电信云网络VxLAN改造需求的分析和实践

Analysis and Practice of VxLAN Transformation of Telecom Cloud Network

史庭祥，田会芹（中兴通讯股份有限公司，江苏 南京 210012）
Shi Tingxiang，Tian Huiqin（ZTE Corporation，Nanjing 210012，China）

摘　要：
随着5G建设周期来临，电信云网络进入大规模部署阶段，VxLAN和SDN成为解决多租户和多DC组网的必备技术。目前业界提供以物理交换机EVPN为核心的硬件VTEP方案以及以虚拟化技术为核心的软件VTEP方案。以VLAN平滑演进到VxLAN为切入点，比较了2种方案的优劣，提出了平滑改造方法，并阐述了相关过程。

关键词：
云计算；VxLAN；电信云；虚拟化；SDN
doi：10.12045/j.issn.1007-3043.2020.05.017
文章编号：1007-3043（2020）05-0074-05
中图分类号：TN919
文献标识码：A
开放科学（资源服务）标识码（OSID）：

Abstract：
With the advent of 5G construction cycle，telecom cloud network has entered the stage of large-scale deployment. VxLAN and SDN become the necessary technologies to solve multi-tenant and multi-DC networking. There are currently two solutions in the industry：hardware VTEP centralized by physical switch and software VTEP based on virtualized technique. Taking the smooth evolution of VLAN to VxLAN as the starting point，it compares the advantages and disadvantages of the two schemes，puts forward the smooth transformation method，and expounds the relevant process.

Keywords：
Cloud computing；VxLAN；Telecom cloud；Virtualization；SDN

引用格式：史庭祥，田会芹．电信云网络VxLAN改造需求的分析和实践[J]．邮电设计技术，2020（5）：74-78.

1 研究背景

起初的数据中心服务于电信网络的互联网业务，包括基于互联网方式的、用户访问数据中心的互联网业务服务器，以及数据中心的服务器通过路由器访问到互联网，这2类流量统称为南北向流量。相应的，数据中心采用传统三层交换机组网，主要用于疏通南北向流量。该数据中心的设备包括：接入层交换机，即TOR（Top of Rack）交换机，提供物理服务器的接入功能；汇聚层交换机，即Spine交换机，对接入层交换机提供数据二层汇聚；Border交换机，即DC-GW，提供跨汇聚和三层交换功能，并作为南北向流量的转发出口。这样设计数据中心的目的是使接入层的二层流量在Spine交换机汇聚，汇聚层通过三层交换在汇聚和Border层之间实现高速转发。

随着云计算和虚拟化技术的发展，应用层网元和业务的跨服务器和虚拟数据中心等更加灵活的部署需求使得数据中心的设计者更加关注东西向流量。东西向流量一般是指数据中心内部、跨内部交换机网络的流量，也包括跨数据中心的流量，也就是说，各种业务和服务要基于大二层环境来部署。

传统的VLAN的作用是将大的广播域隔开成多个小的广播域，减少处理广播包和ARP包的CPU等资源消耗，域内可以直接通信，域间必须通过三层路由转

收稿日期：2020-04-01

发。虚拟化网络也需要使用VLAN构建更多的IP子网，即在现有网络架构上创建虚拟局域网，用12 bit的VLAN ID表示网段名，网段数量被限制在4 096个以内，这是传统VLAN不足以支撑大二层网络的第1点原因。第2点原因是传统TOR交换机存储的MAC地址数受限。虚拟化网络的业务需要跨多个物理节点部署，这使得二层交换需要存储更多的MAC地址，大量的东西向流量导致MAC表项增加。还有一点和虚拟化网络发展密切相关的是，当虚拟网络功能（VNF）的不同组件分布在不同的物理节点、不同的数据中心（DC）或虚拟数据中心（VDC）时，需要在三层网络之上叠加二层网络，以便延续业务原有的通信机制，并满足虚拟机迁移等虚拟化网络的新需求。

那么，为什么VxLAN技术能解决这些问题呢？作为一种大二层的虚拟网络技术，它的技术原理是引入UDP格式的外层隧道（MAC in UDP），作为数据的链路层，既解决VLAN的规模限制问题（VxLAN Header的VNI，即VxLAN网络标识，VLAN规模受限于该标识的容量），又通过打包原有数据报文，使其作为隧道净荷在二三层网络中传送。通过VxLAN技术，将二层VxLAN标记封装于三层数据包，通过路由转发，实现数据中心间的二层互联。由此可见，尽管VxLAN和VLAN都是二层网络服务，但VxLAN比VLAN更具灵活性。

a）VxLAN基于共享网络设施，提供扩展二层网段的可靠解决方案，使多租户的载荷可以在数据中心内跨物理节点传输，并为多租户提供业务隔离功能。

b）VxLAN使用24位的VNI标识符，扩展VLAN ID，消除原有的数量限制。

c）VxLAN数据包基于三层报头，完整地利用三层路由协议优化可用路径选择，如引入三层IP组播来代替以太网广播。

虚拟化网络进入SDN阶段，需要紧密结合VxLAN技术。VxLAN端点通过学习对端地址，保留它的二层协议特征，依靠组播消息发现和获取主机信息，包括IP地址、MAC地址、VNI等，导致数据中心存在大量的泛洪流量。为统一控制和简化管理，VxLAN网络需要独立的控制平面收集和管理设备上线信息，自动搭建和拆除VxLAN隧道以及规划网络环境。而随着5G和边缘计算等业务兴起，传统虚拟化网络技术难以满足云网协同、高性能转发、资源节点的边缘化等方面的需求，SDN正是当前的最佳选择。VxLAN网络有SDN Controller后，就可以向SDN单播获取这些信息，避免大量的组播流量浪费带宽和性能。同时，VxLAN有自学习功能，当收到的UDP报文不是来自已知虚拟机的数据时，VxLAN隧道节点会记录该虚拟机的VNI、IP地址和MAC地址的对应关系，减少组播消息需求。

不仅如此，SDN能通过SNMPv3自动发现网络设备，可以利用Netconf协议下发配置和流表，控制网络设备的转发面，同时还能通过北向接口提供API，向云平台开放可编程接口。SDN的这些功能使得基于VxLAN技术的SDN网络既可控，又具有灵活性。

2 以硬件VTEP为核心的VxLAN技术

VxLAN是数据层协议，需要一种控制层协议来传输二层ARP消息，用于二层网络的不同站点之间MAC地址学习和发布。最典型的控制层协议是以太网虚拟私有网（EVPN）协议，它定义了一种二层网络互联的VPN技术，但需要单独设备计算和传输二层消息，因此，和EVPN相比，SDN作为控制面的Overlay技术是更好的选择。

VxLAN是一种网络虚似化技术，技术标准包括VTEP（VxLAN Tunnel End Point）和VxLAN网关。前者完成VxLAN报文的封装和解封装，后者实现VxLAN虚拟网络之间以及虚拟网络与物理网络之间的通信。如何确定业务虚拟机的VxLAN报文产生的节点位置，即VTEP节点的位置，是电信云网络的VxLAN改造方案首先要解决的问题。当VTEP节点位于交换机网络，VxLAN网络采用的VTEP方案被称为硬件VTEP方案，如图1所示。

VTEP节点置于Leaf交换机是当前主流设计之一，显著的好处是服务器虚拟化的网络配置不必考虑VxLAN设计，不会增加DVS（Distributed Virtual Switch）开销。这样设计的优势，具体描述如下：

a）简化VxLAN组网：接入层的Leaf交换机提供硬件VTEP功能，使机架内虚拟机之间通信实现直通，支持VLAN和VxLAN之间的协议转换，也支持VxLAN网络（VNI不同）之间的通信。服务器不提供VTEP功能，使其网络配置更加简单，既不需要VxLAN设计，不增加DVS额外开销，也不需要承担二层交换功能。如此设计，有利于快速扩容机架内的服务器，达到扩容业务容量的目的；同时该设计在Leaf中引入VTEP节点，不必配置大量MAC地址，节约系统开销。

b）分布式路由L3 GW：传统DC组网的三层交换

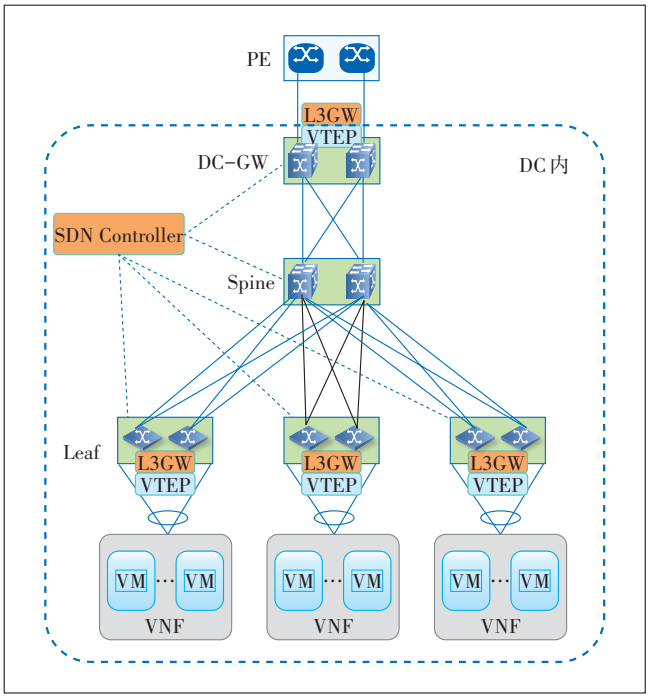

图1 硬件VTEP+分布式路由方案

由DC-GW实现，造成许多问题，如DC-GW资源消耗很大，Leaf和Spine相对负载不足，广播域过大等。采用SDN组网后，通过Leaf启用三层路由功能，使同机架下虚拟机之间的三层流量在Leaf终结，节省DC-GW带宽；同时有利于简化SDN组网配置，将本地节点路由配置固化在Leaf节点。

c）简化SDN组网：服务器不作为VTEP节点，该设计将弱化SDN控制器对服务器的影响，避免SDN控制器出问题时，影响全网通信。电信云网络有安全管控和统一互联网出口等负载均衡的需求，为实现该需求，需要提供vFW，vLB功能，该功能用于主动控制虚拟机流量在Leaf完成二层交换；同时Leaf、Spine和DC-GW由SDN控制器统一管理，Overlay的SDN网络架构和管理更加清晰。

VTEP和L3 GW部署在Leaf，只解决机架内虚拟机之间的通信，当不同机架下的虚拟机之间，即跨Leaf的通信发生时，Spine起到二层汇聚作用；当DC内虚拟机需要和DC外的设备通信时，流量需要从Spine导入到DC-GW，完成三层交换，同时作为DC外网络的交换节点，VTEP和VxLAN GW功能都需要部署在DC，并支持虚拟化网络和物理网络的通信。

3 以虚拟化软件为核心的VxLAN技术

SDN起初以控制和交换分离的思想得到业界认可，通过控制和下发交换机参数配置来弱化对交换机和路由器等网络设备的要求。“SDN+白盒交换机”试图颠覆以Cisco为代表的传统网络设备厂家的优势地位，两者斗争的结果是互让一步，在虚拟化和云计算网络中达成一致，传统厂家保住物理交换机的份额，但需要支持ONF、SNMPv3和NetConf协议，以便接受SDN控制器的控制。

然而对于部署云计算业务、拥有电信云网络的广大业主而言，电信云网络面临业务网元繁多、网络复杂、大流量爆发、传统和虚拟化硬件混杂、安全可靠性要求高等诸多挑战。因此，电信云网络要考虑DC内容纳多个厂家的交换路由设备、多个厂家的服务器，甚至多个SDN控制器和多个云平台管理系统，即异构化的云网络。硬件VTEP方案将面临诸多问题，例如：

a）SDN控制和管理系统与交换网络设备不同厂家的耦合问题。

b）跨DC的大二层互通没有解决方案。

c）交换网络扩展速度和服务器扩容节奏不匹配。

d）DC内不同厂家的交换网络设备的互通互联问题。

e）将VLAN改造成VxLAN，需要重新配置交换网络设备，无法平滑升级。

可见，采用硬件VTEP为核心的VxLAN技术用于解决大规模虚拟化网络扩展时不够完美。此改造方案过于依赖交换网络设备，也违背了SDN以软件定义网络的初衷。

遵循硬件虚拟化原则，如果把VTEP功能内置在虚拟化交换机，对物理交换机网络将没有额外要求，而交换设备可以维持VLAN网络不变，由服务器上的虚拟化软件和云平台管理系统来实现DC内的VxLAN组网，这样的网络设计是一种更“软件化”的Overlay网络。和硬件VTEP方案不同，VMware的网络虚拟化功能没有采用SDN和Openflow技术，而是NSX。按照NSX for vSphere 6.3.1版本的定义，NSX的关键角色如下：

a）VDS：虚拟化分布式交换机。内置在虚拟化管理程序vSphere内核，提供和虚拟机的连接，包括VLAN和VxLAN功能。VDS起到本地逻辑交换机（LS）的作用，也包括VTEP功能。

b）DLR：分布式逻辑路由器。驻留在vSphere内核，提供东西向分布式路由，用于租户的IP地址空间和数据空间分离。DLR使得同一主机上不同VxLAN

子网之间的虚拟机通信不需要跨越传统路由接口。

c）ESG：边界服务网关。运行在虚拟机，提供统一网关业务，包括DHCP、VPN、NAT、动态路由和负载均衡功能，用于把逻辑分开的VxLAN网络连接到共享的上行网络。ESG支持OSPF/BGP协议，连接虚拟化网络和物理网络。

上述3个关键软件组件构成逻辑路由隔离，如图2所示，VNF内部虚拟机之间的通信组成一类VxLAN网络，DC内部跨VNF的通信构成一类VxLAN网络，ESG上行网络是一类VLAN网络。这些网络之间的隔离通过VDS、DLR和ESG实现。NSX构建的这种网络，私有云内（无论是DC内部还是跨DC）的流量都采用VxLAN方式，当需要路由或桥接到外部物理网络（或VLAN网络）时则使用ESG进行隔离。如果VxLAN和VLAN转换不在ESG完成，也可以在Leaf启用二层网桥，但这样就会对Leaf交换机有特殊要求，所以建议采用ESG，以虚拟机方式部署在Edge Cluster，实现该转换。

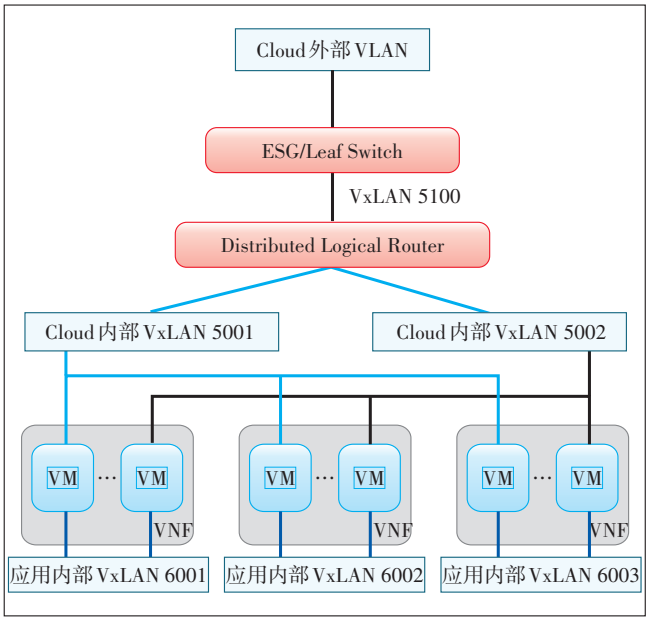

图2　VTEP内置虚拟交换机

由此，解决了硬件VTEP组网方案的上述诸多问题，同时，该方案面向虚拟机进行组网配置，而不像硬件VTEP方案面向物理硬件组网，因此，具备以下优势。

a）简化路由：VDS内置VTEP和DLR，不仅运行效率高，而且服务器内虚拟机之间的通信不必像硬件VTEP方案上行到Leaf节点进行VxLAN打包，降低Leaf下行带宽需求。此外，云内跨DC之间虚拟机通信，采用VDS内置DLR进行路由，无须关注DC间的物理交换机网络，不像硬件VTEP方案的交换机网络和DC需要一一对应，造成跨DC的物理交换机组网和VxLAN组网的交叉影响。

b）管理简化：SDN控制和管理系统只会影响到服务器，不必和交换机网络协同，而且作为Overlay网络，采用软件方式解决路由隔离，对服务器硬件和物理网卡的组网配置没有影响，因而降低了管理难度。

c）丰富的控制手段：SDN直接管理和控制服务器VDS/DLR/ESG的流表，而且可以进一步控制到虚拟机粒度，甚至每个虚拟机的vPort，将控制手段精细化到极致，有利于性能优化和故障定位。

4 VLAN改造成VxLAN网络的实践

目前国内相关云化网络的VxLAN方案是直接部署VxLAN，没有提供从VLAN平滑演进到VxLAN的方案。本文对面向虚拟机的逻辑路由隔离方案和以物理交换机为核心的硬件VTEP方案进行了深入分析和对比，发现前者更有优势，下文将从改造网络和平滑演进的角度，展现其更多的优势和落地价值。

4.1 VxLAN网络的必备要素

VxLAN是大规模云化网络发展的必备技术，SDN是简化VxLAN组网配置的利器，两者的架构设计和演进是云计算发展的重要一环。以VMware NSX系统为例，其必备要素有：

a）NSX Manager：配置NSX Controller，在Hypervisor上安装VIB（vSphere Installation Bundles），以便开启VxLAN、DLR、分布式防火墙（DFW），还提供配置ESG以及负载均衡、VPN和NAT等网络服务。

b）NSX Controller：路由控制平面，实现对转发平面流量的集中式策略控制，本身支持集群配置，并特别针对VxLAN网络环境提供抑制ARP广播包功能，以减少二层网络的广播泛洪。

c）VDS、DLR、ESG：功能如第3章描述，并提供统一分布式路由器，实现跨DC的VxLAN组网服务。

d）安全服务：包括云网络内部租户或虚拟机的安全隔离，与物理网络的互通管控，以及对租户的权限管理等。

4.2 改造方法和相关过程

以虚拟化软件为核心的VxLAN技术，不必改造关联的物理交换机网络，也能完成包括跨DC的云内网

络的VxLAN改造。

VLAN网络逐步改造成VxLAN，需要实现两者的互通和共存，为此，移植计划包括以下工作：

a）下挂主机的Leaf上配置VLAN网络，支持新配置VxLAN网络的南北向流量和跨DC的东西向流量。

b）基于必备要素，创建NSX/VxLAN网络组件：

（a）管理集群（Management Cluster）部署NSX Manager，边缘集群（Edge Cluster）部署NSX Controller。

（b）主机Hypervisor上部署VIB。

（c）主机上部署VDS实现本地虚拟交换机（LS）功能，部署DLR实现主机之间虚拟机的通信。

（d）Edge Cluster部署ESG。

（e）DLR和ESG位于不同主机，跨主机创建OSPF/BGP路由。

c）ESG和下挂主机的Leaf之间创建OSPF/BGP路由。

d）在主机VLAN网络的VDS端口组和VxLAN网络的LS之间创建二层逻辑网桥，如图3所示，除Edge Cluster是新建主机节点外，现有Management Cluster和Resource Cluster（部署应用的服务器节点）所在的VLAN网络也要和“新增的VxLAN网络”建立二层网桥实现互通，以便在移植过程中共存。建议在Hypervisor的VDS上内置网桥，比在主机Leaf上部署更能简化网络。

e）将虚拟机的vNIC从现有VDS端口组迁移到VxLAN网络的LS。

f）将虚拟机的上行到Leaf的交换虚拟接口（SVI）迁移到DLR，从而将跨主机的东西向流量迁移到VxLAN网络。

g）所有虚拟机路由移植到VxLAN网络后，拆除二层网桥。

上述移植过程无须增加额外的应用服务器节点，也不需要重新部署租户业务，除跨主机的东西向流量迁移会有少量业务中断外，VLAN向VxLAN网络移植平滑可控，而且可以逐个主机节点进行迁移，比硬件VTEP方案更加灵活可靠。

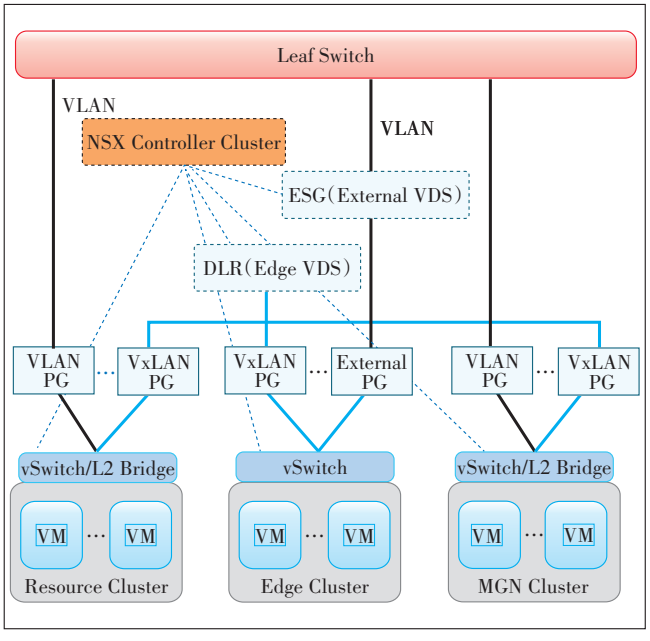

图3　VLAN上移植VxLAN

5 结束语

本文给出的VLAN平滑移植到VxLAN的示例虽然基于VMware NSX系统，但是该方案也可以基于Openstack云平台的DVS技术来构建，只要实现同样的逻辑路由隔离，也会达到同样的效果。

以硬件VTEP为核心的VxLAN方案的最大困扰是被交换机和SDN厂家绑定，而基于Openstack技术的VxLAN组网方案由于面向虚拟机组网，会更加高效，而且虚拟化软件技术更有利于国内IT和电信设备供应商发挥软件定制的本土优势，运营商也相应有更多选择。

参考文献：

[1] 张届新，吴志明. 基于VxLAN组网的云数据中心互联方案[J]. 电信科学，2016(12)：122-128.
[2] 秦杰炜. 基于VxLAN组网的云数据中心互联方案研究[J]. 数码世界，2017(10)：44-44.
[3] 王永建，张健，张富根，等. 基于VxLAN的云数据中心网络研究[J]. 通信技术，2017，50(1)：78-83.
[4] 缪仕福. VxLAN网络技术研究[J]. 科技资讯，2015，13(4)：15-16.
[5] 王锋等. VPC多租户虚拟化网络解决方案[J]. 电信技术，2015(8)：56-58.
[6] 陈佳媛. 中国移动面向5G的电信云基础设施技术研究和实践[J]. 移动通信，2019(1)：57-60.

作者简介：
史庭祥，高级工程师，硕士，主要从事核心网市场规划和管理工作；田会芹，高级工程师，硕士，主要从事业务软件的产品规划工作。