H3C VCF Controller 技术白皮书

目　录

1 概述

SDN 是一种新型网络创新架构，其核心思想是将网络设备的控制层面与转发层面分离，将控制层面逻辑集中后向外开放 API 接口，从而提供一个能够面向业务的新型网络，为新业务快速部署或网络创新提供良好的平台。SDN 控制器是 H3C SDN 解决方案的重要组成部分，它类似一个网络操作系统，为用户提供开发和运行 SDN 应用的平台。可以控制 OpenFlow 网络中的各种资源，并为应用提供接口，应用通过调用控制器提供的接口来实现自己的网络转发需求。

架构先进

采用先进的 OSGi 架构（Open Service Gateway Initiative），可以通过开发 APP 的方式灵活扩展新的功能。

接口丰富

对外提供丰富的 OPEN API 与 REST API 接口，让用户或第三方软件开发商能够非常方便进行 SDN 应用开发。

高可靠性

支持独立运行模式和集群模式，在集群模式下，多台华三 SDN 控制器之间可以组建集群，当集群的部分成员发生故障时，业务不受影响，从而大幅度增强了 SDN 网络的可靠性。

H3C Virtual Converged Framework（VCF）控制器（以下简称 H3C VCF 控制器）是一款 SDN 控制器系统，作为 SDN 解决方案的承载工具，为数据中心网络、公有云、私有云、校园边缘网络等提供了一个软件平台发挥其不同的网络特性。

H3C VCF 控制器是一个开放的软件平台，提供可编程的用户接口，使用 OpenFlow 协议作为网络控制协议实现对物理网络的管理。同时支持 Restful API、Open API 方式北向接口，并提供在线 API 文档。

H3C VCF 控制器是 H3C 推出的 SDN 解决方案的重要组成部分，它类似一个网络操作系统，为用户提供开发和运行 SDN 应用的平台。H3C VCF 控制器可以控制 Overlay 网络中的各种资源，并为应用提供接口，应用通过调用控制器提供的接口来实现自己的网络转发需求。

2 产品特点

2.1 高性能

- 控制通道数量超过 30000，具有 200Kpps 的主动下发流表能力
- 强大的拓扑发现能力，VCFC 能够在能够在 500ms 内完成 19950 条链路的拓扑发现
- 优秀的集群管理能力，对于集群角色分配 、主控制器切换等方面具有完备的应用逻辑

2.2 Overlay网络

- 控制器支持异构厂商 overlay 网络监控、及 VXLAN 配置自动下发
- 控制器支持主机 overlay 与网络 overlay 混合组网。
- VXLAN GW、TOR offload 管理。管理硬件 TOR offload 交换机和硬件 VXLAN GW、主机 VXLAN GW，包括 VXLAN 隧道管理、设备管理等。
- 组网支持多样化的组网部署方式，支持跨域互访。
- 通过 Overlay 组网可实现虚拟机灵活迁移，安全策略动态跟随。
- 服务器内部 vSwitch 管理。管理服务器内部的 vSwitch，包括端口管理、策略管理、子网管理等，构建 VXLAN 隧道，控制地址学习方式等。兼容 H3C S1020V 系列 vSwitch 产品和 Open vSwitch。

2.3 扩展性及兼容性

- 控制器支持 openflow 1.3 协议规范，能够对接通过 OpenFlow v1.3 协议一致性认证测试的网元
- 控制器支持对接虚拟化平台（包括 vsphere、kvm 等）及裸金属平台。
- 云计算接口。向上层云计算系统提供 API 接口和插件，方便云计算系统整合数据中心网络资源，实现“一站式”服务和管理。兼容 H3CloudOS 云计算管理平台、VMware vCenter 和 OpenStack。
- 对外提供丰富的原生 Open API 接口。允许第三方应用程序以控制器内的 OSGi bundle 形式运行，从而实现事件和数据包的高性能处理。这些基于网络底层的 Open API 接口非常强大，使控制器能够按照用户特定环境进行定制和扩展。
- 对外提供丰富的 REST API 接口；可以利用缓存 Cache 来提高响应速度，通讯本身的无状态性可以让不同的服务器的处理一系列请求中的不同请求，提高服务器的扩展性；浏览器即可作为客户端，简化软件需求；相对于其他叠加在 HTTP 协议之上的机制，REST 的软件依赖性更小；不需要额外的资源发现机制；在软件技术演进中的长期兼容性更好。

2.4 安全性

- 符合 ETSI-NFV 框架，支持虚拟防火墙 vFW、虚拟负载均衡 vLB、应用防火墙功能（vWAF）等虚拟网元的管理与服务链功能，支持虚拟设备横向虚拟化集群，控制器支持配置虚拟网元。
- 支持用户权限分级，支持本地用户管理及认证、支持 Radius 服务器认证。
- 控制器支持对接多种 DPI 业务安全类设备，包括 IPS，防病毒和 URL 过滤等，且支持特征库导入与查看。
- 支持安全策略功能和服务链功能，可以根据业务需要灵活选择。
- 支持用户虚拟机 IP 与 MAC 绑定，防止 IP 和 MAC 地址仿冒
- 控制器支持对接不少于 4 个品牌的安全设备及流量牵引，包括主流厂家华为、深信服、天融信、360 等安全品牌等。

- 控制器支持网络资源容量规划管理（整网资源用量统计和实时查询包含 ACL、内存、ARP），及资源消耗的趋势分析、预警、资源耗尽时的主动控制和自我保护，业务变更对资源占用的影响汇总分析，新增业务对现有网络的影响评估等。

- 在单台控制器工作环境中，SDN 网络可靠性较低，存在单点故障的可能性，即当控制器故障时将导致 SDN 网络处于非管理状态。H3C VCF 控制器提供 Team 功能从而提高网络的可靠性，避免单点故障，即当 Team 中的某台设备出现故障不能正常工作时，组内的其它成员控制器会自动接替该故障控制器继续进行工作，保证 SDN 网络的正常运行。同时，Team 还提供集中的控制器配置和监控。

2.5 可维护性

- 控制器支持在线升级并显示 ISSU 升级过程、检查业务运行数据情况
- 大型复杂网络中，通过在 VCF 控制器 Team 中为不同的设备划分为不同的 Region，可以实现分域管理，极大降低管理复杂度
- 支持 Overlay 链路诊断，可以实现端到端的连通性诊断和分段连通性诊断。
- 支持基于虚拟端口、流表、DPID+VNI 和控制器的流量信息统计，还支持基于虚拟网络、虚拟子网、虚拟路由器和虚拟端口的网络信息统计。

2.6 开放架构

- 采用 OSGi 开放式架构，可以通过开发 APP 的方式灵活扩展新的功能；这个框架实现了一个优雅、完整和动态的组件模型。应用程序无需重新引导可以被远程安装、启动、升级和卸载。OSGi 规范是由成员通过公开的程序开发，对公众免费而且没有许可证限制。
- VCFC 集群支持跨三层部署，无需单独为 VCFC 集群提供二层专用网络。

3 H3C SDN Overlay 模型设计

3.1 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 将虚拟网络元素组织为的“资源池”，VCF Controller 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。

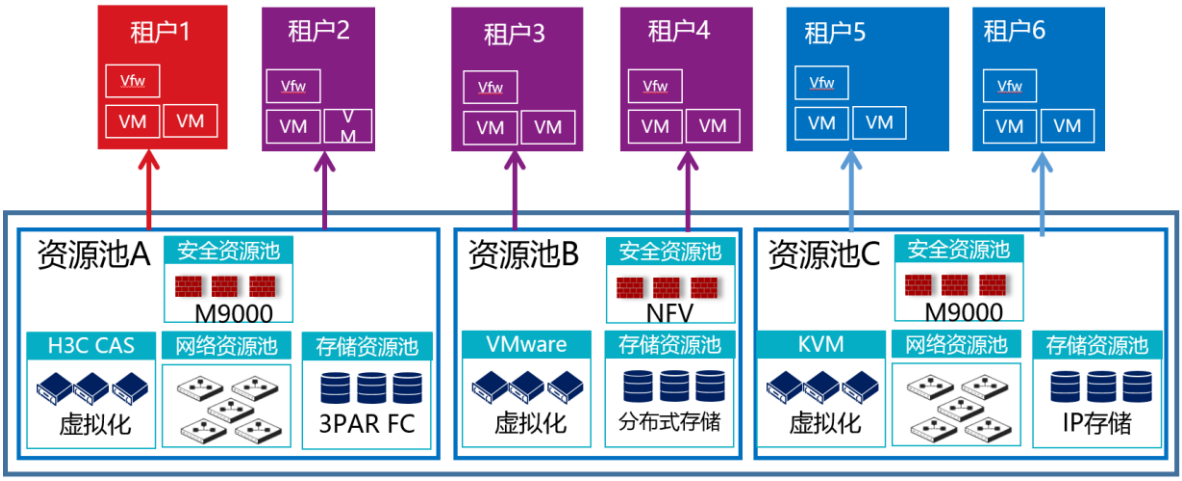


图1 VPC 多租户资源池场景

VCFC 虚拟网络元素的抽象方式与 OpenStack 网络模型兼容，如下图所示：

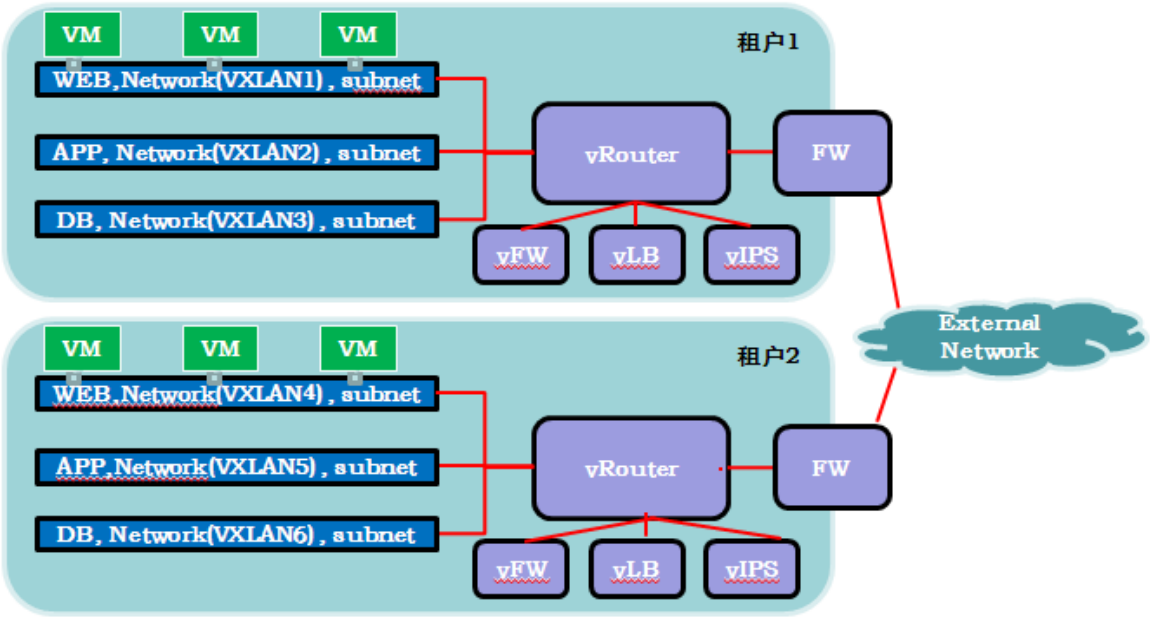


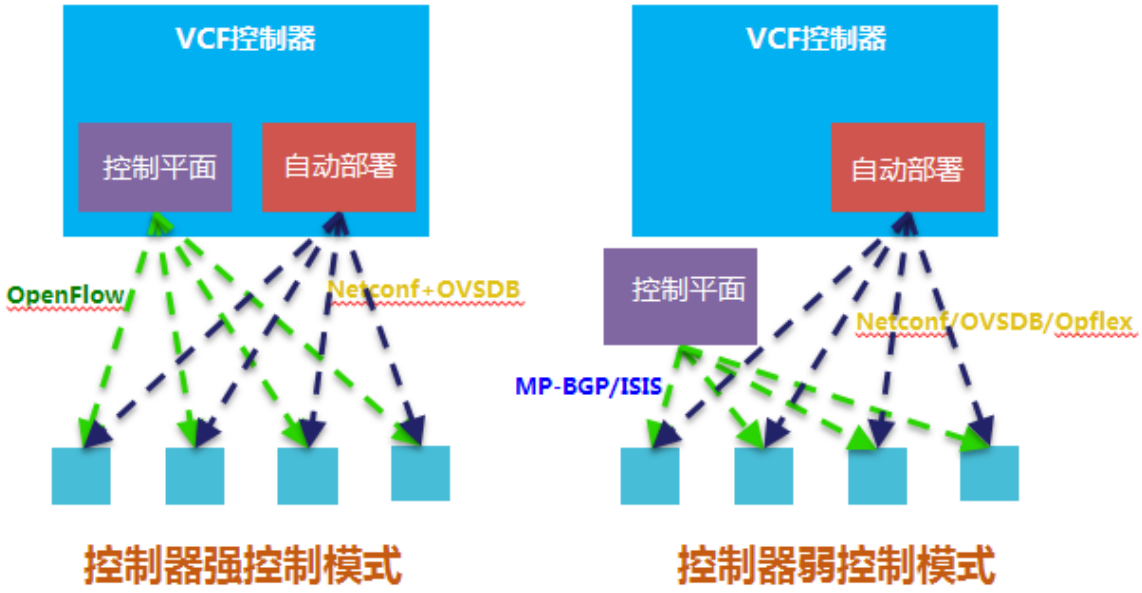
图2 VPC 多租户资源池场景

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的或逻辑的交换机。
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。

3.2 SDN控制器模型介绍



从控制器是否参与转发设备的的转发控制来看，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：

  1、通过控制器集群增加控制器可靠性，避免单点故障

  2、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。

3.3 H3C SDN Overlay组件介绍

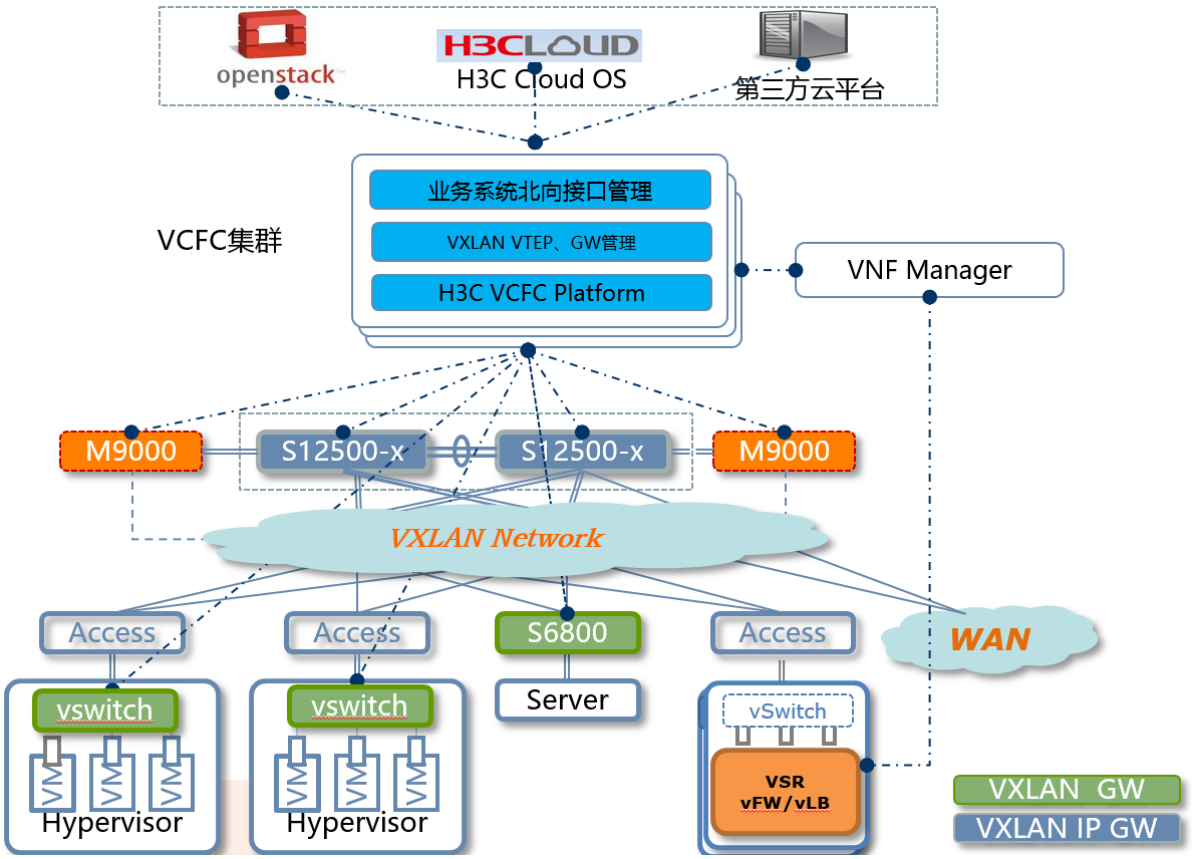


图3 H3C SDN Overlay 组件介绍

如上图所示，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 H3C Cloud 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 ,S68,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 S125-X, S98, 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。

3.4 SDN Overlay网络与云对接

公有云或私有云（VPC）对网络的核心需求是：

- 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。

3.4.1 SDN Overlay 与 openstack 对接

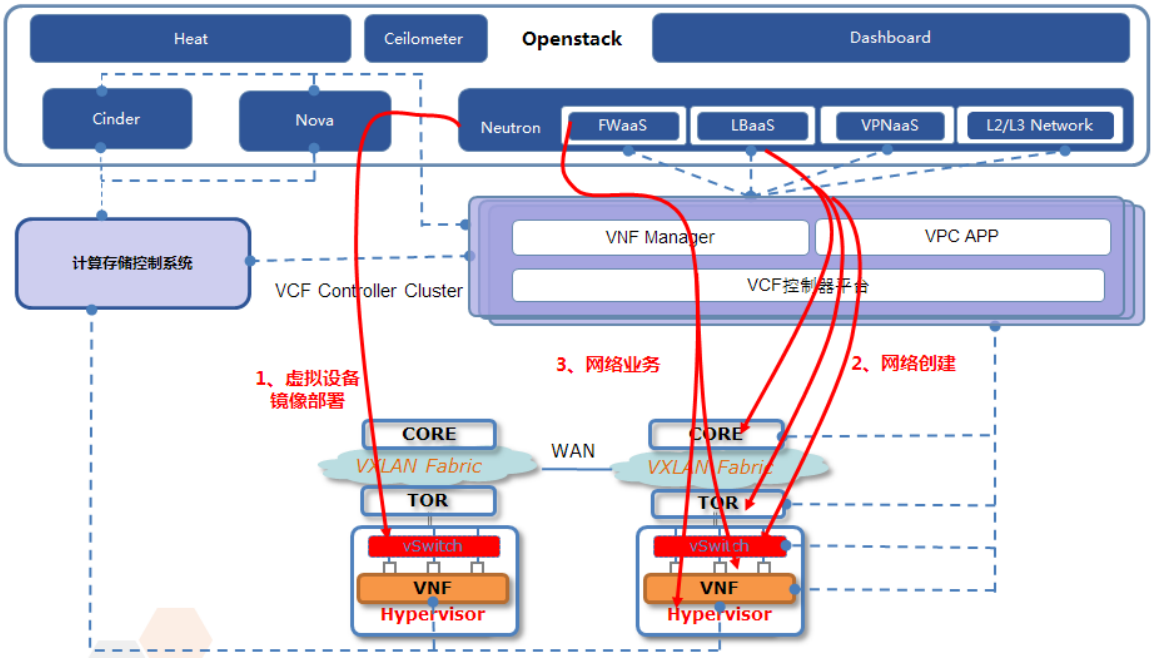


图4 SDN Overlay 与 openstack 对接

如上图所示，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 定义的插件如下表所示：

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 driver，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 H3C CloudOS 对接也是采用 Neutron 插件的方式。

3.4.2 SDN Overlay 与基于 openstack 的增强云平台对接

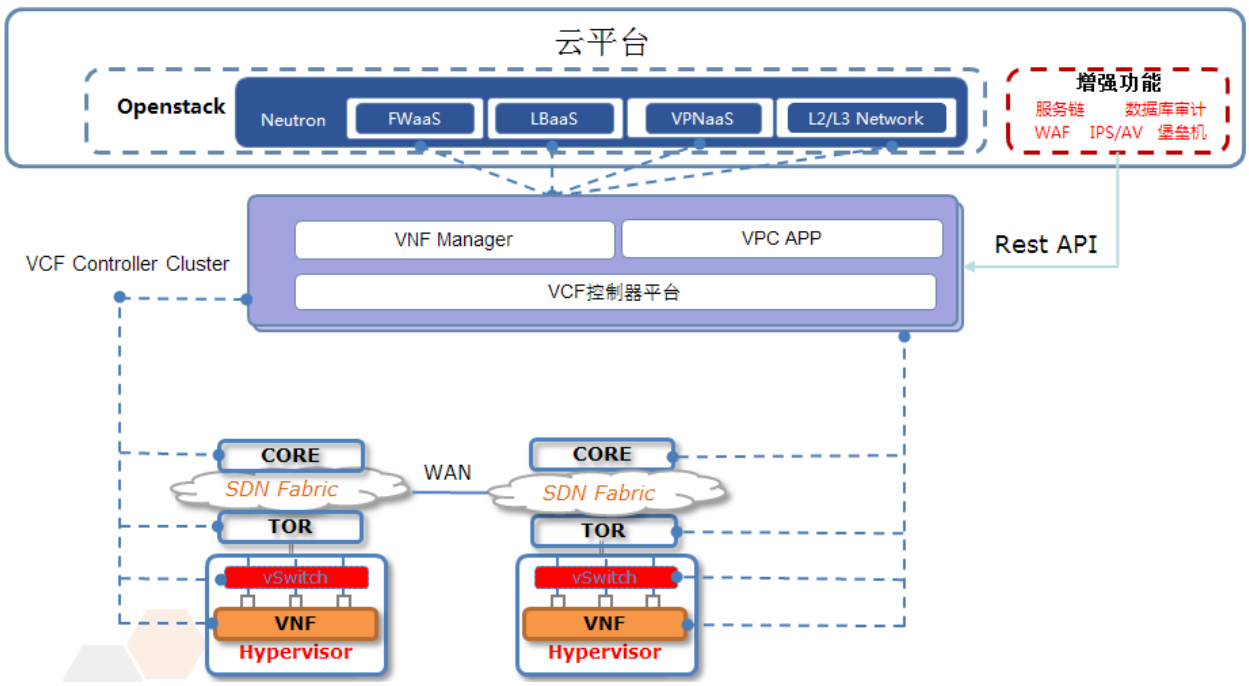


图5 SDN Overlay 与基于 openstack 的增强云平台对接

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。

3.4.3 SDN Overlay 与非 openstack 云平台对接

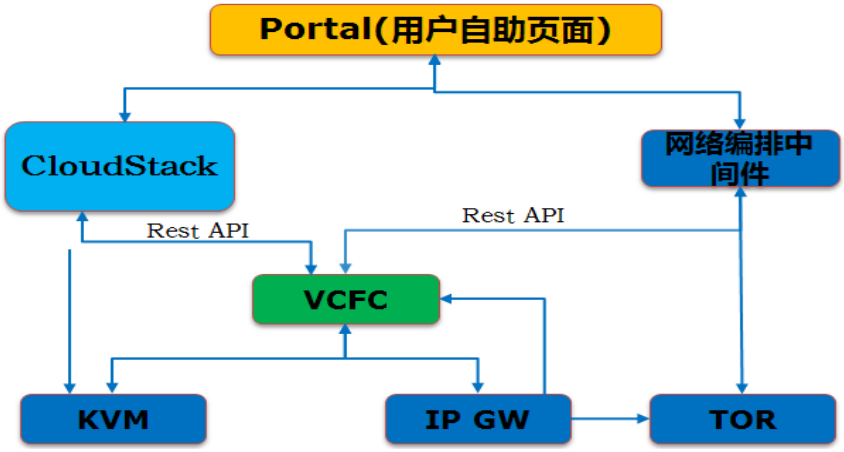


图6 SDN Overlay 与非 openstack 云平台对接

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 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。

4.1 SDN Overlay组网模型：

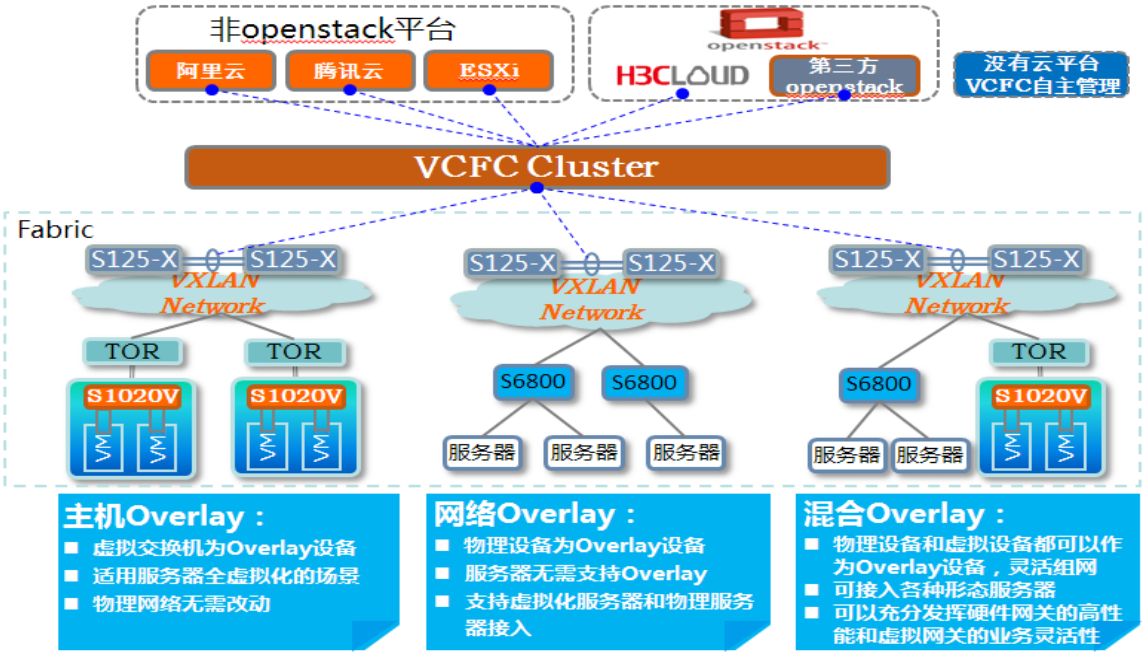


图7 SDN Overlay 组网模型

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。

4.1.1 网络 Overlay

- 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。

- 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。

4.1.2 主机 Overlay

- 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 VMAWRE、KVM 等主流 Hypervisor 平台的 overlay 方案。

- 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。

4.1.3 混合 Overlay

- 定位

混合 Overlay 组网灵活，即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。

- 面向客户

混合 Overlay 主要面向愿意即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。

4.2 H3C SDN Overlay典型组网

4.2.1 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性能转发性能比较高，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：

1、**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，设备会临时切换到自转发状态进行逃生

2、**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。

3. **混合转发模式：** 同时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。

。

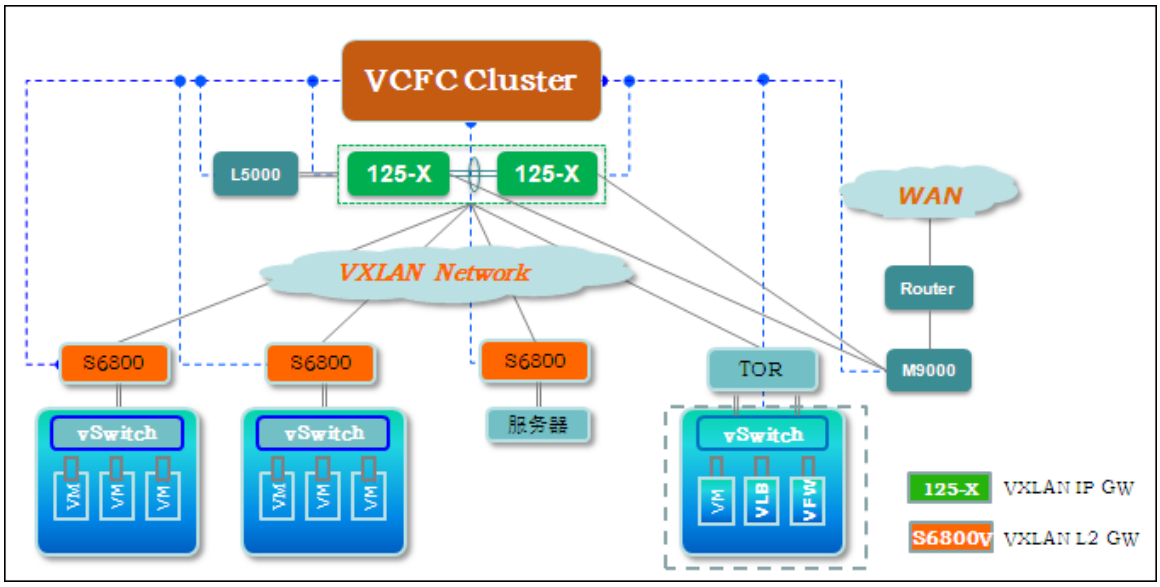


图8 网络 Overlay

在图 14 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 125X/S98 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 125X 串联 M9K 实现 NAT、FW 等服务，125X 旁挂 L5000 提供 LB 服务，由 VCFC 实现引流。

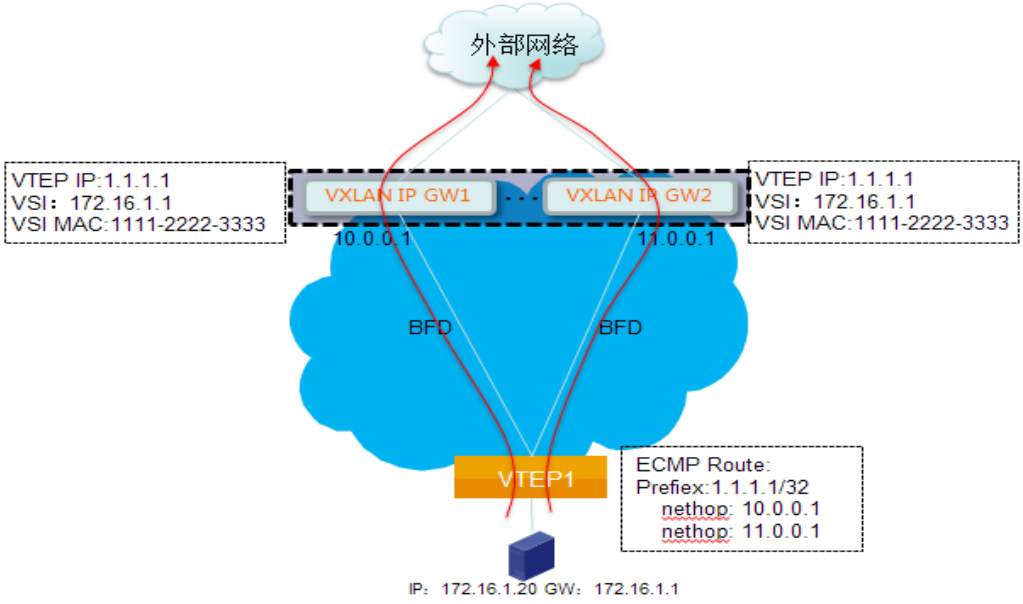


图9 无状态 IP 网关

如上图，在网络 Overlay 的组网模型中，125X/S98 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：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。

4.2.2 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 4 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 5 元组下发规则，可以灵活决定报文是允许还是丢弃。

状态防火墙和安全组的区别是，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。

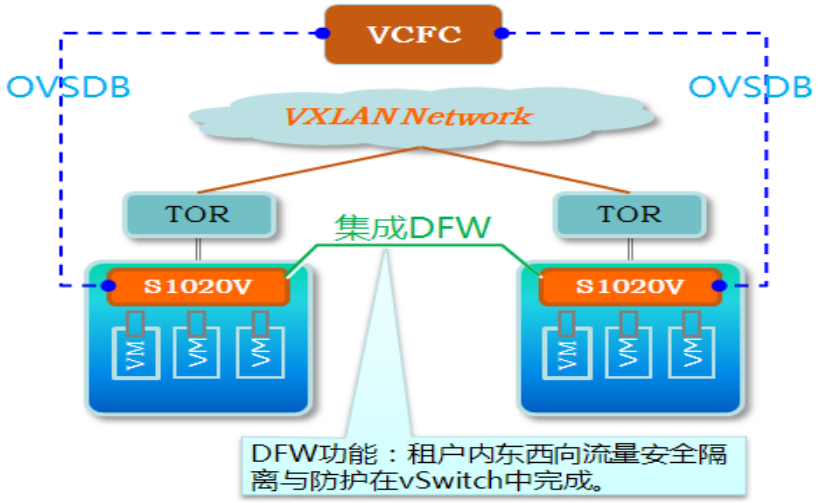


图10 vSwitch 集成状态防火墙

如上图所示，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。
- 配置 DFW 策略后，OVS 的原有转发流程会以黑盒的形式嵌入到 Netfilter 框架的报文处理过程中，接收到报文后依据配置的 DFW 策略在 Netfilter 的对应阶段调用相应的钩子函数实现对应的防火墙功能。
- 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 即承担了 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 17 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。

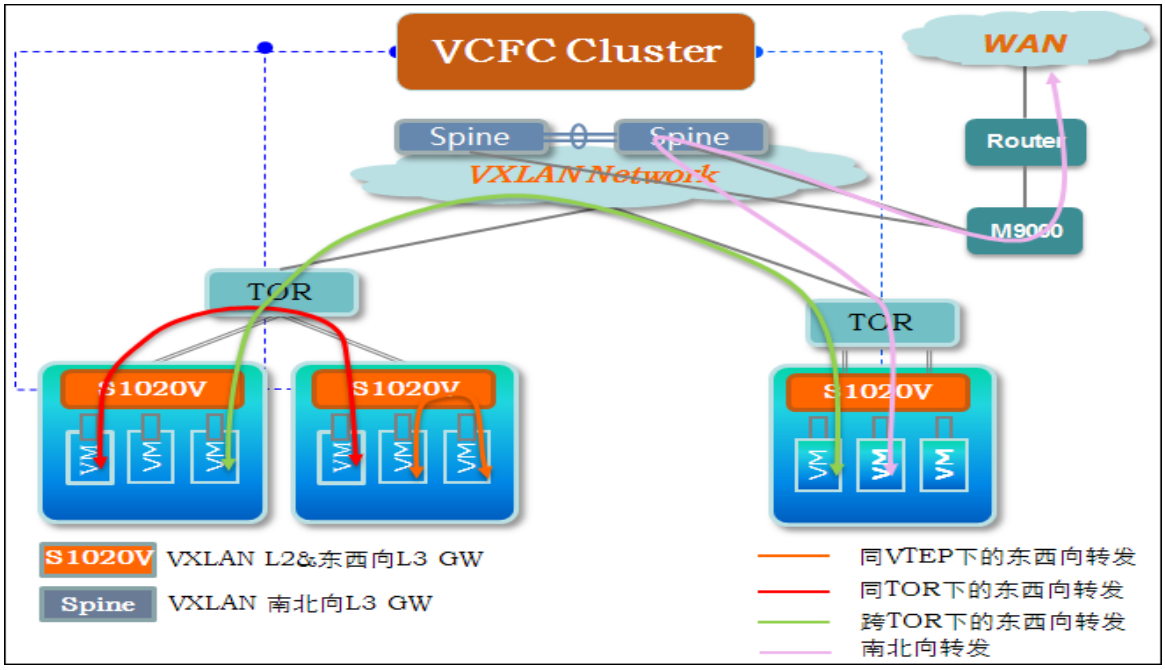


图11 东西向分布式网关方案

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，如下图所示：

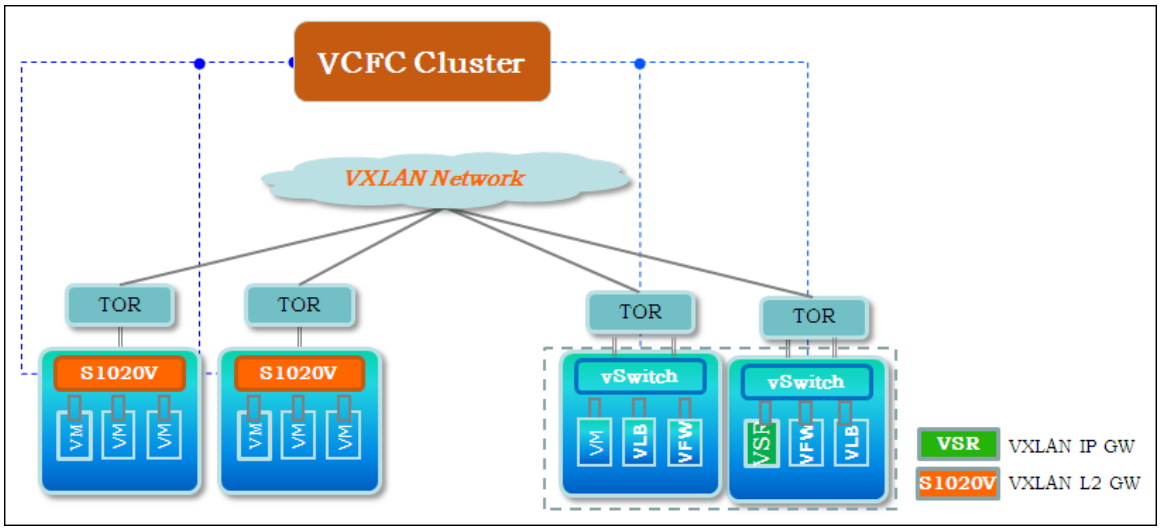


图12 VSR 做网关的主机 Overlay 方案

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 19 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 125-X/98，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 125-X 串联 M9000 实现 NAT、FW 等服务，125-X 旁挂 L5000 提供 LB 服务，由 H3Cloud OS 通过 PBR 实现引流。

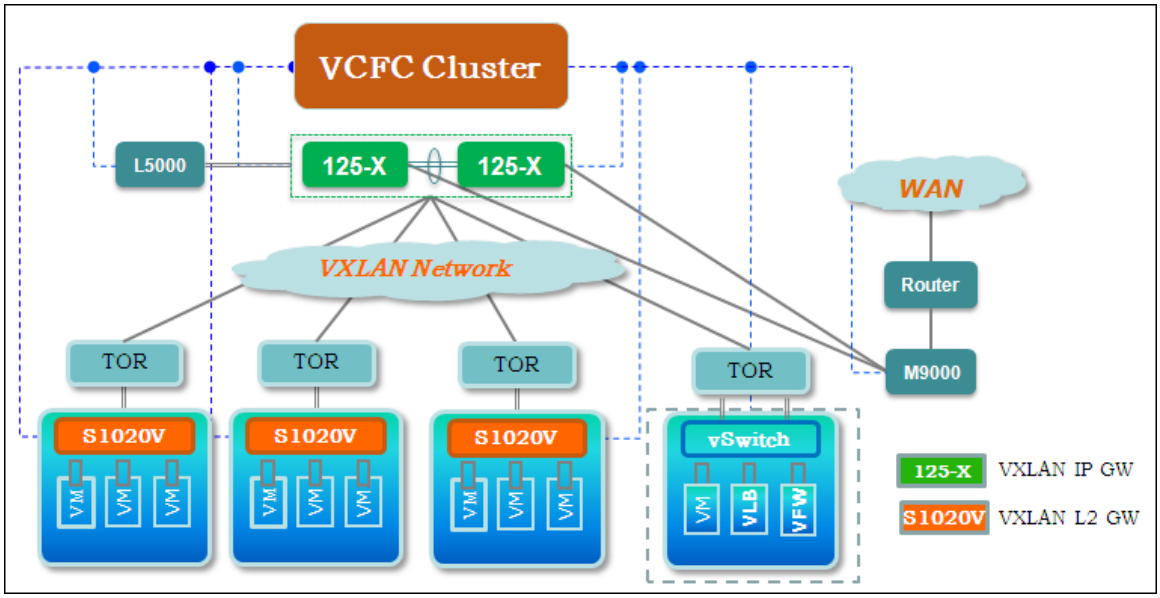


图13 物理交换机做网关的主机 Overlay 方案

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。

4.2.3 混合 Overlay

如图 20 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。

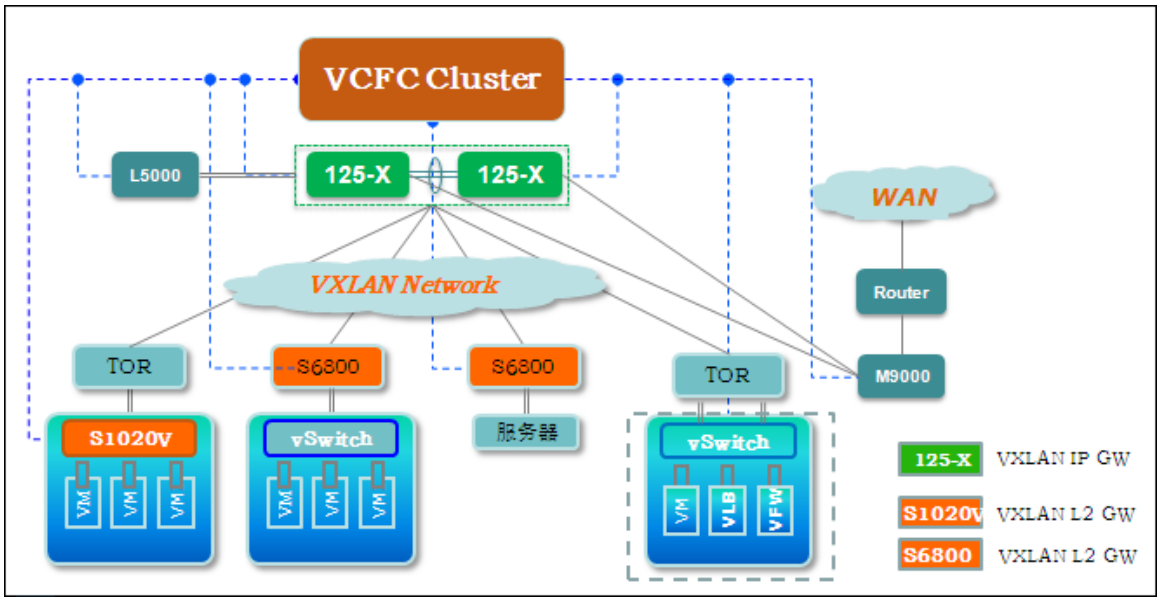


图14 混合 Overlay

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。

125X/S98 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 125X 串联 M9K 实现 NAT、FW 等服务，125X 旁挂 L5000 提供 LB 服务，由 VCFC 实现引流。

4.2.4 Overlay 组网总结

类别	组网	虚拟化平台支持	转发模型	适用场景	服务链方式
主机 overlay	S1020V+VSR	CAS/VMWARE/KVM	流转发	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限条件	南北向 VSR(自带 FW 功能)+VLB，东西向共享南北向 NFV，

				下的私有云	都采用服务链方式
	S1020V+S125	CAS/VMWARE/KVM	流转发	同纯软主机 Overlay 方案相比，主机 Overlay 软硬结合方案使用 125-X 或 10500 做 VXLAN IP GW，跨网段转发性能较高； 跟网络 Overlay 相比，对 TOR 没有要求，不要求 TOR 承担 VTEP 功能	南北向采用 PBR（M9000+L5000)东西向 VFW+VLB 单跳或者多跳服务链
网络 overlay	S68+S125	ALL	流转发/自转发	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+S125 ＋ S1020V	CAS/VMWARE/KVM	流转发	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。

5 H3C SDN 服务链

5.1.1 基本概念

H3C SDN 服务链，基于网络的核心控制部件 SDN 控制器——VCFC （Virtual Converged Framework Controller）进行部署。VCFC 根据租户需求，定义、创建服务链，并部署服务链上每个节点的业务逻辑。VCFC 将需要进入服务链处理的用户报文特征，下发到接入软件/硬件 VTEP，从而将数据报文引入服务链。

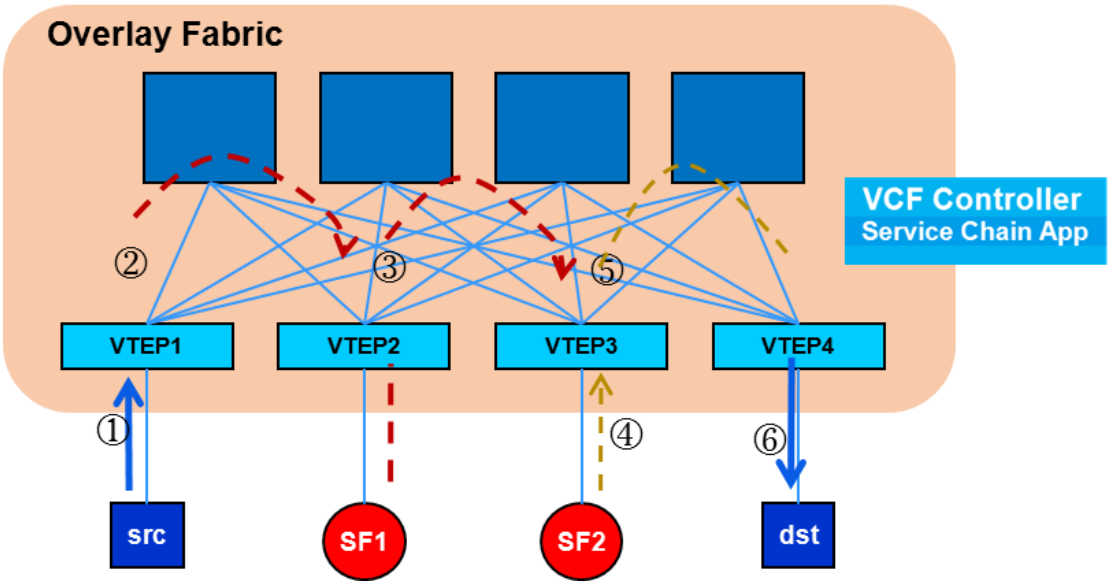
H3C SDN 服务链中具有如下角色：

- 流分类节点(Classification)：也是原始数据报文的接入节点。按照定义的流分类规则匹配数据报文，对报文做服务链的 Overlay 封装,并将其转发到服务链中处理。
- 服务节点（Service Function）：服务节点作为资源被分配使用，它的物理位置可以是任意的，分散的，通过 SDN 对服务链的定义和引流串联，完成预定义的工作。服务节点可以是防火墙（FireWalls）、负载均衡（LoadBalance）、入侵检测（Intrusion Prevention System）等资源/资源池。

- 代理节点（Proxy Node）：对于不支持服务链封装的服务节点，需要通过代理节点剥离服务链封装，将业务策略信息转换成 VLAN 等，转交给服务节点处理。
- 控制平面（Control Plane）：负责管理服务链域内的设备，创建服务链，将服务节点的配置定义，下发到各个相关节点上。在 H3C SDN 网络中，通过 VCFC 实现。

5.1.2 转发流程

如下图所示，是服务链的典型示意图：



其中各对应角色及其处理为：

- VCFC：H3C SDN 控制器。作为网络资源池的唯一控制点，VCFC 控制了虚拟化网络，并且通过对虚拟网络进行抽象和编排，定义服务链特征；VTEP 和服务节点上的转发策略都由控制器下发。
- 服务链流分类节点（VTEP1）：原始报文通过 VTEP1 接入 VXLAN 网络，并直接进行流分类，以确定报文是否需要进入服务链：如果需要进入服务链，则将报文做 VXLAN+服务链 ID 的封装，转到服务链首节点处理。
- 服务链首节点（SF1）：进行服务处理后，将数据报文继续做服务链封装，交给服务链下一个服务节点。
- 服务链尾节点（SF2）：进行服务处理后，服务链尾节点需要删除服务链封装，将报文做普通 VXLAN 封装，并转发给目的 VTEP。如果 SF2 不具备根据用户报文寻址能力，则需要将用户报文送到网关(VTEP3)，VTEP3 再查询目的 VTEP 进行转发。

报文转发说明如下：

- ①⑥Native以太报文，IP(src)---->IP(dst)
- ② VXLAN+业务链报文，外层: IP(VTEP1)---->IP(SF1)
- ③ VXLAN+业务链报文，外层: IP(SF1)---->IP(SF2)

- ④ VXLAN报文，外层: IP(SF2)---->IP(VTEP3)
- ⑤ VXLAN报文，外层: IP(VTEP3)---->IP(VTEP4)

具体的匹配转发流程描述如下：

- VM首包上送控制器处理时，在VCFC上解析packet in报文，根据报文目的地址确定是虚拟网络内的东西向流量还是通往传统网络的南北向流量，对于南北向流量则将报文转发到网关设备，报文后续的处理由网关设备负责；
- 对于东西向流量，从收到的packet in报文中提取源端口，并根据源端口确定源subnet、network、router信息；并根据packet in报文的目的IP获取目的VM连接的目的端口，并根据目的端口确定目的subnet、network、router信息；
- 对于东西向流量，根据报文特征进行服务链匹配，首先使用源端口和目的端口的属性与服务链配置进行匹配，如果找到匹配的服务链，则下发导流流表；如果没找到匹配的服务链，就下发东西向卸载的流表项（即非服务链转发）；
- 如果存在匹配的服务链，确定服务链所在VTEP的VTEP IP，VCFC向VTEP和后续处理节点下发流表项，流表项格式如下：

  Match：port & vlan & 五元组　　普通报文进入服务链

  　　或 tunnel & vni & 五元组　　vxlan 报文 进入服务链

  Action：vni & service chain id & order counter & tunnel，指向服务链首节点所在的服务节点，order counter =1。
- 当匹配到多条服务链时，按照最精确匹配的原则确定实际使用的服务链配置。上述4种维度按照精确程度从低到高排序依次为：Routers, Networks, Subnets, Ports。

5.1.3 服务链流分类节点的类型

H3C SDN 服务链支持如下几种流分类节点：

- 支持业务链的 vSwitch：vSwitch 收到虚机报文后，直接做流分类；在 vSwitch 上进行服务链 Overlay 封装。例如 H3C s1020v。
- 硬件交换机接入普通 vSwitch：虚机通过普通 vSwitch 接入，vSwitch 仅作二层交换使用，上送硬件交换机后，由硬件交换机进行流分类，进行服务链 Overlay 封装。例如 H3C s6800 交换机。
- 硬件交换机接入物理设备：硬件交换机直接接入物理设备，对物理设备发送的数据报文做流分类，进行服务链 Overlay 封装。例如 H3C s6800 交换机。
- 硬件交换机接入普通 VXLAN 报文：普通 VXLAN 报文上送到硬件交换机，由硬件交换机进行流分类，进行服务链 Overlay 封装。例如 H3C s6800 交换机。

5.1.4 服务链服务节点的类型

H3C SDN 服务链支持如下几种服务节点：

- H3C NFV 服务节点：支持 VXLAN 和服务链功能。可以直接进行 VXLAN 封装/解封装处理以及业务处理。
- H3C 硬件安全设备：支持 VXLAN 和服务链功能。可以直接进行 VXLAN 封装/解封装处理以及业务处理。
- 传统物理服务节点：第三方厂家的传统防火墙、LB 等无法支持服务链的安全功能节点，通过服务链代理节点（例如 H3C s6800 交换机）外挂。
- 服务链尾节点：删除服务链封装，根据用户报文的目的地址，找到最终目的 VTEP 的 IP 地址：

  1) 服务链尾节点自行向 VCFC 请求解析目的 VTEP IP，做普通 VXLAN 转发。

  2) 没有解析处理能力，则做普通 VXLAN 封装，转交其他网关处理。

5.1.5 服务链在 Overlay 网络安全中的应用

Overlay 网络服务链节点描述

图15 overlay 网络服务链节点描述

如上图所示，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。

服务链在 overlay 网络安全中的应用

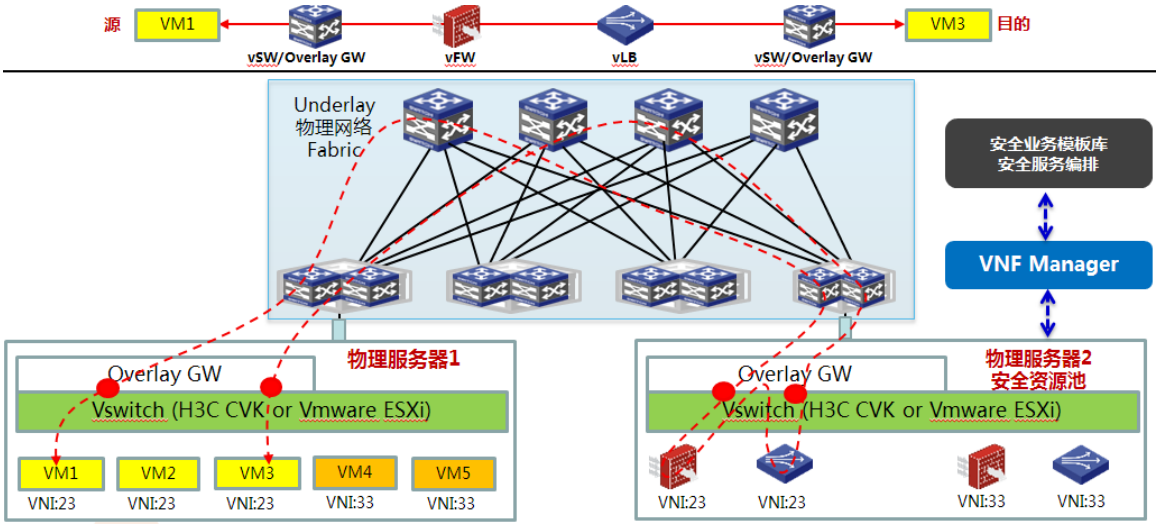


图16 overlay 网络服务链流程描述

上图是一个基于 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，必须通过中间这样 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。

6 H3C SDN 服务链部署模式

6.1 虚拟路由器VSR做网关的服务链应用

该模式组网下，H3C VSR 充当 VXLAN IP GW，提供 Overlay 网关功能；H3C S1020v 充当 L2 VTEP，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。

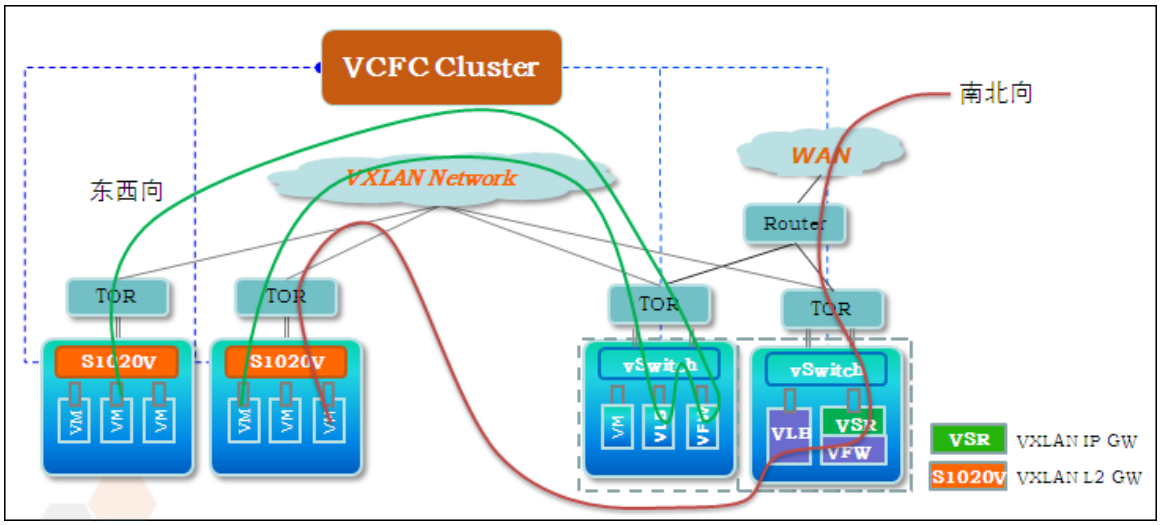
Service 安全节点包括 VSR、VFW、VLB 等设备，通过 H3C VCF 控制器集中控制和编排实现东西向和南北向服务链服务节点的功能。服务链能够支持 VM vport、vRouter、Network、Subnet、IP 地址进行服务链分流配置服务链，服务节点的经过顺序是先 VFW 再 VLB。

在此场景下，依据安全服务功能的配置位置，又可以分为灵活服务链模型和 Openstack 服务链模型两种。

6.1.1 灵活服务链模型

如下图所示，服务链可以直接在 VCFC 上配置实现灵活服务链部署。

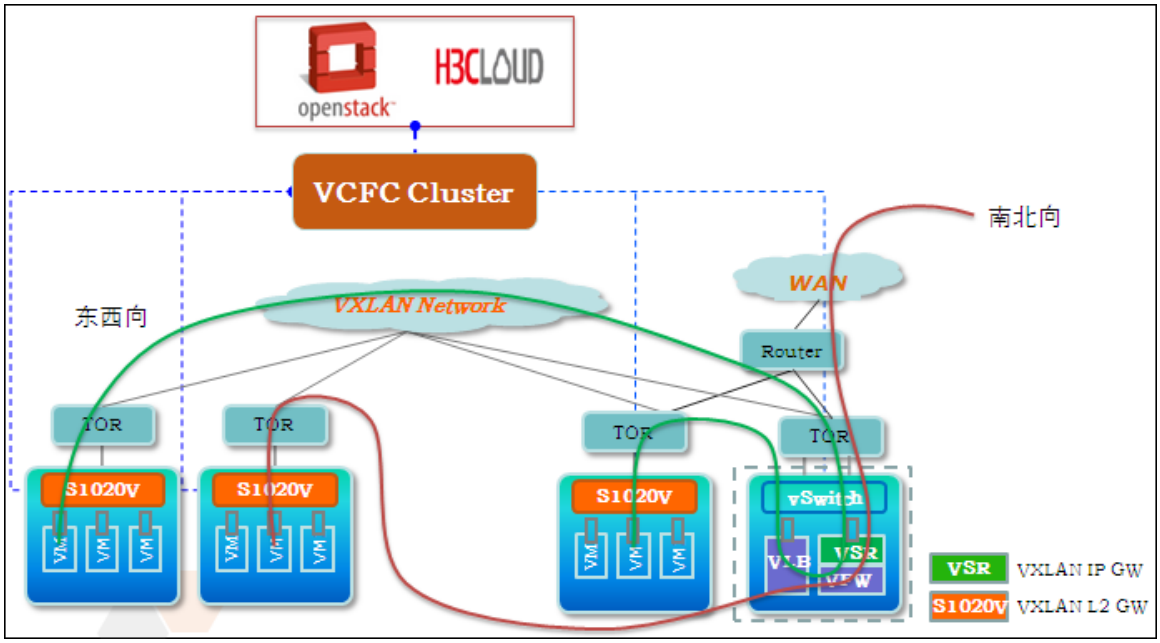
这种场景下，南北向服务链由 VSR 集成 VFW，而 VLB 可以独立部署；东西向 VFW、VLB 也可以独立部署。



6.1.2 Openstack 服务链模型

如下图所示，在 Openstack、H3Cloud OS 等云平台上配置安全服务，VCFC 配合实现安全服务功能，安全服务功能在云平台上配置。

这种场景下南北向由 VSR 集成 VFW 功能，VLB 可以独立部署；东西向跨网段 VFW 业务集成在 VSR 上，VLB 可以独立部署。



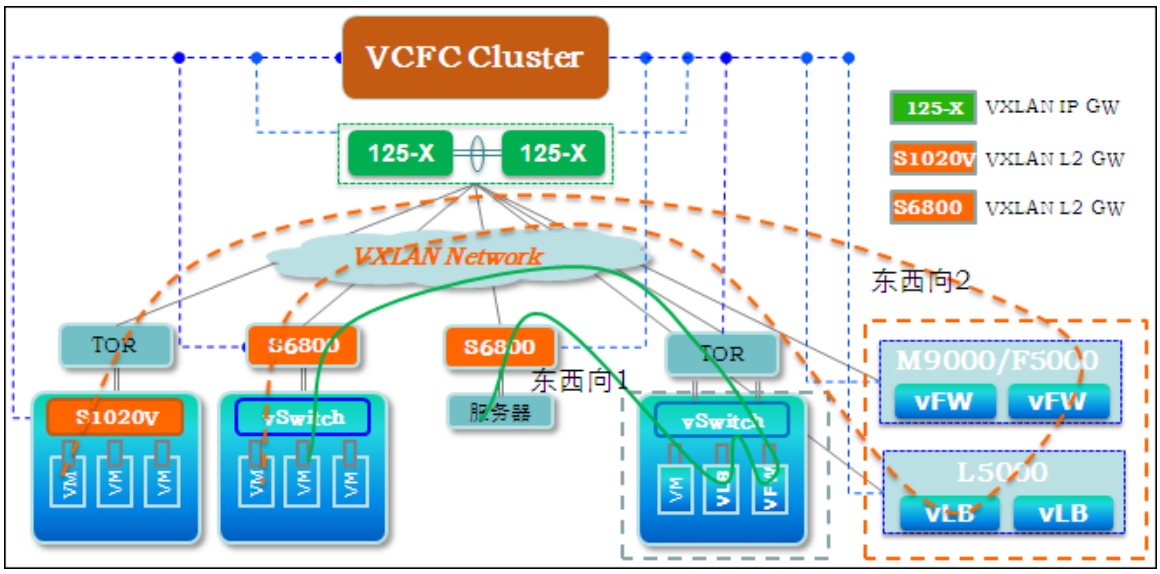
6.2 物理网络设备做VXLAN网关的服务链应用

该模式组网下，125X/98 充当 VXLAN IP GW，提供 Overlay 网关功能；H3C S1020V、S6800 充当 L2 VTEP，实现虚拟机或物理服务器接入到 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 安全节点包括 VSR、VFW、VLB、M9000/F5000、L5000 等设备。同 3.1 类似，也分为灵活服务链模型和 Openstack 模型。

6.2.1 灵活服务链模型

如下图所示，在 VCFC 配置实现东西向服务链提供安全服务。其中，服务节点支持硬件形态或 NFV 形态。

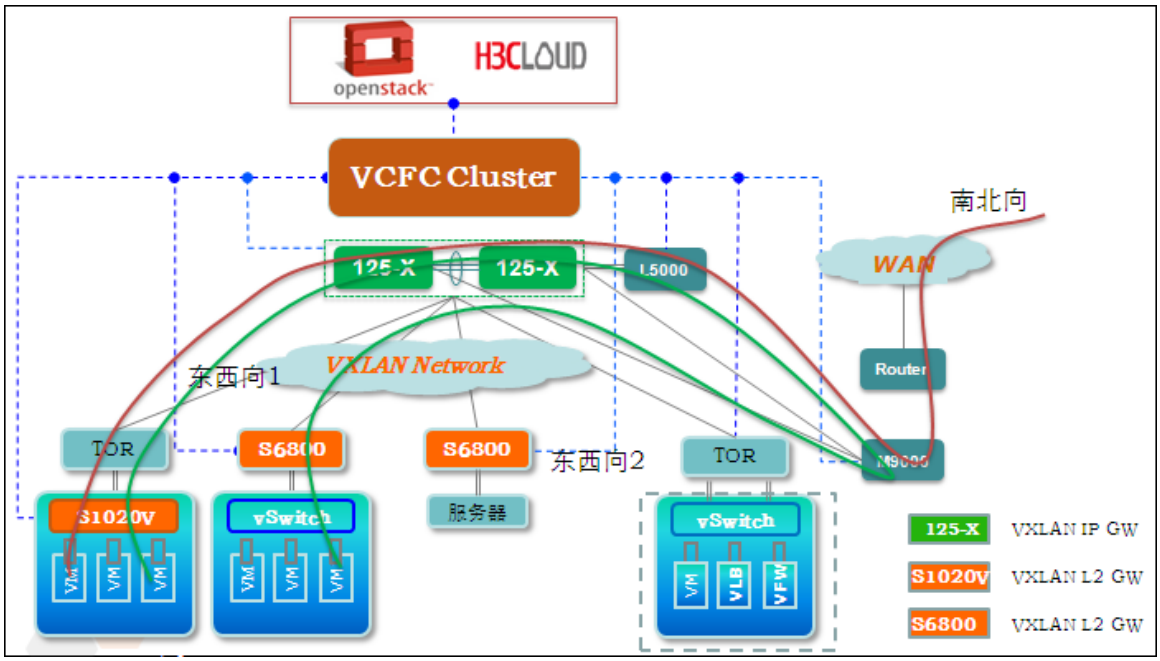


该模型中，南北向服务链无需专门编排，可以通过 125X 串联 M9K 实现 NAT、FW 等服务，125X 旁挂 L5000 提供 LB 服务；由 VCFC 通过 PBR、静态路由等技术实现引流。

6.2.2 Openstack 模型

在 Openstack、H3Cloud OS 等云平台上配置安全服务，VCFC 配合实现安全服务功能，安全服务功能在云平台上配置。

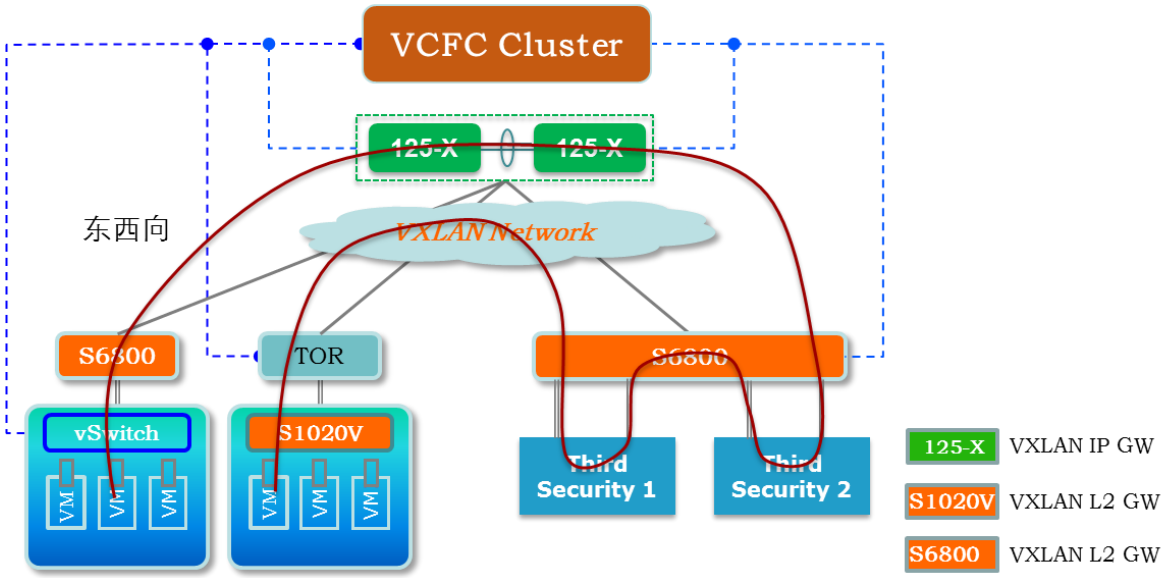
如下图所示，南北向 s125-X、s9800 做 VXLAN 终结，在 OpenStack、H3Cloud OS 进行相关安全资源配置后，由 VCFC 向 M9000、L5000 等安全设备自动下发 PBR、静态路由等配置，实现对安全流量的引导；对于跨网段的东西向流量，也可以共享南北向的 M9000、L5000 等安全设备，由 VCFC 自动下发 PBR、静态路由等配置，将流量引流到 M9000、L5000。



6.3 第三方安全设备服务链代理应用

传统的安全设备不支持 VXLAN 和服务链功能。H3C 通过服务链代理设备，可以将传统的甚至是第三方的安全设备引入 H3C SDN 服务链，从而共享 SDN 服务链先进技术，为客户提供更加多样化的选择。

如下图所示，该模式组网下，传统安全设备通过 s6800 的 AC 口（可以是 s6800 的堆叠、聚合后的逻辑接口）222 接入 SDN VXLAN 网络；s6800 作为服务链的服务代理节点，进行 SDN 服务链的报文特征识别和解析。S6800 与安全设备二层连接，其同一个接口可以接入多台虚拟化安全设备，并以 VLAN 进行区分。



第三方安全设备服务链代理的应用，可以支持第三方安全设备通过 s6800 的单臂模式和双臂模式两种接入模式，实现了网络中东西向流量的安全自动防护，也为客户的安全设备选型提供了更加多样的选择。在这种模式下，VCFC 不再识别具体的安全业务类别，只需识别业务所在 s6800 的接口，并负责导流到该接口。

7 第三方安全设备对接纳管

H3C SDN 解决方案涵盖了对第三方安全设备对接纳管功能，可实现对第三方安全设备的对接和流量牵引

VCFC 的安全纳管提供了 VLAN 及 VXLAN VPC 下 GW，FW 等服务的硬件资源虚拟化方案。实际的 OpenStack 项目中，有需求希望第三方厂商的防火墙设备能够接入 VCFC 的安全纳管方案，为用户的云业务提供安全服务。

7.1 实施准备

VCFC 已有的安全纳管方案，控制器需要向安全设备下发：

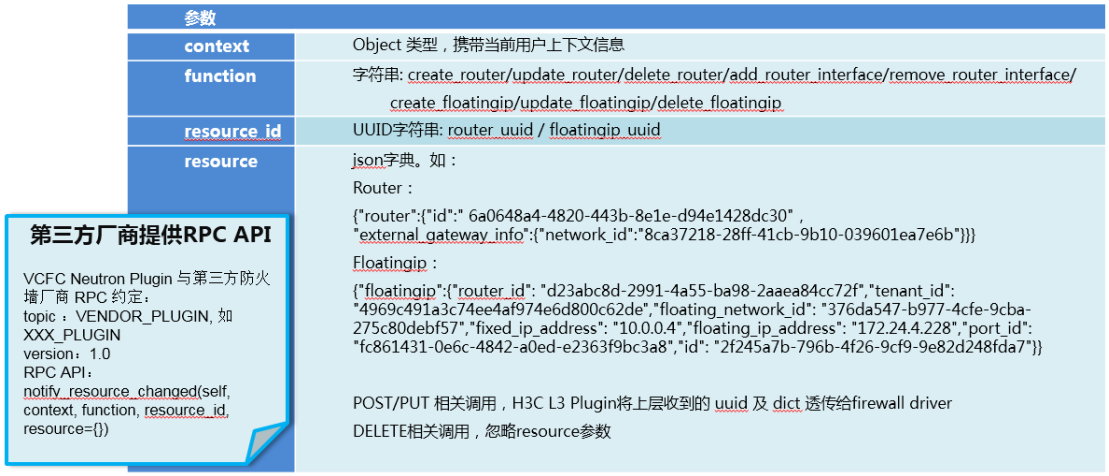
1）防火墙服务的安全配置

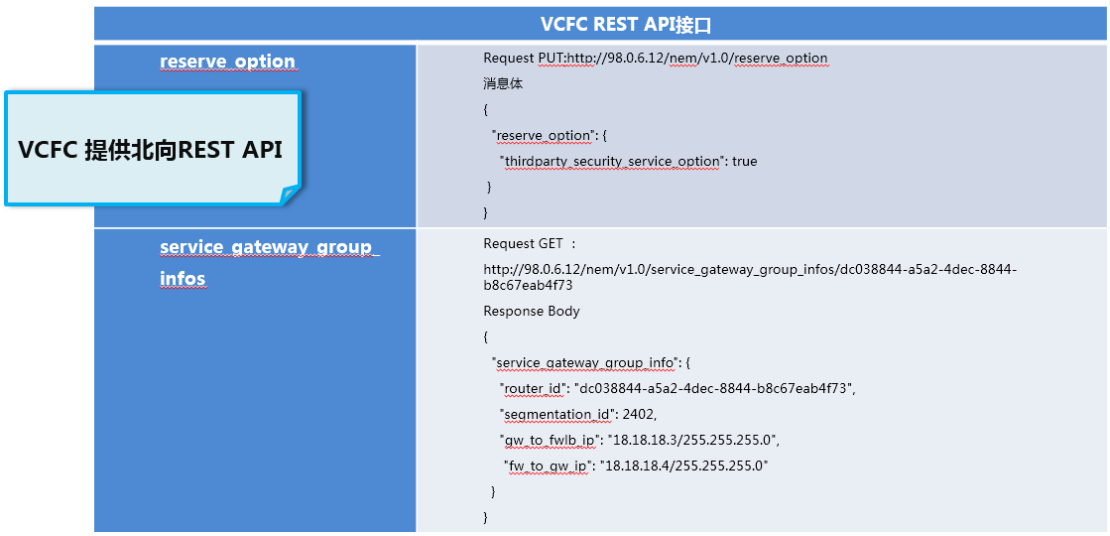
2）控制器为 Router 分配的租户承载网及安全内网的 VLAN ID 接口及网段地址

3）南北向业务指导转发的 Router 各网段的网段路由

4）南北向业务指导转发的 SNAT 及 DNAT 的相关配置

7.2 原理介绍





VCFC Neutron Plugin 实现：

- 新增加 vendor_rpc_topic 配置项，配置为第三方厂商 RPC TOPIC。对接时配置为 Plugin 和第三方防火墙使用 RPC 机制进行 L3 扩展数据的通信，上下文独立 ，耦合性低
- L3 Plugin 在 L3 所有对外提供的接口增加对 vendor_rpc_topic RPC TOPIC 的事件通知机制，实现上使用异步 CAST 调用，忽略通知过程中遇到的异常信息

7.3 第三方厂商的参考实现

一、防火墙服务的安全配置

需求实现：1）第三方厂商提供符合 OpenStack Neutron FW 要求的 FWaaS Driver； 2）配置 OpenStack Neutron 防火墙 Driver 为第三方厂商，OpenStack Neutron FW Plugin 会将防火墙消息通告给第三方厂商 Driver，再由 Driver 向设备下发安全配置

二、控制器为 Router 分配的租户承载网及安全内网的 VLAN ID 接口及网段地址

需求实现：1） 控制器提供获取为 Router 分配的租户承载网，安全内网的 VLAN ID 及网段 IP 的北向 REST API；2)第三方厂商 Driver 提供 RPC 监听机制，由 VCFC L3 Plugin 向第三方安全厂商 Driver 提供 Router 创建及删除的完整消息。Driver 监听到后向 VCFC 获取 1）的 REST API 消息，并向设备下发与网关设备互联的接口配置

三、南北向业务指导转发的 Router 各网段的网段路由

需求实现：1）第三方厂商提供 RPC 监听机制，当收到 Router 绑定解绑 Subnet 消息时，通知 FW 设备下发或者删除相应的 Subnet 网段路由； 2）VCFC L3 Plugin 向第三方厂商的 RPC TOPIC 发送 Router 绑定及解绑 Subnet 事件 通知消息

四、南北向业务指导转发的 SNAT 及 DNAT 的相关配置

需求实现：1)第三方厂商提供 RPC 监听机制，当收到 Router 绑定网关，FloatingIP 创建，更新，删除消息时，通知 FW 设备下发或删除相应的 NAT 配置； 2）VCFC L3 Plugin 向第三方厂商的 RPC TOPIC 发送 Router 绑定及解绑网关，FloatingIP 创建，更新，删除事件通知消息

7.4 优点介绍

- VCF Plugin 提供了 h3c-agent(准开源 l3-agent)进程用来接收 FW RPC 消息，第三方防火墙厂商 仅需要提供防火墙 driver，省去了第三方防火墙厂商的部分开发工作
- VCF Plugin 和第三方防火墙使用 RPC 机制进行 L3 扩展数据的通信，上下文独立，耦合性低
- VCF Plugin 启动时加载配置项值作为和第三方厂商扩展数据 RPC 通信的 TOPIC，可维护性好
- VCF Plugin 本次定义的 RPC 扩展数据通信方法抽象度高，后续扩展消息内容不需要新定义接口

- VCFC 可以支持第三方安全厂商防火墙设备无缝接入安全纳管方案

8 SDN 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：**
  - 支持智能化诊断：全面覆盖的故障自动探测,雷达仿真,故障定位和自动修复；
  - 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。