



基于 VxLAN 组网的云数据中心互联方案

张届新,吴志明

(中国电信股份有限公司上海分公司,上海 200120)

摘要:从 VxLAN 技术实现现状、云数据中心底层承载网络现状等维度,多视角地对基于 VxLAN 组网的云数据中心之间的互联互通方案进行了研究与测试验证,并分析了各种方案的优缺点及适用场景。通过本文的研究,以期对运营商的云数据中心发展提供相关技术参考。

关键词:云数据中心;MP-BGP EVPN;overlay;VxLAN

中图分类号:TN919.21

文献标识码:A

doi: 10.11959/j.issn.1000-0801.2016301

Interconnection scheme of VxLAN based cloud datacenter networks

ZHANG Jiexin, WU Zhiming

Shanghai Branch of China Telecom Co., Ltd., Shanghai 200120, China

Abstract: Several interconnection and interworking schemes of cloud datacenter networks based on VxLAN technique were studied in prospective of VxLAN technique and underlay network status. And the respective advantages and disadvantages of these schemes were analyzed. Finally, the applicable scenarios of each scheme were summarized. It is in the hope that the study would provide some useful technique references for the development of carriers' datacenters.

Key words: cloud datacenter, MP-BGP EVPN, overlay, VxLAN

1 引言

作为一种具有高扩展、高冗余、支持大二层组网的 overlay(叠加网)网络技术,VxLAN(virtual extensible local area network)越来越多地应用到电信运营商以及互联网服务商的云数据中心中,以实现系统容量和逻辑通道的扩展、网络的高效自愈、面向云的端到端网络运维效率提高以及跨广域网的二层可靠互联。

随着基于 VxLAN 技术组网的云数据中心建设日益增多,这些云数据中心之间实现互联互通以满足虚拟机迁

移、灾备等业务需求的工作凸显紧迫且重要。

本文结合 VxLAN 的技术、设备实现现状、云数据中心互联的底层网络现状等信息,对基于 VxLAN 组网的云数据中心之间互联互通方案进行了研究与验证。同时,云数据中心之间多节点互联,包括业务网络互联和存储网络互联等。本文主要探讨云数据中心之间的业务网络互联方案。

2 VxLAN 云数据中心互联方案

基于 VxLAN 构建的云数据中心之间互联方案,按照 overlay 网络及 underlay(底层承载)网络的技术实现可以有

多种方案。

在 overlay 网络层面,由于 IETF RFC7348 定义的基于多播实现的 VxLAN 方案,不仅对现有 IP 承载网络有较高要求,而且不同厂商实现细节不同,导致异厂商 VxLAN 设备互通存在问题^[1],因此,部分厂商提出了基于 SDN (software defined networking, 软件定义网络) 控制器作为 ARP (address resolution protocol, 地址解析协议) 代理的 VxLAN 方案和基于 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) 作为控制平面的 VxLAN 方案。在 underlay 网络层面,可以有二层网络和三层网络方案。

2.1 SDN 控制器作为 ARP 代理

SDN 控制器作为 ARP 代理的 VxLAN 云数据中心互联方案根据 underlay 网络可以分为二层互联和三层互联。

2.1.1 二层互联

两个 VxLAN 云数据中心之间通过数据中心出口的 VxLAN/VLAN 双栈交换机或路由器基于 VLAN 互联。此时,互联的 VxLAN 云数据中心之间只有 underlay 层,基于 VLAN 实现多节点互联和业务转发。SDN 控制器作为 ARP 代理的 VxLAN 云数据中心通过 VLAN 二层网络互联如图 1 所示。

如图 1 所示,虚拟机通过 VLAN 方式接入 vSwitch (虚拟交换机)。根据 vSwitch 对 VxLAN 的支持能力,VxLAN 交换机/路由器分别采用对数据透传 (vSwitch 支持 VxLAN 封装) 或基于 VxLAN 封装方式 (如 vSwitch 不支持 VxLAN 封装而仅支持传统 VLAN 封装) 处理 vSwitch 进入的数据。由

于目前城域二层网络绝大多数仍然采用传统局域网络,仅支持 VLAN 进行逻辑通道隔离,因此,云数据中心的出口 VxLAN/VLAN 双栈交换机或路由器将从数据中心内部接收到的数据转换为 VLAN 封装,并通过二层网络传送到对端云数据中心。对端云数据中心则执行反向操作。

基于 VLAN 二层网络的 VxLAN 云数据中心互联方案实现简单,不受互联的云数据中心内所采用的具体技术方案限制,支持任意厂商解决方案的互联互通。但是,方案中需要把数据中心内运行的 VxLAN 终结,转换为传统 VLAN 后再进行跨节点互联,无法很好地支持跨广域网的多个云数据中心之间大二层互联。

由于当前 SDN 控制器东西向互通性较差^[2],因此,VxLAN 云数据中心之间通过 VLAN 互联适用于底层网络为二层及以下,且各云数据中心内采用不同厂商的 SDN 控制器作为控制平面的 VxLAN 方案组网场景。

2.1.2 三层互联

在实际部署中,由于远距离光纤成本高、VLAN 二层网络先天的局域网特性等因素限制,云数据中心之间大多数情况下基于 IP 网络实现互联互通。此时,VxLAN 云数据中心之间互联,IP 网络作为 underlay 层,而 VxLAN 隧道作为 overlay 层实现多节点互联及业务转发。SDN 控制器作为 ARP 代理的 VxLAN 云数据中心通过 IP 网络互联如图 2 所示。

目前主流厂商具体技术实现有一定差异,尤其 SDN 控制器东西向兼容性较差,异厂商 VxLAN 交换机/路由器

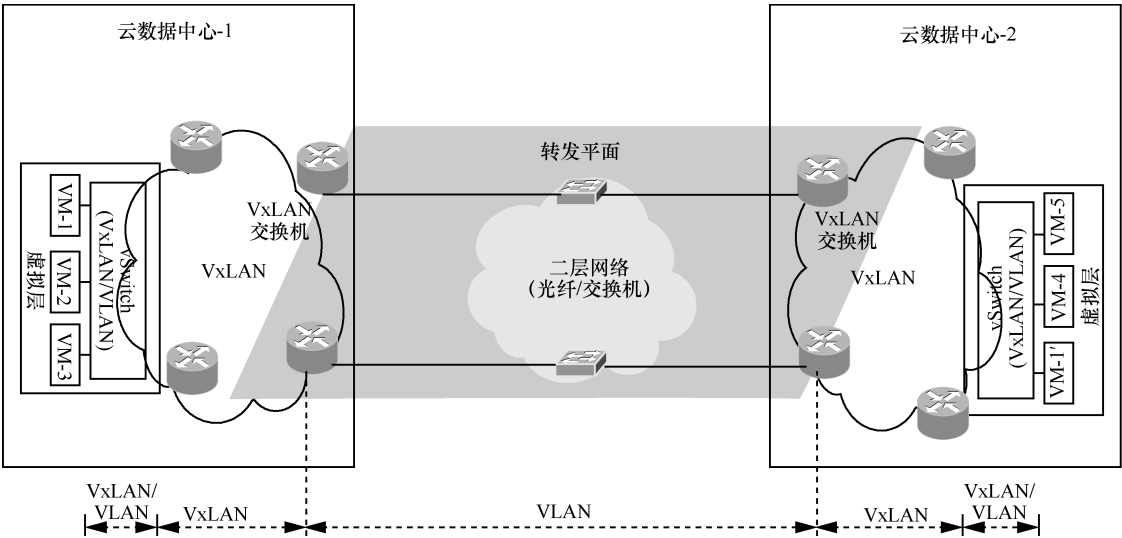


图 1 SDN 控制器作为 ARP 代理的云数据中心二层互联

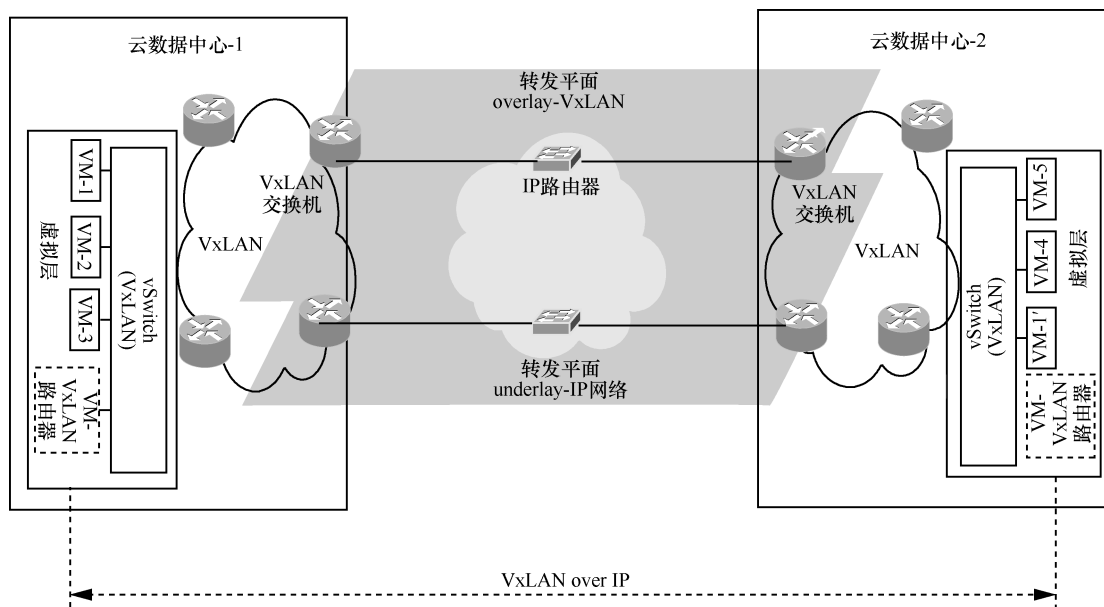


图2 SDN控制器作为ARP代理的云数据中心通过IP网络互联

之间直接构建 VxLAN 隧道存在困难。为此,可以利用支持 VxLAN 功能的 vRouter (虚拟路由器) 实现异厂商 VxLAN 交换机/路由器组网的云数据中心节点间 VxLAN 隧道的建立与互联互通,其中,选择一家厂商的 VxLAN vRouter 作为虚拟机部署在云数据中心节点的计算资源池内,构建基于同厂商的 VxLAN 隧道,规避异厂商硬件解决方案组网的 VxLAN 隧道互通性问题。

SDN 控制器作为 ARP 代理的 VxLAN 云数据中心通过 IP 网络互联具体有两种实现方式,对 IP 网络有不同的技术要求。

(1) 多播方案^[2]

云数据中心之间的 IP 网络全程端到端开启多播或多播 VPN。多播方案是最早提出的 VxLAN 方案,但是,在实施部署时有较大的限制:一是如果采用多播方案,全网开启多播将带来城域网流量大幅增加;二是如果采用多播 VPN 方案,则对城域网 IP 设备的技术要求较高,不仅某些设备不支持多播 VPN 功能,而且异厂商设备之间多播 VPN 互通也会存在不可预知的问题。

(2) 单播方案

单播方式是利用头端复制(head-end replication)功能实现 VxLAN 互通,2 个云数据中心的 VxLAN vRouter 之间配置在同一个 peer 域内,2 个 VxLAN vRouter 之间三层可达即可。单播方案对 IP 承载网本身没有额外要求,但对构建 VxLAN 隧道的 VxLAN vRouter 要求较高:一方面所必

须的头端复制功能并非所有厂商的 VxLAN vRouter 都支持,另一方面,VxLAN vRouter 作为一种 VNF (virtual network function,虚拟网络功能),其性能指标目前与硬件 VxLAN 交换机有一定差距,同时,与虚拟化平台性能耦合较紧。考虑到 VNF 的性能受虚拟化平台限制,具体实施时,建议单播方案可以每租户方式部署 VxLAN vRouter。

SDN 控制器作为 ARP 代理的 VxLAN 云数据中心通过 IP 网络互联,可以很好地实现跨广域网的云数据中心之间大二层互联,同时,由于全程是 VxLAN 隧道,可以基于每个互联端口提供 4 000×4 000 个二层逻辑实例 VNI (VxLAN network identifier,VxLAN 网络标识)。但是,目前大多数厂商不支持跨不同 SDN 控制器部署 VxLAN 隧道;同时,如采用基于多播的 VxLAN 技术实现方案,需要 underlay 的 IP 网络全程开启多播或多播 VPN,对现网设备要求较高且需要更改现网配置。

SDN 控制器作为 ARP 代理的 VxLAN 云数据中心通过 IP 网络互联,需要各云数据中心内采用同厂商 VxLAN 设备或同厂商 VxLAN vRouter 组网。其中,多播方案适用于已有 IP 网络已开启多播或多播 VPN 功能,且对大流量穿网不敏感的城域网;单播方案需要 vRouter 支持单播 VxLAN 功能,当前虚拟化实现性价比比较低,因此,单播方案适用于对成本不敏感的电信运营商。

2.2 MP-BGP EVPN 作为控制平面

MP-BGP EVPN 作为控制平面的 VxLAN 云数据中心

互联如图 3 所示。两个基于 VxLAN 组网的云数据中心分别部署支持 MP-BGP EVPN 功能的 VxLAN 网关设备,实现两方面的功能:一是对内,基于 VxLAN 实现数据中心内无控制平面的 VxLAN 网络转换为互通性较好的 EVPN VxLAN;二是对外,基于 MP-BGP EVPN 通过 underlay 承载网络实现与对端数据中心 EVPN VxLAN 网关的互联互通。

MP-BGP EVPN 控制平面实现 peer 发现、路由学习(包括本地学习和远端学习)和分发、根据终端接入信息构建用户标识并基于标识进行路由转发。IETF RFC7432^[3]定义 MP-BGP EVPN 控制平面采用 MPLS(multiple protocol label switch,多协议标签交换)作为转发标识,但目前考虑到实现简单,大多数厂商采用 VNI 作为控制平面的二类路由转发标识。当控制平面根据 MP-BGP 建立好路由转发表后,业务流量就可以沿着建立好的路径进行转发。

MP-BGP EVPN 标准主要由 IETF 制定,其中关于 MP-BGP EVPN 的二类路由转发机制有 2 个草案,即 draft-ietf-bess-evpn-overlay-04^[4]和 draft-ietf-bess-evpn-prefix-advertisement-02^[5]。两者的关键区别在于:前者定义 MP-BGP EVPN 的二类路由采用内、外两层 MPLS 标签标识用户并进行转发,而后者只定义了单层 MPLS 标签标识并转发。具体实现而言,这 2 个草案分别对应于 MP-BGP EVPN 二类路由转发机制中的对称转发和非对称转发。

(1) 对称转发

每个 EVPN VRF(virtual routing forwarding,虚拟路由转发)域内都需要设置一个专门用于路由转发的专用 VNI。当数据流进入 VRF 域,入口 EVPN VxLAN 网关将其

转发到专用 VNI,再由专用 VNI 将其路由转发到出口 EVPN VxLAN 网关。因此,对称转发将执行两次路由转发,需要两层标签,但不需要 VRF 中每个 VxLAN 交换机/路由器都存储 VRF 内所有 VNI 信息。因此,减少了 VxLAN 交换机/路由器的资源消耗,也降低了对 VxLAN 交换机/路由器硬件性能的要求。

(2) 非对称转发

EVPN VRF 域入口的 EVPN VxLAN 网关进行基于 VNI-a 的全局路由查找和转发,然后在出口 VxLAN 交换机/路由器仅执行二层转发。非对称转发无需为每个 VRF 建立一个专用路由的 VNI,且只需要一层标签。但是,非对称转发需要 VRF 中的每个 VxLAN 交换机/路由器都存储本 VRF 中所有 VNI 信息,对 VxLAN 交换机/路由器硬件资源消耗较大,对 VxLAN 交换机/路由器设备性能要求高。

目前,主流厂商的 MP-BGP EVPN VxLAN 解决方案基本遵循了这两个草案,设备形态主要是高端交换机和多业务路由器两种。

MP-BGP EVPN 作为控制平面的 VxLAN 云数据中心互联方案可以较好地实现跨广域网的异厂商 VxLAN 云数据中心互联,从而实现端到端的 VxLAN 大二层组网,组网扩展性强,可基于每互联端口提供 4 000×4 000 个二层逻辑通道实例。

3 实验模拟测试

中国电信股份有限公司上海分公司(以下简称上海电

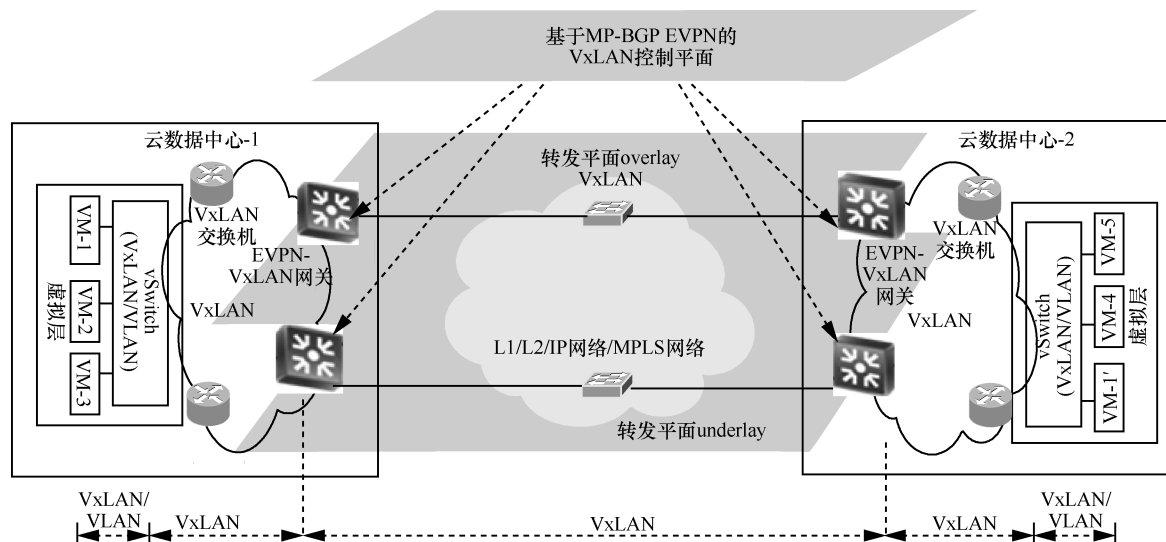


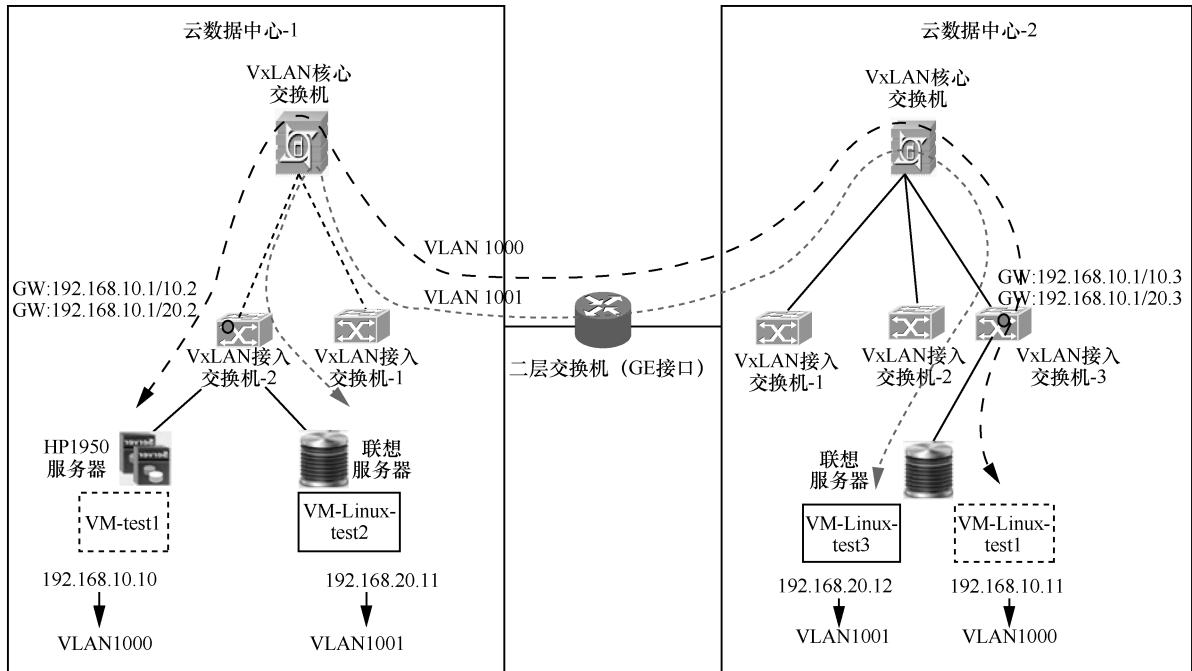
图 3 MP-BGP EVPN 作为控制平面的云数据中心互联



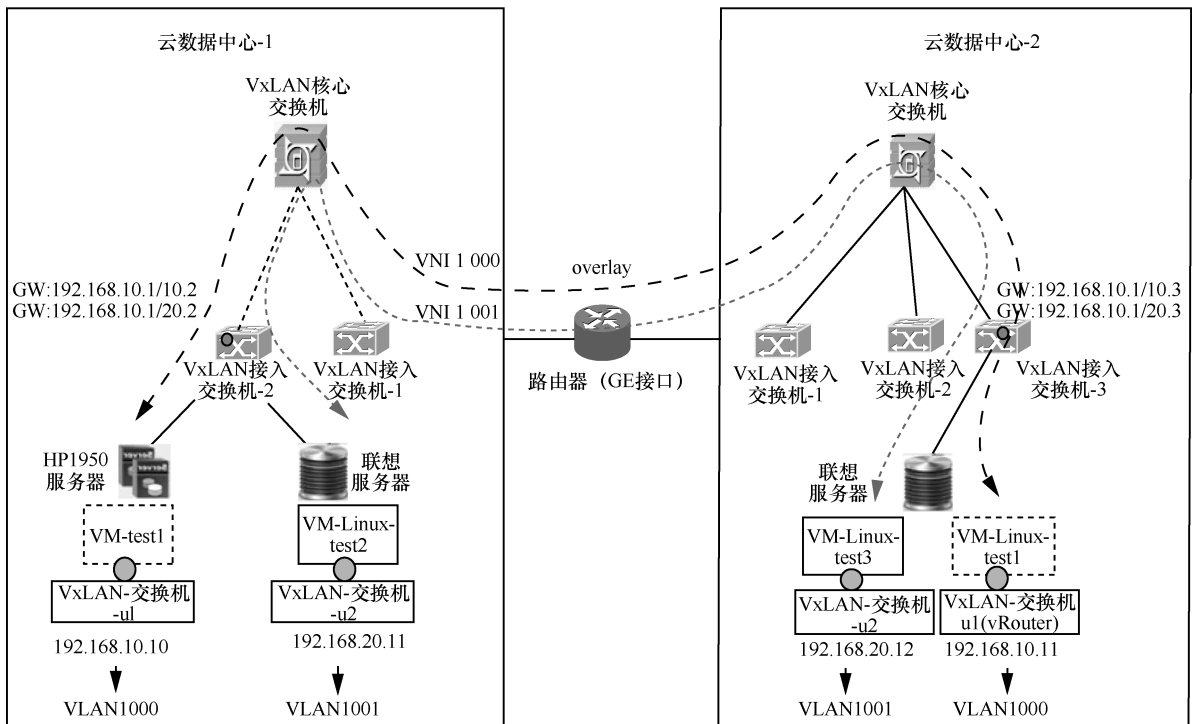
信)在2个云数据中心搭建了试验环境,分别对SDN控制器作为ARP代理的VxLAN云数据中心互联和MP-BGP EVPN作为控制平面的VxLAN云数据中心互联方案进行了测试。

(1)SDN控制器作为ARP代理的VxLAN云数据中心互联测试验证

如图4所示,将不同厂商模拟构建的云数据中心按照第2.1节方案分别通过二层网络和IP网络互联。2个云数



(a) SDN控制器作为ARP代理的云数据中心二层互联测试拓扑



(b) SDN控制器作为ARP代理的云数据中心通过IP网络互联测试拓扑

图4 SDN控制器作为ARP代理的VxLAN云数据中心互联测试

据中心的虚拟机可以互相 ping 通,且可以实现虚拟机实时迁移。

(2)MP-BGP EVPN 作为控制平面的 VxLAN 云数据中心互联测试验证

如图 5 所示,不同厂商的 VxLAN 网络分别部署在 2 个数据中心,所有 EVPN VxLAN 网关部署在同一个 AS (autonomous system,自治域)内,并配置一台路由器作为 RR(route reflector,路由反射器)。

测试中,控制平面分别采用 MP-BGP 二层 EVPN 和三层 EVPN,进行路由学习和转发;数据转发平面分别采用 VxLAN 隧道和 MPLS 技术,验证了基于 MP-BGP EVPN 的 VxLAN 互联互通效果。

- 从控制平面来看:由于参测厂商参照相同的 MP-BGP 二层 EVPN 的 2 类路由转发标准,因此,可以实现异厂商 VxLAN 数据中心互通;而由于参测厂商参照不同的 MP-BGP 三层 EVPN 的 2 类路由转发标准,导致部分厂商无法实现互通。
- 从转发平面来看:只要控制平面 MP-BGP EVPN 路由转发成功,转发平面不论是 VxLAN over IP 的 overlay 数据转发方式还是基于 MPLS VPN 转发,都

可以实现业务流的正常通信,只是此时,两种不同的转发方式对底层承载的 IP 网络要求不同。

4 结束语

随着 VxLAN 技术在云数据中心的部署日益广泛,基于 VxLAN 技术组网的云数据中心之间互联互通方案也提上了日程。电信运营商在选择跨 VxLAN 云数据中心的互联方案时,需要秉承技术标准化以规避被厂商绑架、技术成熟度以规避现网引入风险、技术经济性以实现提质增效的原则,综合考虑所选用设备的技术实现能力、互联节点间的底层承载网络现状、运营商的建设运维成本要求等情况,分类别、分场景地进行部署。同时,还要根据测试评估、现网应用、业务需求等推进相关技术进一步标准化、相关设备进一步成熟。

参考文献:

[1] 张届新,傅志仁,吴志明,等. VxLAN 在云数据中心组网的应用[J]. 电信科学, 2015, 31(9): 163-169.
ZHANG J X, FU Z R, WU Z M, et al. VxLAN application in cloud datacenter network[J]. Telecommunications Science, 2015, 31(9): 163-169.

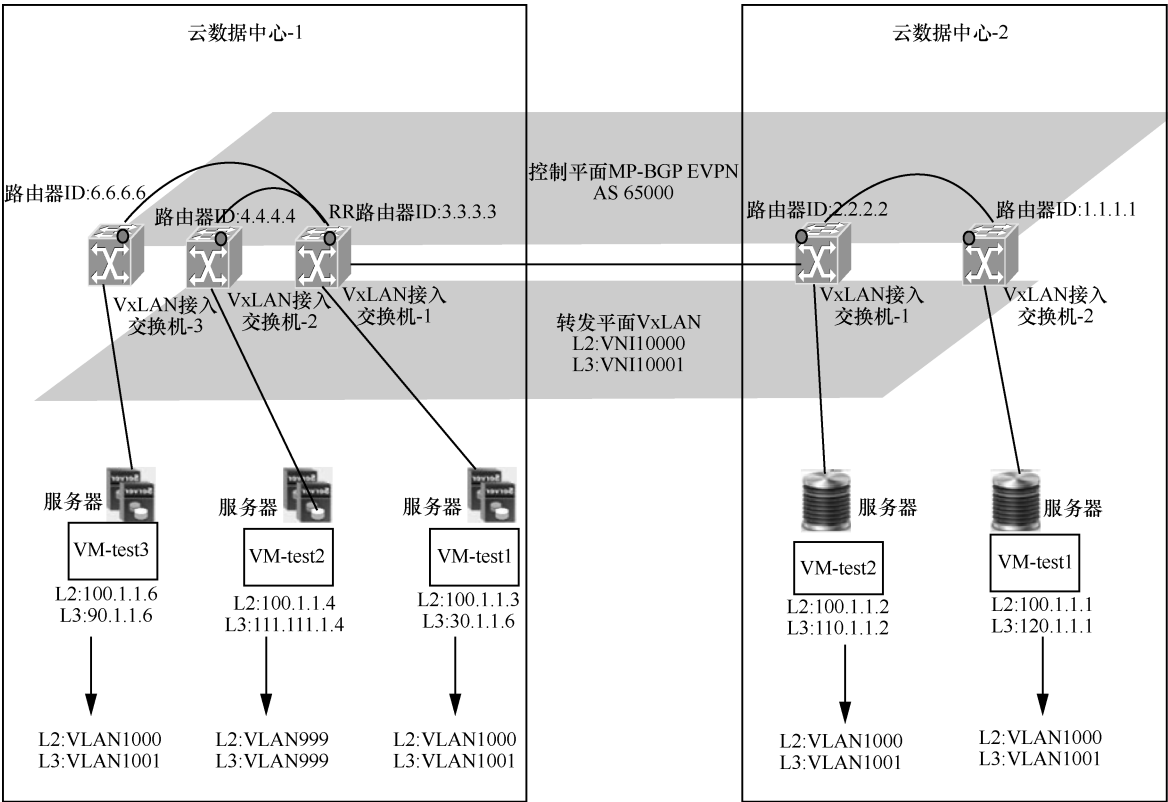


图 5 MP-BGP EVPN 作为控制平面的云数据中心互联测试拓扑



- [2] IETF. Virtual eXtensible local area network (VxLAN): a framework for overlaying virtualized layer 2 networks over layer 3 networks: IETF RFC7348 [EB/OL]. (2014-08-01)[2016-06-22]. <https://tools.ietf.org/html/rfc7348>.
- [3] IETF. BGP MPLS-based ethernet VPN: IETF RFC7432[EB/OL]. (2015-02-01) [2016-06-22]. <https://datatracker.ietf.org/doc/rfc7432/>.
- [4] Draft-ietf-bess-evpn-overlay-04, 2016 [EB/OL]. (2016-05-01) [2016-06-22]. <http://ietf.10.n7.nabble.com/bess-Comments-on-draft-ietf-bess-evpn-overlay-04-wrt-Inter-AS-Option-B-td500222.html> draft-ietf-bess-evpn-prefix-advertisement-02.
- [5] IP prefix advertisement in EVPN draft-ietf-bess-evpn-prefix-advertisement-02[EB/OL]. (2015-04-19)[2016-06-22]. <https://tools.ietf.org/html/draft-ietf-bess-evpn-prefix-advertisement-02>.

[作者简介]



张届新(1975-),女,中国电信股份有限公司上海分公司高级工程师,主要研究方向为信息网络、云数据中心组网等。



吴志明(1965-),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。