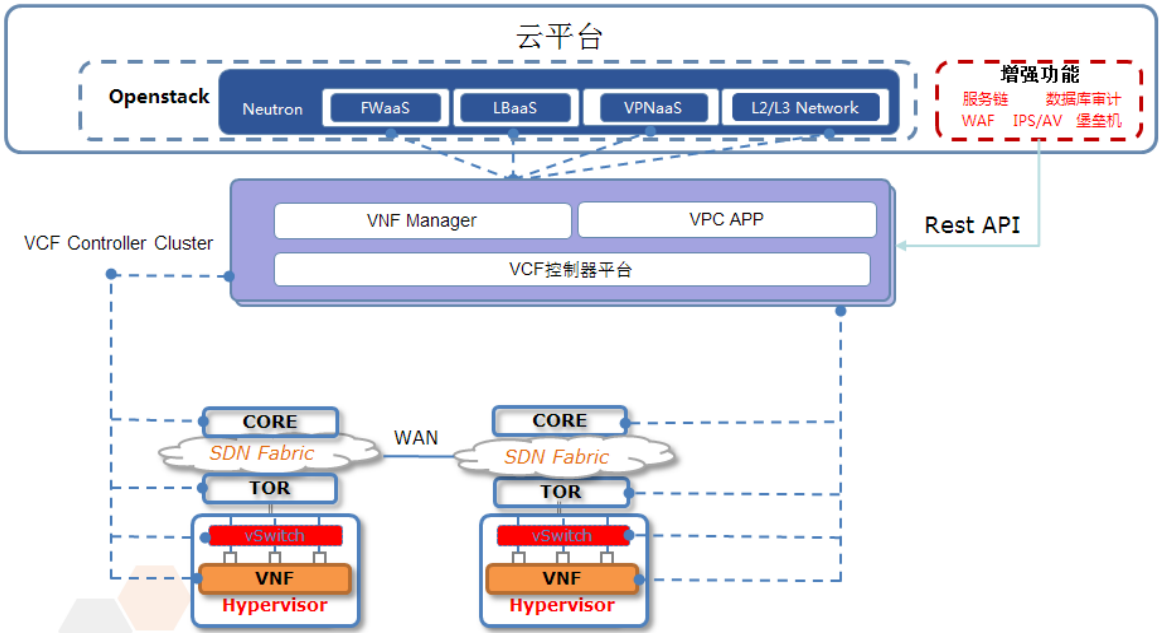
OpenStack 插件类似于一个硬件驱动，以网络组件 Neutron 为例，Neutron 本身实现抽象的虚拟网络功能，Neutron 先调用插件把虚拟网络下发到 VCFC，然后由 VCFC 下发到具体的设备上。插件可以是核心组件也可以是一项服务：核心插件实现“核心”的 Neutron API——二层网络和 IP 地址管理。服务插件提供“额外”的服务，例如三层路由、负载均衡、VPN、防火墙和计费等。

H3C VCFC 实现了上述插件，在插件里通过 REST API 把 Nuetron 的配置传递给 VCFC，VCFC 进行网络业务编排通过 OpenFlow 流表等手段下发到硬件交换机、NFV 以及 vSwitch 上，以实现相应的网络和服务功能。

VCFC 与 H3Cloud OS 对接也是采用 Neutron 插件的方式。

### 4.4.2 SDN Overlay与基于OpenStack的增强云平台对接

图10 SDN Overlay 与基于 OpenStack 的增强云平台对接

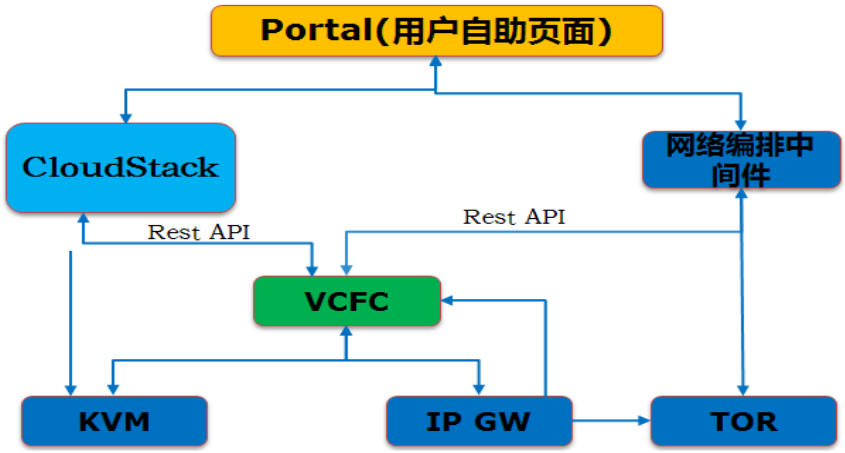



考虑到 OpenStack 标准版本不一定都能满足用户的需求，很多基于 OpenStack 开发的云平台都在 Oenstack 基础之上进行了增强开发，以满足自己特定的需求。

与这类增强的 OpenStack 版本对接时，基础的网络和安全服务功能仍通过插件形式对接；标准 OpenStack 版本的 Nuetron 组件未定义的增强功能，如服务链、IPS/AV 等等，通过 Rest API 对接。

### 4.4.3 SDN Overlay与非OpenStack云平台对接

图11 SDN Overlay 与非 OpenStack 云平台对接




以 CloudStack 为例，VCFC 与非 OpenStack 云平台的对接通过 Rest API 进行，H3C 提供了完整的用于实现虚拟网络及安全功能的 Rest API 接口。云平台调用这些接口来实现 VM 创建、删除、上线等一系列流程。

## 4.5 服务链在Overlay网络安全中的应用

### 4.5.1 什么是服务链

数据报文在网络中传递时，首先按特定策略进行流分类，再按照一定顺序经过一组抽象业务功能的节点，完成对应业务功能处理，这种打破了常规网络转发逻辑的方式，称为服务链。

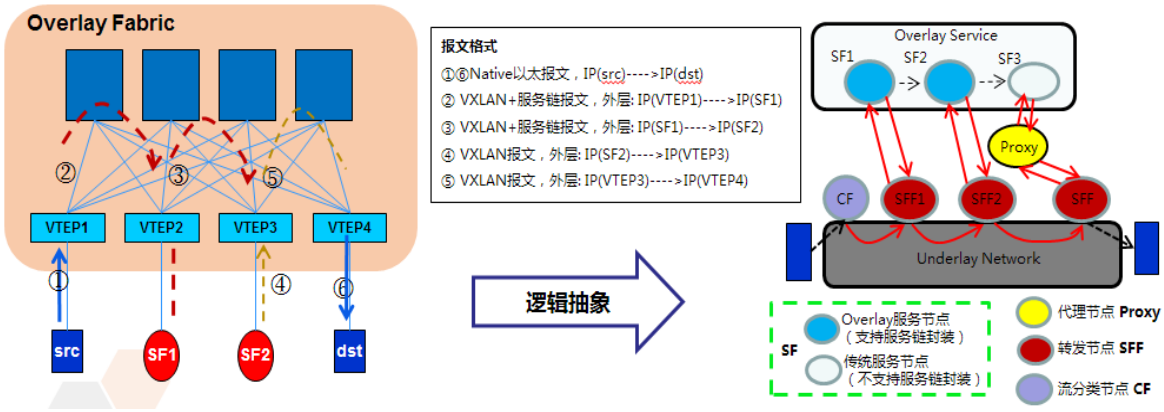
服务链常见的服务节点（Service Node）：防火墙（FW）、负载均衡（LB）、入侵检测（IPS）、VPN 等。

H3C VCF 控制器支持集中控制整个服务链的构建与部署，将 NFV 形态或硬件形态的服务资源抽象为统一的服务资源池，实现服务链的自定义和统一编排。

服务链在实现 Overlay 网络安全方面有独到的优势，服务链方案/VxLAN 终结方案除了能够满足 OpenStack FWaaS、LBaaS 定义外，还能提供更灵活的 FW/LB 编排方案。

### 4.5.2 Overlay网络服务链节点描述

图12 Overlay 网络服务链节点描述




如 图 12 所示，Overlay网络中的服务链主要由如下几个部件组成：

- 控制器（Controller）：VTEP 和 ServiceNode 上的转发策略都由控制器下发。
- 服务链接入节点（VTEP1）：通过流分类，确定报文是否需要进入服务链。需要进入服务链，则将报文做 VXLAN+服务链封装，转到服务链首节点处理。
- 服务链首节点（SN1）：服务处理后，将用户报文做服务链封装，交给服务链下一个节点。
- 服务链尾节点（SN2）：服务处理后，服务链尾节点需要删除服务链封装，将报文做普通 VXLAN 封装，转发给目的 VTEP。如果 SN2 不具备根据用户报文寻址能力，需要将用户报文送到网关 VTEP3，VTEP3 再查询目的 VTEP 发送。

### 4.5.3 服务链在Overlay网络安全中的应用

图13 Overlay 网络服务链流程

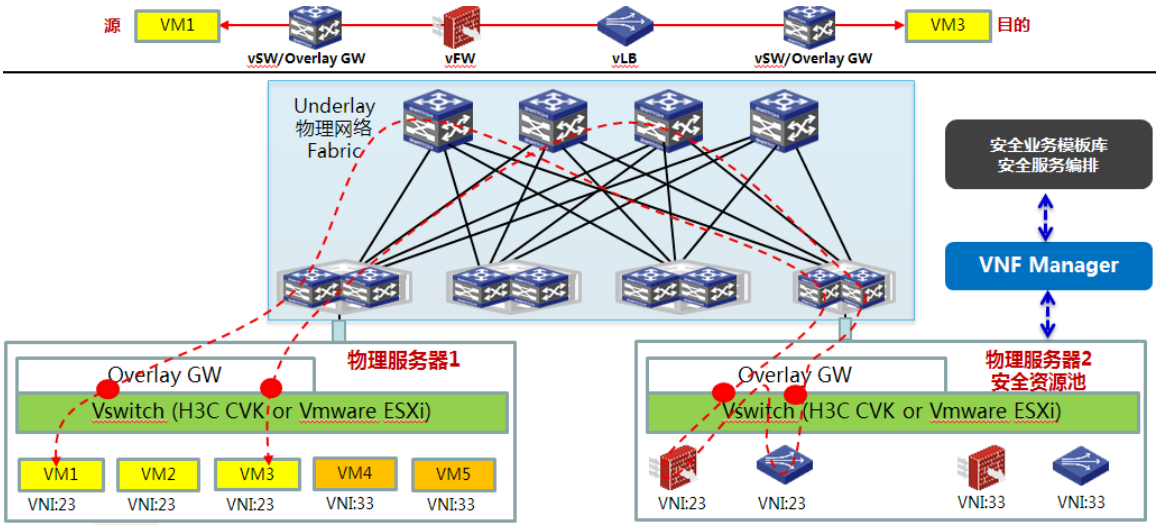


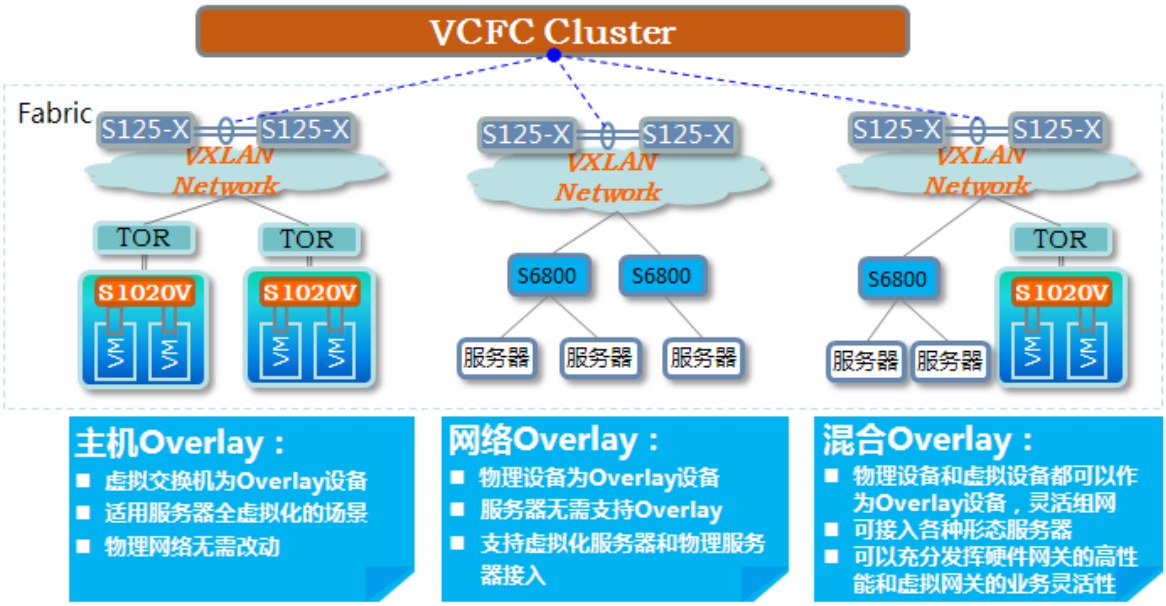
图 13 是一个基于SDN的服务链流程。SDN Controller实现对于SDN Overlay、NFV设备、vSwitch的统一控制；NFV提供虚拟安全服务节点；vSwitch支持状态防火墙的嵌入式安全；同时SDN Controller提供服务链的自定义和统一编排。我们看一下，假设用户自定义从VM1 的VM3 的业务流量，必须通过中间的FW和LB等几个环节。通过SDN的服务链功能，业务流量一开始就严格按照控制器的编排顺序经过这组抽象业务功能节点，完成对应业务功能的处理，最终才回到VM3，这就是一个典型的基于SDN的服务链应用方案。

# 5 SDN Overlay组网方案设计

Overlay 控制平面架构可以有多种实现方案，例如网络设备之间通过协议分布式交互的方式。而基于 VCF 控制器的集中式控制的 SDN Overlay 实现方案，以其易于与计算功能整合的优势，能够更好地使网络与业务目标保持一致，实现 Overlay 业务全流程的动态部署，在业界逐步成为主流的 Overlay 部署方案。

## 5.1 SDN Overlay组网模型

图14 SDN Overlay 组网模型




如上图所示，H3C 的 SDN Overlay 组网同时支持网络 Overlay、主机 Overlay 和混合 Overlay 三种组网模型：

- 网络 Overlay：在这种模型下，所有 Overlay 设备都是物理设备，服务器无需支持 Overlay，这种模型能够支持虚拟化服务器和物理服务器接入；
- 主机 Overlay：所有 Overlay 设备都是虚拟设备，适用服务器全虚拟化的场景，物理网络无需改动；
- 混合 Overlay：物理设备和虚拟设备都可以作为 Overlay 边缘设备，灵活组网，可接入各种形态服务器，可以充分发挥硬件网关的高性能和虚拟网关的业务灵活性。

三种 Overlay 商用模型都通过 VCF 控制器集中控制，实现业务流程的下发和处理，应该说这三种 Overlay 模型都有各自的应用场景。用户可根据自己的需求从上述三种 Overlay 模型和 VLAN VPC 方案中选择最适合自己的模型。

### 5.1.1 网络Overlay

#### 1. 定位

网络 Overlay 组网里的服务器可以是多形态，也无需支持 Overlay 功能，所以网络 Overlay 的定位主要是网络高性能、与 Hypervisor 平台无关的 Overlay 方案。

#### 2. 面向客户

网络 Overlay 主要面向对性能敏感而又对虚拟化平台无特别倾向的客户群。该类客户群的网络管理团队和服务器管理团队的界限一般比较明显。

### 5.1.2 主机Overlay

#### 1. 定位

主机 Overlay 不能接入非虚拟化服务器，所以主机 Overlay 主要定位是配合 VMware、KVM 等主流 Hypervisor 平台的 Overlay 方案。

#### 2. 面向客户

主机 Overlay 主要面向已经选择了虚拟化平台并且希望对物理网络资源进行利旧的客户。

### 5.1.3 混合Overlay

#### 1. 定位

混合 Overlay 组网灵活，既可以支持虚拟化的服务器，也可以支持利旧的未虚拟化物理服务器，以及必须使用物理服务器提升性能的数据库等业务，所以混合 Overlay 的主要定位是 Overlay 整体解决方案，它可以为客户提供自主化、多样化的选择。

#### 2. 面向客户

混合 Overlay 主要面向愿意既要保持虚拟化的灵活性，又需要兼顾对于高性能业务的需求，或者充分利旧服务器的要求，满足客户从传统数据中心向基于 SDN 的数据中心平滑演进的需求。

## 5.2 H3C SDN Overlay典型组网

### 5.2.1 网络Overlay

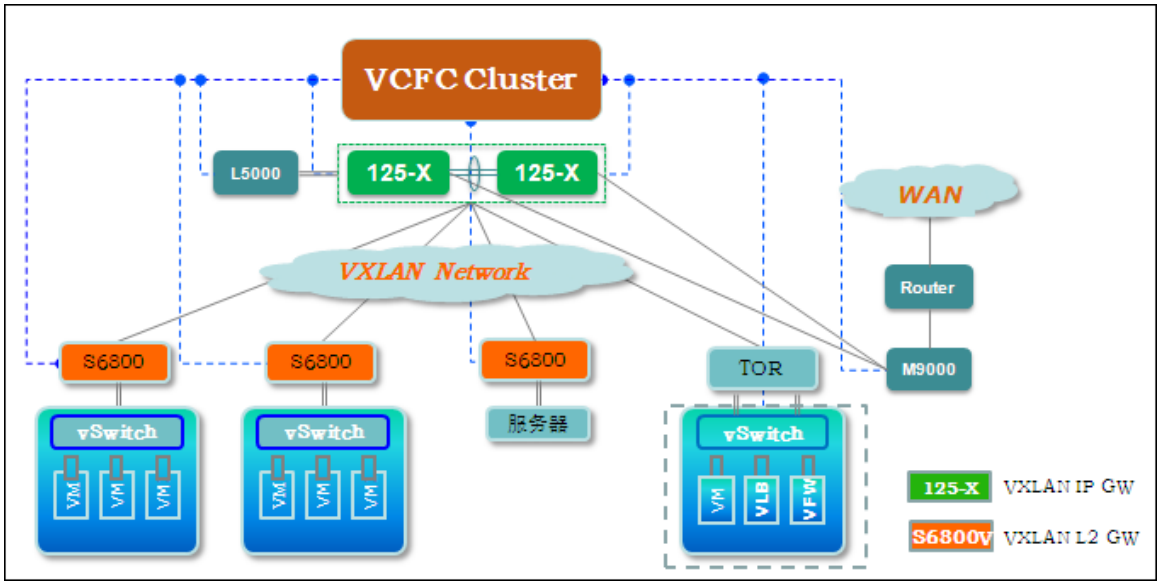
网络 Overlay 的隧道封装在物理交换机完成。这种 Overlay 的优势在于物理网络设备的转发性能比较高，可以支持非虚拟化的物理服务器之间的组网互通。

H3C 提供的网络 Overlay 组网方式，支持以下转发模式：

- 控制器流转发模式：控制器负责 Overlay 网络部署、主机信息维护和转发表项下发，即 VXLAN L2 GW 上的 MAC 表项由主机上线时控制器下发，VXLAN IP GW 上的 ARP 表项也由控制器在主机上线是自动下发，并由控制器负责代答和广播 ARP 信息。这种模式下，如果设备和控制器失联，设备会临时切换到自转发状态进行逃生。
- 数据平面自转发模式：控制器负责 Overlay 网络的灵活部署，转发表项由 Overlay 网络交换机自学习，即 VXLAN L2 GW 上自学习主机 MAC 和网关 MAC 信息，VXLAN IP GW 上可以自学习主机 ARP 信息并在网关组成员内同步。

- 混合转发模式：控制器也可以基于主机上线向 VXLAN IP GW 上下发虚机流表，如果 VXLAN IP GW 上自学习 ARP 和控制器下发的虚机流表信息不一样，则以 VXLAN IP GW 上自学习 ARP 表项为主，交换机此时触发一次 ARP 请求，保证控制器和交换机自学习主机信息的正确性和一致性；数据平面自转发模式下 ARP 广播请求报文在 VXLAN 网络内广播的同时也会上送控制器，控制器可以做代答，这种模式是华三的一种创新，实现了 Overlay 网络转发的双保险模型。

图15 网络 Overlay

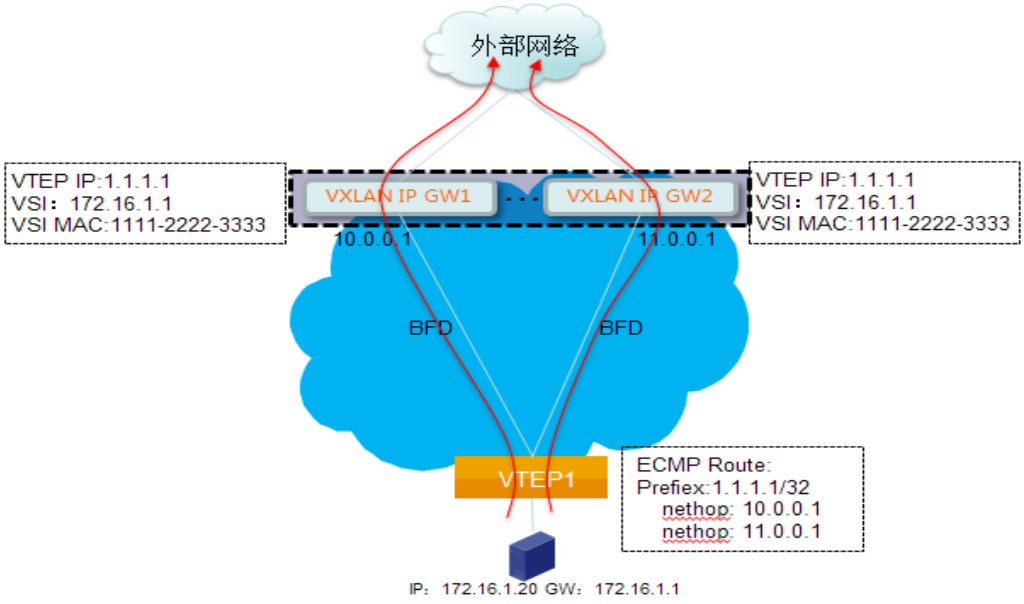



在 图 15 的组网中，VCFC集群实现对整个VXLAN网络的总体控制，以及对VNF的生命周期管理和服务链编排；VCFC可以同OpenStack、VMware vCenter、H3Cloud OS等其他第三方云平台，通过插件方式或REST API方式进行对接。

物理交换机 S12500-X/S9800 充当 VXLAN IP GW，提供 Overlay 网关功能，实现 VXLAN 网络和经典网络之间的互通，支持 Overlay 报文的封装与解封装，并根据内层报文的 IP 头部进行三层转发，支持跨 Overlay 网络之间的转发，支持 Overlay 网络和传统 VLAN 之间的互通以及 Overlay 网络与外部网络的互通；H3C S6800 充当 VTEP，支持 Overlay 报文的封装与解封装，实现虚拟机接入到 VXLAN 网络中。

Service 安全设备属于可选项，包括 vFW、vLB、M9000、L5000 等设备。东西向支持基于 vFW、vLB 的服务链；南北向可以由 S12500-X 串联 M9000 实现 NAT、FW 等服务，S12500-X 旁挂 L5000 提供 LB 服务，由 VCFC 实现引流。

图16 无状态 IP 网关




如 图 16 所示，在网络Overlay的组网模型中，S12500-X/S9800 作为Overlay网关功能，考虑到网关的扩容功能，可以采用无状态IP网关方案：

- VXLAN IP GW 实现 VXLAN 网络与传统网络的互联互通。
- 网关组内的 VXLAN IP GW 设置相同的 VTEP IP 地址，设置相同的 VNI 接口 IP 地址及 MAC 地址，VTEP IP 地址通过三层路由协议发布到内部网络中。
- 支持多台 VXLAN IP GW 组成网关组。

无状态网关的业务流向如下：

- 北向业务：VTEP 设备通过 ECMP（HASH 时变换 UDP 端口号）将 VXLAN 报文负载均衡到网关组内的不同网关上处理。
- 南向业务：每个网关都保存所有主机的 ARP，并在外部网络上将流量分流给各网关。
- 路由延迟发布确保网关重启和动态加入时不丢包。

网络 Overlay 组网方案有以下优点：

- 更高的网卡和 VXLAN 性能。
- 通过 TOR 实现 QoS、ACL，可以实现线速转发。
- 不依赖虚拟化平台，客户可以有更高的组网自由度。
- 可以根据需要自由选择部署分布式或者集中式控制方案。
- 控制面实现可以由 H3C 高可靠的 SDN Controller 集群实现，提高了可靠性和可扩展性，避免了大规模的复杂部署。
- 网关组部署可以实现流量的负载分担和高可靠性传输。

### 5.2.2 主机Overlay

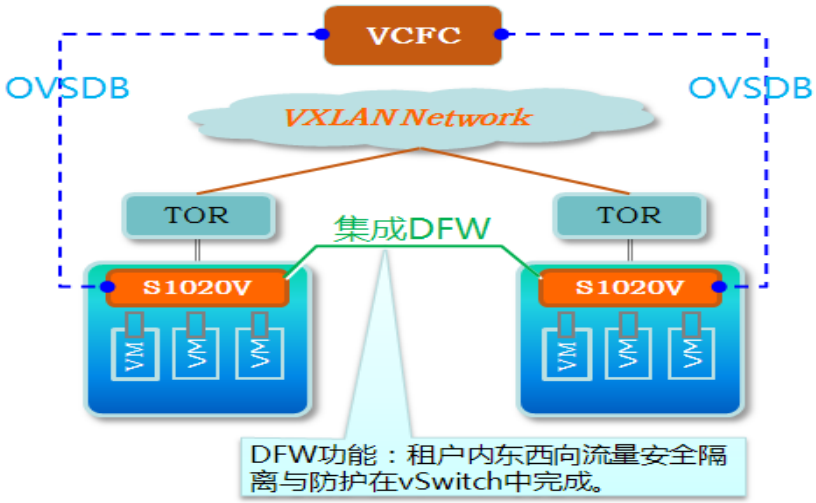
主机 Overlay 将虚拟设备作为 Overlay 网络的边缘设备和网关设备，Overlay 功能纯粹由服务器来实现。主机 Overlay 方案适用于服务器虚拟化的场景，支持 VMware、KVM、CAS 等主流 Hypervisor 平台。主机 Overlay 的网关和服务节点都可以由服务器承担，成本较低。

H3C vSwitch（即 S1020V）以标准的进程和内核态模块方式直接运行在 Hypervisor 主机上，这也是各开源或者商用虚拟化平台向合作伙伴开放的标准软件部署方式，性能和兼容性可以达到最佳。

S1020V 上除了实现转发功能，还集成了状态防火墙功能，防火墙功能可以支持四层协议，如 TCP/UDP/IP/ICMP 等协议。可以基于源 IP、目的 IP、协议类型（如 TCP）、源端口、目的端口的五元组下发规则，可以灵活决定报文是允许还是丢弃。

状态防火墙（DFW）和 ACL 是实现 OpenStack 安全组的两种方式，状态防火墙是有方向的，比如 VM1 和 VM2 之间互访，状态防火墙可以实现 VM1 能访问 VM2，VM2 不能访问 VM1 这样的需求。

图17 vSwitch 集成状态防火墙




如 图 17 所示，vSwitch功能按下述方式实现：

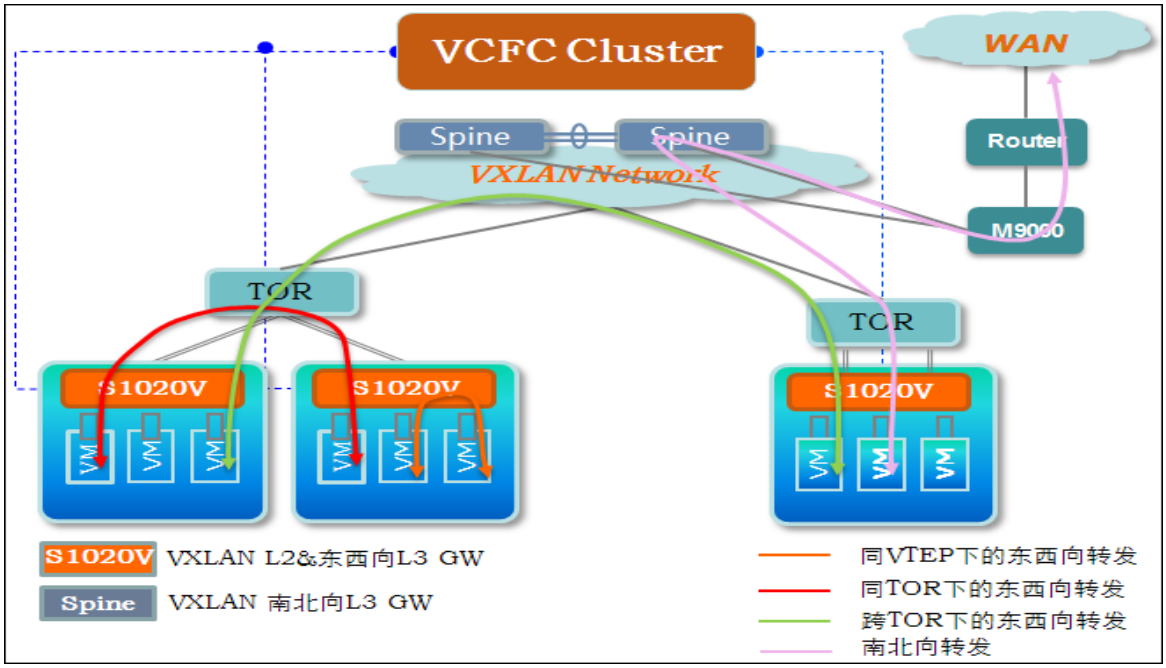
- VCFC 通过 OVSDB 通道将 DFW 策略下发给 S1020V。
- S1020V 集成 DFW 功能，依据下发的防火墙策略对端口报文做相应处理。
- 在虚机迁移或删除时，VCFC 控制下发相关防火墙策略随即迁移，实现整个数据中心的分布式防火墙功能。

在主机 Overlay 情况下，H3C vSwitch 既承担了 VTEP（即 VXLAN L2 GW）功能，也可以承担东西向流量三层网关的功能。三层网关同时亦可以由 NFV、物理交换机分别承担。vSwitch 功能也可以实现 Overlay 网络内虚机到虚机的跨网段转发。按照 VXLAN 三层转发实现角色的不同，可以分为以下几个方案：

(1) 东西向分布式网关转发方案

如 图 18 所示，在分布式网关情况下，采用多个vSwitch逻辑成一个分布式三层网关，东西向流量无需经过核心设备Overlay层面的转发即可实现东西向流量的跨VXLAN转发，以实现跨网段最短路径转发；南北向的流量仍然会以核心Spine设备作为网关，虚机访问外网时，vSwitch先把报文通过VXLAN网络转发到Spine设备上，Spine设备进行VXLAN解封装后再根据目的IP转发给外部网络。

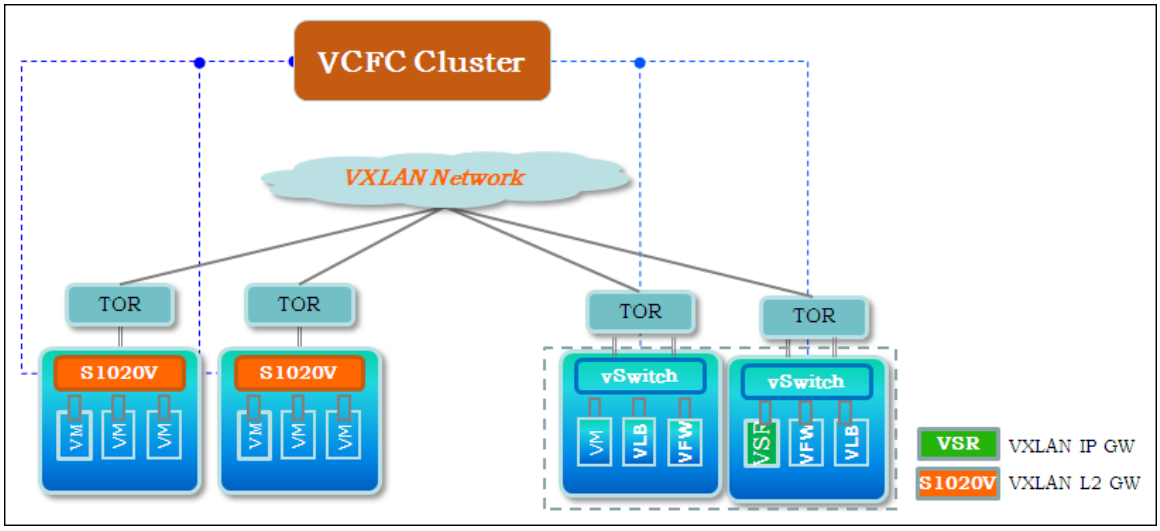
图18 东西向分布式网关方案




(2) NFV 设备 VSR 做网关方案

VSR做网关的情况下，VXLAN IP GW、VXLAN L2 GW、服务节点都由服务器来实现，如 图 19 所示。

图19 VSR 做网关的主机 Overlay 方案




VCFC 集群实现对整个 VXLAN 网络的总体控制，以及对 VNF 的生命周期管理和服务链编排；VCFC 可以同 OpenStack、VMware vCenter、H3Cloud OS 等其他第三方云平台，通过插件方式或 REST API 方式进行对接。

NFV 设备 VSR 充当 VXLAN IP GW，提供 Overlay 网关功能，实现 VXLAN 网络和经典网络之间的互通，支持 Overlay 报文的封装与解封装，并根据内层报文的 IP 头部进行三层转发，支持跨 Overlay

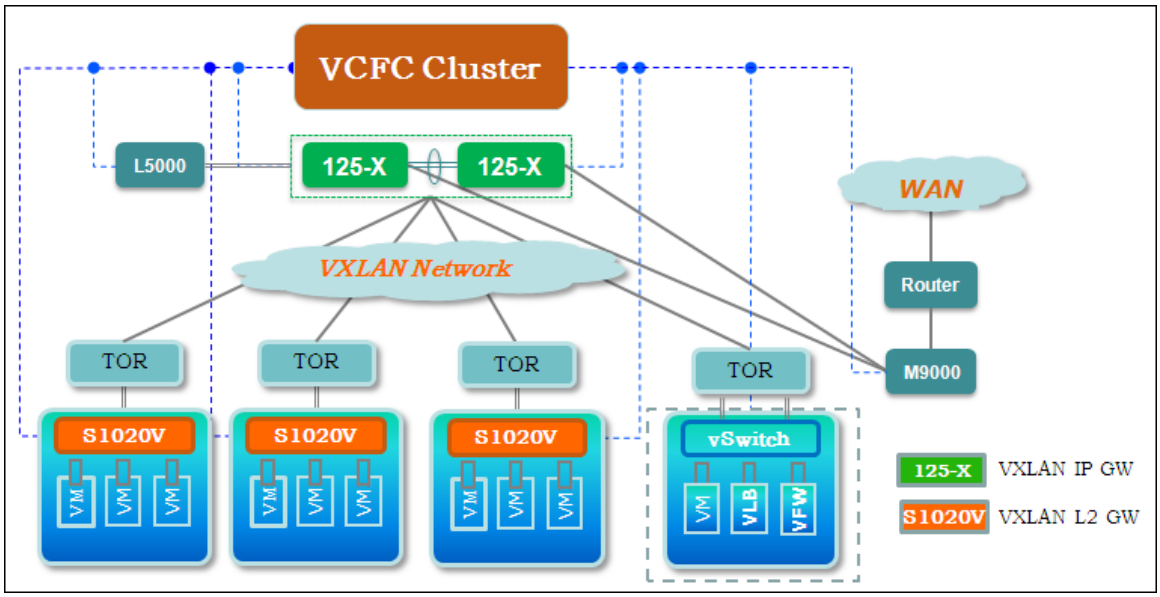
网络之间的转发，支持 Overlay 网络和传统 VLAN 之间的互通以及 Overlay 网络与外部网络的互通；H3C S1020V 充当 L2 VTEP，支持 Overlay 报文的封装与解封装，实现虚拟机接入到 VXLAN 网络中，其中 H3C S1020V 支持运行在 ESXi、KVM、H3C CAS 等多种虚拟化平台上。

Service 安全设备属于可选项，包括 VSR、vFW、vLB 等设备，实现东西向和南北向服务链服务节点的功能。

(3) 物理交换机做网关方案

如 图 20 所示，同纯软主机Overlay方案相比，软硬结合主机Overlay方案使用Spine设备做VXLAN IP GW。Spine设备可以使用S12500-X/S9800，也可以使用S10500，在使用S10500 和S1020V组合的情况下可以实现更低的使用成本。Service安全设备属于可选项，包括vFW、vLB、M9000、L5000等设备。东西向支持基于vFW、vLB的服务链；南北向可以由S12500-X串联M9000 实现NAT、FW等服务，S12500-X旁挂L5000 提供LB服务，由VCFC通过服务链实现引流。

图20 物理交换机做网关的主机 Overlay 方案



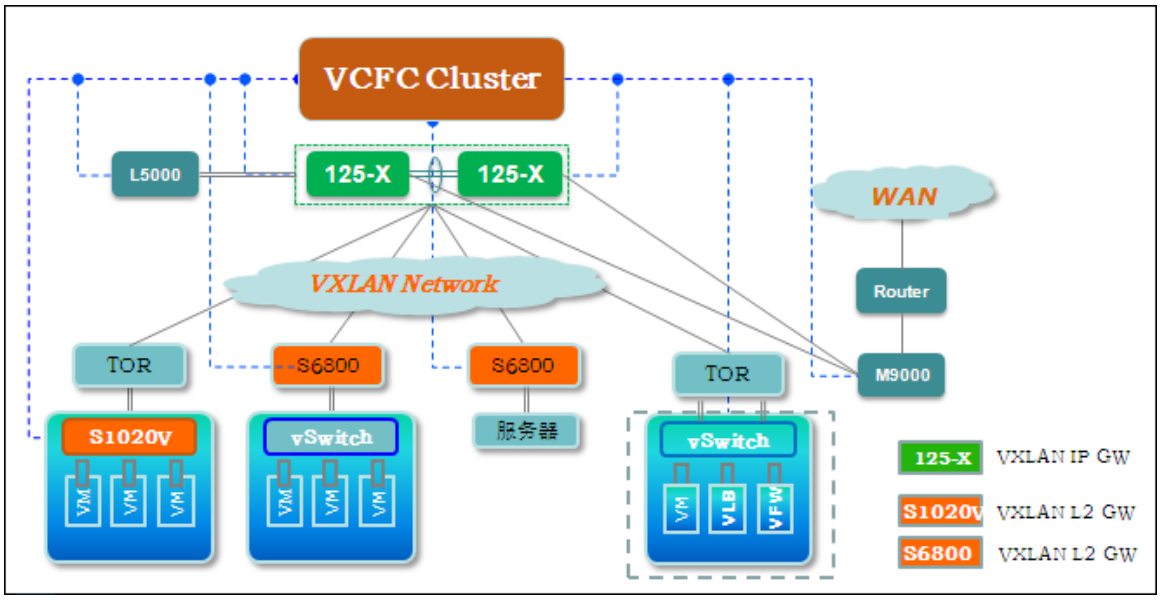

主机 Overlay 组网方案总体来说有以下优点：

- 适用于服务器虚拟化的场景，成本较低。
- 可以配合客户已有的 VMware、Microsoft 等主流 Hypervisor 平台，保护客户已有投资。
- 可以根据需要自由选择部署分布式或者集中式控制方案。
- 控制面实现可以由 H3C 高可靠的 SDN Controller 集群实现，提高了可靠性和可扩展性，避免了大规模的复杂部署。
- 物理交换机做网关的情况下，也同网络 Overlay 一样可以使用多网关组功能，网关组部署可以实现流量的负载分担和高可靠性传输。
- vSwitch 作为东西向 IP 网关时，支持分布式网关功能，使虚机迁移后不需要重新配置网关等网络参数，部署简单、灵活。

### 5.2.3 混合Overlay

如 图 21 所示，混合Overlay是网络Overlay和主机Overlay的混合组网，可以支持物理服务器和虚拟服务器之间的组网互通。它融合了两种Overlay方案的优点，既可以充分利用虚拟化的低成本优势，又可以发挥硬件GW的转发性能、将非虚拟化设备融入Overlay网络，它可以为客户提供自主化、多样化的选择。

图21 混合 Overlay




VCFC 集群实现对整个 VXLAN 网络的总体控制，以及对 VNF 的生命周期管理和服务链编排；VCFC 可以同 OpenStack、VMware vCenter、H3Cloud OS 等其他第三方云平台，通过插件方式或 REST API 方式进行对接。

S12500-X/S9800 充当 VXLAN IP GW，提供 Overlay 网关功能，实现 VXLAN 网络和经典网络之间的互通，支持 Overlay 报文的封装与解封装，并根据内层报文的 IP 头部进行三层转发，支持跨 Overlay 网络之间的转发，支持 Overlay 网络和传统 VLAN 之间的互通以及 Overlay 网络与外部网络的互通；H3C S6800、H3C S1020V 充当 VTEP，支持 Overlay 报文的封装与解封装，实现服务器和虚拟机接入到 VXLAN 网络中。

Service 安全设备属于可选项，包括 vFW、vLB、M9000、L5000 等设备。东西向支持基于 vFW、vLB 的服务链；南北向可以由 S12500-X 串联 M9000 实现 NAT、FW 等服务，S12500-X 旁挂 L5000 提供 LB 服务，由 VCFC 实现引流。

### 5.2.4 Overlay组网总结

表2 Overlay 组网总结

| 类别 | 组网 | 虚拟化平台支持 | 适用场景 | 服务链方式 |
| --- | --- | --- | --- | --- |
| 主机 Overlay | S1020V+VSR | CAS/VMware/KVM | 适合海量租户，但单租户对转发性能要求不高的场景，如公有云，网络设备利旧或成本受 | 南北向VSR（自带FW功能）+vLB<br>东西向共享南北向 |

| 类别 | 组网 | 虚拟化平台支持 | 适用场景 | 服务链方式 |
| --- | --- | --- | --- | --- |
| | | | 限条件下的私有云 | NFV<br>都采用服务链方式 |
| | S1020V+S12500-X | CAS/VMware/KVM | 同纯软主机Overlay方案相比，主机Overlay软硬结合方案使用S12500-X或S10500做VXLAN IP GW，跨网段转发性能较高；<br>与网络Overlay相比，对TOR没有要求，不要求TOR承担VTEP功能 | 南北向采用服务链引流到（M9000+L5000）<br>东西向vFW+vLB单跳或者多跳服务链 |
| 网络Overlay | S68+S12500-X | ALL | 适合于要求高网络转发性能的场景，以及大规模网络的私有云应用场景 | |
| 混合Overlay | S68+S12500-X＋S1020V | CAS/VMware/KVM | 混合业务场景，有部分业务要求高转发性能，如数据库，存储等 | |

上述几种 Overlay 组网均支持和 OpenStack K 版本对接。

# 6 SDN Overlay转发流程描述

## 6.1 SDN Overlay流表建立和发布

我们以流转发为例介绍 SDN Overlay 的转发流程。

### 6.1.1 流表建立流程对ARP的处理

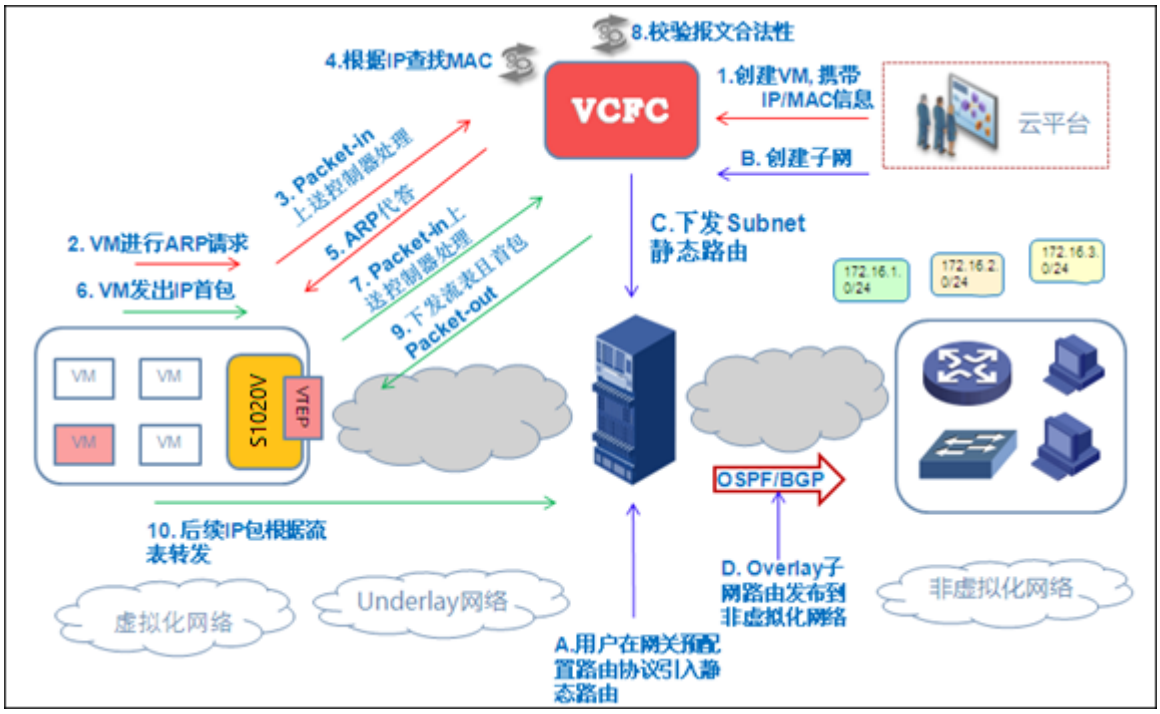
对于虚拟化环境来说，当一个虚拟机需要和另一个虚拟机进行通信时，首先需要通过 ARP 的广播请求获得对方的 MAC 地址。由于 VXLAN 网络复杂，广播流量浪费带宽，所以需要在控制器上实现 ARP 代答功能。即由控制器对 ARP 请求报文统一进行应答，而不创建广播流表。

ARP 代答的大致流程：控制器收到 S1020V 上送的 ARP 请求报文，做 IP-MAC 防欺骗处理确认报文合法后，从 ARP 请求报文中获取目的 IP，以目的 IP 为索引查找全局表获取对应 MAC，以查到的 MAC 作为源 MAC 构建 ARP 应答报文，通过 Packetout 下发给 S1020V。

### 6.1.2 Overlay网络到非Overlay网络

Overlay网络到非Overlay网络的流表建立和路由发布如 图 22 所示：

图22 Overlay 网络到非 Overlay 网络的流表建立和路由发布




创建 VM 的时候，会同时分配 IP 和 MAC 信息。然后 VM 发送 ARP 请求报文，该报文会通过 Packet-in 被上送到控制器。控制器做 IP-MAC 防欺骗处理确认报文合法后，通过 ARP 代答功能构建 ARP 应答报文并通过 Packet-out 下发。

VM 收到 ARP 应答报文后，封装并发送 IP 首包。S1020V 收到 IP 首报后发现没有对应流表，就将该 IP 首包通过 Packet-in 上送控制器。控制器通过 OpenFlow 通道收到 Packet-in 报文后，判断上

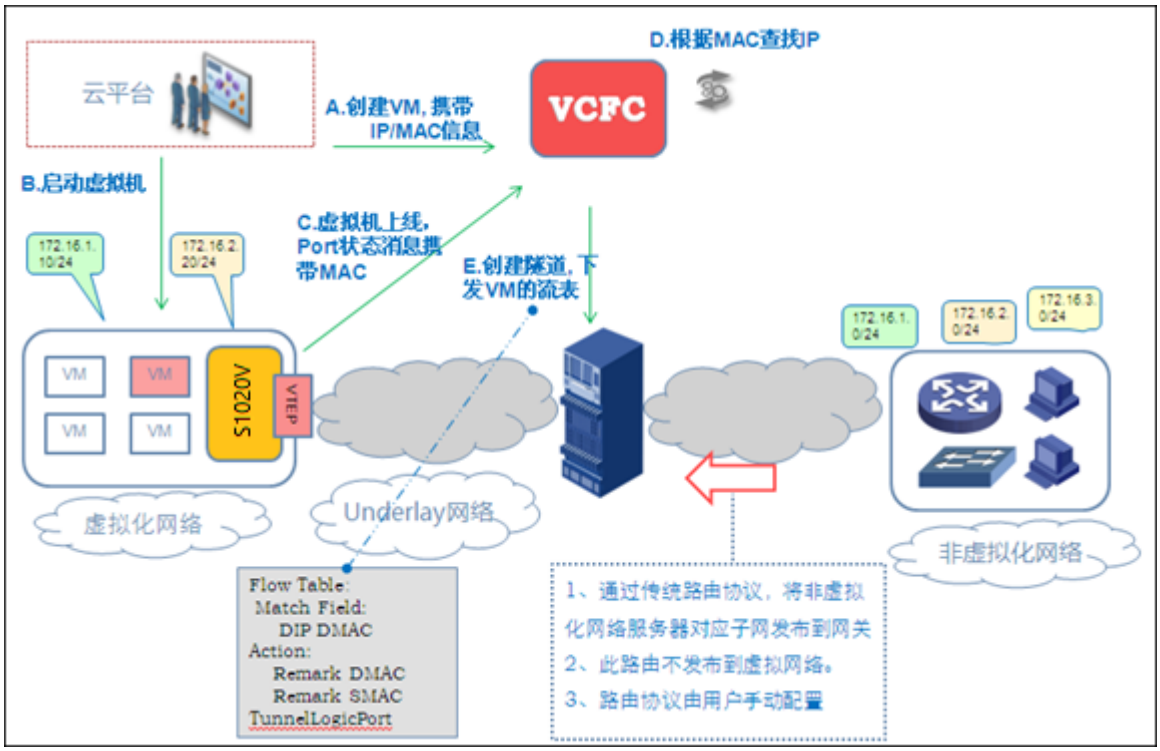
送的 IP 报文的 IP-MAC 为真实的。然后根据报文中的目的 IP 查询目的端口，将 IP 首包直接发送到目的端口，同时生成相应流表下发到 S1020V。

流表下发到 S1020V 后，而后续的 IP 报文就会根据 S1020V 上的流表进行转发，而不再需要上送控制器。

### 6.1.3 非Overlay网络到Overlay网络

非Overlay网络到Overlay网络的流表建立和路由发布如 图 23 所示：

图23 非 Overlay 网络到 Overlay 网络的流表建立和路由发布




创建 VM 的时候，会同时分配 IP、MAC 和 UUID 等信息。VM 上线时会触发 S1020V 发送 Port Status 消息上送控制器，该消息携带 VM MAC 信息。控制器根据 VM MAC 查找 IP 等相关信息，然后携带 VM 的相关信息通知 GW 虚机上线。

控制器根据 VM 上线消息中携带的数据，构造物理机向 VM 转发报文时使用的流表表项，并下发到 VM 所在 VNI 对应网关分组中的所有 GW。

## 6.2 Overlay网络转发流程

#### 1. 识别报文所属VXLAN

VTEP 只有识别出接收到的报文所属的 VXLAN，才能对该报文进行正确地处理。

- VXLAN 隧道上接收报文的识别：对于从 VXLAN 隧道上接收到的 VXLAN 报文，VTEP 根据报文中携带的 VNI 判断该报文所属的 VXLAN。

- 本地站点内接收到数据帧的识别：对于从本地站点中接收到的二层数据帧，VTEP 通过以太网服务实例（Service Instance）将数据帧映射到对应的 VSI，VSI 内创建的 VXLAN 即为该数据帧所属的 VXLAN。

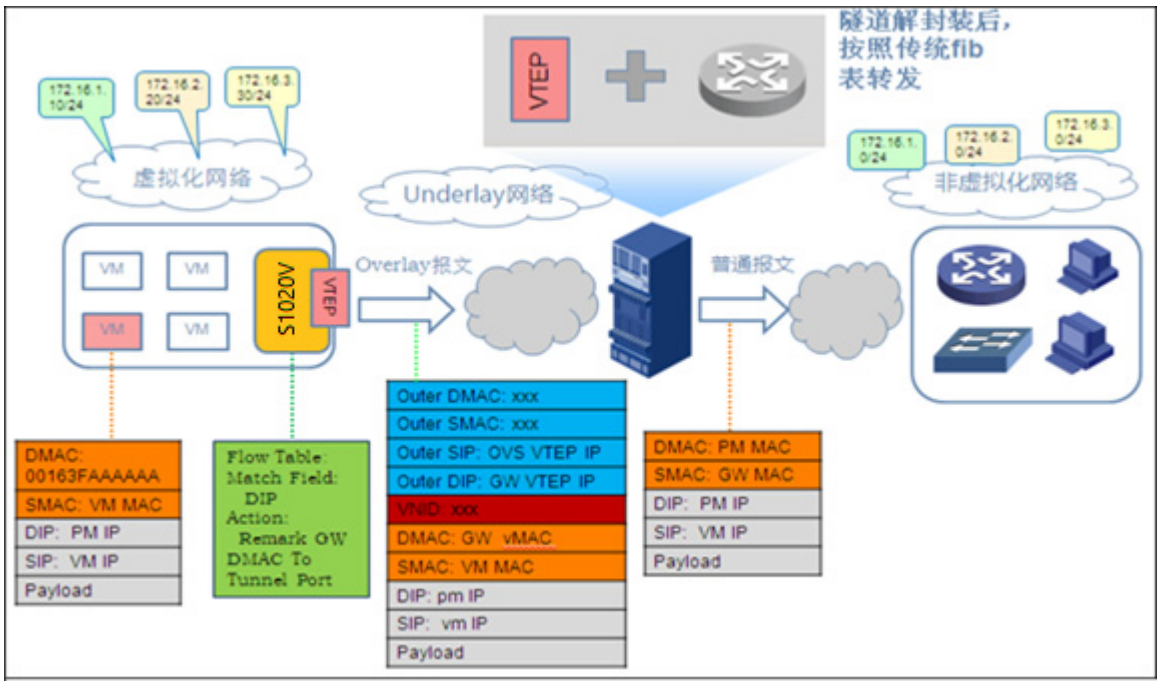
2. MAC地址学习

- 本地 MAC 地址学习：指本地 VTEP 连接的本地站点内虚拟机 MAC 地址的学习。本地 MAC 地址通过接收到数据帧中的源 MAC 地址动态学习，即 VTEP 接收到本地虚拟机发送的数据帧后，判断该数据帧所属的 VSI，并将数据帧中的源 MAC 地址（本地虚拟机的 MAC 地址）添加到该 VSI 的 MAC 地址表中，该 MAC 地址对应的出接口为接收到数据帧的接口。
- 远端 MAC 地址学习：指远端 VTEP 连接的远端站点内虚拟机 MAC 地址的学习。远端 MAC 学习时，VTEP 从 VXLAN 隧道上接收到远端 VTEP 发送的 VXLAN 报文后，根据 VXLAN ID 判断报文所属的 VXLAN，对报文进行解封装，还原二层数据帧，并将数据帧中的源 MAC 地址（远端虚拟机的 MAC 地址）添加到所属 VXLAN 对应 VSI 的 MAC 地址表中，该 MAC 地址对应的出接口为 VXLAN 隧道接口。

### 6.2.1 Overlay网络到非Overlay网络

Overlay网络到非Overlay网络的转发流程如 图 24 所示：

图24 Overlay 网络到非 Overlay 网络的转发流程




虚拟机构造发送到物理机的报文，目的 MAC 为 S1020V 的 MAC，目的 IP 为要访问的物理机的 IP，报文从虚拟机的虚拟接口发出。

S1020V 接收到虚拟机发送的报文，根据报文中的目的 IP 匹配 S1020V 上的流表表项。匹配到流表表项后，修改报文的目的 MAC 为 VXLAN-GW 的 MAC，源 MAC 为 S1020V 的 MAC，并从指定的隧道接口发送。从指定的隧道接口发送报文时，会在报文中添加 VXLAN 头信息，并封装隧道外层报文头信息。

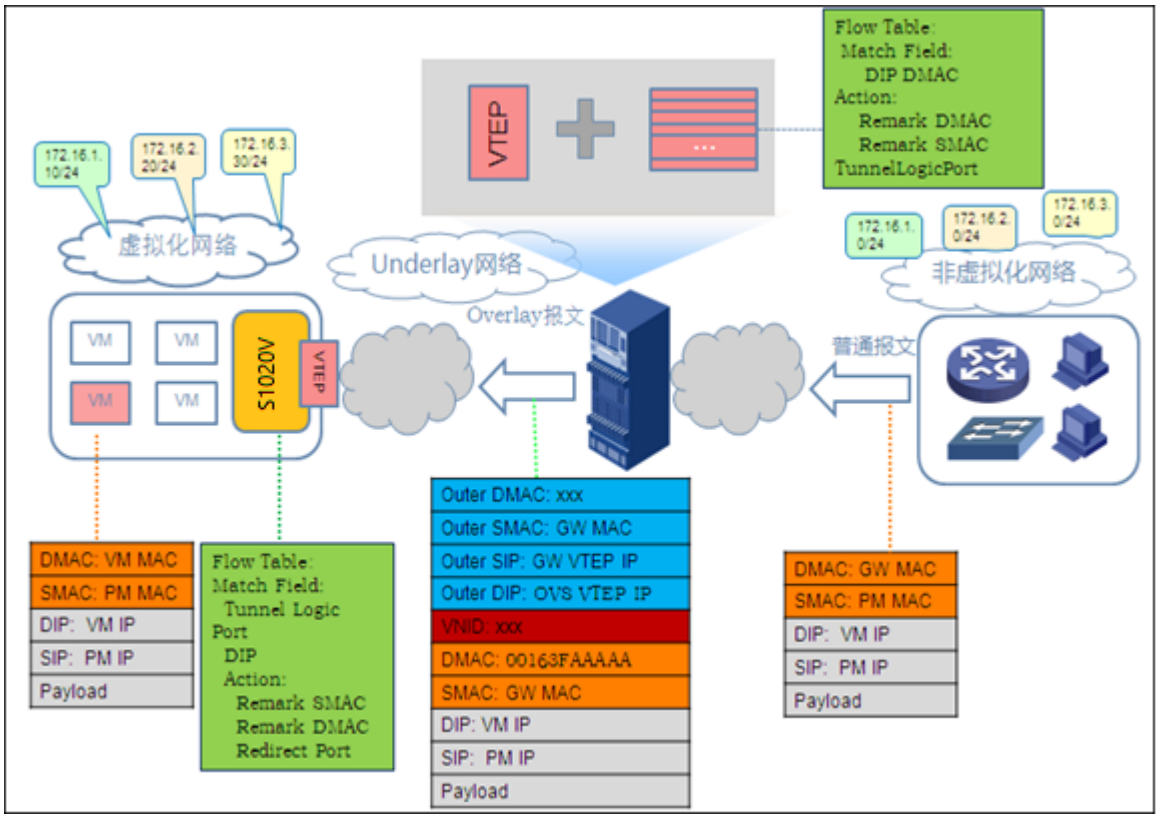
VXLAN-GW 从隧道口接收到 VXLAN 隧道封装报文，隧道自动终结，得到内层报文。然后根据内层报文的目的 IP 按照 FIB（非流表）进行报文三层转发。

报文按照传统网络的转发方式继续转发。物理机接收到 VXLAN-GW 转发的报文，实现虚拟机到物理机的访问。

### 6.2.2 非Overlay网络到Overlay网络

非Overlay网络到Overlay网络的转发流程如 图 25 所示：

图25 非 Overlay 网络到 Overlay 网络的转发流程




物理机构造发送到虚拟机的报文，在传统网络中通过传统转发方式将报文转发到 VXLAN-GW。VXLAN-GW 接收该报文时，报文的目的 MAC 为 VXLAN-GW 的 MAC，目的 IP 为虚拟机的 IP 地址，从物理机发送出去的报文为普通报文。

VXLAN-GW 接收报文，根据报文的入接口 VPN，目的 IP 和目的 MAC 匹配转发流表。然后从指定的 VXLAN 隧道口发送。从隧道口发送报文时，根据流表中的信息添加 VXLAN 头信息，并对报文进行隧道封装。从 GW 发送报文为封装后的 Overlay 报文。

S1020V 接收到报文后，隧道自动终结。根据报文 VNI 和目的 IP 匹配转发流表。匹配到流表后，从指定的端口发送。从 S1020V 发送的报文为普通报文。

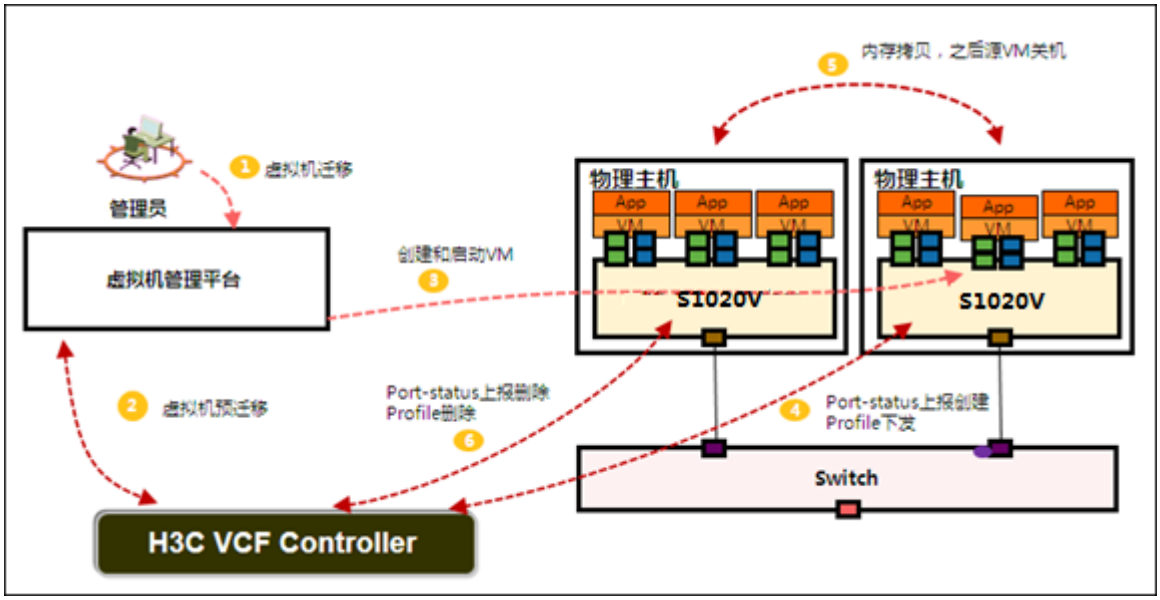
根据报文的目的 MAC，虚拟机接收到物理机发送的报文，实现物理机到虚拟机的访问。

## 6.3 Overlay网络虚机迁移

在虚拟化环境中，虚拟机故障、动态资源调度功能、服务器主机故障或计划内停机等都会造成虚拟机迁移动作的发生。虚拟机的迁移，需要保证迁移虚拟机和其他虚拟机直接的业务不能中断，而且虚拟机对应的网络策略也必须同步迁移。

虚拟机迁移及网络策略跟随如 图 26 所示：

**图26 虚拟机迁移及网络策略跟随**




网络管理员通过虚拟机管理平台下发虚拟机迁移指令，虚拟机管理平台通知控制器预迁移，控制器标记迁移端口，并向源主机和目的主机对应的主备控制器分布发送同步消息，通知迁移的 vPort，增加迁移标记。同步完成后，控制器通知虚拟机管理平台可以进行迁移了。

虚拟机管理平台收到控制器的通知后，开始迁移，创建 VM 分配 IP 等资源并启动 VM。启动后目的主机上报端口添加事件，通知给控制器，控制器判断迁移标记，迁移端口，保存新上报端口和旧端口信息。然后控制器向目的主机下发网络策略。

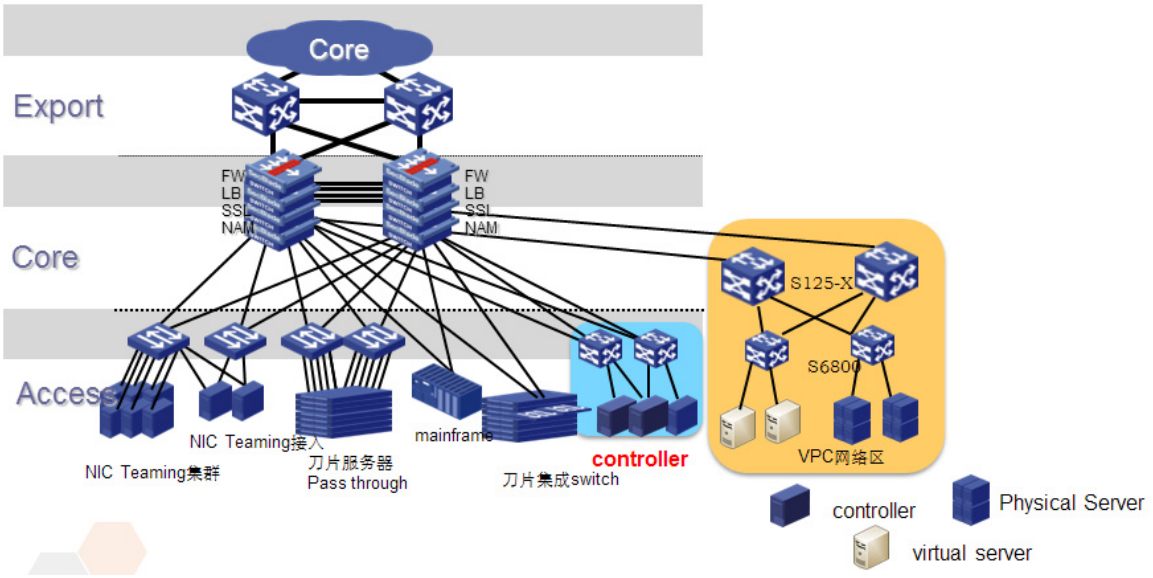
源 VM 和目的 VM 执行内存拷贝，内存拷贝结束后，源 VM 关机，目的 VM 上线。源 VM 关机后，迁移源主机上报端口删除事件，通知给控制器，控制器判断迁移标记，控制器根据信息删除旧端口信息并同时删除迁移前旧端口对应的流表信息。

主控制器完成上述操作后在控制器集群内进行删除端口消息的通知。其他控制器收到删除端口信息后，也删除本控制器的端口信息，同时删除对应端的流表信息。源控制器需要把迁移后新端口通知控制器集群的其他控制器。其他控制器收到迁移后的端口信息，更新端口信息。当控制器重新收到 Packet-in 报文后，重新触发新的流表生成。

## 6.4 SDN Overlay升级部署方案

### 6.4.1 SDN Overlay独立分区部署方案

图27 DC 增量部署，SDN Overlay 独立分区

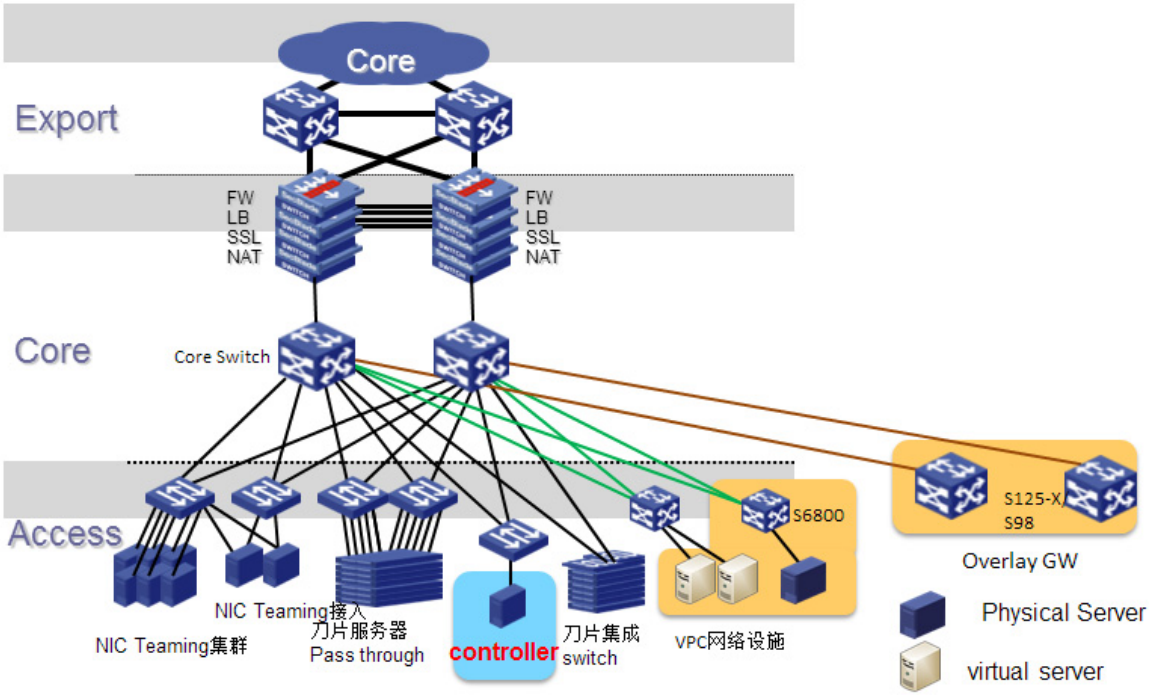



基于对原有数据中心改动尽量少的思路下，可以把 SDN Overlay 部署在一个独立分区中，作为 VXLAN IP GW 的核心交换机作为 Underlay 出口连接到原有网络中，对原有网络无需改动，南北向的安全设备和原有 DC 共享。

场景：在现有数据中心的独立区域部署，通过原有网络互联。

### 6.4.2 IP GW旁挂部署方案

图28 DC 增量部署，IP GW 旁挂

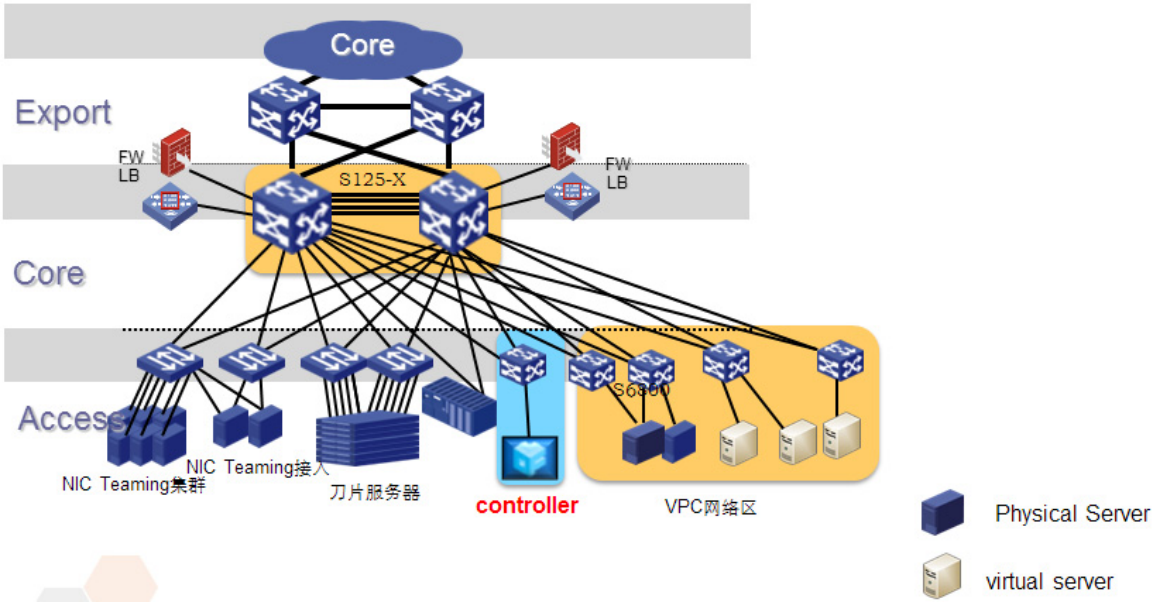



考虑到尽量利用原有数据中心空间部署 VXLAN 网络的情况下，可以采用物理交接机（S12500-X/S9800）作为 VXLAN IP GW 旁挂的方案，与经典网络共用核心；而 VXLAN 网络作为增量部署，对原有网络改动小。

场景：利用现有数据中心剩余空间增量部署。

### 6.4.3 核心升级，SDN Overlay独立分区

图29 核心利旧升级，SDN Overlay 独立分区




核心设备升级为支持 VXLAN IP GW 的 S12500-X，同时作为传统和 Overlay 网络的核心，原有网络除核心设备外保持不变，充分利旧，保护用户原有投资。安全设备物理上旁挂在核心 S12500-X 上，通过 VCFC 把 VPC 流量引流到安全设备进行安全防护。

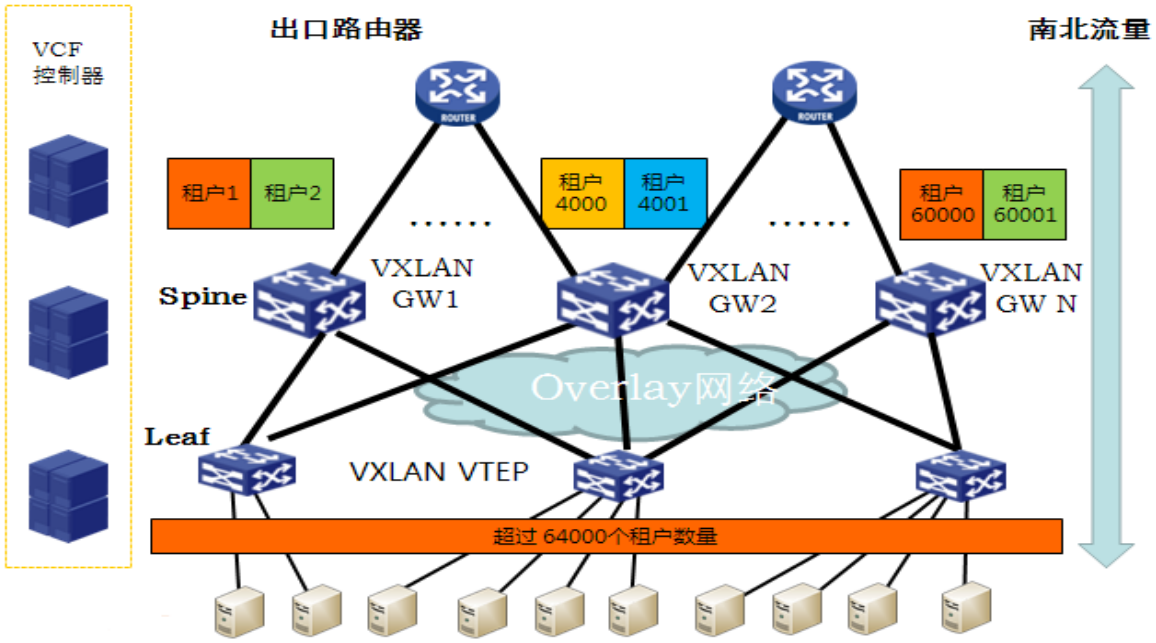
场景：全新建设数据中心区域，或者升级现有中心的网络核心，原有服务器和网络设备重复利用。

### 6.4.4 Overlay网关弹性扩展升级部署

受制于芯片的限制，单个网关设备支持的租户数量有限，控制器能够动态的将不同租户的隧道建立在不同的 Overlay 网关上，支持 Overlay 网关的无状态分布，实现租户流量的负载分担。

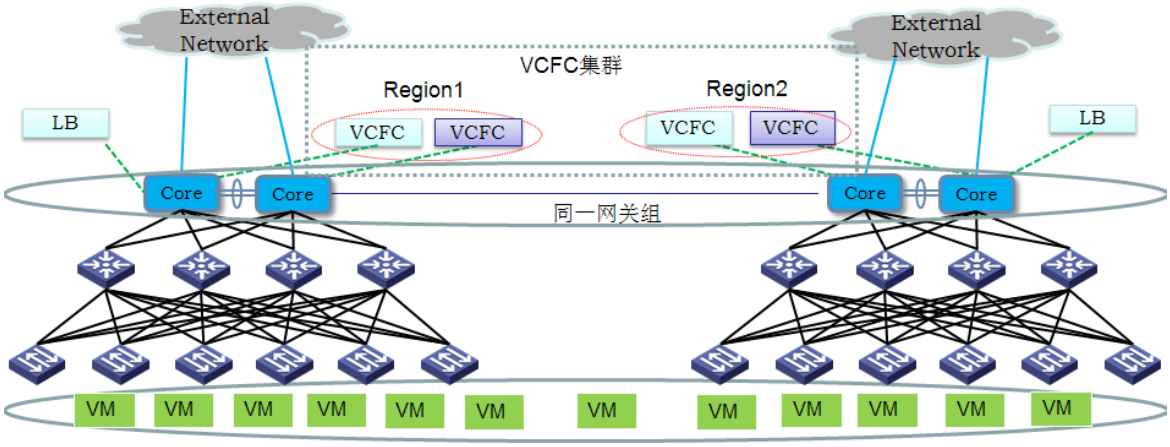
如 图 30 所示，Overlay网络可以支持Overlay网关随着租户数量增加的扩充，从而提供一个具有弹性扩展能力的Overlay网络架构。

图30 Overlay 网络弹性扩展




## 6.4.5 多数据中心同一控制器集群部署

图31 多数据中心同一控制器集群部署




控制器跨数据中心部署在多个数据中心，把多个数据中心逻辑上连接为一个数据中心：

- 任一个 GW 上有全网所有的虚拟机信息，任意一个 GW 上都可以正确通过 Overlay 隧道转发到正确的虚拟机。
- 一个网关组发布相同的 VTEP IP 地址，每一个数据中心会自动根据最短路径算法，将选择本数据中心的核心设备作为网关，实现本地优先转发。
- 4 台控制器的部署，推荐使用 2Leader+2Member 的主备模式，每个中心各一台 Leader 和一台 Member，4 台以上推荐采用多数派/少数派模式。

# 7 SDN Overlay方案优势总结

- 网络架构方面具有下述明显优势：
  - 应用与位置解耦，网络规模无限弹性扩展；
  - 网络虚拟化，实现大规模多租户和业务隔；
  - 支持多种 Overlay 模型，满足场景化需求；
  - 跨多中心的网络资源统一池化，随需分配。
- 网络安全方面具有下述特点：
  - 各种软硬件安全设备灵活组合，形成统一安全资源池；
  - 丰富的安全组合功能，可以充分满足云计算安全合规要求；
  - 针对主机，南北和东西向流量，可以实现精细化多层次安全防护；
  - 通过服务链，可以实现安全业务的灵活自定义和编排。
- 网络业务发放具有下述优点：
  - 支持 VPC 多租户虚拟网络：基于 OpenStack 模型，租户相互隔离、互不干扰，各租户可提供独立 FW/LB/NAT 等服务；
  - 网络灵活自定义：租户虚拟网络根据自身需求可灵活自定义，实现对于 SDN 和 NFV 的融合控制；
  - 网络自动化：业务流程全自动发放，配置自动化下发，业务部署从数天缩短到分钟级；
  - 与云无缝对接融合：实现网络、计算与存储的无缝打通，实现云计算业务的自助服务。
- 在网络运维上能充分满足客户需求：
  - 支持流量可视化：应用、虚拟、网络拓扑的统一呈现、资源映射、流量统计、路径和状态感知；
  - 支持自动化运维：用户能够自定义网络运维管理能力，实现 DC 内自定义流量调度、动态流量自动监控分析。