



面向 vDC 的 overlay 网络关键技术研究

张届新, 吴志明, 徐文华

(中国电信股份有限公司上海分公司 上海 200085)

摘 要: 云计算时代, 数据中心趋于虚拟化演进。分析了新时期数据中心虚拟化组网演进的需求, 对面向虚拟数据中心组网的基于 overlay 网络架构的多种关键技术进行了对比研究, 并提出了电信运营商虚拟数据中心组网仍需关注的问题。

关键词: 虚拟数据中心; 隧道; NVGRE; overlay; STT; SPB-M; VXLAN

doi: 10.3969/j.issn.1000-0801.2014.05.020

Study on the Key Overlay Networking Techniques in vDC

Zhang Jiexin, Wu Zhiming, Xu Wenhua

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

Abstract: The new demand from virtual data center was analyzed. Several key overlay SDN techniques that would be used for vDC interconnection were discussed, and several key problems of vDC interconnection should be focused were put forward.

Key words: vDC, tunnel, network virtualization GRE, overlay, stateless TCP transmission, shortest path bridging-MAC in MAC, virtual extensible VLAN

1 引言

随着我国云计算进入应用落地的快速发展期, 作为云计算基础的数据中心也面临从传统的部署管理模式向云计算数据中心转型的问题。云计算从互联网演变而来, 本质是通过网络将计算力进行集中和存储, 并通过网络提供服务。如果没有网络, 计算力集中规模、服务的种类和可获得性就会受到极大的限制。可见, 面向云计算的数据中心包括计算、存储和网络 3 种资源。又由于这些资源采用物理形式或虚拟形式存在, 且用户所需的整体系统资源可能分布在用户端、电信运营商的一个或多个数据中心内; 反之, 电信运营商的一个数据中心内的资源可以虚拟为多个独立的虚拟环境, 分别为不同用户提供逻辑上完全隔离的 DC 微环境, 这种模式的数据中心架构称为虚拟数据中心(vDC)。

现阶段, 电信运营商的 vDC 发展面临以下 3 方面的驱动。

一是业务更新驱动, 包括电信运营商内外部双方的驱动。就外部环境而言, 企业的 IT 系统主要为企业实现减少成本、创造价值和提高效率这 3 种价值可能性。基于云计算虚拟化、集中化运营的 IT 外包日益成为趋势, 越来越多的企业将部分或全部 IT 系统部署在电信运营商的数据中心, 这就需要电信运营商为群体用户提供大量的计算和存储资源, 同时又要保障个体用户的应用体验。就内部环境而言, 电信运营商在积极开发创新型业务实现开源的同时, 还需尽可能降低运营成本, 如保护既有投资、引入可以提升运维效率和降低总体建设成本的新技术等。

二是网络发展驱动, IT 技术的变革带来高密度的虚拟机, 也带来高密度的 MAC 和 ARP, 需要提供大量虚拟资源

隔离,因此需要二层网络突破 4 096 个 VLAN 资源的限制问题。同时,用户 IT 应用模式的改变和 IT 技术本身的变革,需要电信运营商的数据中心提供更加高可靠的二层网络环境,如规避环路以支持虚拟机实时迁移等。而对于电信运营商而言,面对大量虚拟机接入所带来的高带宽接入、高质量迁移需求,电信运营商需要以较低成本提供更高的带宽资源利用率及带宽通道,才能保证降本、提质、增效。

三是技术演进驱动,云计算风起云涌,一方面 IT 虚拟化技术不断发展,如 VM 等;另一方面,承载 IT 并提供业务互通的数据中心网络组网技术也在快速发展。如可以突破 VLAN 数量限制,且可以较好地利用现有网络组网技术——overlay 网络技术,如 VXLAN (virtual extensible VLAN,VLAN 虚拟化扩展)、STT(stateless TCP transmission,无状态 TCP 传送)、NVGRE (network virtualization GRE,网络虚拟化的 GRE)、SPB-M (shortest path bridging-MAC in MAC,基于 MAC in MAC 的最短路径桥接协议)等。

因此,在尽可能保护既有网络投资(包括用户接入网络、数据中心内组网等)、降低网络改造成本和风险的前提下,如何接应云计算时代的 IT 服务变革需求,选择哪种技术进行 vDC 组网,是一个亟需研究的问题。本文就 vDC 组网的 overlay 网络关键技术及其组网的难点进行研究、对

比,以为电信运营商的 vDC 组网提供参考。

2 vDC 应用需求及组网架构

从 vDC 定义可以看出,vDC 包括 CPN (customer premise network,用户驻地网)和电信运营商的 DC,涉及组网包括用户接入网络、DC 内网络、DC 互连网络,如图 1 所示。

vDC 的主要应用需求及网络互通要求见表 1。

因此,用户接入网络可以利用已有的二层或三层网络,如光接入网、VPN(virtual private network,虚拟专用网络)等;而对于 DC 内网络,二层和三层两种网络技术都可以实现异地 DC 互联,以实现数据中心灾备、业务弹性扩展、就近接入和按需迁移等价值,但两者有差异,表现在以下两方面。

- 三层互联技术:组网简单灵活、经济。不支持大二层网络,DC 间虚拟机互访需要进行三层 IP 路由转发,需要 ARP 的广播报文,三层多播报文开销较大,主要技术为 L3 VPN。
- 二层互联技术:支持大二层网络,DC 间建立端到端二层网络,可靠性高,时延小于三层组网,报文开销也比三层网络小,承载率高。主要技术为 L2 VPN (如 VPLS)等。

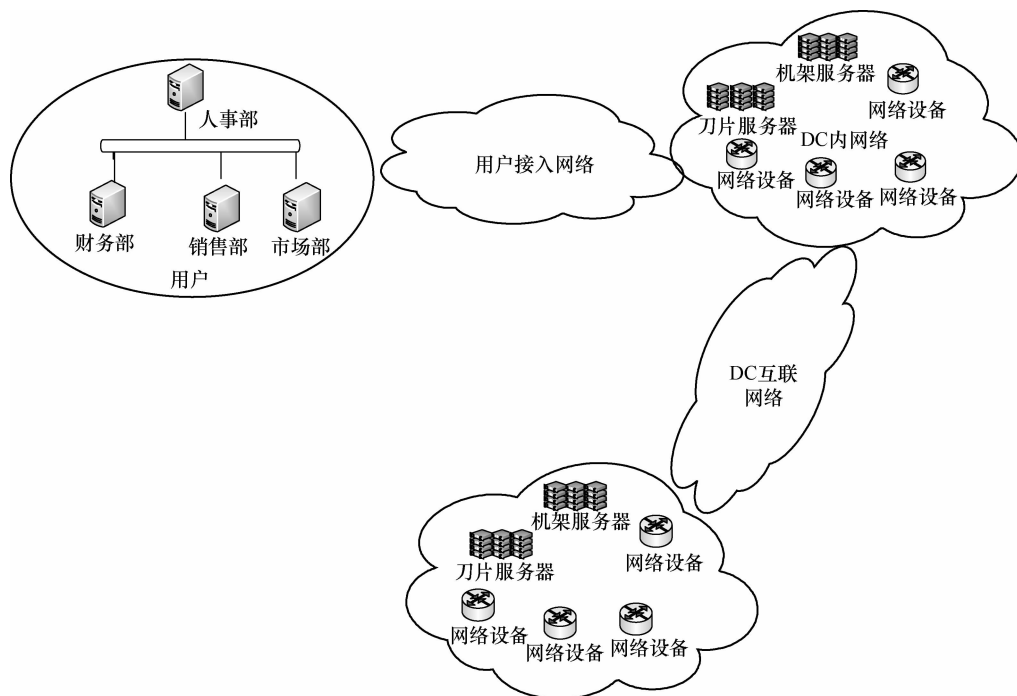


图 1 vDC 组网涉及网络示意图



表 1 vDC 的主要应用需求及网络互通要求对应

应用场景	应用需求	网络互通要求
用户—DC 流量	内容访问	二层接入、三层 VPN 接入或“公网+SSL VPN”等方式接入
	虚拟机实时迁移	二层互联
	无盘启动	二层互联
	数据同步	二层或三层互联,视用户需求而定
	数据异地容灾备份	二层或三层互联,视用户需求而定
DC 内流量	...	
	虚拟机实时迁移	二层互联
	服务器集群	二层互联
	资源整合(当前主要是计算资源整合)	二层互联
DC 间流量	...	
	服务器集群	二层互联
	数据同步	二层或三层互联,视用户需求而定
	资源整合	二层互联
	虚拟机实时迁移	二层互联
	数据异地容灾备份	二层或三层互联,视用户需求而定
	协同计算	可三层互联
	...	

对于 DC 互联网络,传统 DC 内既有二层组网也有三层组网。但云计算的发展,带来了 DC 内虚拟机数量日益增大、虚拟机间必须安全隔离、虚拟机需要实时迁移等新要求,需考虑引入可以解决这些问题的 overlay 网络。

3 面向 vDC 的 overlay 网络关键技术

所谓 overlay 网络,是指在现有网络上叠加一层网络

进行数据承载与传送。相对于实现目标尚远、标准尚未完善、电信级商用设备尚未成熟的 SDN (software defined networking,软件定义网络),overlay 网络技术可以看作 SDN 应用于虚拟数据中心的前导模式,基于 overlay 网络的 vDC 组网不仅可以突破传统二层网络 4 096 个 VLAN 数量的限制,而且可以与现有二层或三层网络较好地共存、融合,满足虚拟机隔离、实时迁移等需求,技术实现相对简单,引入部署对现网的改动相对较小。

目前,基于 overlay 网络的技术主要有 4 种,分别是 VXLAN、STT 协议、NVGRE 协议和 SPB-M,这 4 种 overlay 技术的本质区别体现在封装方式和转发机制两方面。

3.1 VXLAN 技术

VXLAN 是由思科和 VMware 为主导提出的 IETF 草案,支持方主要是网络设备制造商、芯片厂商和软件服务提供商,除思科和 VMware 之外,还有红帽、思杰(Citrix)、Arista Networks 和博通等。

(1)封装方式

VXLAN 是一种 MAC over IP 的协议,使用 UDP 封装内层以太帧,封装报文头共 50 byte。VXLAN 使用 24 bit VXLAN 网络标识符 VNI (virtual network identifier,逻辑网络标识),如图 2 所示,一个 VXLAN 管理域最大支持 16 000 000 个逻辑网络(VNI),而一般一个 VNI 最多支持 4 096 个 VLAN。

VXLAN 封装包括如下部分:

- VNI,同一个 VNI 内的主机才可互相通信;
- RRRR IRRR 中的 I 置 1,表示是一个有效的 VNI;
- 目的端口号为 4789;
- 源端口号可根据内部的以太头散列算法生成,这样路径上的设备可根据隧道的 UDP 信息实现路

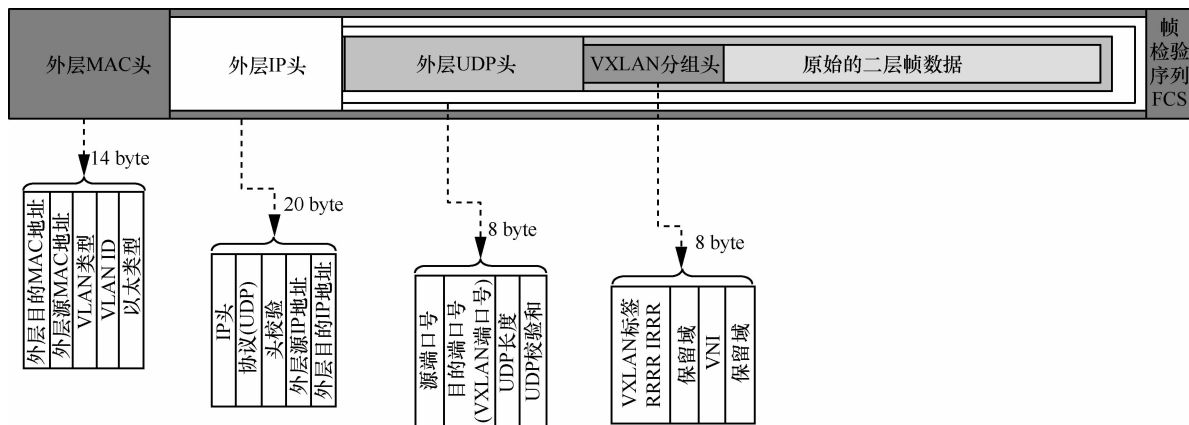


图 2 VXLAN 帧封装格式示意

径均衡;

- 外层源 IP 地址为源 VTEP (virtual tunnel end point, 虚拟通道终端) 的 IP 地址;
- 外层目的 IP 地址为目的 VTEP 的 IP 地址。

在将以太帧封装到隧道时,一般应把 VLAN 剥掉;在接收端,如隧道内的以太帧带 VLAN,一般把这个帧扔掉。实际应用中,也可以通过配置,改变处理规则。

(2)转发机制

在 VXLAN 网内,广播或多播的报文被封装在 IP 多播中,如 ARP 报文,在初始化阶段,采用 IP 多播进行 ARP 学习;也可以采用 IP 单播方式。而在 ARP 老化时间内通过 MAC:VNI 寻找目的地址数据分组。

VXLAN 利用 ECMP(equal-cost multi-path routing, 等价多路径路由)对不同数据流使用不同的 UDP 源端口,实现等价多路径的多链路负载均衡,有效提升带宽利用率。

VXLAN 可跨越物理三层网络,使用 UDP 封装能在三层物理网络上建立二层逻辑网络。在三层组网时,VXLAN 网关进行 MAC in UDP 封装,定义 UDP 隧道的目的 IP 地址及目的 UDP 端口号 4798 进行寻径。其中,需要注意的是,VXLAN 的 UDP 隧道是个不可靠的通信信道,如果需要可靠通信,则需要终端虚拟机的 TCP 保证。

3.2 STT 技术

无状态 TCP 传送技术主要为 VMware 旗下的 Nicira 倡导,并提出 IETF 草案,支持厂商较少。

(1)封装方式

STT 是一种 MAC over IP 的协议,在 IP 报文的有效载荷(payload)中,除了虚拟网络的二层分组以外,还要依次封装一个 TCP 头和一个 STT 头,如图 2 所示。但 STT 是一种无状态 TCP,采用 TCP 封装形式,将 TCP 各字段意义重新定义,无需 3 次握手建立 TCP 连接。

STT 帧封装格式如图 3 所示。其中,4 层偏移量(L4 offset)表示 STT 头+TCP 头的长度;STT frame(即 STT 分组头+有效载荷)是按照最大分段尺寸(max segment size, MSS)的大小分段的,分段之后每个段会加上 TCP 分组头,利用 TSO(TCP segment offload, TCP 分段卸载)机制交给网卡;PCP(priority ceiling protocol, 优先级上限协议)表示优先级,可以不设;VIVLAN 是 VLAN 相关的设置,略过;context ID 表示租户 ID,每个租户不同。

(2)转发机制

STT 的 TCP 源端口号和目的端口号都是随机选的,源

版本	标识	4层偏移量	保留
最大分段尺寸		PCP	V VLAN
context ID (租户ID, 64 bit)			
填充	数据		

图 3 STT 帧封装格式示意

端口号范围建议在 49 152~65 535,目的端口号范围建议在 1 024~49 151。对于一条隧道/通道两端固定的端点而言,两者之间的通信应该用固定的源/目的端口。

在数据帧转发过程中,SEQ/ACK 也不会用作窗口重传或者拥塞控制,而是用作一个 STT 帧的分段/合并,对于被分段的同一个 STT 帧,其 TCP 头的 ACK 字段是相同的。

STT 端点需要给 STT 预留 TCP 端口,所有发往这个端口的 TCP 分组会被认为是 STT 分组,从而不会走正常的 TCP 协议栈路径。

同 VXLAN 类似,STT 隧道也是不可靠通信信道,如要可靠通信,则需要终端虚拟机的 TCP 协议栈保证。

3.3 NVGRE 技术

NVGRE 是一种基于 GRE 优化的面向虚拟数据中心组网的 overlay 技术。目前 NVGRE 为 IETF 草案,主要支持者为软件服务提供商和芯片厂商,如微软、惠普、Arista Networks、英特尔、戴尔和博通等。

(1)封装方式

RFC1701 定义的 GRE 帧封装格式如图 4 所示。NVGRE 封装是对 GRE 的改造和再利用,利用 GRE 封装中 key 域的低 24 bit 标识隧道 VNI。其中,协议类型为 2 byte,表示 GRE 报文内的净荷协议类型,如 IP 或 IPX 等非 IP;校验和为 2 byte,表示 GRE 头部和 IP 净荷的校验和;key 为 4 byte,表示隧道内的单个业务流,同一业务流采用相同的 key,隧道拆除点根据 key 域的值判断业务流归属;流序列号为 4 byte,表示隧道内的报文顺序。

5					8			13			16			31 bit			
C	R	K	S	s	递归控制	标识			版本			协议类型					
校验和(可选)									偏移量(可选)								
key(可选)																	
序列号(可选)																	
路由(可选)																	
净荷																	

图 4 RFC1701 定义的 GRE 帧封装格式示意

(2)转发机制

NVGRE 在 GRE 基础上,引入控制平面(尚未完全定义),如利用 GRE 关键域进行 ECMP。



3.4 SPB-M 技术

首先,介绍一下 PBB(provider backbone bridge,运营商骨干桥接)技术。PBB 是 IEEE 802.1ah 标准草案定义的 MAC in MAC(下文简称为 MinM)技术;采用 MinM 封装,基于运营商 MAC 地址,而不是用户 MAC 地址转发流量。PBB 分为骨干网和接入网两个层次。接入网是标准的二层网络,可以采用 STP 或其他二层技术管理控制。2006 年左右在业界有部分商用,用于进行基础数据网络的二层专线迁移。但随着标准制定的旗手北电网络公司倒闭,PBB 后来的标准和网络发展一度停滞。

SPB-M 技术是基于 PBB 技术,在其控制平面进行优化而形成的一个新的 IEEE 标准,封装格式与 PBB 帧相同,如图 5 所示。

SPB-M 的数据转发平面基于 PBB 技术,但控制平面对 PBB 进行了优化,增加了等价多路径计算等。SPB-M 管理二层网络的拓扑、环路、计算路径等,通过 IS-IS(intermediate system to intermediate system,中间系统到中间系统)协议学习骨干网的拓扑,计算最短转发路径和等价路径,指导报文在骨干网上进行最优的二层转发,有效地减少转发跳数,提高转发性能;PBB 根据 SPB-M 计算的路径进行转发。

3.5 overlay 网络技术对比

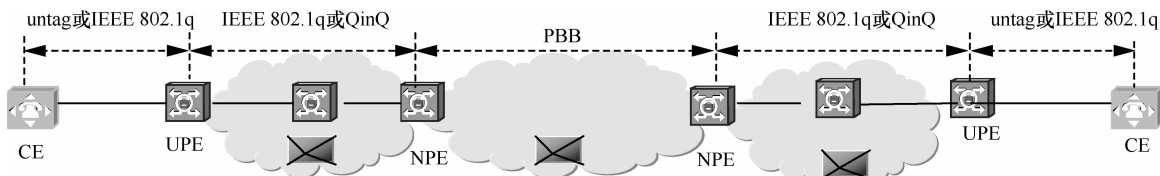
从封装方式及转发机制等理念上看,VXLAN、STT、NVGRE 与 SPB-M 分属于两种不同类型的 overlay 大二层

组网技术。VXLAN、STT、NVGRE 是采用 MAC in IP 封装,一般在服务器上安装插件即可支持隧道建立,VXLAN 与 STT 关系最为紧密,体现在:一是基于无状态隧道进行二层数据传送,可通过软件或硬件支持;二是等价多路径算法基本相同;三是两者都是同一个厂商 VMware 的子产品。相对而言,SPB-M 是 MinM 协议,不仅封装和转发机制完全不同,而且还需要专门的硬件交换机支持隧道的起止,目前无法直接在虚拟化服务器上通过软件支持。

因此,技术对比分两个层面展开:一是以 VXLAN 为代表的基于 MAC in IP 封装的隧道,与基于 MinM 封装的 SPB-M 隧道进行对比,见表 2;二是在 3 种 MAC in IP 封装的隧道技术中进行进一步细化比较,比较结果见表 3。

综上所述,通过全方位的技术对比,针对当前技术设备及设备实现,总结如下。

- VXLAN 与 STT 封装方式类似且转发机制相同,但 VXLAN 相对于 STT 更易于现网部署。虽然 VXLAN 也尚未完全标准化,但 VXLAN 基于现有成熟协议 UDP 实现,对服务器要求低;而理论上 STT 优于 VXLAN 技术的“控制平面”仍尚未定义,且 STT 基于的无状态 TCP 不仅非标准,而且对现有协议进行了改造,对服务器有较高要求,需要支持 TSO 和 TOE 功能。
- 与 VXLAN 相比,NVGRE 最具优势的控制平面也尚未完全定义,且 NVGRE 对现网设备的硬件处理能



净荷		净荷	净荷
以太类型		以太类型	以太类型
用户 VLAN(CVLAN)		用户 VLAN(CVLAN)	用户 VLAN(CVLAN)
以太类型		以太类型	以太类型
用户外层 VLAN(SVLAN)		用户外层 VLAN(SVLAN)	用户外层 VLAN(SVLAN)
以太类型		以太类型	以太类型
用户源 MAC 地址		用户源 MAC 地址	用户源 MAC 地址
用户目的 MAC 地址		用户目的 MAC 地址	用户目的 MAC 地址
IEEE 802.1	IEEE 802.1q	IEEE 802.1ad	骨干网服务实例(I-SID)
			以太类型
			骨干网 VLAN(BVLAN)
			以太类型
			骨干网源 MAC 地址(外层地址)
			骨干网目的 MAC 地址(外层地址)
			IEEE 802.1ah

图 5 PBB/SPB-M 帧封装格式示意

表 2 VXLAN 与 SPB-M 技术对比

	VXLAN	SPB-M
实例标识	VNI	I-SID
实例空间	24 bit(4 KB×4 KB)	一般为 24 bit (4 KB×4 KB); 可将 VLAN 或 QinQ (也称 stacked VLAN 或 double VLAN)封装其中
VLAN 处理	VXLAN 封装一般将 VLAN 剥离,在目的地重新封装 VLAN	可将 VLAN 或 QinQ 封装其中
转发平面	通信双方在一个二层域中时,基于 MAC 地址转发;通信双方需要跨三层网络时,可将 MAC 地址封装入 UDP 隧道,基于目的 VTEP 隧道的 IP 地址进行寻径,在三层网络上进行转发;或将 VXLAN 封装入 L2 VPN 或终结 VXLAN 后封装入 L3 VPN 进行转发;需端到端支持 IP 多播	二层网络中,基于运营商 MAC 地址进行寻径;如需要跨三层网络,可进行 MinM in L2 VPN 封装后转发,或终结 SPB-M 后封装入 MPLS (multi-protocol label switching, 多协议标签交换) L3 VPN 后转发
控制平面	MP-BGP/IS-IS/OSPF/PIM,可基于 SDN 控制器进行控制	主要基于 IS-IS 进行最优等价多路径计算
等价多路径	支持	支持
实例标识分配	由 VTEP 自动为服务器 Hyper-V 中的 vSwitch(虚拟交换机)端口分配 VNI,而 VTEP 一般在 Hyper-V 上自动支持	需要在服务器外部部署支持 SPB-M 协议的网络设备,为虚拟机进行 VLAN 到 I-SID 业务实例标识映射
优点	VTEP 天然与 VMware 等服务器虚拟环境契合较好,实现较为简单,支持厂商较多	实例空间更大 1 倍;无需 IP 多播环境;可同时用于 WAN 组网
局限及不足	一般需要网络端到端开启 IP 多播;如采用 UDP 隧道模式穿越三层网络,则隧道数量对网关性能要求较高,且在 IP 层可能发生分组丢失,需要 VM 终端启用 TCP 重传机制予以可靠性保证;标准尚未完全成熟;实例空间比 SPB-M 小 1 倍	SPB-M 目前无法与 VMware 等虚拟环境直接契合,需要在服务器外部部署 SPB-M 网络设备,组网则需要端到端部署 SPB-M 设备,实现较为复杂
标准化情况	目前处于 IETF 草案阶段	已为 IEEE 标准发布

力要求高,改造成本大。

- SPB-M 相对于 VXLAN 具有标准化的优势,同时有较好的技术储备和商用基础——不仅其转发平面机制 PBB 技术本身曾有多多个商用案例,而且 SPB-M 也有商用,如索契冬奥会网络采用 Avaya 的 SPB-M 产品构建。但当前服务器虚拟化与 SPB-M 的契合度尚欠缺,SPB-M 无法直接部署在服务器的虚拟平台上,因而导致基于 SPB-M 的 vDC 组网实现相对复杂,vDC 中部署成本相对 VXLAN 较高。

4 结束语

vDC 网络涉及用户接入 DC、DC 内组网和异地 DC 间互联 3 个层面;而云计算仍在发展阶段,每个 vDC 部分都可能物理资源和虚拟资源共存。因此,vDC 整体组网环境比较复杂。

近期,用户接入 DC 可以充分利用已有光网城市网络或 IP/MPLS 网络(视用户需求及用户类型而定);DC 间互联也可利用已有 IP/MPLS 三层网络或 OTN 等二层网络。但 DC 内的 overlay 组网架构具体实现以及面向 vDC 全网的虚拟机迁移与资源访问等应用需求所涉及的用户接入

与 DC 内 overlay 网络、DC 内 overlay 网络跨 IP 城域网、其他 DC 互联的平滑衔接,尚需等待相关技术进一步发展成熟和相关设备能力进一步完善后,再进行组网测试验证。

现网部署时,需要结合技术成熟度、技术经济性、技术扩展性、现有资源保护性等因素综合考虑,选择对现网影响最小、运营效率最高、风险最低的技术进行 vDC 组网。

参考文献

- 1 STT 协议解析. <http://blog.csdn.net/majieyue/article/details/7941878>,2012
- 2 IETF. A Stateless Transport Tunneling Protocol for Network Virtualization (STT). draft-davie-stt-05,2014
- 3 IETF. VXLAN: a Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks. draft-mahalingam-dutt-dcops-vxlan-08, 2014
- 4 李明. 网络虚拟化技术在云计算数据中心的应用. 电信工程技术与标准化, 2011(4)
- 5 雷鸣, 刘玲. 云计算数据中心组网技术. 电信工程技术与标准化, 2012(4)



表3 VXLAN 与 STT、NVGRE 技术对比

	VXLAN	STT	NVGRE
共性	三者都引入了远大于 VLAN 空间的实例空间;使用通道在网络和子网边界传输数据分组;使用 IP 多播;基本都支持 ECMP;尚未标准化,都处于 IETF 草案阶段;都支持 vSwitch,即隧道的起点和终点一般都在 vSwitch 上,而不像 SPB-M 那样必须在硬件网络设备上		
	理论上,看,VXLAN 和 STT 之间有很多相似之处,如:隧道端点一般由 Hyper-V vSwitch 提供;隧道实例的长度都是 24 bit,可通过控制传输源报头发挥多路径优势等;VXLAN 基于 MAC in UDP,而 STT 基于 MAC in 无状态 TCP,均存在 IP 网络分组丢失的可能,为了保证可靠性,需要 VM 终端启用 TCP 重传机制		
个性(隧道的构造不同)	实例标识 VNI(24 bit)	context ID(64 bit)	TNI(24 bit),使用 GRE key 域的低 24 bit 作为网络租户的标识符
存储目标地址的定位	定义 UDP 隧道的目的 IP 地址及目的 UDP 端口号 4798 进行寻径	目的端口号为 TCP 端口范围:1 024~49 151,且对于隧道两个固定的端点而言,通信应用固定的源及目的端口号	IETF 尚未定义
ECMP 支持能力及实现机制	VXLAN 通过随机分配源 UDP 端口号分散负载:将原始以太网数据头(MAC 地址、IP 地址、4 层端口号等)的散列值作为 UDP 的号,以任意源多播方式传送	与 VXLAN 类似,但采用 TCP 端口号标识:通过将原始以太网数据头(MAC 地址、IP 地址、4 层端口号等)的散列值作为无状态 TCP 的源端口号;未知目的、广播、多播等网络流量均被封装为多播转发	建议使用 GRE 扩展字段关键域中保留的 8 bit——flow ID(流标识),要求物理网络能识别到 GRE 隧道的扩展信息,并以流标识进行流量分担;未知目的、广播、多播等网络流量均被封装为多播转发
封装机制	MAC in UDP,一般在 vSwitch 上进行 MAC in UDP 封装,也可在硬件交换机上实现	MAC in TCP 封装:在 IP 报头内使用无状态 TCP 类报头,允许端系统的隧道端点利用驻留在服务器网卡上的 TOE(TCP 卸载引擎)和 TSO(TCP 分段卸载)功能,把数据帧先进行分割再封装,充分利用网卡的硬件加速功能提升效率	一般在 vSwitch 上进行 MAC in IP 封装
技术成熟度及经济性	VXLAN 利用现有通用的 UDP 传输,成熟性较高,易于实现	STT 对于 TCP 有较大修改,隧道模式接近 UDP 性质,隧道构造技术属于革新技术,且复杂度较高	在一定程度上利用 GRE 扩展字段关键域预留字段,成熟度较低,有一定实现难度
对现网设备的改造难度	改造成本/难度较小;VXLAN 对于现网设备对流量均衡要求较低,一般的网络设备都能对 L2~L4 数据参数进行等价路由的流量均衡	整体改造成本相对较小,包括两方面:同 VXLAN;需要服务器部署支持 TSO 的专用网卡	现网网络设备改造成本较大;NVGRE 需要网络设备对 GRE 扩展头感知并对流标识进行散列,但现网大多数网络设备不会解析硬件的 GRE 报头,从而影响性能;同时,在多路径数据中心 LAN 中解析流量分发的五元组散列表也可能产生问题,因此 NVGRE 部署需要硬件升级
最大不足	端到端网络设备需维护大量多播状态,导致 VXLAN 扩展性受到影响;XLAN 隧道为不可靠隧道	尚无利用 SDN 控制平面进行地址学习的具体方案;对服务器有较高要求;STT 隧道为不可靠隧道	目前 NVGRE 尚无利用控制平面进行地址学习的具体方案;现网设备改造成本较高;GRE 隧道不可靠,但可以辅加以加密协议来提升隧道可靠性
代表厂商	网络设备制造商和芯片厂商为主:思科、VMware、思杰、红帽、Arista Networks 和博通	VMware (Nicira-VMware 旗下另一产品平台)	以软件和芯片厂商为主:微软、惠普、Arista Networks、英特尔、戴尔、博通等
注:VMware 分别在其两个产品平台上提供两种 overlay 组网机制			

[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。

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、高级工程师,主要研究方向为云数据中心组网、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、数据中心组网等。

(收稿日期:2014-03-31)