# 基于多租户的云计算Overlay网络

文/刘新民

云计算已经成为当前企业IT建设的常规形态，而在云计算中大量采用和部署的虚拟化几乎成为一个基本的技术模式。

服务器虚拟化技术的广泛部署，极大增加了数据中心的计算密度，而且，因为虚拟机本身不受物理计算环境的约束，基于业务的灵活性变更要求，需要在网络中无限制地迁移到目的物理位置，（如图1所示）虚机增长的快速性以及虚机迁移成为一个常态性业务。


图1 虚拟化的快速增长及带来的密集迁移效应

## 1 云计算虚拟化网络的挑战与革新

在云中，虚拟计算负载的高密度增长及灵活性迁移在一定程度上对网络产生了压力，然而当前虚拟机的规模与可迁移性受物理网络能力约束，云中的业务负载不能与物理网络脱离。

### 虚拟机迁移范围受到网络架构限制

由于虚拟机迁移的网络属性要求，其从一个物理机上迁移到另一个物理机上，要求虚拟机不间断业务，则需要其IP地址、MAC地址等参数维保持不变，如此则要求业务网络是一个二层网络，且要求网络本身具备多路径多链路的冗余和可靠性。传统的网络生成树

(STPSpaning Tree Protocol)技术不仅部署繁琐，且协议复杂，网络规模不宜过大，限制了虚拟化的网络扩展性。基于各厂家私有的的IRF/vPC等设备级的（网络N:1）虚拟化技术，虽然可以简化拓扑简化、具备高可靠性的能力，但是对于网络有强制的拓扑形状限制，在网络的规模和灵活性上有所欠缺，只适合小规模网络构建，且一般适用于数据中心内部网络。而为了大规模网络扩展的TRILL/SPB/FabricPath/VPLS等技术，虽然解决了上述技术的不足，但对网络有特殊要求，即网络中的设备均要软硬件升级而支持此类新技术，带来部署成本的上升。

**虚拟机规模受网络规格限制**

在大二层网络环境下，数据流均需要通过明确的网络寻址以保证准确到达目的地，因此网络设备的二层地址表项大小（(即MAC地址表)），成为决定了云计算环境下虚拟机的规模的上限，并且因为表项并非百分之百的有效性，使得可用的虚机数量进一步降低，特别是对于低成本的接入设备而言，因其表项一般规格较小，限制了整个云计算数据中心的虚拟机数量，但如果其地址表项设计为与核心或网关设备在同一档次，则会提升网络建设成本。虽然核心或网关设备的MAC与ARP规格会随着虚拟机增长也面临挑战，但对于此层次设备能力而言，大规格是不可避免的业务支撑要求。减小接入设备规格压力的做法可以是分离网关能力，如采用多个网关来分担虚机的终结和承载，但如此也会带来成本的上升。

**网络隔离/分离能力限制**

当前的主流网络隔离技术为VLAN（或VPN），在大规模虚拟化环境部署会有两大限制：一是VLAN数量在标准定义中只有12个比特单位，即可用的数量为4000个左右，这样的数量级对于公有云或大型虚拟化云计算应用而言微不足道，其网络隔离与分离要求轻而易举会突破4000；二是VLAN技术当前为静态配置型技术（只有EVB/VEPA的802.1Qbg技术可以在接入层动态部署VLAN，但也主要是在交换机接主机的端口为常规部署，上行口依然为所有VLAN配置通过)，这样使得整个数据中心的网络几乎为所有VLAN被允许通过(核心设备更是如此)，导致任何一个VLAN的未知目的广播数据会在整网泛滥，无节制消耗网络交换能力与带宽。

对于小规模的云计算虚拟化环境，现有的网络技术如虚拟机接入感知(VEPA/802.1Qbg)、数据中心二层网络扩展(IRF/vPC/TRILL/FabricPath)、数据中心间二层技术(OTV/EVI/TRILL)等可以很好的满足业务需求，上述限制不成为瓶颈。然而，完全依赖于物理网络设备本身的技术改良，目前看来并不能完全解决大规模云计算环境下的问题，一定程度上还需要更大范围的技术革新来消除这些限制，以满足云计算虚拟化的网络能力需求。在此驱动力基础上，逐步演化出Overlay的虚拟化网络技术趋势。

## 2 Overlay虚拟化网络的技术标准及比较

## 2.1 Overlay技术形态

Overlay在网络技术领域，指的是一种网络架构上叠加的虚拟化技术模式，其大体框架是对基础网络不进行大规模修改的条件下，实现应用在网络上的承载，并能与其它网络业务分离，并且以基于IP的基础网络技术为主（如图2所示）。其实这种模式是以对传统技术的优化而形成的。早期的就有标准支持了二层Overlay技术，如RFC3378(Ethernet in IP)，就是早期的在IP上的二层Overlay技术。并且基于Ethernet over GRE的技术，H3C与Cisco都在物理网络基础上发展了各自的私有二层Overlay技术——EVI(Ethernet Virtual Interconnection)与OTV(Overlay Transport Virtualization)。EVI与OTV都主要用于解决数据中心之间的二层互联与业务扩展问题，并且对于承载网络的基本要求是IP可达，部署上简单且扩展方便。

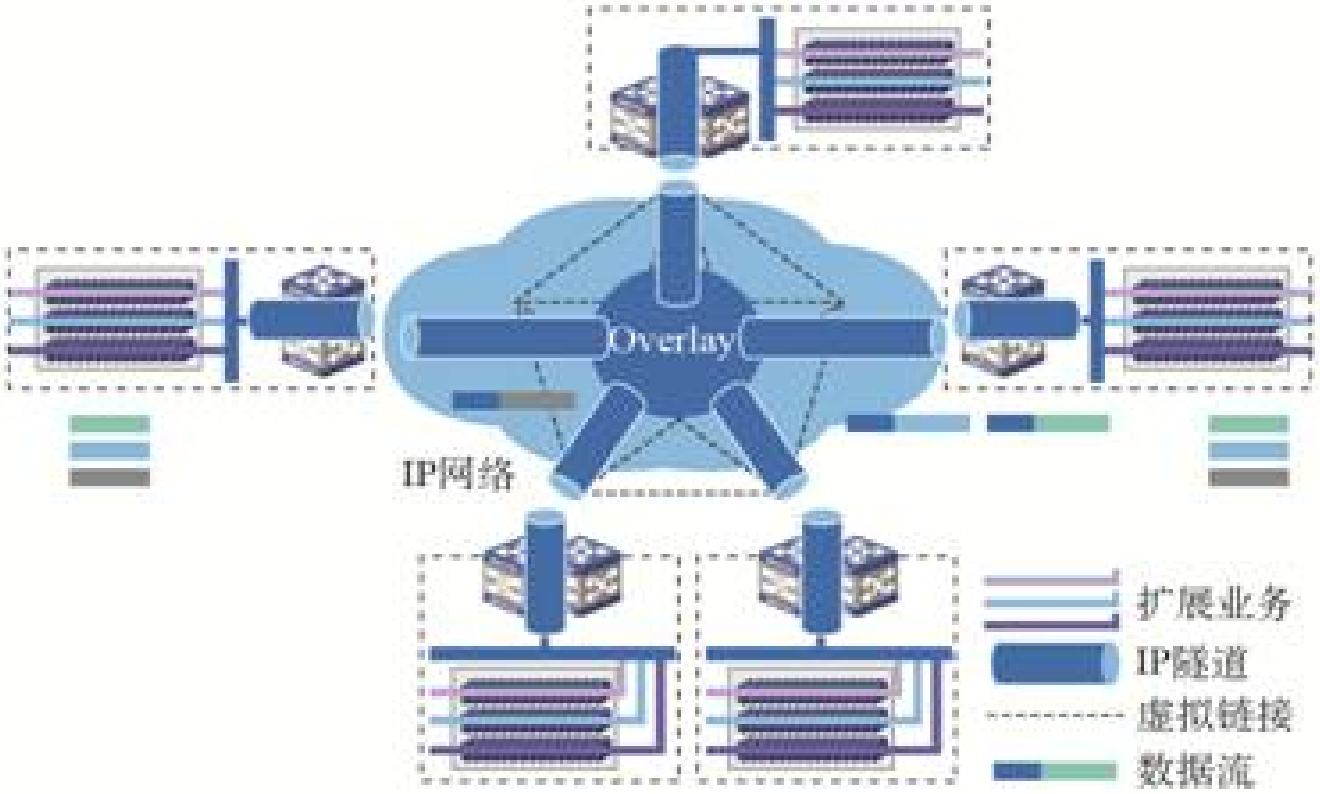



图2 Overlay网络模型

随着云计算虚拟化的驱动，基于主机虚拟化的Overlay技术出现，在服务器的Hypervisor内vSwitch上支持了基于IP的二层Overlay技术，从更靠近应用的边缘来提供网络虚拟化服务，其目的是使虚拟机的部署与业务活动脱离物理网络及其限制，使得云计算的网络形态不断完善。（如图3所示）主机的vSwitch支持基于IP的Overlay之后，虚机的二层访问直接构建在Overlay之上，物理网不再感知虚机的诸多特性，由此，Overlay可以构建在数据中心内，也可以跨越数据中心之间。

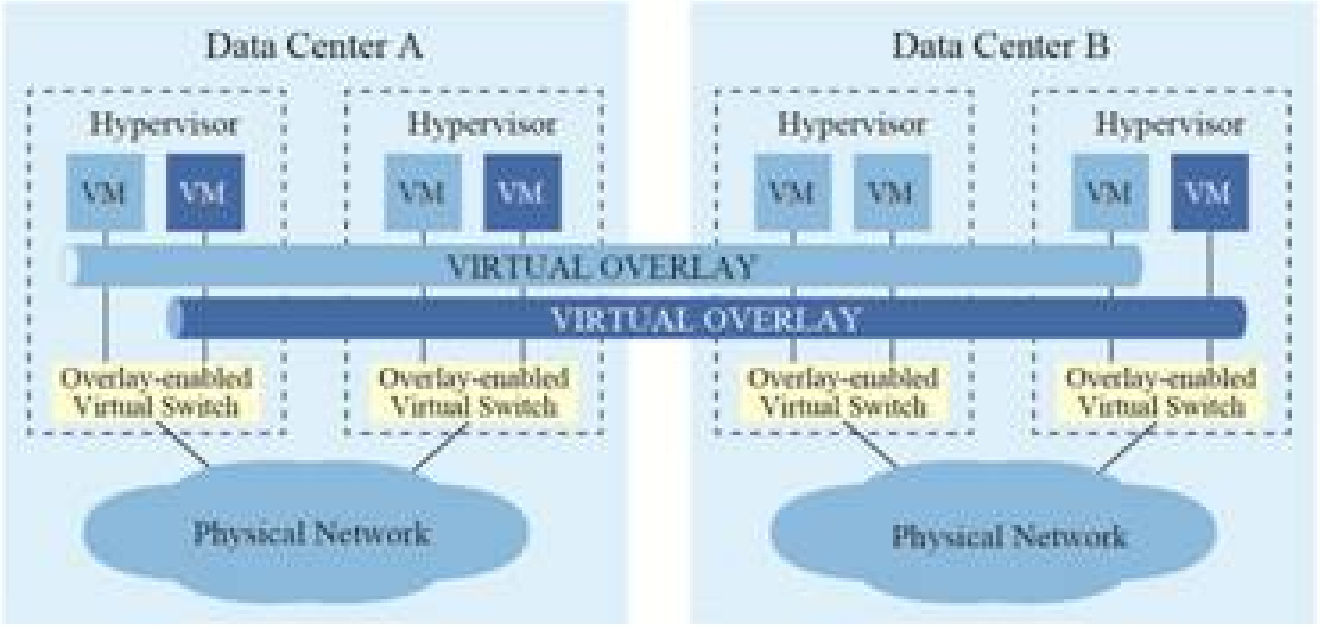



图3 hypervisor支持的二层Overlay

## 2.2 Overlay如何解决当前的主要问题

针对前文提出的三大技术挑战，Overlay在很大程度上提供了全新的解决方式。

**针对虚机迁移范围受到网络架构限制的解决方式**

Overlay是一种封装在IP报文之上的新的数据格式，因此，这种数据可以通过路由的方式在网络中分发，而路由网络本身并无特殊网络结构限制，具备良性大规模扩展能力，并且对设备本身无特殊要求，以高性能路由转发为佳，且路由网络本身具备很强的的故障自愈能力、负载均衡能力。采用Overlay技术后，企业部署的现有网络便可用于支撑新的云计算业务，改造难度极低(除性能可能是考量因素外，技术上对于承载网络并无新的要求)。

**针对虚机规模受网络规格限制的解决方式**

虚拟机数据封装在IP数据包中后，对网络只表现为封装后的的网络参数，即隧道端点的地址，因此，对于承载网络（特别是接入交换机），MAC地址规格需求极大降低，最低规格也就是几十个（每个端口一台物理服务器的隧道端点MAC）。当然，对于核心/网关处的设备表项(MAC/ARP)要求依然极高，当前的解决方案仍然是采用分散方式，通过多个核心/网关设备来分散表项的处理压力。(另一种更分散的方式便是虚拟网络路由服务方式，详见后文描述)。

**针对网络隔离/分离能力限制的解决方式**

针对VLAN数量4000以内的限制，在Overlay技术中引入了类似12比特VLAN ID的用户标识，支持千万级以上的用户标识，并且在Overlay中沿袭了云计算“租户”的概念，称之为Tenant ID(租户标识)，用24或64比特表示。针对VLAN技术下网络的TRUANK ALL(VLAN穿透所有设备)的问题，Overlay对网络的VLAN配置无要求，可以避免网络本身的无效流量带宽浪费，同时Overlay的二层连通基于虚机业务需求创建，在云的环境中全局可控。

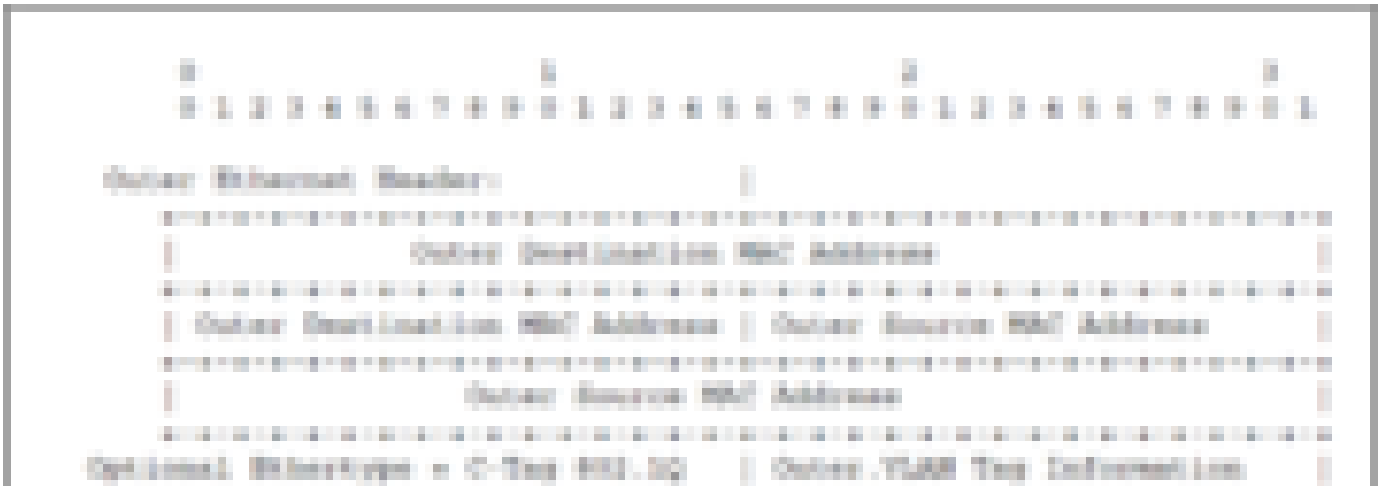
## 2.3 Overlay主要技术标准及比较

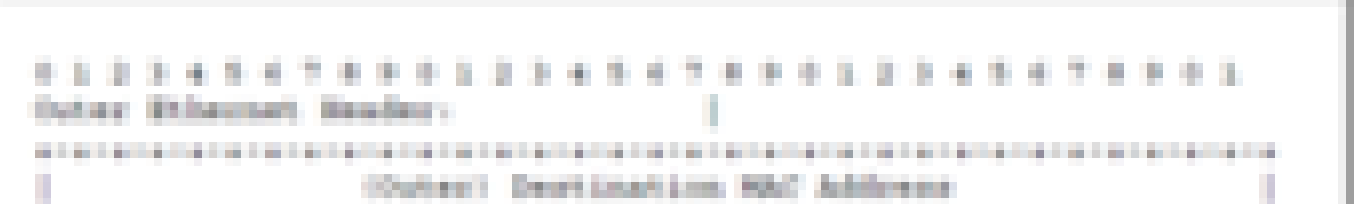
目前，IETF在Overlay技术领域有如下三大技术路线正在讨论，为简单起见，本文只讨论基于IPv4的Overlay相关内容（如图4所示）。

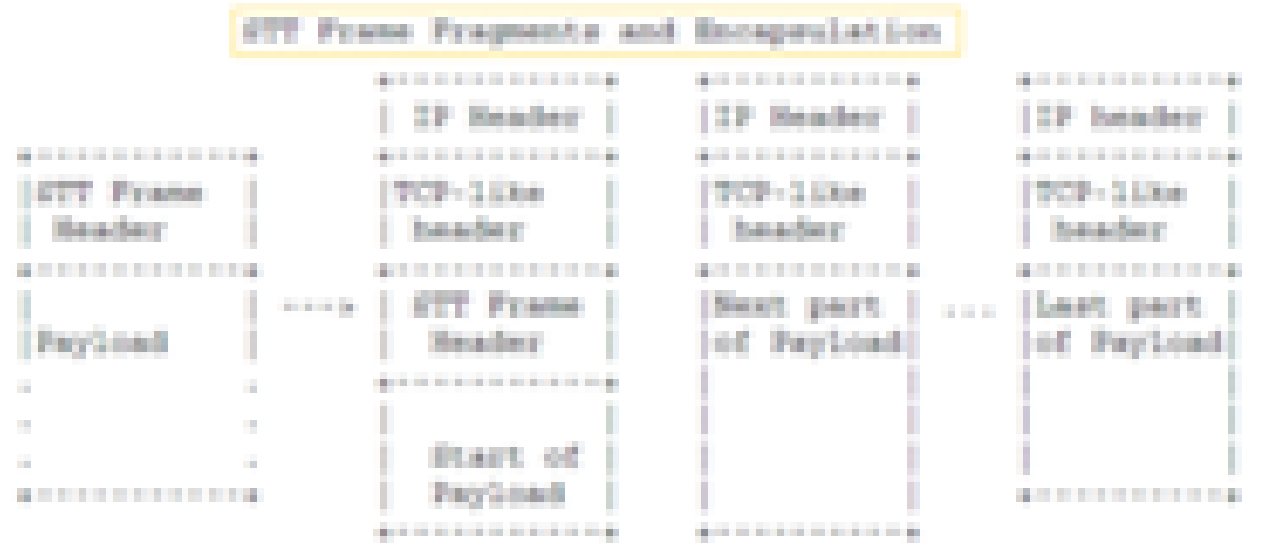
**VXLAN。**VXLAN是将以太网报文封装在UDP传输层上的一种隧道转发模式，目的UDP端口号为4798；为了使VXLAN充分利用承载网络路由的均衡性，VXLAN通过将原始以太网数据头(MAC、IP、四层端口号等)的HASH值作为UDP的号；采用24比特标识二层网络分段，称为VNI(VXLAN Network Identifier)，类似于VLAN ID作用；未知目的、广播、组播等网络流量均被封装为组播转发，物理网络要求支持任意源组播(ASM)。

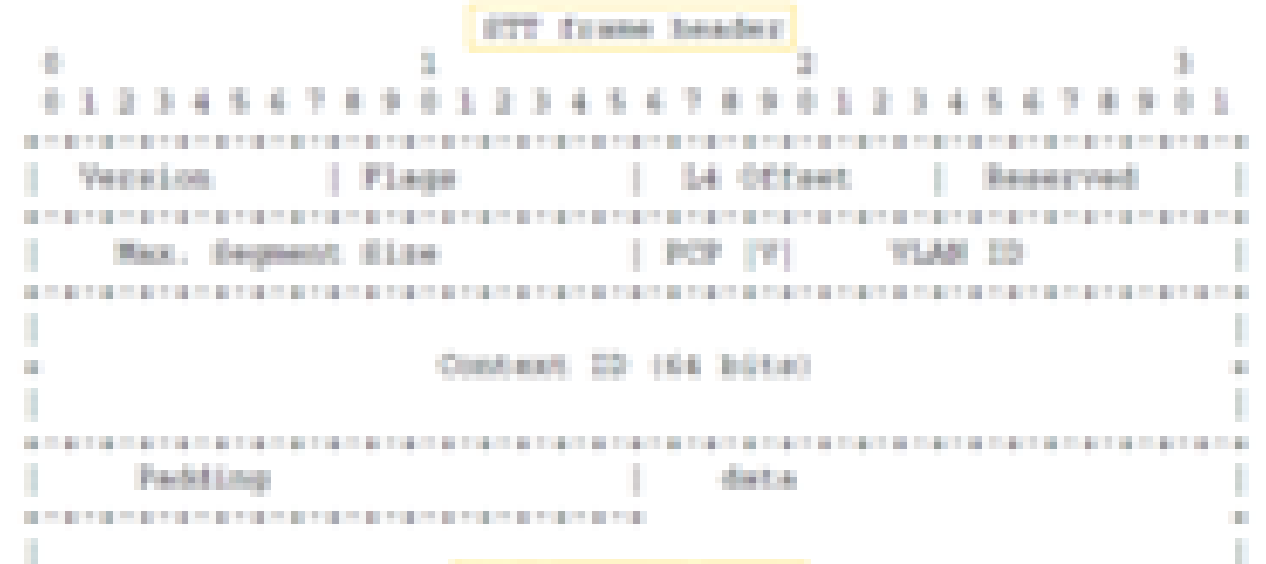
**NVGRE。**NVGRE是将以太网报文封装在GRE内的一种隧道转发模式；采用24比特标识二层网络分段，称为VSI(Virtual Subnet Identifier)，类似于VLAN ID作用；为了使NVGRE利用承载网络路由的均衡性，NVGRE在GRE扩展字段flow ID，这就要求物理网络能够识别到GRE隧道的扩展信息，并以flow ID进行流量分担；未知目的、广播、组播等网络流量均被封装为组播转发。

**STT。**STT利用了TCP的数据封装形式，但改造了TCP的传输机制，数据传输不遵循TCP状态机，而是全新定义的无状态机制，将TCP各字段意义重新定义，无需三次握手建立TCP连接，因此称为无状态TCP；以太网数据封装在无状态TCP；采用64比特Context ID标识二层网络分段；为了使STT充分利用承载网络路由的均衡性，通过将原始以太网数据头(MAC、IP、四层端口号等)的HASH值作为无状态TCP的源端口号；未知目的、广播、组播等网络流量均被封装为组播转发。






GTP Frame Fragments and Encapsulation
IP Header
TCP-like Header
GTP Frame Header
Start of Payload
GTP Frame Header
Payload
IP Header
TCP-like Header
Next part of Payload
IP header
TCP-like Header
Last part of Payload
Original data frame is encapped with GTP Header
GTP Frame is segmented and transmitted as a set of TCP segments (MAC headers not shown)


GTP frame header
0 1 2 3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
Version
Flags
L4 Offset
Reserved
Max. Segment Size
PCP (Y)
VLAN ID
Context ID (64 bits)
Padding
Data


GTP Segment Format
0 1 2 3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
Source Port
Destination Port
Sequence Number(*)
Acknowledgment Number(*)
Data Offset
Reserved
U A P R S F
R C S S Y I
G K H T N N
Window
Checksum
Urgent Pointer

VXLAN NVGRE SST

图4 三种数据详细封装

这三种二层Overlay技术，大体思路均是将以太网报文承载到某种隧道层面，差异性在于选择和构造隧道的不同，而底层均是IP转发。如表1所示为这三种技术关键特性的比较：VXLAN和STT对于现网设备对流量均衡要求较低，即负载链路负载分担适应性好，一般的网络设备都能对L2-L4的数据内容参数进行链路聚合或等价路由的流量均衡，而NVGRE则需要网络设备对GRE扩展头感知并对flow ID进行HASH，需要硬件升级；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性，且复杂度较高，而VXLAN利用了现有通用的UDP传输，成熟性极高。总体比较，VLXAN技术相对具有优势。

| 技术名称 | 支持者 | 支持方式<br>简述 | 网络虚拟化<br>方式 | 数据新增<br>报头长度 | 链路 HASH<br>能力 | 数据封装 |
|---|---|---|---|---|---|---|
| VXLAN | Cisco/VMWare/<br>Ctrix/<br>Red Hat<br>/Broadcom | L2 over UDP | VXLAN 报头<br>24 bit VNI | 50Byte(+ 原数据) | 现有网络可进行<br>L2-L4 HASH | |
| NVGRE | HP/ 微软 /<br>Broadcom/<br>Dell/Emulex/Intel | L2 over GRE | NVGRE 报头<br>24 bit VSI | 42Byte(+ 原数据) | GRE 头的 HASH<br>需要网络升级 | |
| STT | VMWare<br>(Nicira) | L2oTCP( 无状态<br>TCP，即 L2 在类<br>似TCP的传输层) | STT 报头<br>64 bit Context ID | 58~76Byte(+ 原数<br>据) | 现有网络可进行<br>L2-L4 HASH | |

表1 IETF三种Overlay技术的总体比较

IETF讨论的Overlay技术，主要聚焦在数据转发层面的实现上，控制层面并无涉及，因此在基本实现上依赖于不同厂家的控制层面设计，IETF讨论稿《draft-pfaff-ovsdb-proto-02.pdf》则针对Open vSwitch提供了一种控制管理模型的建议（如图5所示），但在细节实现上仍不是很明确。

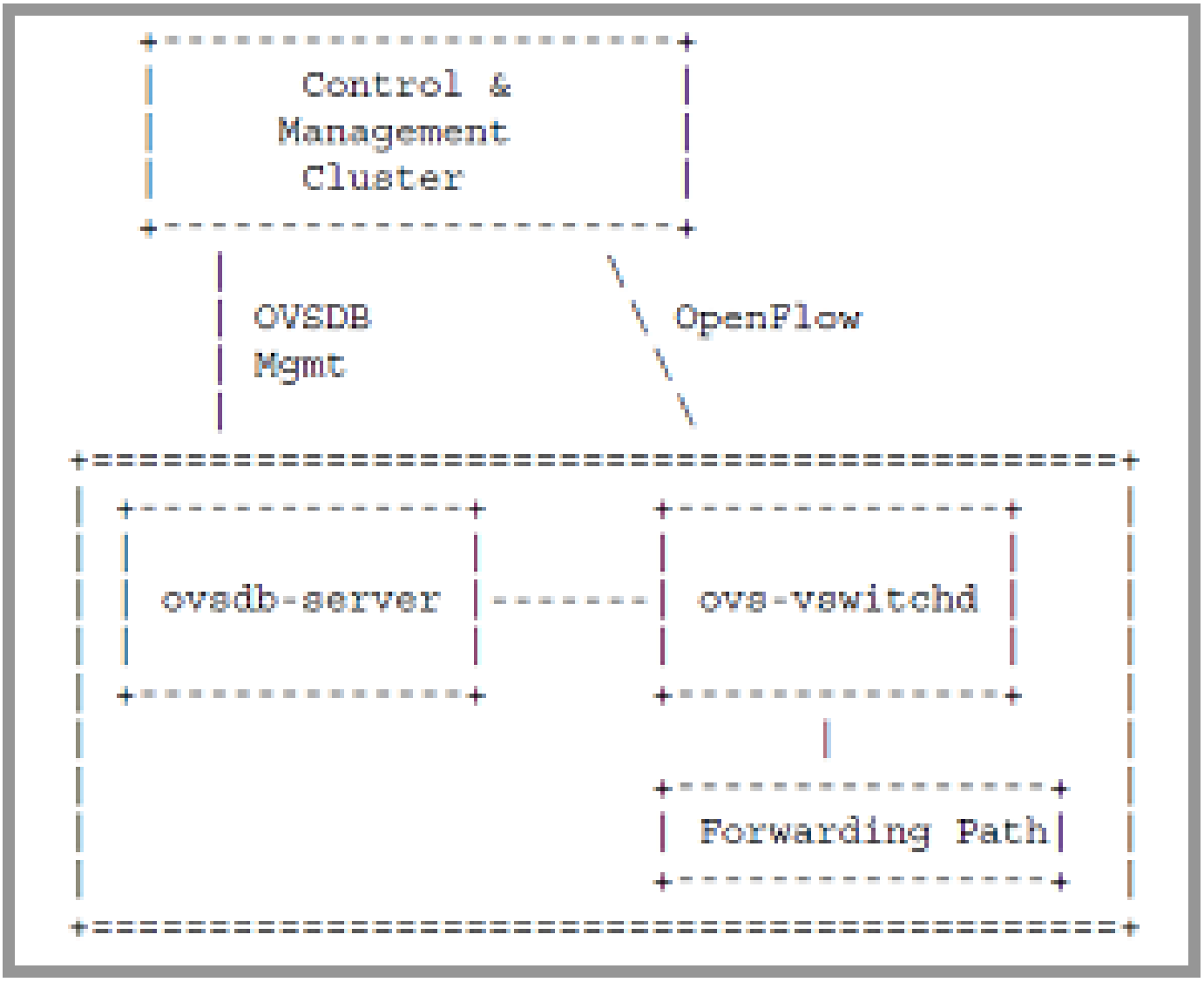



图5 IETF draft讨论的OVS管理方式

## 3 多租户的Overlay网络架构

### 3.1 数据中心虚拟化网络的发展阶段

随着虚拟化技术在数据中心、云计算中的不断深入应用，伴随着网络技术的发展，数据中心的二层组网结构出现了阶段性的架构变化（如图6所示）。

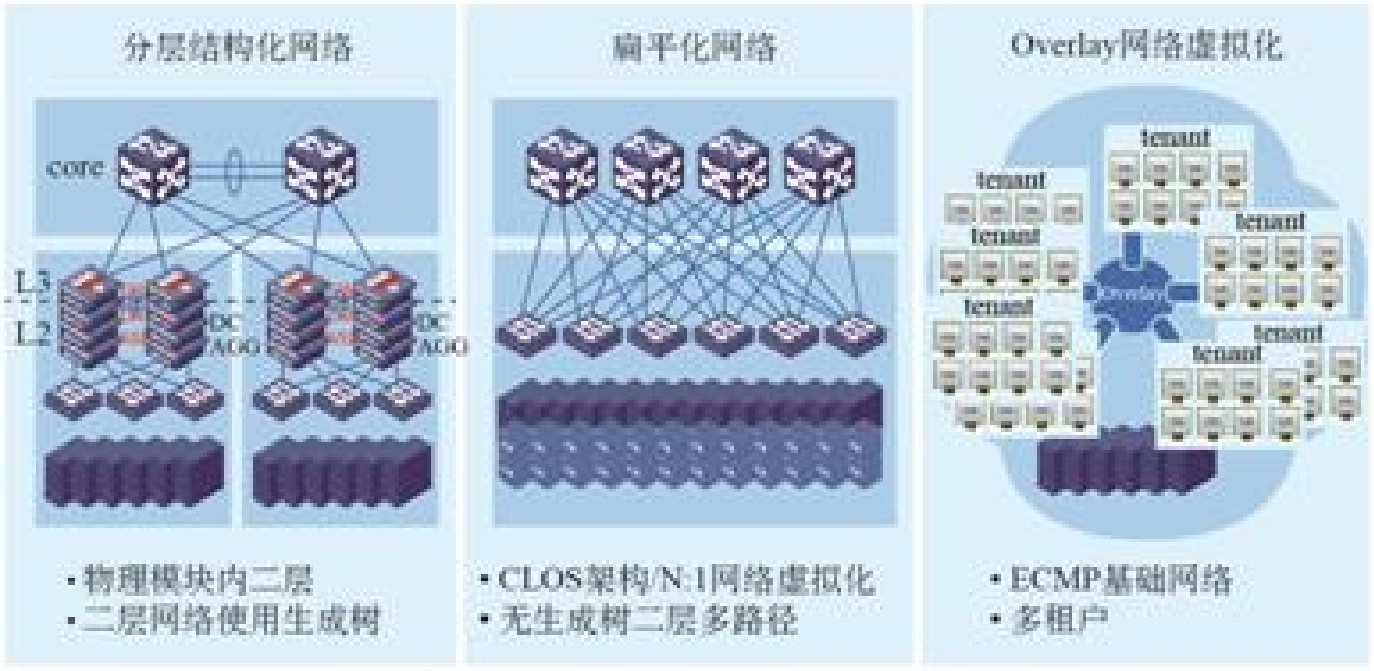



图6 阶段性网络与虚拟化的匹配

**分层结构化网络**

早期的数据中心网络，虚拟化需求非常少，并没有强烈的大二层技术要求，多是面向一定的业务应用系统构建网络模块，并且规模一般不大，性能要求也不高。数据中心使用多层架构，网关层面比较低，业务的二层访问基本可以在网络模块内解决，只需要通过基础的生成树技术来支撑模块内的二层网路可靠性运行即可。

**扁平化网络**

随着虚拟化在X86架构服务器上的流行及广泛部署，模块化的数据中心网络结构已经不能满足虚拟机大范围迁移要求，而生成树协议的复杂性也严重影响大规模网络的稳定运行。由此网络本身技术出现适应虚拟化的变革，包含TRILL/FabricPath/VSS/vPC/IRF等新的技术出现并大量部署，同时为了使得网络进一步感知虚让你因为机的业务生命周期，IEEE还制订了802.1Qbg（即VEPA技术）与802.1BR来配合二层网络技术增强对虚拟机的感知能力。为了保证网络的高性能业务要求，出现了应对高密虚拟化云计算环境的CLOS网络架构。

**Overlay网络虚拟化**

当进入云计算时代，云的业务需求与网络之间出现了前文提到的挑战，网络技术再次发生变革，以Overlay的虚拟化方式来支撑云与虚拟化的建设要求，并实现大规模的多租户能力，网络进入Overlay虚拟化架构阶段。

## 3.2 Overlay网络的组成模式

Overlay的本质是L2 Over IP的隧道技术，在服务器的vSwitch、物理网络上技术框架已经就绪，并且从当前的技术选择来看，虽然有多种隧道同时实现，但是以L2 over UDP模式实现的VXLAN技术具备较大优势，并且在ESXi和Open vSwitch、当前网络的主流芯片已经实现，预计会成为主流的Overlay技术选择，因此后文的Overlay网络均参考VXLAN相关的技术组成描述，其它NVGRE、STT等均类似。

Overlay网络架构有多种实现，就纯大二层的实现，可分为主机实现方式和网络实现方式；而在最终实现Overlay与网络外部数据连通的连接方式上，则更有多种实现模式，并且对于关键网络部件将有不同的技术要求。

### 3.2.1 基于主机的Overlay虚拟化网络

（如图7所示）目前的虚拟化主机软件在vSwitch内支持VXLAN，使用VTEP(VXLAN Tunnel End Point)封装和终结VXLAN的隧道。

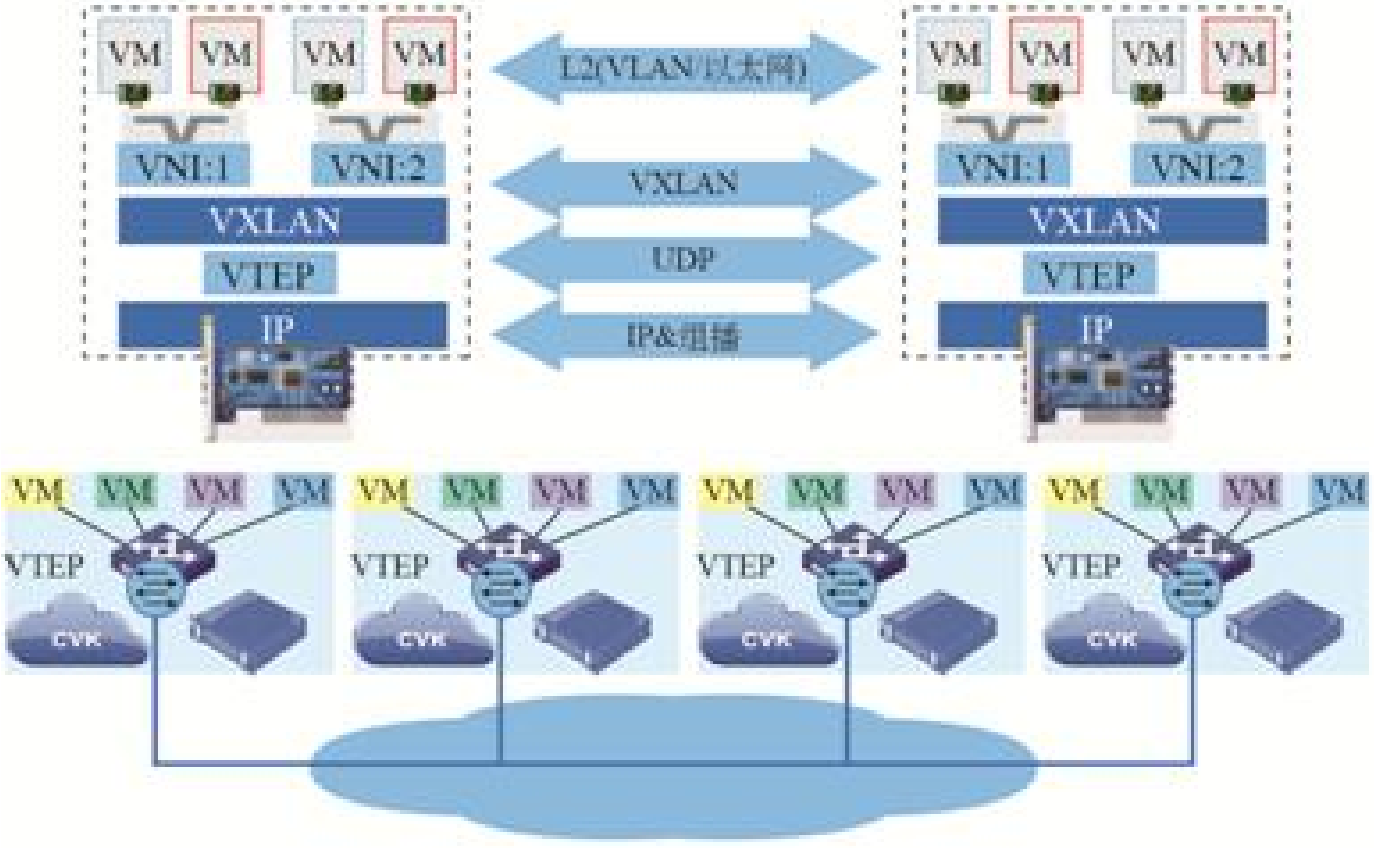


图7 基于主机的Overlay虚拟化网络

VXLAN运行在UDP上，物理网络只要支持IP转发，则所有IP可达的主机即可构建一个大范围二层网络。这种vSwitch的实现，屏蔽了物理网络的模型与拓扑差异，将物理网络的技术实现与计算虚拟化的关键要求分离开来，几乎可以支持以太网在任意网络上的透传，使得云的计算资源调度范围空前扩大。

特别的，为了使得VXLAN Overlay网络更加简化运行管理，便于云的服务提供，各厂家使用集中控制的模型，将分散在多个物理服务器上的vSwitch构成一个大型的、虚拟化的分布式Overlay vSwitch（如图8所示），只要在分布式vSwitch范围内，虚拟机在不同物理服务器上的迁移，便被视为在一个虚拟的设备上迁移，如此大大降低了云中资源的调度难度和复杂度。

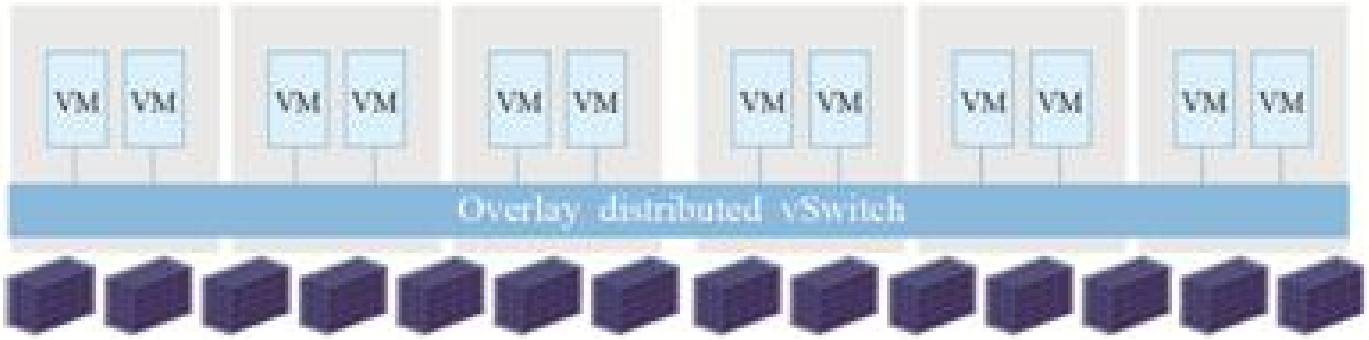



图8 分布式Overlay vSwitch

基于主机的Overlay网络数据流量出入物理网络，需要实现VXLAN的Overlay流量与传统的以太网数据流量之间的封装与解封装过程，而执行这个过程操作的功能点，被称为Overlay/VXLAN Gateway（如图9所示）。因为VXLAN网络的VTEP功能点本身就是VXLAN的封装与解封装隧道点，因此VXLAN Gateway首先需要具备VTEP功能，形态可以是vSwitch、物理交换机等，只是对于网络中的虚机或其它设备地址表项的处理有所差异。

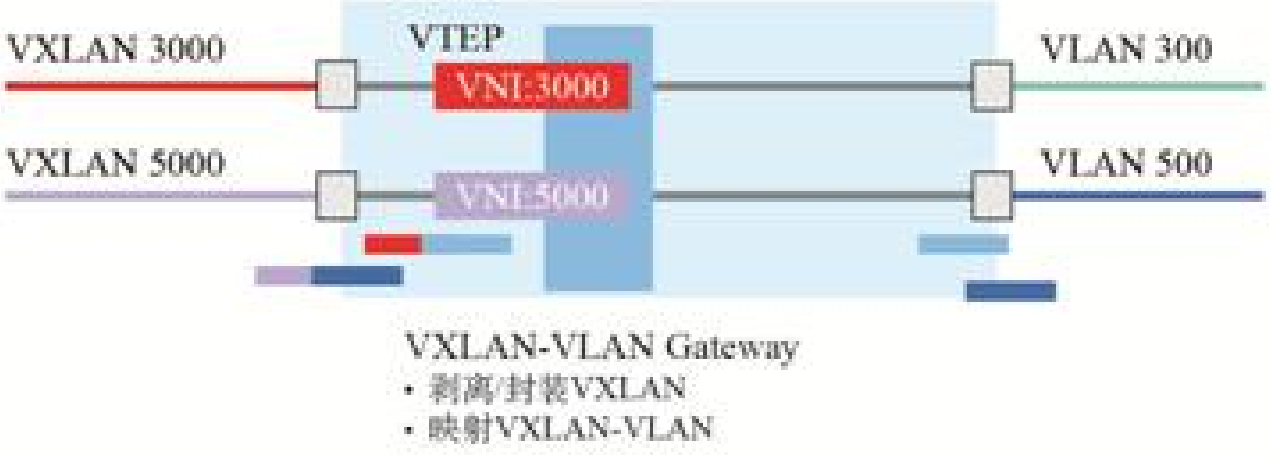



图9 Overlay Gateway（VXLAN）

VXLAN Ovelay网络与物理网络连通有以下三种组网方案（如图10所示）。

**方案一：Overlay虚拟化网络+vSwitch GW+vRouter**

Overlay的vSwitch本身具备隧道的封装与解封装能力，因此，H3C提供一种虚拟路由器来配合这种基本方式。将在硬件路由器中运行的软件vSR（基于H3C Comware平台的路由软件包)作为虚拟机运行在主机中，提供Overlay网络的虚拟路由功能（即vRouter能力），vRouter的接口同时连接到VXLAN的网络和VLAN基本功能的物理网络。从vSwitch接收到的VXLAN数据包被解除封装后，进入vRouter路由接口，可以被路由到外部网络；反之，vRouter接收到外部网络的数据可以进入VXLAN网络。

该方案的好处是：涉及Overlay的功能均在主机虚拟化环境vSwitch实现，并且虚拟路由功能使得Overlay网络部署更加灵活，极大降低外部物理网络要求。

**方案二：Overlay虚拟化网络+vSwitch GW+pRouter**

本方案使用服务器中的vSwitch专门用作VXLAN的Gateway功能，而数据的路由功能，则由外部网络物理路由器承担。该方案除了不具备虚拟路由能力，Overlay的功能也都在主机虚拟化环境vSwitch实现，同时可以支持虚机与非虚拟化的物理服务器之间的二层数据通信要求。

**方案三：Overlay虚拟化网络+pSwitch GW+pRouter**

当物理交换机支持VXLAN功能，则pSwitch与vSwicth可以实现Overlay的统一虚拟化组网，物理交换机执行VXLAN Gateway功能，不仅可以实现Overlay网络在物理网络上的终结，也可以支持虚机与非虚拟化服务器的混合组网业务，因为基于物理实现（物理交换机+物理路由器），整体网络可以达到极高性能。

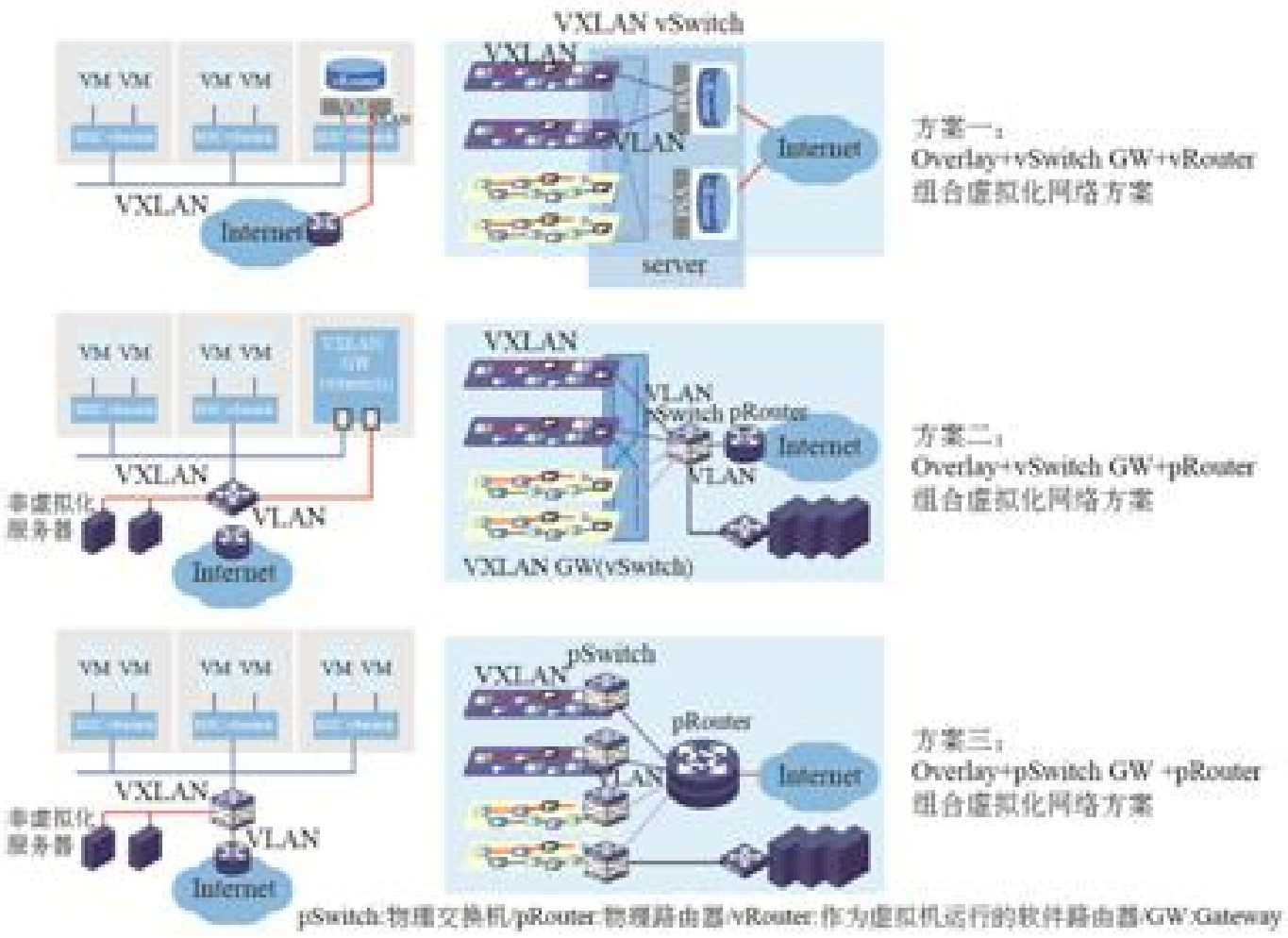



图10 VXLAN Overlay网络与物理网络的连通方案

### 3.2.2 基于物理网络的Overlay虚拟化

（如图11所示）该方案在网络架构上与TRILL/FabricPath等技术类似，但是因为对于非VTEP要求的网络只需要IP转发，它比TRILL/FabricPath构建的成本更低，技术要求也更加简单，同时也容易构建多个数据中心之间的网络连接。

为了解决网络对虚拟机的感知与自动化控制，结合IEEE的802.1Qbg/VEPA技术，可以使得网络的Overlay与计算虚拟化之间产生关联，这样既可以保持服务器内部网络的简化，使用基本的VEPA，利用外部网络强化来保证高能力的控制要求，又在物理网络Overlay的虚拟化基础上增强了虚机在云中大范围调度的灵活性。

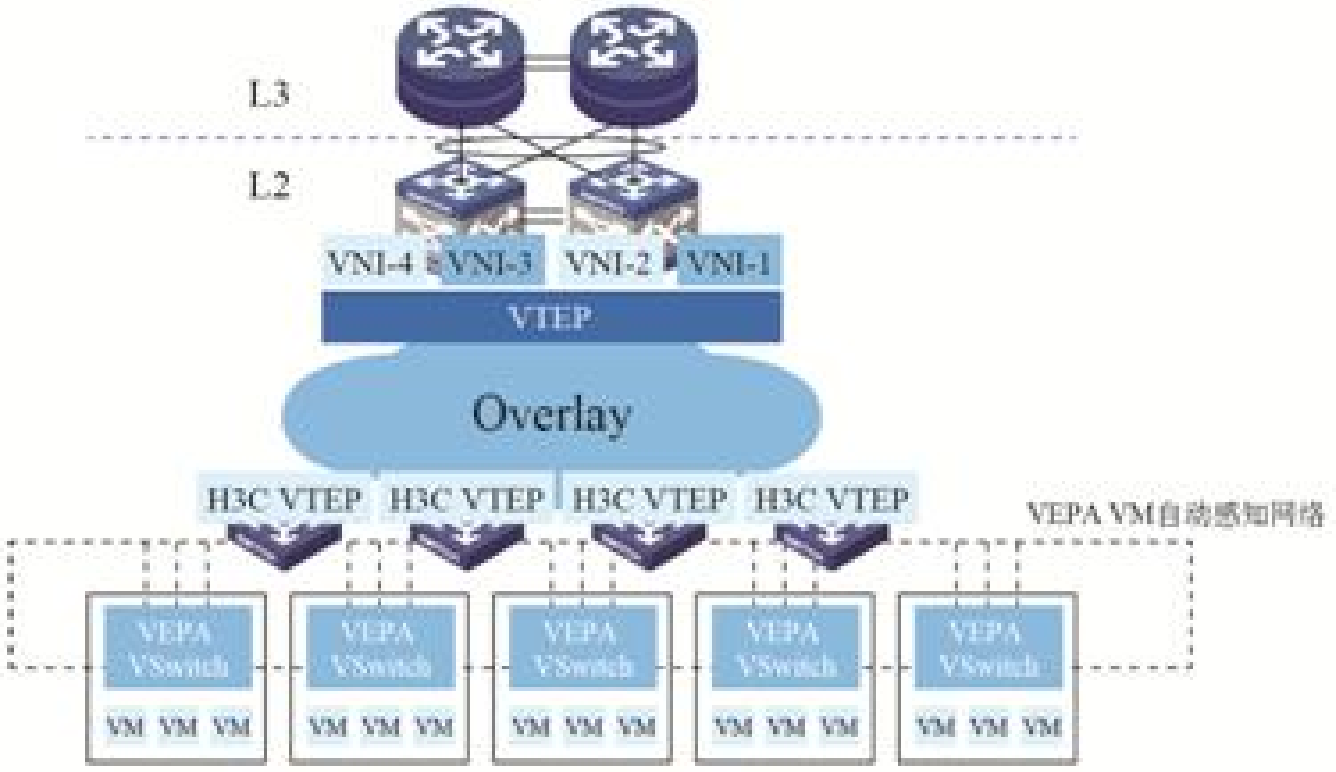



图11 物理网络的Overlay+VEPA

## 3.3 基于Overlay网络的多租户与网络服务——H3Cloud云网融合路线

对于计算资源丰富的数据中心，Overlay网络使得虚拟机不再为物理网络所限制，但是对于网络的L4-L7深度服务，在云计算环境下需求更为强烈，资源动态调度的计算环境，需要动态可调度的网络服务支撑。因此将传统的L4-L7服务转换为可云化的、可动态调度的服务资源，成为Overlay网络环境下的必须集成框架（如图12所示）。H3C基于Comware V7软件平台的L4-L7产品系列，可在虚拟化网络环境下资源化，以虚拟服务单元运行，分别提供对应路由器、防火墙、负载均衡、深度防御的vSR/vRouter、vFW、vLB、vIPS，利用数据中心计算资源与Overlay网络集成，提供等同于传统的L4-L7网络服务能力。

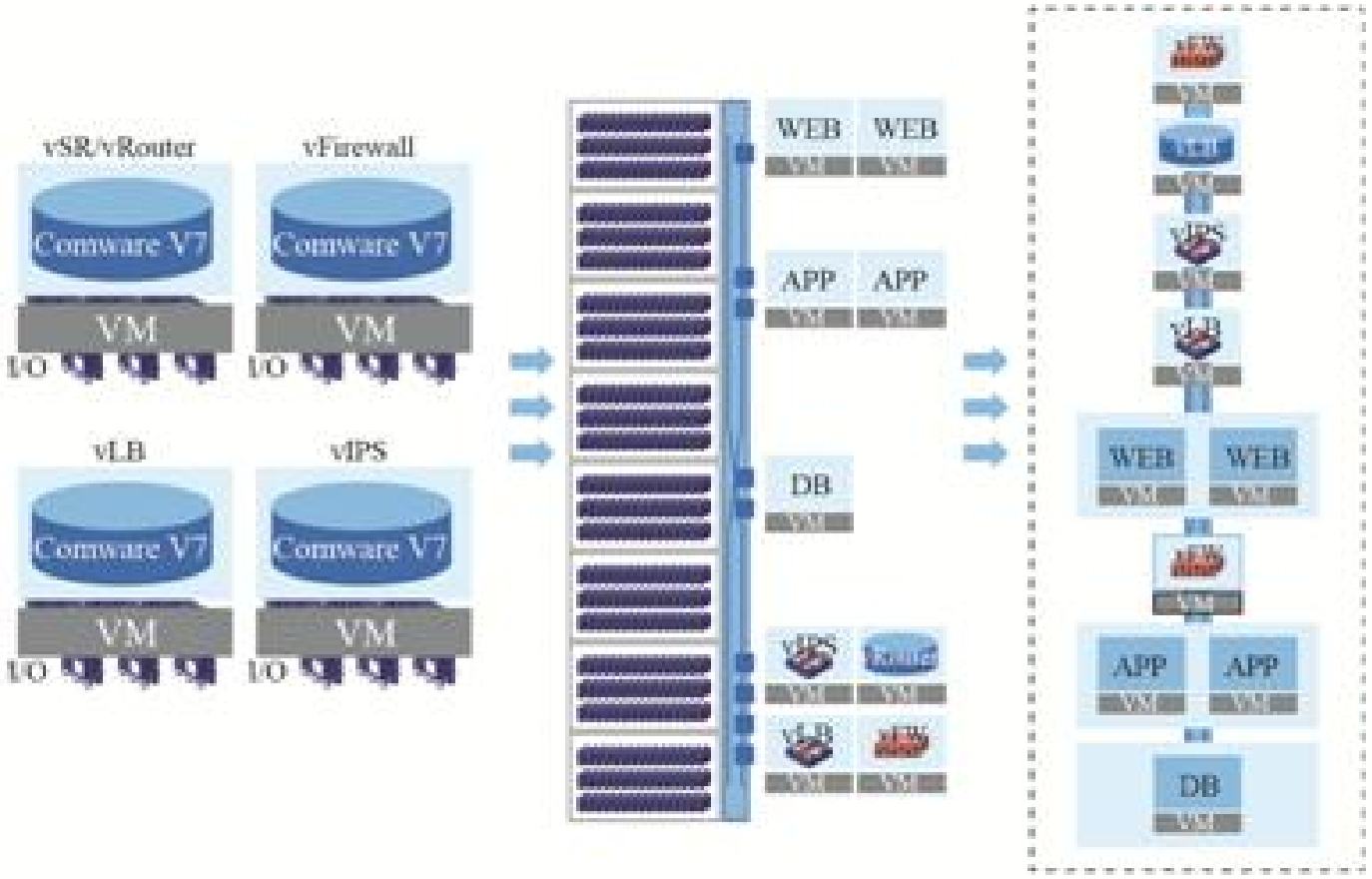



图12 Overlay网络集成服务虚拟化

在Overlay环境下的虚拟化网络服务，必须具备灵活的可分配性、可扩展性及可调度性，因此，自动化的编排组织能力显得非常重要（如图13所示）。除了将虚拟服务资源化，还要具备业务逻辑的组合关联，这些与物理网络L4-L7服务的固定配置管理不同，所有的网络服务资源也是在计算池中，要实现相应的业务关联和逻辑，难以通过物理网络实现，在Overlay网络的连通与编排下则相对易于实现。这种集成思路也将是H3C服务虚拟化的主要模式。

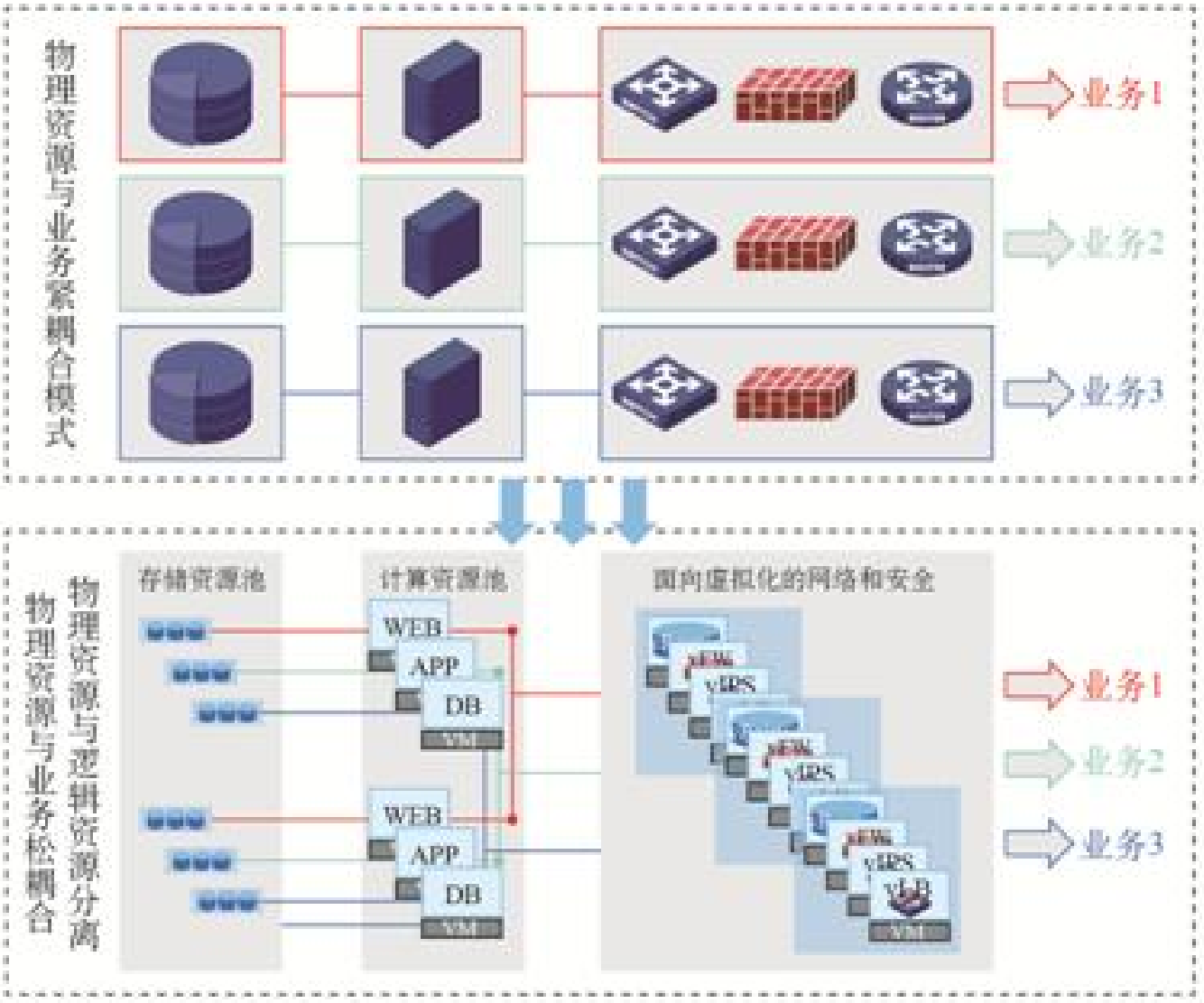



图13 基于Overlay虚拟化网络的服务编排

H3C云计算解决方案核心的架构是云网融合。在当前的特色方案系列均充分利用了云计算与物理网络融合、结合的特点，如：

VEPA——提供网络计算自动化的感知关联与自动化部署；

FCoE——统一交换的计算、存储网络；

DRX——基于H3C LB与虚拟机管理平台关联的虚机资源动态扩展；

PoC（Point of Cloud）——借助云端网络连通云计算中心与路由器集成X86虚拟化计算部件的统一云分支扩展；

分级云——通过层级化网络构建纵向层次化云架构。

针对云计算即将进入Overlay阶段，H3C的技术路线目标是在H3Cloud架构中进一步实现Overlay虚拟化网络的融合（如图14所示）。新一阶段的云网融合，不仅包含分布式的Overlay、多租户的虚拟化网络，还将不断集成逐步产品化的虚拟网络服务部件(目前已经可提供vSR/vFW的服务部件)，而对于原有的VEPA/FCoE/DRX/PoC/分级云等方案，将在基于Overlay的虚拟化网络架构下重新构建和集成，实现技术演进和延续的完整性。

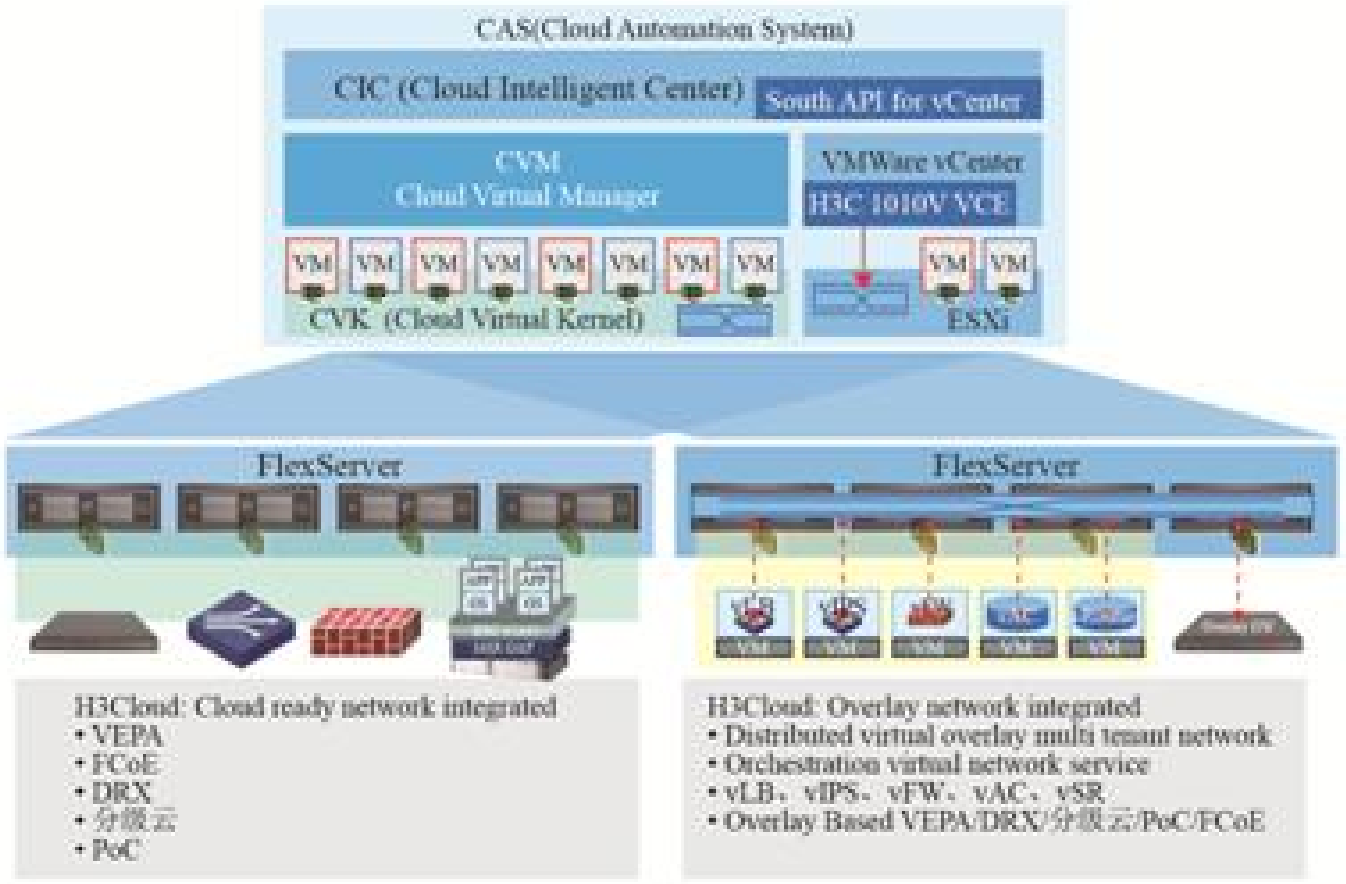



图14 云网融合：H3Cloud从物理网络到Overlay虚拟网络的集成

## 4 结束语

Overlay的网络架构是物理网络向云和虚拟化的深度延伸，云的资源化能力可以脱离物理网络的多种限制，但两个网络本身却是需要连通交互才能实现云的服务能力，随着技术的发展，主机的Overlay技术也将向硬件化发展，并逐步会成为物理网络的一部分。但Overlay尚没有深度成熟，还需要较长时间发展和应用，其中的问题会逐步暴露和解决。

感谢您对本刊物的关注，如果您在阅读时有何感想，请点击 我要评论 (/aspx/voteforms/frm50.aspx?doctitle=%u57FA%u4E8E%u591A%u79DF%u6237%u7684%u4E91%u8BA1%u7B97overlay%u7F51%u7EDC&magaz 反馈。