



VxLAN 在云数据中心组网的应用

张届新,傅志仁,吴志明,徐文华,徐海峰

(中国电信股份有限公司上海分公司 上海 200085)

摘要:梳理了数据中心云化演进对新技术、新架构的需求,对当前 VxLAN 主流实现方案进行了分析。研究了 VxLAN 技术在云数据中心的几种典型组网应用,并总结了 VxLAN 在云数据中心规模商用尚需解决的问题。

关键词:云数据中心;overlay 网络;脊-叶;Spine;SDN;VxLAN

doi: 10.11959/j.issn.1000-0801.2015238

VxLAN Application in Cloud Datacenter Network

Zhang Jiexin, Fu Zhiren, Wu Zhiming, Xu Wenhua, Xu Haifeng

(Shanghai Branch of China Telecom Co., Ltd., Shanghai 200085, China)

Abstract: Cloud datacenter demand on new techniques and new architecture were summarized firstly. And also the mainstreams of VxLAN technical schemes were analyzed. Then typical VxLAN applications in cloud datacenter network were studied. Finally, several VxLAN problems were discussed in case of scale applications in cloud datacenter.

Key words: cloud datacenter, overlay network, spine-leaf, Spine, software defined networking, VxLAN

1 引言

据 IDC(International Data Corporation, 国际数据公司)分析,社交化、移动化、大数据、云计算等成为第三代 ICT 的关键词。电信运营商可提供 ICT 服务的主要基地——数据中心需要接应挑战,适应新时期业务需求的变化。

在新时期,数据中心的业务需求从面向传统的机架出租、带宽出租以及简单的增值服务等应用越来越多地向与云计算、大数据和移动化相关的需求转型,包括以下 4 个方面。

- 终端多、流量大:海量虚拟机接入及其带来的数据中心内、数据中心之间大量的东西向流量。
- 应用可靠性高、体验好:实现虚拟机和物理机的无

缝迁移。

- 业务发放敏捷性高、可管可控能力强:按需快速实现计算、存储和网络虚拟化资源的快速部署以及端到端可视化运营管理。
- 混合云业务为趋势:随着公有云接受度的逐渐提高,企业 IT 运营成本不断降低,混合云成为越来越多的企业选择。

但目前数据中心网络架构仍属于传统组网,将不能很好地适应新时期的业务发展,主要体现在以下 6 个方面。

- 数据中心网络架构不够扁平:仍是传统的三层网络架构,主要用于支撑传统互联网访问的南北向流量,而不适应云计算引入后日益增多的虚拟机访问或迁移带来的东西向流量的快速调整。

收稿日期:2015-06-23;修回日期:2015-09-09

论文引用格式:张届新,傅志仁,吴志明等. VxLAN 在云数据中心组网的应用. 电信科学,2015238

Zhang J X, Fu Z R, Wu Z M, et al. VxLAN application in cloud datacenter network. Telecommunications Science, 2015238

- 网络系统容量有限:面向越来越多的海量虚拟机部署,传统数据中心网络设备的二层 MAC 地址表容量日显不足;云引入大量租户,将带来大量 IP 地址重叠,甚至 MAC 地址重叠。
- 二层逻辑通道有限:现有数据中心一般采用 VLAN 作为用户的二层逻辑通道隔离标识,难以满足今后大量租户多终端多业务的需求。另一方面,由于 QinQ 无法实现跨域网的物理网络和虚拟网络端到端透传与统一标识,所以传统网络中采用 QinQ 进行二层逻辑通道扩展的情况并不适用于云数据中心。
- 数据中心网络柔性不足:现有网络设备主要由专用硬件构成,网络中各个网元与业务耦合紧密,整体呈刚性,不利于业务提供的灵活性和新业务开发的敏捷性。
- 面向云的端到端网络运维效率不高:在云计算引入数据中心的早期,传统数据中心的 IP 网络/以太网尚未来得及与计算虚拟化网络很好地对接,无法实现基于租户逻辑通道颗粒度的端到端网络流量、流向可视化。
- 跨机房二层网络互联成本较高:为了支持跨地域虚拟机实时迁移,需要实现跨地域大二层组网。而基于裸光纤互联、二层专线互联等传统二层组网投入成本高。

VxLAN (virtual extensible local area network) 作为业界当前主流的 overlay (叠加网) 技术,具有 spine (脊)-leaf (叶) 扁平组网、二层逻辑通道数量大、实现大二层互联对现网改造影响小、主流厂商设备支持较广泛、支持厂商涉及面广 (从网络设备厂商到虚拟化平台厂商) 等优点,越来越多的互联网服务商及电信运营商开始关注 VxLAN 技术并在其数据中心中引入,以期解决上述数据中心向云化演进过程中所面临的问题。另一方面,SDN (software defined networking, 软件定义网络) 作为网络演进的主流方向之一,且具有转发平面与控制平面分离从而实现低成本网络部署、灵活便捷的业务提供等优势,也是广大互联网服务商和电信运营商日益关注的技术焦点。但城域网组网复杂、设备种类繁多,因此,相对简单、封闭且需要适应云时期新需求的数据中心是 SDN 最佳引入场点。综合 SDN 和 VxLAN 的优势,很多厂商将 SDN 和 VxLAN 技术进行结合,以实现数据中心更加灵活、可靠、扩展的组网和业务。

2 VxLAN 技术实现

2.1 VxLAN 主流技术实现原理

目前 VxLAN 主要有 2 种实现方式,其区别本质在于控制平面机制不同。

2.1.1 无控制平面而基于多播的 VxLAN

基于多播的 VxLAN 是 IETF RFC7348 定义的没有控制平面的 VxLAN 方案,即基于数据驱动的泛洪再学习的方式,利用多播来传送 VxLAN 中的 BUM (broadcast, unknown unicast, multicast, 广播分组、未知单播分组、多播分组) 流量。当网络中任意主机发起 ARP 请求时,网络中所有 VTEP (VxLAN tunneling end point, VxLAN 隧道终端) 都会收到该请求。因此,带来较大的泛洪流量,消耗网络带宽资源,也带来安全隐患。

通过引入 SDN 控制器作为 VxLAN 的 ARP proxy (代理),避免 VxLAN 在通信开始必须全网开启多播以实现 MAC 地址学习的需求,从而降低对 underlay (底层承载网) IP 网的要求 (不需要全网开启多播或多播 VPN),并避免了 IP 网络中因此而产生的大量多播流量。

如图 1 所示,SDN 控制器作为 ARP 代理,代答所有 VxLAN 交换机 (包括物理交换机和虚拟交换机) 发出的 ARP 请求,并形成流表下发给 VxLAN 交换机。一般会形成 2 张流表,一张是单播流进行转发参照的流表;另一张是对广播分组和未知单播分组进行识别后,通过多播头端复制方式在相关二层域内广播发送这些分组的流表。

2.1.2 基于 MP-BGP EVPN 的 VxLAN

IETF RFC7432 等相关标准和草案定义了基于 MP-BGP EVPN (multiprotocol border gateway protocol ethernet virtual private network, 多协议边界网关协议—以太虚拟专网) 控制平面。其中,MP-BGP 是 EVPN 的路由协议,通过 VRF (virtual routing forwarding, VPN 路由转发表,以 RD/RT 为唯一标识) 构建多租户环境,利用标准新定义的地址组 EVPN 发布 EVPN 路由 (EVPN 路由以“MAC 地址+IP 地址”标识)。

基于 MP-BGP EVPN 的 VxLAN 方案中,有控制平面和数据平面两个平面。其中,控制平面实现 peer 发现及路由学习 (包括本地学习和远端学习) 和分发;数据平面实现 overlay 的 L2/L3 单播流转发及 BUM 流量的转发。

具体实现而言,基于 MP-BGP EVPN 的 VxLAN 技术实现主要有 2 种,分别如图 2 和图 3 所示。区别在于控制平

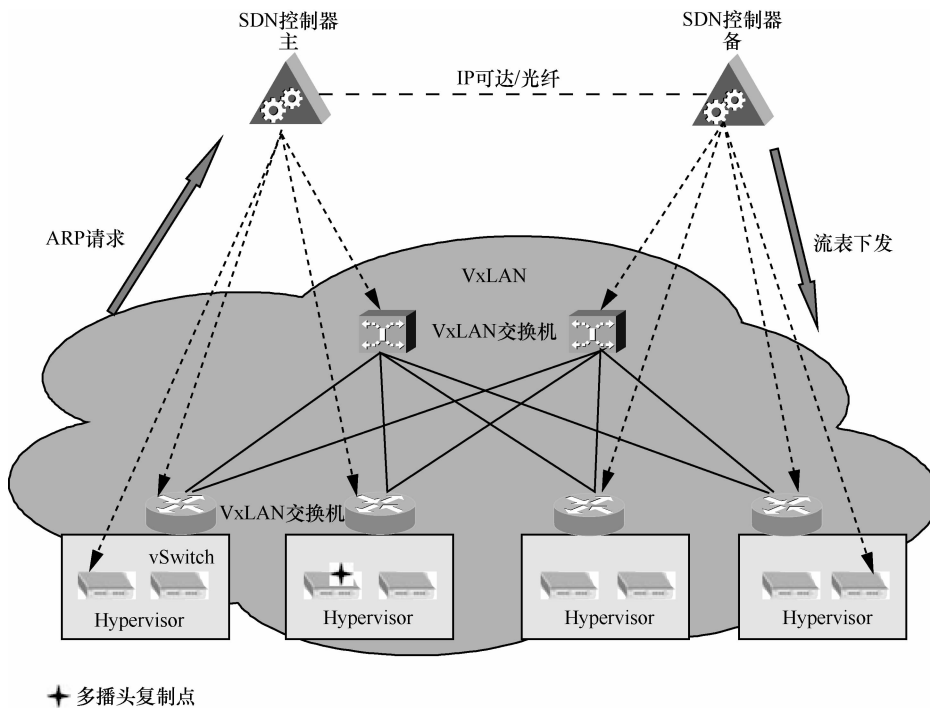


图1 基于SDN控制器的VxLAN技术实现示意

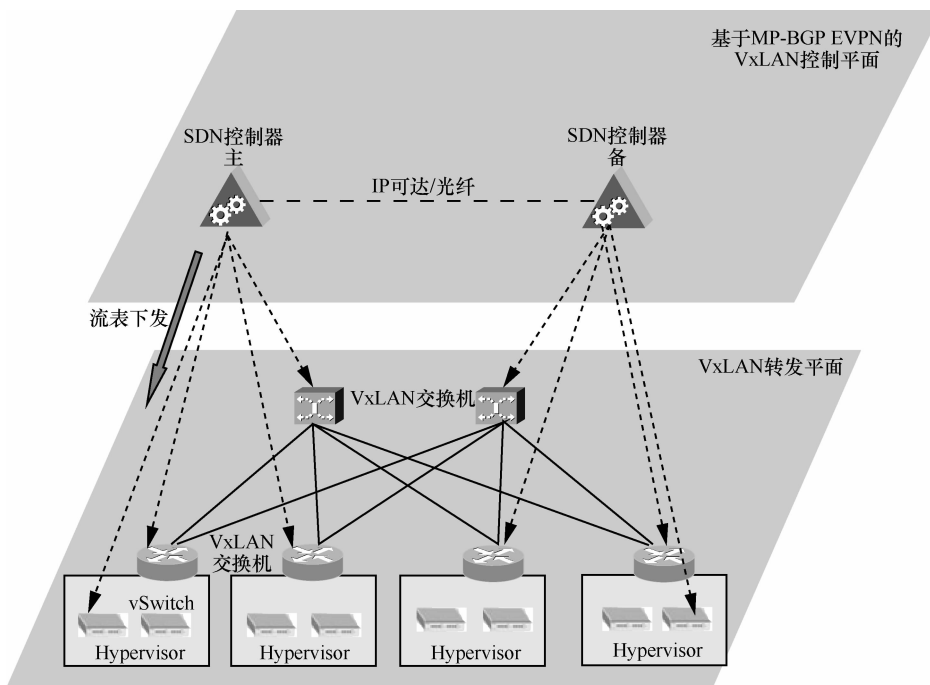


图2 SDN控制器作为MP-BGP EVPN控制平面的VxLAN技术实现示意

面是否VxLAN设备自身支持。

在图2中,MP-BP EVPN控制平面由SDN控制器实现,且RR(route reflector,路由反射器)由SDN控制器承担;而根据流表的数据转发则由VxLAN交换机实现。这种方案适合于早期VxLAN解决方案,即引入SDN控制器的

场景。但由于需要SDN控制器支持RR功能,对SDN控制器有较高要求。同时,VxLAN网络的数据转发与SDN控制器紧耦合。

在图3中,VxLAN的MP-BP EVPN控制平面和数据转发均由VxLAN交换机实现。这种方案适合于早期VxLAN

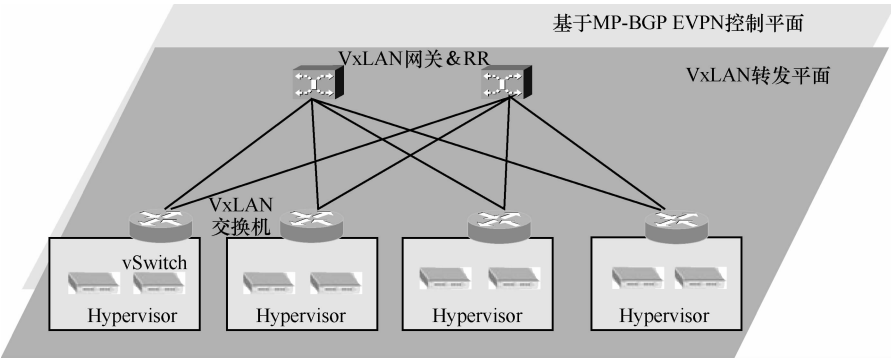


图3 VxLAN 交换机自身支持双平面的 VxLAN 技术实现示意

解决方案，即与 SDN 控制器松耦合的场景。此方案对 VxLAN 交换机要求较高,不仅要支持 VxLAN 基本功能,还需支持 MP-BGP EVPN 完整功能。

2.1.3 两种技术对比

两种 VxLAN 技术实现方式的比较见表 1。

通过表 1 可以看出，与基于多播实现的 VxLAN 方案相比,基于 MP-BGP EVPN 控制平面的 VxLAN 实现方案具有以下优点：

- 由于 MP-BGP EVPN 利用了 L3 VPN 的 VRF 特性来传递 VxLAN,因此天生支持多租户组网；
- 由于利用 MP-BGP EVPN 发布主机的 MAC 地址和 IP 地址,因此是一种协议驱动的地址学习,从而支持很高的组网扩展性；
- 当主机移动时，主机新接入的 VTEP 将通过 MP-BGP 马上更新路由表，通知 VNI (VxLAN network identifier,VxLAN 逻辑通道标识) 内所有其他 VTEP 这个主机的新位置,从而实现了网络失败或主机移动时的网络快速收敛；
- 通过 MP-BGP 认证实现了 VTEP peer 的认证，从而提升了 VTEP 的安全性，可有效防范 VTEP 伪冒等

网络安全隐患。

2.2 VxLAN 技术实现现状

目前主流设备厂商的 VxLAN 解决方案百家争鸣。

按照控制平面实现方式划分：一种是无控制平面,但可以通过 SDN 控制器作为 ARP 代理来减少泛洪影响,从而实现组网优化,如华三、华为 VxLAN 解决方案;另一种是基于 MP-BGP EVPN 的控制平面来消除泛洪,如思科、阿朗 MP-BGP EVPN 的 VxLAN 解决方案。

按照组网连接拓扑划分,主要有 spine(脊)-leaf(叶)架构和 Mesh(网状)架构。由于 spine-leaf 架构可以实现任意两点之间的一跳可达,具有高可靠、高扩展、高度扁平组网的特性，越来越多厂商的 VxLAN 解决方案已开始支持 spine-leaf 架构。

按照网络功能实现载体分,主要有 3 种方式：一是解决方案全部基于纯硬件网络设备实现,如 Arista;二是解决方案全部基于纯软件实现,如 VMware (NSX);三是解决方案包括硬件网络设备和软件 vSwitch 实现，包括思科 NEXUS 系列(包括专用虚拟交换机)、华三 S 系列交换机和专用虚拟交换机、阿朗的 VTEP 交换机和 VSR 虚拟机交换机等。

表 1 VxLAN 两种实现方式比较

比较项	基于多播的 VxLAN	基于 MP-BGP EVPN 的 VxLAN
overlay 承载的业务	L2 及 L3	L2 及 L3
underlay 网络	IP 网络及 ECMP(equal cost multipath,等价多路径)	IP 网络及 ECMP
协议封装	MAC in UDP	MAC in UDP
peer 发现	数据驱动的泛洪与学习	MP-BGP
peer 认证	无	MP-BGP
主机路由学习	本地主机:数据驱动的泛洪与学习 远端主机:数据驱动的泛洪与学习	本地主机:数据驱动的泛洪与学习 远端主机:MP-BGP
主机路由发布	无	MP-BGP
L2/L3 单播流转发	单播封装	单播封装
BUM 流转发	多播复制;单播入向复制	多播复制;单播入向复制



3 VxLAN 在云数据中心的组网应用

VxLAN 在云数据中心组网应用主要有 3 个场景:一是作为企业用户本地 CPN 内应用及系统与云数据中心部署的 ICT 系统通过大二层互联,提供虚拟机实时迁移、数据灾备等服务;二是数据中心内部实现网络虚拟化与计算虚拟化的进一步无缝对接(计算虚拟化平台也开始支持 VxLAN 技术);三是分别部署在多个云数据中心,以现有城域网作为 underlay 底层架构,通过基于 VxLAN 构建的 overlay 网络提供虚拟机实时迁移与数据灾备的大二层通道。

3.1 VxLAN 在云数据中心组网应用的技术要求

由于云数据中心将面临海量虚拟机和物理机接入、大量虚拟机实时迁移以及虚拟机访问的承载需求,因此,云数据中心对 VxLAN 交换机(支持 VTEP 建立/终结功能的交换机)的要求,除了需要传统交换机的功能外(如安全功能、IPv6 功能等),还需要与云/虚拟化相关的一些个性化的技术要求。

(1) 分布式路由

大量的东西向流量要求物理 VxLAN 交换机和虚拟 VxLAN 交换机都可以支持分布式路由,实现流量本地转发,且流表的稳定性与 SDN 控制器弱相关,甚至解耦。

(2) VxLAN bridge(桥接)和 VxLAN router(路由)

满足跨子网、跨二层域的访问需求。

(3) 系统高扩展性

一是支持 L2 和 L3 的网络虚拟化,包括 MAC 地址和 IP 地址重用(overlapping),为今后全网虚拟化奠定网络基础,也要求物理 VxLAN 交换机和虚拟 VxLAN 交换机都可以支持地址重用,从而可以灵活按需接入物理服务器和虚拟机;二是需要支持大量的二层逻辑链路标识,且支持新引入的 VxLAN 标识 VNI 与传统二层逻辑通道标识 VLAN 的相互映射,便于与接入光网互通;三是 MAC 地址和 IP 地址转发表项足够大,且系统支持按节点线性扩展相关能力。

(4) 系统高可靠性

一是要求 SDN 控制器失效时不影响已下发建立的流表稳定运行;二是支持 ECMP(等价多路径),尽可能实现网络带宽的高效利用以及网络链路的高度冗余。

(5) 系统开放性和可演进性

参照业界主流标准,随着设备能力完善,逐步支持与第三方 NFV 互通、与第三方 VxLAN 交换机互通。

(6) 端到端可视化

需要物理网络和虚拟网络(虚拟交换机)的端到端可

管理、可配置。

(7) 分权分域管理

支持面向运营商、租户以及租户内不同级别用户的多维度网络和业务的差异化、个性化管理。

3.2 基于 VxLAN 的云数据中心组网解决方案

VxLAN 在云数据中心组网主要有如下 3 种应用场景。

(1) 基于 VxLAN 网关的混合云组网

用户 CPN 部署的 VxLAN 网关通过城域网 underlay 网络与 DC(data center, 数据中心)中的 VxLAN 交换机互联,由于基于 VxLAN 构建了跨用户驻地网和电信运营商 DC 的大二层网络,因此可以提供用户驻地网中重要应用在电信运营商 DC 内的灾备以及虚拟机实时迁移。不仅虚拟机应用可以互通,同时,对用户驻地网与云数据中心之间现有 IP 网络也几乎无影响。另一方面,对于总部与多个分支机构互联组网模式,由于企业内网一般采用私网地址组网,各分支机构的 IP 地址可能重叠,随着今后虚拟机的大量引入,MAC 地址也可能重叠,因此,在用户总部所在 DC 内采用 VxLAN 组网,将可以允许 IP/MAC 地址重用而不会影响用户分支已组建好的网络。

图 4 中用户驻地网(CPN)侧的 VxLAN 网关可以是硬件形态,也可以是安装在虚拟机上的软件形态。

(2) DC 内 VxLAN 组网

如图 5 所示,用户驻地网侧可以仍然为传统 VLAN 组网。在 DC 内通过 VxLAN 网关实现 VLAN 到 VxLAN 映射,从而在 DC 内部实现从虚拟化层网络到物理网络的端到端 VxLAN 组网。

DC 内采用 VxLAN 组网比传统的 VLAN 组网有更多优势:其一可以支持大量虚拟机、服务器的 MAC 地址和 IP 地址重叠,并解决二层逻辑通道 4 096 个 VLAN 标识限制问题,实现 DC 网络高扩展性;其二,VxLAN 先天支持 ECMP(equal cost multipath, 等价多路径),可以解决传统 LAN 中的生成树环路问题,并实现基于流的精细化颗粒度负载均担,充分利用网络资源并提供网络高冗余性;其三,DC 采用 VxLAN 组网,可以与已支持 VxLAN 的计算虚拟化层无缝融合组网,实现虚拟网络、物理网络端到端逻辑路径可视化。

(3) 跨城域网的 DC 间互联组网

基于 VxLAN 实现跨 DC 互联组网,不仅可以满足新时期业务需求,而且可以降低运营成本。如图 6 所示,由于 VxLAN 是一种 MAC in UDP 的 overlay 网络技术,因此,可以在对跨现有 DC 间 IP 网络几乎无影响的情况下构建跨

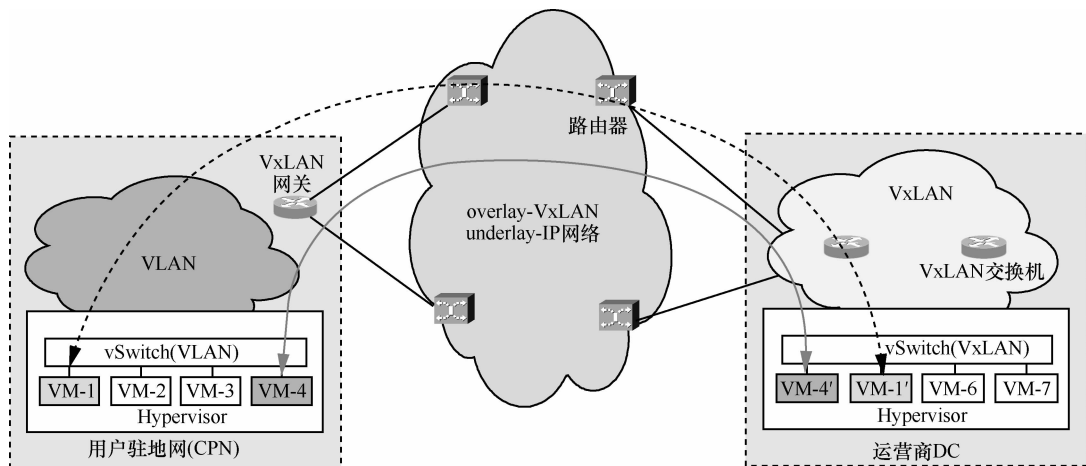


图 4 基于 VxLAN 网关的混合云组网示意

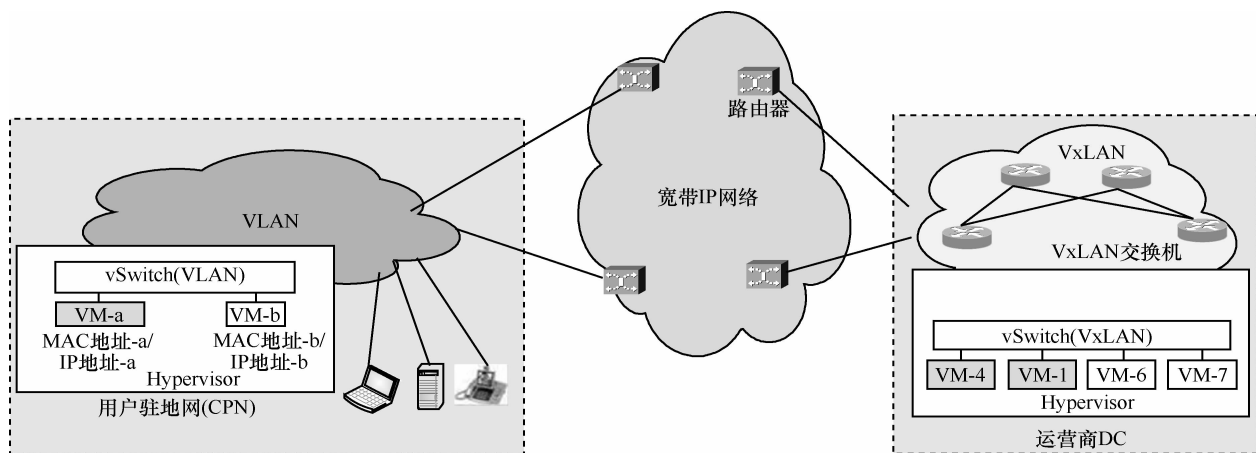


图 5 基于 VxLAN 的 DC 内组网示意

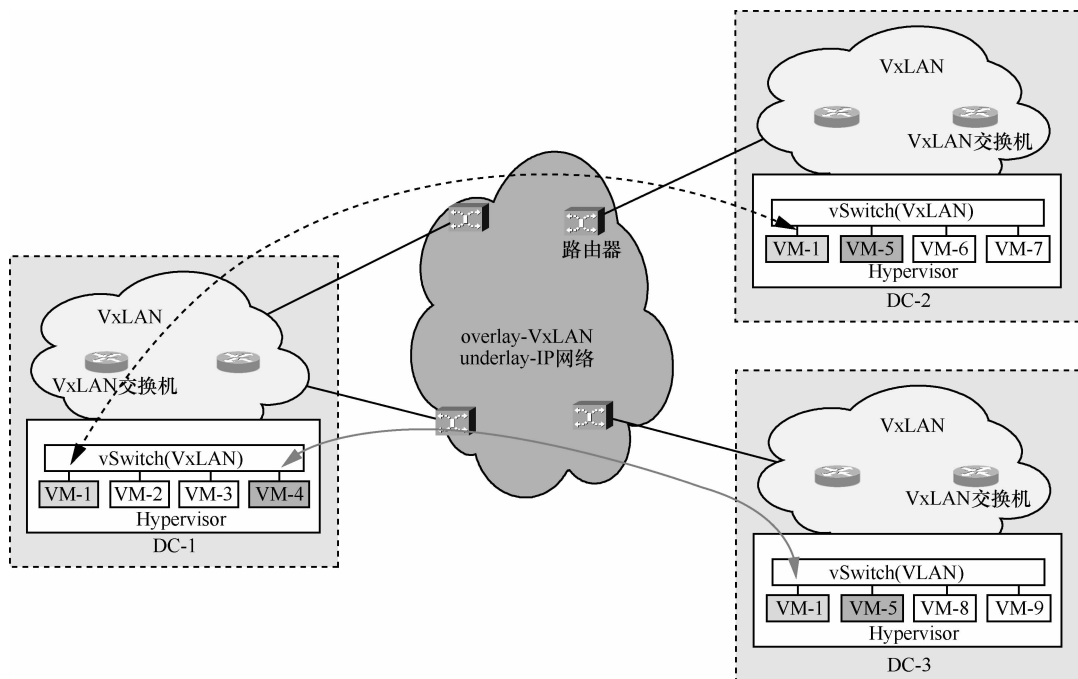


图 6 基于 VxLAN 的 DC 间大二层互联组网示意



地域大二层网络,满足大量虚拟机实时迁移导致的東西向流量对广域网低时延等要求。同时,相对于传统基于裸光纤互联、构建新型 OTN 实现大二层组网具有较高的经济性;而相对于传统基于 IP 网络的 MPLS L2 VPN 二层专线网络,不需要全网开启 MPLS,且运维管理相对简单。

4 结束语

VxLAN 技术逐步成熟,知名互联网服务商和主流电信运营商、大型企业已逐步试验或商用部署。但是,受限于标准化进展以及厂商不同的实现能力,VxLAN 技术在云数据中心大规模、多场景商用尚有一些技术问题需要进一步深入研究,并推动解决。

一是 VxLAN 协议开销对现网的影响。VxLAN 网关建立 VxLAN 隧道时,由于 VxLAN 封装会将原有 IP 分组增加 50 byte 开销,因此,需保证 VxLAN 隧道沿途所有 underlay IP 网络设备的 MTU 值设置得大于 VxLAN 报文。或者,可以在 VxLAN VTEP 之间先开启 GRE 隧道,然后再进行 VxLAN 隧道封装,此时只要 GRE 隧道两个端点将 MTU 值适当调大即可。但 GRE 隧道承载 VxLAN 隧道也存在一定问题,因为 GRE 隧道是点对点连接,每一个 VxLAN 隧道都需为之配置一条点对点 GRE 隧道,将带来一定的运维复杂度。

二是 SDN 控制器与 VxLAN 交换机之间南向接口(协议)的标准性和开放性有待提高。目前,SDN 控制器与交换机之间南向接口协议主要有 OpenFlow 和 OpFlex 两种,OpenFlow 为 ONF(Open Networking Foundation,开放网络基金会)所定义,而 OpFlex 为 IETF 定义。但是,由于各厂商仍有私有实现机制在标准定义范畴之外,因此目前商用解决方案中异厂商 SDN 控制器与 VxLAN 交换机的互通性很差。

三是 SDN 控制器东西向互通方案需要标准化。目前,不同厂商的 SDN 控制器之间均无法实现通信,且各厂商的 SDN 控制器灾备方案也不尽相同,如有采用群集技术实现方案、有采用 MP-BGP EVPN 实现方案等。

四是基于 MP-BGP EVPN 作为控制平面的 VxLAN 技术实现虽已部分标准化,但已有解决方案的各厂商不仅实现架构不尽相同,而且互通性存在问题。

参考文献

- 1 IETF RFC7348. Virtual eXtensible Local Area Network (VxLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks, 2014

- 2 IETF RFC7432. BGP MPLS-Based Ethernet VPN, 2015
- 3 吴志明,傅志仁,张宏斌等. LSN: 面向 ICT 服务的电信网络架构及其技术实现. 电信科学, 2014, 30(8):112~130
Wu Z M, Fu Z R, Zhang H B, et al. An ICT service oriented telecom network infrastructure architecture and implementation. Telecommunications Science, 2014, 30(8):112~130
- 4 张届新,吴志明,徐文华. 面向 vDC 的 overlay 网络关键技术及组网研究. 电信科学, 2014, 30(5):138~144
Zhang J X, Wu Z M, Xu W H. Study on the key overlay networking techniques in vDC. Telecommunications Science, 2014, 30(5):138~144

[作者简介]



张届新,女,中国电信股份有限公司上海分公司信息网络部高级工程师,主要研究方向为数据网络、数据中心组网。



傅志仁,男,博士,中国电信股份有限公司上海分公司信息网络部总经理、教授级高级工程师,主要研究方向为云数据中心组网、信息网络。



吴志明,男,博士,中国电信股份有限公司上海分公司信息网络部副总经理、教授级高级工程师,主要研究方向为互联网络技术、云计算、数据中心、信息网络等。



徐文华,男,中国电信股份有限公司上海分公司信息网络部副总工程师、高级工程师,主要研究方向为数据网络、数据中心组网。

徐海峰,男,中国电信股份有限公司上海分公司信息网络部工程师、CCIE,主要研究方向为数据网络、数据中心组网。