



overlay SDN 实现异构兼容的关键技术

黄志兰,樊勇兵,陈楠,何晓武

(中国电信股份有限公司广州研究院,广东 广州 510630)

摘要:云计算的规模化发展对数据中心网络提出了新的挑战,要求应用 SDN 技术对数据中心网络进行虚拟化,以满足多租户环境下灵活调整的动态网络需求,overlay SDN 是实现数据中心网络虚拟化的核心技术。研究了 overlay SDN 网络虚拟化的关键技术,提出了商业 overlay SDN 存在的兼容性问题,分析了 overlay SDN 实现异构兼容的主要方法,并提出了基于开源和白盒交换机实现异构兼容的 overlay SDN 架构。

关键词:overlay SDN;异构兼容;白盒交换机

中图分类号:TP315

文献标识码:A

doi: 10.11959/j.issn.1000-0801.2016287

Key technologies of heterogeneous compatibility in overlay SDN

HUANG Zhilan, FAN Yongbing, CHEN Nan, HE Xiaowu

Guangzhou Research Institute of China Telecom Co., Ltd., Guangzhou 510630, China

Abstract: Cloud computing scale development provides a new challenge to data center network. Data center network is required virtualizing by SDN in order to meet flexible dynamic network requirement under multi-tenant environment. SDN is the core technology to realize data center virtualization. The key technology of overlay SDN network virtualization was studied, the compatibility problems remained in commercial overlay SDN products was discussed, the main methods to solve this problem was analyzed, and finally a new overlay SDN architecture based on open source software and white-box switch that could work in heterogeneous hypervisors environment was proposed.

Key words: overlay SDN, heterogeneous compatibility, white-box switch

1 引言

传统数据中心静态网络面临 VLAN 数量不足(4 096 个)、MAC 表容量有限、资源利用率偏低(STP 冗余链路)、网络配置复杂等问题^[1],已经不能适应云计算灵活多变的业务需求,需要引入 SDN 技术实现数据中心网络的虚拟化和自动化,使得网络变成可池化的资源,能够与计算、存储等资源一起按需提供给客户。

目前,市面上主流的 SDN 解决方案以厂商销售的

overlay SDN 方案为主,如 VMware(威睿)/Nicira 的 NSX/NVP、阿尔卡特朗讯的 Nuage VSP、Juniper(瞻博网络)的 Contrail、华为的 FusionNetwork 以及华三的 VCF 等。这些商业解决方案多为单厂商封闭方案,面临兼容性、开放性不足的问题,具体表现在:对异构 Hypervisor 的兼容性不足,虽然部分厂商宣称能够同时支持 VMware ESXi、KVM 和 Xen Server,但大多数厂商的 SDN 产品只能支持自己的 Hypervisor 产品;对异构设备兼容性不足,部分厂商方案可能将网关设备交由硬件实现以提高产品性能,纯硬件或者

软硬结合的方案对硬件设备的要求普遍比较高,通常只支持厂商自身的设备;控制器私有方式实现,普遍存在开放程度低、北向接口缺乏标准化、采用私有南向协议等问题,商业方案整体开放性不足,并且不能与其他厂商方案互通,一方面增加了厂商锁定的风险,另一方面也限制了用户网络应用创新的能力。

开放、兼容性差制约了商业 SDN 产品在云数据中心的规模部署和应用。随着虚拟化技术的不断发展和云计算业务的深化,云数据中心的 ICT 环境变得更加复杂,表现在:多虚拟化平台共存,例如多数数据中心同时具有 VMware 资源池和 KVM 资源池,部分数据中心还设有独立的物理资源区;异构资源增多,数据中心的每次扩容都可能带入新的设备和软件;用户差异化、等级化的云服务需求,要求将资源分布在不同级别的云平台,并能实现跨平台的互通和管理。云数据中心复杂的异构环境,对 SDN 产品的开放、兼容性提出了更高的要求,要求能够兼容异构虚拟化平台,并能实现不同云平台间的网络互通。

针对上述问题,研究了 overlay SDN 的关键技术,分析了商业 overlay SDN 产品开放兼容性差的主要原因,总结了 overlay SDN 支持异构虚拟化平台主要方法,提出了基于开源和白盒交换机实现异构兼容的通用 overlay SDN 架构。

2 overlay SDN 关键技术

overlay SDN 通过隧道技术在现有物理网络架构上叠加虚拟的逻辑网络,实现网络虚拟化^[2]。在 overlay SDN 中,租户网络的流量以隧道的方式封装并在底层 IP 网络上透传,隧道建立等复杂的控制逻辑由 SDN 控制器集中处理。控制器向上开放北向接口,供用户编程网络转发行

为,用户可以根据需求灵活定义逻辑网络、调整网络策略;向下则通过南向协议管理和控制隧道端点,通过控制隧道端点来编程逻辑网络。可见,overlay SDN 是通过软件方式来改造 IP 网络,实现现有 IP 网络开放可编程的软件定义网络技术。overlay SDN 概念如图 1 所示。

overlay SDN 常用的隧道协议有 VxLAN、STT、NvGRE、Geneve^[3]。这些隧道协议最显著的特点是,采用 MAC-in-IP 的方式进行封装,能够跨越三层网络上实现大二层网络互通,并且这些协议大多保留超过 24 bit 标识用于区分租户网络,使得租户规模能够高达 16 MB,从而突破 VLAN 4 KB 限制。表 1 总结了 4 种隧道协议的关键特点。

从标准成熟度和产业支持来看,VxLAN 发展最好,也因此成为了 overlay SDN 的主流协议,被众多商业方案所采用。

overlay 将网络分为核心和边缘两层,网络智能存在于边缘设备中,负责隧道封装/解封装、提供业务接入及网络服务等,边缘层可以位于服务器或 TOR 网络设备,通常由虚拟接入交换机、Linux 网桥等软件设备构成。核心层为底层物理传输网络,只负责高效数据传输,因此完全不需要修改底层网络架构。

除了控制器、虚拟网络边缘(virtual network edge, VNE)节点外,overlay SDN 方案还有二层网关和三层网关两个重要的功能实体。其中,二层网关也称 VxLAN 网关,负责实现 overlay 网络与外部网络的 L2 互通,三层网关则在二层网关的基础上增加路由的功能,实现 overlay 网络与外部网络的 L3 互通。两种网关都可以采用软件或者硬件的形态部署。

根据隧道端点、网关的实现形态不同,overlay SDN 可

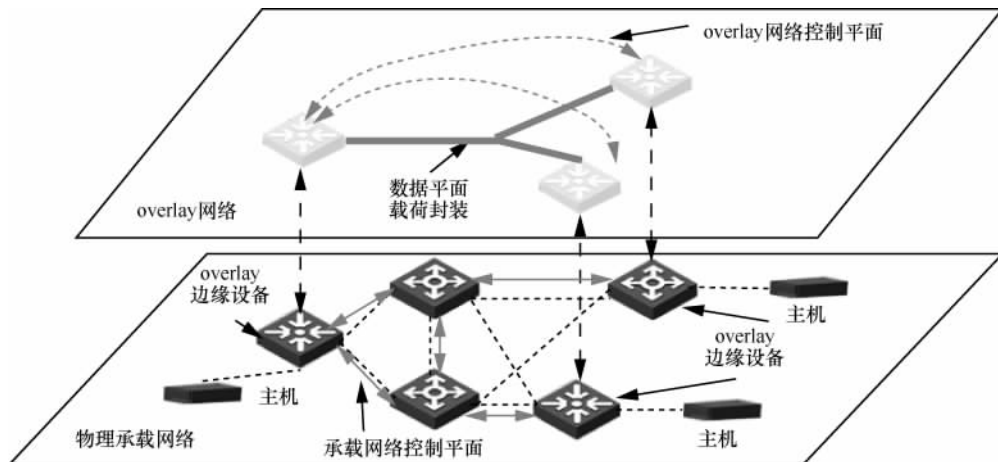


图 1 overlay SDN 概念



表 1 overlay 隧道协议对比

协议类型	网络标识数量	封装方式	封装分组头长度	标准情况	支持者	优缺点
VxLAN	24 bit	UDP (端口号 4789)	50 byte	RFC7348	思科、VMware、博通等	优点:支持 ECMP/LB 缺点:数据分组较大,要求传输网络支持大型帧
NvGRE	24 bit	GRE	42 byte	IETF 草案	微软、惠普、Arista 等	优点:支持非 IP,支持多播地址复用和多路单播 缺点:GRE 分组不被传统的负载均衡/防火墙设备支持
STT	64 bit	类 TCP	54 byte	IETF 过期草案	VMware/Nicira	优点:支持网卡的 TSO (TCP segmentation offload,TCP 分段卸载)功能,对 CPU 消耗更小;支持 ECMP/LB 缺点:缺乏控制面定义;技术复杂度高
Geneve	24 bit	UDP	可变, 最小 8 byte	ITEF 草案	VMware、微软、红帽、英特尔	优点:综合其余 3 种隧道协议的优势,自带控制信息,扩展性好; 缺点:尚不成熟

以有 3 种不同的部署方案,分别是纯硬方案、纯软方案和软硬结合方案。3 种方案的对比见表 2。

3 种部署方案各有优缺点,适用场景也不同,主流商业方案通常会提供一种或者多种实现方式。

3 overlay SDN 兼容性问题

无论采用何种部署形态,对于商业 overlay SDN 而言,

都应具备兼容异构产品和解决方案的能力,包括,兼容第三方开源或商业的虚拟化平台;兼容异构的网关设备;能够与第三方 SDN 方案互通。但是目前来看,绝大多数的商业 overlay SDN 方案均不满足上述兼容性要求,反而在虚拟化、网关设备和控制器各层绑定自家产品,形成整体封闭的方案。表 3 整理了主流商业 overlay SDN 产品的对其他产品的兼容能力。

表 2 overlay SDN 的部署方案对比

方案	方案特色	优缺点	适用场景
纯硬件方案	VTEP、VxLAN 网关全部由硬件设备组成,VxLAN 3 层网关集中部署	优点:性能高、与第三方虚拟化平台易异构 缺点:成本高,硬件设备需支持 VxLAN,规格受设备限制,存在厂商设备锁定风险	适用于构建对租户规格要求不高,对服务、性能要求较高的大型企业私有云
纯软件方案	VTEP 及 VxLAN 网关均为软件形态,VTEP 由 vSwitch 实现,均部署在 x86 服务器上	优点:成本低,部署灵活,租户规格不受限 缺点:南北向流量性能有限,兼容第三方虚拟化平台部署方案复杂	适用于构建提供大规模租户规格的公有云
软硬结合方案	VTEP 为软件/硬件形态,由 vSwitch/TOR 实现,VxLAN 网关由物理交换机组成	优点:成本适中,兼备纯软、纯硬组网的部分优点 缺点:租户规格受南北向硬件设备限制	适用于构建对南北向出口有高性能诉求的公有云租户网络(如大型企业公有云、电商公有云、其他特定业务租户)

表 3 商业 overlay SDN 的兼容性

overlay SDN 方案	支持的 Hypervisor	支持的网关设备	其他厂商方案互通的能力
VMware NSX	ESXi	edge(软件)	-
阿尔卡特朗讯 Nuage VSP	ESXi,KVM,Xen	VRS-G(软件),7750(硬件)	7750 网关支持 EVPN 协议
Juniper Contrail	KVM,Xen	MX 系列路由器(硬件)	-
华为 FusionNetwork	FusionCompute(基于 Xen)	network node(软件),CE 系列路由器(硬件)	-
华三 VCF	CSA(基于 KVM)	VSR(软件),6800,9800(硬件)	6800 网关支持 EVPN 协议
中兴通讯	ZXCPP(基于 KVM)	M6000(硬件)	M6000 网关支持 EVPN 协议

可见,主流商业 overlay SDN 产品的兼容性整体较差。兼容性的不足也限制了这些商业产品在云数据中心的规模应用。

首先,对厂商 Hypervisor 的支持不足,限制了 overlay SDN 在异构混合云环境下的应用。云计算技术是渐进式发展的,用于构建云资源池的计算虚拟化技术最初由 VMware 的一家独大,发展到思杰、红帽、微软、华为等厂商方案的百花齐放以及开源 KVM 技术的异军突起。不同解决方案有其独特的优点,适用于不同的业务场景。在这种情况下,大多数运营商的云数据中心都会部署一种或多种虚拟化解决方案^[4],构建不同类型的资源池用于承载不同的应用,典型的场景是面向高端用户的 VMware vSphere 资源池和面向公有云的开源 KVM 资源池并存。这就要求网络虚拟化能够适配计算和存储,支持多种云资源池构建技术,并能实现异构资源池的网络互通。

其次,在 OCP(Open Compute Project,开放计算项目)^[5]等组织的积极推动下,云数据中心的硬件整体朝标准化、通用化、软件化的趋势发展^[6]。随着白盒交换机和 NFV 技术的日趋成熟,云数据中心将更多部署支持开放标准的通用网络设备以及软件化的网元设施,例如采用白盒交换机作为虚拟网络的接入点、采用软件路由器作为三层网关等。商业 overlay SDN 方案在网关层面绑定自家设备显然有悖于这一趋势,并且面临厂商锁定的风险。因此,商业 overlay SDN 应能兼容新形态、异厂商的网络设备。

最后,对运营商而言,云资源池的建设是分省、分区进行的,出于业务能力、功能需求和价格等多方面因素考虑,不同区域的云资源池可能采用不同的 overlay SDN 方案,要求不同云资源池的 overlay SDN 具备网络互通的能力。

4 实现异构兼容的主要方法

计算与网络紧耦合、采用私有协议和硬件、控制器私有实现是导致商业 overlay SDN 方案兼容性不足的主要原因。针对这些问题,分析和总结了 overlay SDN 实现异构兼容的主要方法。

4.1 解耦计算与网络,兼容异构 Hypervisor

计算与网络紧耦合,是导致商业 overlay SDN 无法兼容异构虚拟化平台的主要问题。大多数厂商方案都选择将网络边缘节点设置在 Hypervisor 层的虚拟交换机之上,由于商业 Hypervisor 的封闭特性,大多数厂商的 Hypervisor 仅支持安装自己的虚拟交换机,第三方虚拟交换机无法渗

透,也因此导致 overlay SDN 无法兼容第三方 Hypervisor。解耦计算与网络是实现异构 Hypervisor 兼容的关键。

根据虚拟网络边缘节点与 Hypervisor 的相对位置,可以将计算与网络解耦的技术路线分为 4 类,分别如下。

- 虚拟机模式:在宿主主机上部署虚拟机形态的 VNE,其他虚拟机的流量被重定向到 VNE 进行隧道封装。
- 计算/网络一体模式:VNE 设置在宿主主机的 Hypervisor 层,由虚拟交换机实现隧道封装等工作,需要 Hypervisor 层深度开放接口支持。
- 智能网卡模式:宿主主机上的物理网卡作为 VNE,直接由网卡完成隧道的封装工作。
- 物理交换机模式:使用连接到宿主机的物理网络设备如 TOR 作为 VNE,虚拟机流量经物理链路进入 overlay 网络。根据物理交换机的实现方式,又可以进一步细分为基于 SDN 商用交换机的方案和基于白盒交换机的方案。

4 种模式各有优略,详细对比结果见表 4。

4.2 采用开放标准的协议和硬件,实现异构设备兼容

目前,厂商方案在网关层面绑定自家设备的主要原因在于标准化的缺失,具体表现在以下方面。

(1) 网关设备实现缺乏标准化

对于 overlay SDN 而言,网关的功能相对明确,主要用于 overlay 网络与外部传统网络的 L2 和 L3 互通。通常认为 L2 网关需要实现 VxLAN 和 VLAN 的转换,L3 网关在 L2 网关的基础上增加路由功能。但是,大多数厂商方案都会在网关上叠加其他功能,以提升 overlay SDN 方案的整体优势,导致不同厂商间网关设备的实现千差万别,缺乏标准化。

(2) 南向协议缺乏标准化

虽然 Netconf 已经成为控制器配置非 OpenFlow 设备的主流协议,多数厂商会采用 Netconf+XML 或者 Netconf+YANG 的方式配置网关设备。但是,Netconf 仅定义了传输层面的协议标准,并没有定义数据层面的语义标准,因此即使都支持 Netconf 协议,不同厂商的网关设备间仍缺乏通用性。OVSDB 近年来发展迅速,已经成为 OpenFlow 事实上的配置管理标准协议,但是鉴于网关设备对 OpenFlow 协议私有扩展实现,即使采用标准 OVSDB 协议,在管理和配置异构厂商网关设备时,仍面临功能丢失的风险。

可见,推动网关设备的通用标准化和南向配置协议的标准化是实现异构厂商设备兼容的主要手段。OCP 正在制定



表 4 计算与网络解耦的技术路线

技术路线	优势	劣势
虚拟机模式	· VNE 以虚机方式部署,部署便捷; · 无需新增或替换现有设备; · 资源灵活扩展、软件升级便捷; · 对计算虚拟化接口对接要求较低	· 受限于虚拟接口性能,转发性能有较大损耗; · 需占用宿主机计算资源; · 网络界面不清晰
计算/网络一体模式	· 转发性能较好; · 部署便捷,无需更改物理网络	· 对计算虚拟化存在较高的接口开放要求; · 占用服务器计算资源; · 网络界面不清晰
智能网卡模式	· 转发性能较好; · 无需占用服务器计算资源	· 与 Hypervisor 的对接支持需二次开发; · 部署较复杂,需更换服务器物理配件; · 能力升级、扩展困难
物理交换机模式	商业交换机 · 转发性能较好; · 无需占用服务器计算资源; 白盒交换机 · 网络界面清晰	· 部署较复杂,需改变现有物理网络; · 设备扩容需变更硬件设备; · 对计算虚拟化接口对接要求较高

数据中心硬件的开源和开放标准,白盒设备有望成为数据中心硬件采购的新趋势,基于白盒设备的软硬分离、开放可编程和定制化(软、硬件)特性,可以支持设备功能和协议接口的企业级规范定义。Netconf+YANG 已经成为非 OpenFlow 网络设备配置的事实标准,被开源控制器项目如 ODL、ONOS 等普遍支持,推动网关设备 YANG 模型语义层面的标准化,实现设备与控制器的解耦,有助于实现对异构厂商网关设备的兼容。

4.3 控制器协同,实现异构 SDN 方案互通

当前,大多数厂商的 overlay SDN 都不具备异厂商方案互通的能力,部分厂商正尝试通过网关支持 EVPN (ethernet VPN, 以太网 VPN) 协议^[7],实现异厂商互通。EVPN 抽象和隔离了网络的控制平面和数据平面,可用作 overlay 网络的控制平面。EVPN 采用 MP-BGP 作为控制面的路由协议,用于在 VTEP 设备间通告主机的 MAC/IP 路由以及自动发现 VTEP 间的对等关系。采用 EVPN 协议实现异厂商方案互通的主要问题是:协议复杂,目前标准仍在定义中,尚未完善;控制逻辑仍由网络设备实现,不能发挥软件定义的灵活性;增加 overlay 网络方案的复杂性。

更合理的选择是在控制器层面进行 overlay 网络控制信息交换。IETF 正在制定 SDNi (software defined network interface, 软件定义网络接口)规范^[8],用于定义控制器间的东西向交互的协议标准。开源 SDN 控制器 OpenDaylight 也有 SDNi App 项目^[9],致力于实现 SDN 控制器间的交互。SDNi App 收集网络的控制信息如拓扑、统计、主机信息等,然后通过 BGP 在控制器间共享。整体而言,SDNi 的发

展前景仍不明朗,IETF 标准草案已经过期,ODL 的 SDNi App 项目也还不具备落地应用的能力。

国内也有部分运营商正在推动“协同控制器”或者“超级控制器”概念,目标是在业务系统和厂商 SDN 控制器之间进行网络协同。协同控制器概念如图 2 所示。

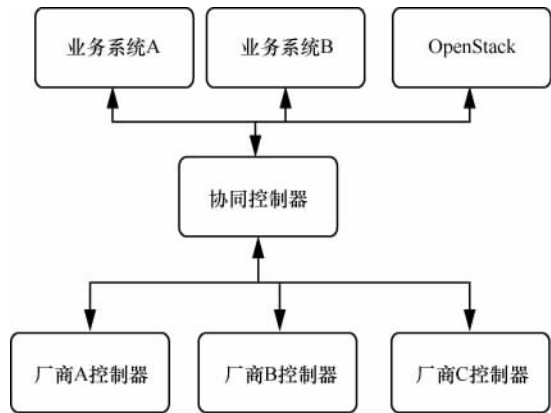


图 2 协同控制器实现不同厂商方案互通

协同控制器将上层业务系统的网络需求经过抽象、编排后,自动下发配置给底层厂商控制器,由厂商 SDN 方案实现具体的网络功能。目前,协同控制器仍处于概念阶段,暂未有成熟的商业产品。

5 基于开源和白盒交换机实现异构兼容的 overlay SDN

根据第 4 节所述,有多种方法可以解决 overlay SDN 的异构兼容问题。设计了基于开源和白盒交换机的通用 overlay SDN 方案,应用以下几种技术实现异构兼容。

- 采用开源 OpenStack 作为上层资源编排层,用于对云数据中心的计算、存储、网络资源进行统一的管理和调度,同时采用 Neutron 提供的租户网络模型作为 overlay SDN 的虚拟网络模型。OpenStack 已经成为开源云计算领域的行业事实标准,对异构虚拟化平台的兼容性最好。
- 采用开源 SDN 控制器实现对异构设备的兼容:开源技术已经成为推动云计算和 SDN/NFV 发展的重要力量,越来越多的企业选择基于开源进行二次开发,相应地各厂商也在其软硬件产品和设备中增加对主流开源软件的支持。以 OpenDaylight 和 ONOS 为代表的 SDN 控制器获得了越来越多厂商的支持,采用它们作为 overlay SDN 的控制器,能够实现对异构设备的支持。
- 采用混合方式解耦计算与网络,实现对异构 Hypervisor 的兼容:对于开源虚拟化平台,可将 VNE 设置在 Hypervisor 之上的虚拟交换机,以获得更加丰富的网络接入功能;对于闭源的商业虚拟化平台,则将 VNE 设置在外部的 TOR 交换机,由 TOR 交换机实现隧道的封装/解封装,虚拟化平台以传统 VLAN 的方式接入 TOR。
- 采用白盒交换机作为闭源虚拟化平台的隧道端点,避免厂商锁定:相对于商业交换机而言,白盒交换机具有廉价、基于通用标准硬件、转发性能高、软硬

分离、可定制等优点,能够满足 overlay SDN 网络承载需求,同时避免厂商锁定。

- 采用软件路由器(vRouter)作为三层网关,避免网关层面的厂商锁定。
- 采用标准协议和接口,避免任意层面的厂商锁定:转发面采用 VxLAN 协议,控制面采用 OpenFlow 协议,管理面则采用 OVSDB、Netconf 等标准协议。在北向接口方面,采用标准 REST API。

系统的整体架构如图 3 所示。

6 结束语

云计算的规模化发展对数据中心网络提出了巨大的挑战,要求网络能够适配业务、计算/存储新技术和新型流量模型需求。overlay SDN 通过隧道技术在现有网络上叠加虚拟逻辑网络,实现了业务与网络的解耦,使得网络变为可池化的资源并能作为服务提供出去,成为数据中心网络云化的主要手段。出于商业策略考虑,主流商业 overlay SDN 方案多为单厂商封闭实现,在 Hypervisor、网关设备和控制器等方面绑定自有产品,不具备异厂商产品兼容的能力,不能满足异构混合云环境复杂 ICT 系统的建设需求。本文重点研究 overlay SDN 的兼容性问题,并提出相应的解决方案,对于制定具有异构兼容能力的通用 SDN 虚拟化方案具有重要的指导意义。

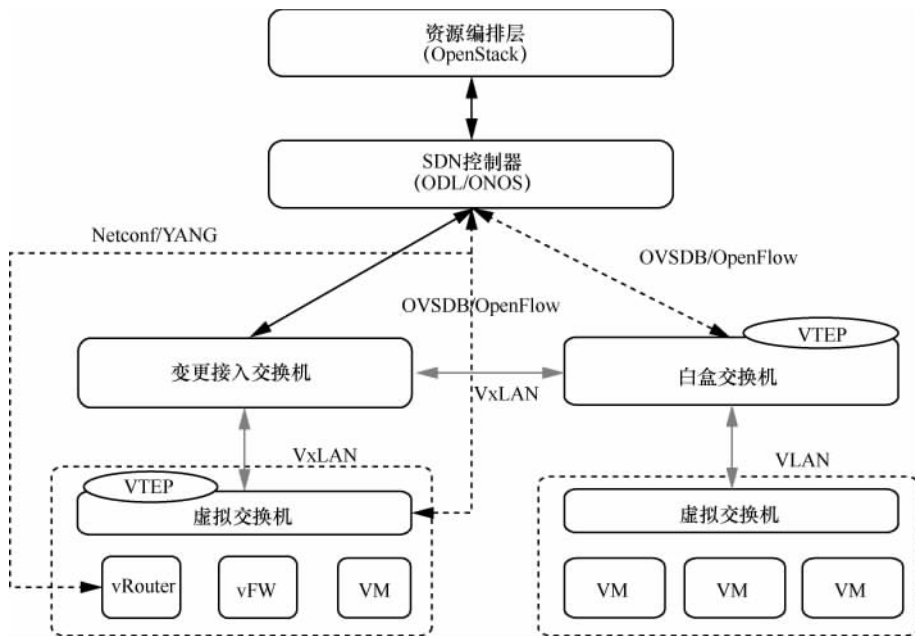


图 3 基于开源和白盒交换机的通用 overlay SDN 架构



参考文献:

- [1] 陈楠, 樊勇兵, 何晓武, 等. 面向应用的云数据中心网络技术研究[J]. 电信科学, 2014, 30(9): 128-132.
CHEN N, FAN Y B, HE X W, et al. Research on the technology of application-oriented cloud data center network[J]. Telecommunications Science, 2014, 30(9): 128-132.
- [2] 樊勇兵, 陈楠, 黄志兰, 等. 解惑 SDN [M]. 北京: 人民邮电出版社, 2015.
FAN Y B, CHEN N, HUANG Z L, et al. Disabuse SDN [M]. Beijing: Posts and Telecommunications Press, 2015.
- [3] H3C SDN overlay 技术白皮书[EB/OL]. [2016-08-01]. http://www.h3c.com.cn/Products_Technology/Products/New_network/SDNCatalog/VCF_Controller/VCF_Controller/White_Paper/201603/920927_30003_0.htm.
H3C SDN overlay technology white paper [EB/OL]. [2016-08-01]. http://www.h3c.com.cn/Products_Technology/Products/New_network/SDN/Catalog/VCF_Controller/VCF_Controller/White_Paper/201603/920927_30003_0.htm.
- [4] Multiple hypervisor strategies [EB/OL]. [2016-08-01]. <http://www.networkworld.com/article/2223807/virtualization/multiple-hypervisor-strategies.html>.
- [5] OCP[EB/OL]. [2016-08-01]. <http://www.opencompute.org/projects/networking>.
- [6] 陈平平. 软硬件分离趋势及开放网络发展[EB/OL]. [2016-08-01]. <http://chuansong.me/n/906603>.
CHEN P P. The trend of software and hardware separation and the development of open network [EB/OL]. [2016-08-01]. <http://chuansong.me/n/906603>.
- [7] ANATOL B. EVPN ethernet virtual private network [M]. Berlin: Springer, 2016.
- [8] TSOU T, ARANDA P, XIE H, et al. SDNi: a message exchange protocol for software defined networks (SDNS) across multiple domains [EB/OL]. [2016-08-01]. <https://www.tools.ietf.org/html/draft-yin-sdn-sdni-00.txt>.
- [9] OpenDaylight SDNi App [EB/OL]. [2016-08-01]. http://www.wiki.opendaylight.org/view/Project_Proposals:ODL-SDNi_App.

[作者简介]



黄志兰 (1982-), 女, 现就职于中国电信股份有限公司广州研究院, 主要研究方向为 SDN、云计算、IPv6。



樊勇兵 (1970-), 男, 博士, 现就职于中国电信股份有限公司广州研究院, 中国数据中心产业发展联盟云计算服务专家委员会第一批专家委员, 主要从事中国电信云计算、IDC 等方面的研究工作。



陈楠 (1981-), 男, 博士, 现就职于中国电信股份有限公司广州研究院, ITU-T Q14/SG11 联合报告人, Q18/SG13 编辑人, 主要研究方向为 SDN、NFV、服务器虚拟化、数据中心网络、管理平台技术等。



何晓武 (1987-), 男, 现就职于中国电信股份有限公司广州研究院, 主要从事虚拟化、云计算、NFV、数据网络等方面的研究工作。