HCRSE115-SDN VXLAN 原理

SDN （Soft ware Defined Network ） 软件定义网络
VXLAN (Virtual eXtensible Local Area Network) 虚拟可扩展局域网
POD（Point of Delivery ）交付单元
TRILL （Transparent Interconnection of lots of links ）多链路透明互联
IT （ Information Technology ） 信息技术
CT （Communication Technology ） 通信技术
DCN (Data Center Network ） 数据中心网络
IDC (Internet Data Center) 互联网数据中心

VTEP（VXLAN Tunnel EndPoint）VXLAN 隧道端点
NVE（Network Virtualization Edge）网络虚拟边缘节点
VNI（VXLAN Network Identifier）VXLAN 网络标识

SDN 软件定义网络

SDN（software defined network），即软件定义网络，其核心技术是通过将网络设备控制面和数据面分离开来，从而实现了网络流量的灵活控制，为核心网络及应用的创新提供了良好的平台。

SDN 是一种新型网络创新架构，是网络虚拟化的一种实现方式，其核心技术 OpenFlow 通过将网络设备控制面与数据面分离开来，从而实现了网络流量的灵活控制，使网络作为管道变得更加智能。

传统 IT 架构中的网络，根据业务需求部署上线以后，如果业务需求发生变动，重新修改相应网络设备（路由器、交换机、

防火墙）上的配置是一件非常繁琐的事情。在互联网/移动互联网瞬息万变的业务环境下，网络的高稳定与高性能还不足以满足业务需求，灵活性和敏捷性反而更为关键。SDN 所做的事是将网络设备上的控制权分离出来，由集中的控制器管理，无须依赖底层网络设备（路由器、交换机、防火墙），屏蔽了来自底层网络设备的差异。而控制权是完全开放的，用户可以自定义任何想实现的网络路由和传输规则策略，从而更加灵活和智能。

进行 SDN 改造后，无需对网络中每个节点的路由器反复进行配置，网络中的设备本身就是自动化连通的。只需要在使用时定义好简单的网络规则即可。如果你不喜欢路由器自身内置的协议，可以通过编程的方式对其进行修改，以实现更好的数据交换性能。

假如网络中有 SIP、FTP、流媒体几种业务，网络的总带宽是一定的，那么如果某个时刻流媒体业务需要更多的带宽和流量，在传统网络中很难处理，在 SDN 改造后的网络中这很容易实现，SDN 可以将流量整形、规整，临时让流媒体的“管道”更粗一些，让流媒体的带宽更大些，甚至关闭 SIP 和 FTP 的“管道”，待流媒体需求减少时再恢复原先的带宽占比。

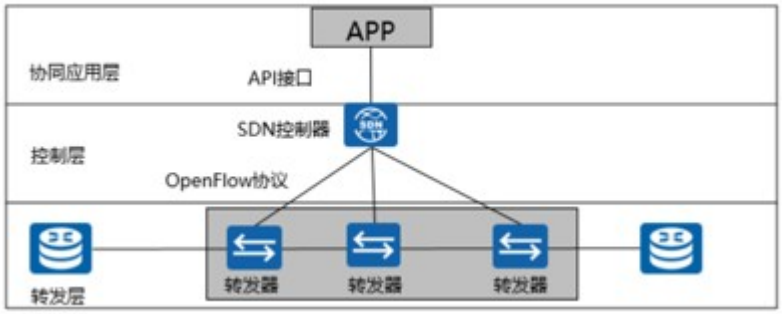
SIP（Session Initiation Protocol，会话初始协议）是由 IETF（Internet Engineering Task Force，因特网工程任务组）制定的多媒体通信协议。它是一个基于文本的应用层控制协议，用于创建、修改和释放一个或多个参与者的会话。例如 Internet 电话

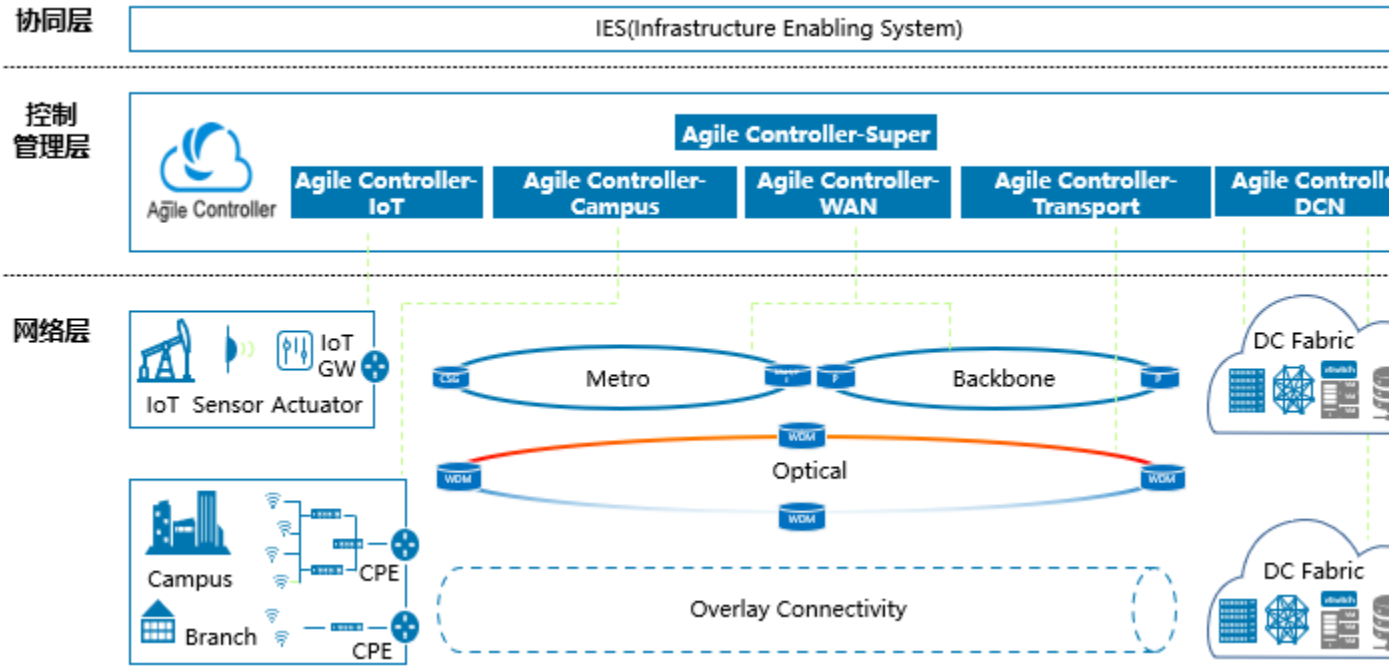
正是因为这种业务逻辑的开放性，使得网络作为“管道”的发展空间变为无限可能。如果未来云计算的业务应用模型可以简化

为“云—管—端”，那么 SDN 就是 “管”这一环的重要技术支撑。

SDN 的三个主要特征：
转控分离：网元的控制平面在控制器上，负责协议计算，产生流表；而转发平面只在网络设备上。
集中控制：设备网元通过控制器集中管理和下发流表，这样就不需要对设备进行逐一操作，只需要对控制器进行配置即可。
开放接口：第三方应用只需要通过控制器提供的开放接口，通过编程方式定义一个新的网络功能，然后在控制器上运行即可。

华为 SDN 解决方案全景图
SDN 架构分为三层：协同层、控制管理层和网络层。

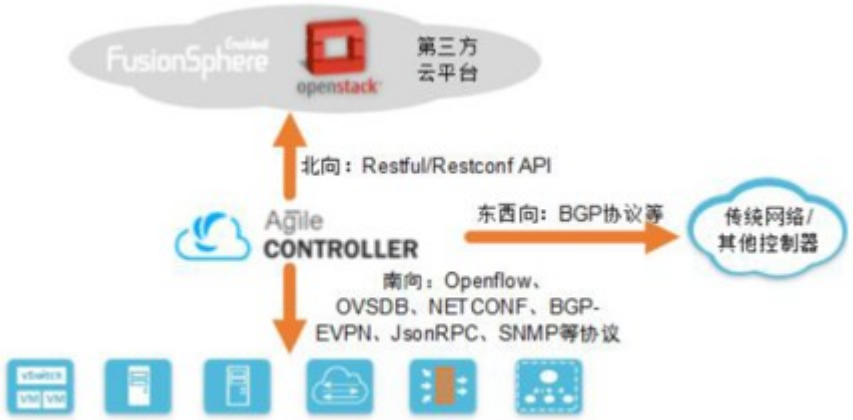




接口开放

Agile Controller-DCN 控制器以开放的软件平台为基础，采用组件松耦合架构，可对外提供丰富的北向 API 承接网络业务、南向接口控制网络设备或计算资源，以及东西向兼容互通传统网络的能力。

1 Agile Controller-DCN 北向通过标准 Restful 接口实现与开源 OpenStack 云平台、华为 FusionSphere 云平台、HP/Redhat/Mirantis/移动大云等第三方云平台和应用的无缝对接。
2 Agile Controller-DCN 南向通过标准 OPENFLOW、OVSDB、NETCONF、BGP-EVPN、JsonRPC、SNMP 等协议管理物理和虚拟网络设备。
3 Agile Controller-DCN 东西向通过跨路由协议、BGP 协议等与传统网络或者其他控制器控制面通
信，以达到互通的目的。



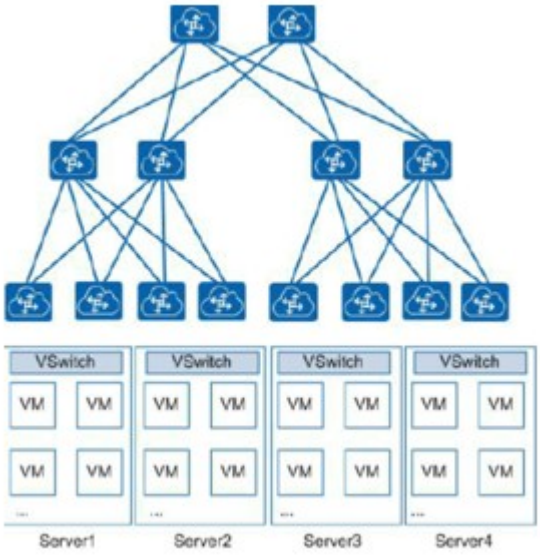
VXLAN (Virtual eXtensible Local Area Network) 虚拟可扩展局域网

VXLAN 采用 NVO3 技术，通过 MAC in UDP 的报文封装方式，实现基于 IP overlay 的虚拟局域网。

NVO3(Network Virtualization Over Layer 3)，基于三层 IP overlay 网络构建虚拟网络技术统称为 NVO3

任何技术的产生，都有其特定的时代背景与实际需求，VXLAN 正是为了解决云计算时代虚拟化中的一系列问题而产生的一项技术。

云计算，凭借其在系统利用率高、人力/管理成本低、灵活性/可扩展性强等方面表现出的优势，已经成为目前企业 IT 建设的新形态；而在云计算中，大量的采用和部署虚拟化是一个基本的技术模式。
服务器虚拟化技术的广泛部署，极大地增加了数据中心的计算密度；同时，为了实现业务的灵活变更，虚拟机 VM（Virtual Machine）需要能够在网络中不受限迁移。实际上，对于数据中心而言，虚拟机迁移已经成为了一个常态性业务。



一台服务器可虚拟多台虚拟机，而一台虚拟机相当于一台主机。主机的数量发生了数量级的变化.

1.虚拟网络问题

（1）虚拟机规模受网络规格限制:

在传统二层网络环境下，数据报文是通过查询 MAC 地址表进行二层转发，而 MAC 地址表的容量限制了虚拟机的数量。

（2）网络隔离能力限制:

当前主流的网络隔离技术是 VLAN，在大规模的虚拟化网络中部署存在如下限制：

1 由于 IEEE 802.1Q 中定义的 VLAN Tag 域只有 12 比特，仅能表示 4096 个 VLAN，无法满足大二层网络中标识大量租户或租户群的需求。

2 传统二层网络中的 VLAN 无法满足网络动态调整的需求。

（3）虚拟机迁移范围受网络架构限制:

虚拟机启动后，可能由于服务器资源等问题（如 CPU 过高，内存不够等），需要将虚拟机迁移到新的服务
器上。为了保证虚拟机迁移过程中业务不中断，则需要保证虚拟机的 IP 地址、MAC 地址等参数保持不变，这就要求业务网

络是一个二层网络，且要求网络本身具备多路径的冗余备份和可靠性。

2.VXLAN 的优点

（1）针对虚拟机规模受网络规格限制：

VXLAN 将虚拟机发出的数据包封装在 UDP 中，并使用物理网络的 IP、MAC 地址作为外层头进行封装，

对网络只表现为封装后的参数。因此，极大降低了大二层网络对 MAC 地址规格的需求。

（2）针对网络隔离能力限制：

VXLAN 引入了类似 VLAN ID 的用户标识，称为 VXLAN 网络标识 VNI（VXLAN Network Identifier），

由 24 比特组成，支持多达 16M 的 VXLAN 段，从而满足了大量的用户标识。

（3）针对虚拟机迁移范围受网络架构限制：

VXLAN 通过采用 MAC in UDP 封装来延伸二层网络，将以太报文封装在 IP 报文之上，通过路由在网络中传输，无需关注虚拟机的 MAC 地址。且路由网络无网络结构限制，具备大规模扩展能力、故障自愈能力、负载均衡能力。通过路由网络，虚拟机迁移不受网络

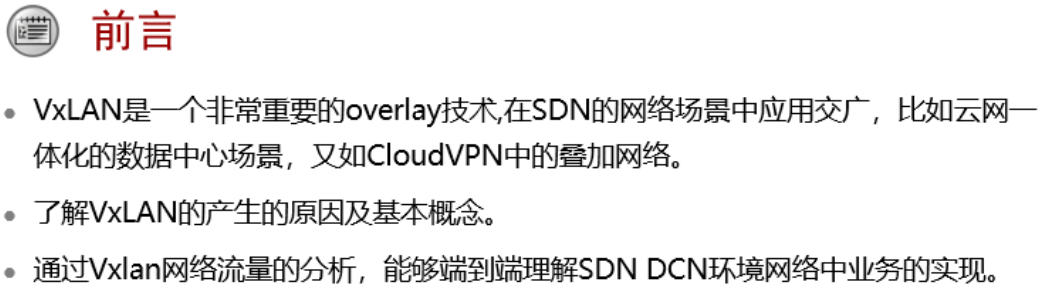
VXLAN 基本概念

VXLAN 是 NVO3 中的一种网络虚拟化技术，通过将虚拟机发出的数据包封装在 UDP 中，并使用物理网络的 IP、MAC 作为 outer-header 进行封装，然后在 IP 网络上传输，到达目的地后由隧道终结点解封装并将数据发送给目标虚拟机。

NVO3 标准技术之一，采用 MAC in UDP 封装方式，将二层报文用三层协议进行封装，可对二层网络在三层范围进行扩展，

同时支持 24bits 的 VNI ID（16M 租户能力），满足数据中心大二层 VM 迁移和多租户的需求。

通过 VXLAN，虚拟网络可接入大量租户，且租户可以规划自己的虚拟网络，不需要考虑物理网络 IP 地址和广播域的限制，降低了网络管理的难度。



- 概括来说，SDN2006 年诞生于园区网，2012 年可谓是 SDN 商用元年,发生了 google 部署 sdn 等重要事件将 SDN 推向了全球瞩目的焦点；同时 2012 年延展到电信网络。
- 下面我们结合重大事件加以介绍（内容较多，每一事件

需记住关键点）。

- 2006 年，SDN 诞生于美国 GENI 项目资助的斯坦福大学 Clean Slate 课题，斯坦福大学 Nick McKeown（尼克 麦吉翁）教授为首的研究团队提出了 Openflow 的概念用于校园网络的试验创新，后续基于 Openflow 给网络带来可编程的特性，SDN 的概念应运而生。Clean Slate 项目的最终目的是要重新发明英特网，旨在改变设计已略显不合时宜，且难以进化发展的现有网络基础架构。
- 2007 年，斯坦福大学的学生 Martin Casado（马丁 卡萨多）领导了一个关于网络安全与管理的项目 Ethane，该项目试图通过一个集中式的控制器，让网络管理员可以方便地定义基于网络流的安全控制策略，并将这些安全策略应用到各种网络设备中，从而实现对整个网络通讯的安全控制。
- 2008 年，基于 Ethane 及其前续项目 Sane 的启发， Nick McKeown 教授等人提出了 OpenFlow 的概念，并于当年在 ACM SIGCOMM 发表了题为《OpenFlow: Enabling Innovation in Campus Networks》的论文，首次详细地介绍了 OpenFlow 的概念。该篇论文除了阐述 OpenFlow 的工作原理外，还列举了 OpenFlow 几大应用场景。
- 基于 OpenFlow 为网络带来的可编程的特性，Nick McKeown 教授和他的团队进一步提出了 SDN（Software Defined Network，软件定义网络）的概念。2009 年，SDN 概念入围 Technology Review 年度十大前沿技术，自此获得了学术界和工业界的广泛认可和大力支持。
- 2009 年 12 月，OpenFlow 规范发布了具有里程碑意义的可用于商业化产品的 1.0 版本。如 OpenFlow 在 Wireshark 抓包分析工具上的支持插件、OpenFlow 的调试工具（liboftrace）、OpenFlow 虚拟计算机仿真（OpenFlowVMS）等也已日趋成熟。OpenFlow 规范已经经历了 1.1、1.2 以及 1.3 等版本。

OpenFlow 1.4 标准已经在 ONF 内部审阅，预计 2013 年 8 月初将获得批准发布。

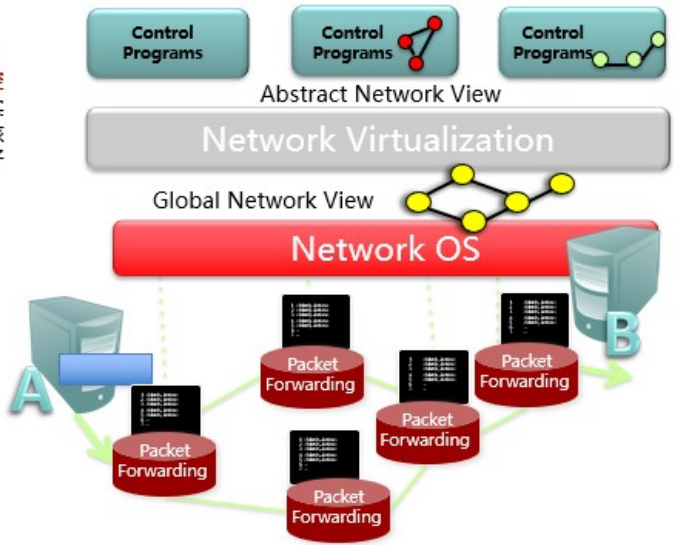
- 2011 年 3 月，在 Nick Mckeown 教授等人的推动下，开放网络基金会 ONF 成立，主要致力于推动 SDN 架构、技术的规范和发展工作。ONF 成员 96 家，其中创建该组织的核心会员有 7 家，分别是 Google、Facebook、NTT、、Verizon、德国电信、微软、雅虎。
- 2011 年 5 月，NEC 推出第一台可商用的 OpenFlow 交换机
- 2012 年 4 月，谷歌宣布其主干网络已经全面运行在 OpenFlow 上，并且通过 10G 网络链接分布在全球各地的 12 个数据中心，使广域线路的利用率从 30%提升到接近饱和。从而证明了 OpenFlow 不再仅仅是停留在学术界的一个研究模型，而是已经完全具备了可以在产品环境中应用的技术成熟度。
- 2012 年 7 月，软件定义网络(SDN)先驱者、开源政策网络虚拟化私人控股企业 Nicira 以 12.6 亿被 VMware 收购。 Nicira 是一家颠覆数据中心的创业公司，它基于开源技术 OpenFlow 创建了网络虚拟平台（NVP）。OpenFlow 是 Nicira 联合创始人 Martin Casado 在斯坦福攻读博士学位期间创建的开源项目，Martin Casado 的两位斯坦福大学教授 Nick McKeown 和 Scott Shenker 同时也成为了 Nicira 的创始人。VMware 的收购将 Casado 十几年来所从事的技术研发全部变成了现实——把网络软件从硬件服务器中剥离出来，也是 SDN 走向市场的第一步。
- 2012 年底，AT&T、英国电信(BT)、德国电信、Orange、意大利电信、西班牙电信公司和 Verizon 联合发起成立了网络功能虚拟化产业联盟(Network Functions Virtualisation，NFV)，旨在将 SDN 的理念引入电信业。由 52 家网络运营商、电信设备供应商、IT 设备供应商以及技术供应商组建。
- 2013 年 4 月，思科和 IBM 联合微软、Big Switch、博科、

思杰、戴尔、爱立信、富士通、英特尔、瞻博网络、微软、NEC、惠普、红帽和 VMware 等发起成立了 Open Daylight，与 LINUX 基金会合作，开发 SDN 控制器、南向/北向 API 等软件，旨在打破大厂商对网络硬件的垄断，驱动网络技术创新力，使网络管理更容易、更廉价。这个组织中只有 SDN 的供应商，没有 SDN 的用户——互联网或者运营商。Open Daylight 项目的范围包括 SDN 控制器，API 专有扩展等，并宣布要推出工业级的开源 SDN 控制器。

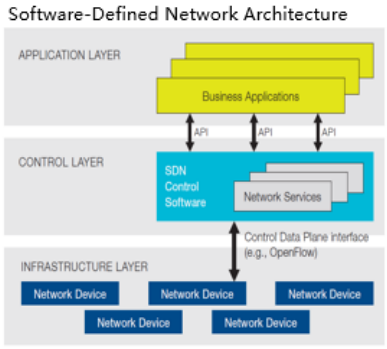
- 简单再补充点背景知识
- Clean Slate 计划
- 痛点：现有网络架构不断修补，难以解决根本问题，重新定义网络架构也许是根本解决方案，推倒重来可行么？
- CleanSlate 项目的最终目的是要重新发明因特网，旨在改变设计已略显不合时宜，且难以进化发展的现有网络基础架构。
- 引出广义和狭义 Clean Slate 项目概念
- 广义：泛指各种各样的下一代网络（NGN）项目
- 狭义：斯坦福大学尼克 麦吉翁（Nick McKeown）教授牵头的实验室研究计划（SDN 诞生处）
- Ethane 项目（CleanSlate 项目计划的子课题）
- 斯坦福的学生马丁.卡萨多（尼克-麦吉翁是马丁的老师）领导了一个关于网络安全与管理的项目 Ethane，该项目试图通过一个集中式的控制器，让网络管理员可以方便地定义基于网络流的安全控制策略，并将这些安全策略应用到各种网络设备中，从而实现对整个网络通讯的安全控制。
- 受此项目启发，Martin 和他的导师 Nick McKeown 教授提出了 OpenFlow 的概念。

SDN的定义

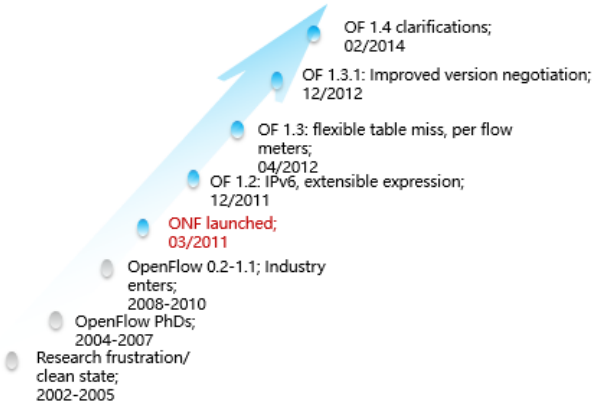
SDN（Software Defined Network），即软件定义网络，其核心技术是通过将网络**设备控制面**与**数据面**分离开来，从而实现了网络流量的灵活控制，为核心网络及应用的创新提供了良好的平台。



ONF定义的SDN基本架构

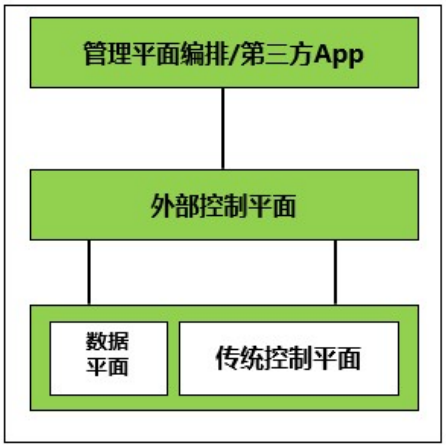


Source: ONF white paper, April 13, 2012



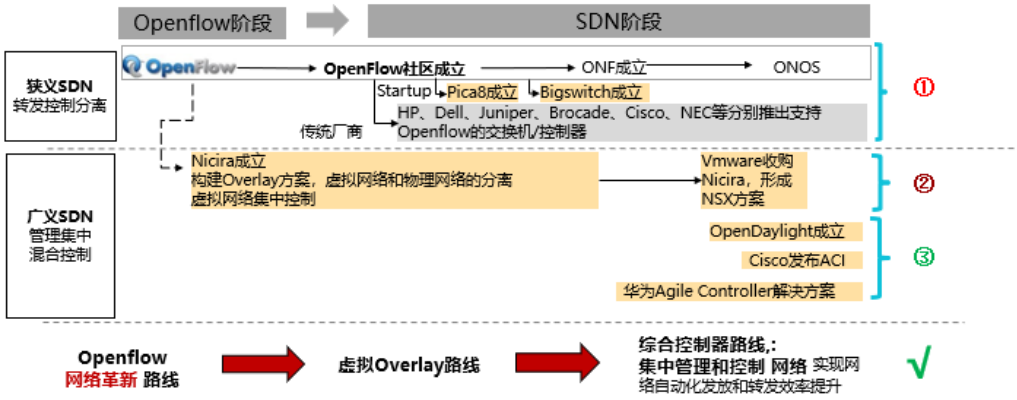
ONF强调OpenFlow-based SDN，强调控制与转发分离以实现转发设备的标准化，重点是OF协议标准化。

传统CT厂商眼中的SDN架构

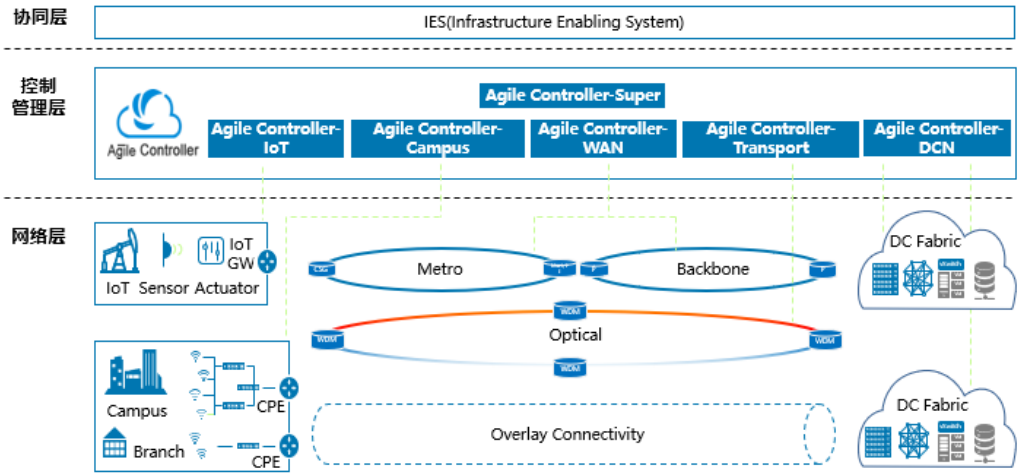


SDN主要技术路线

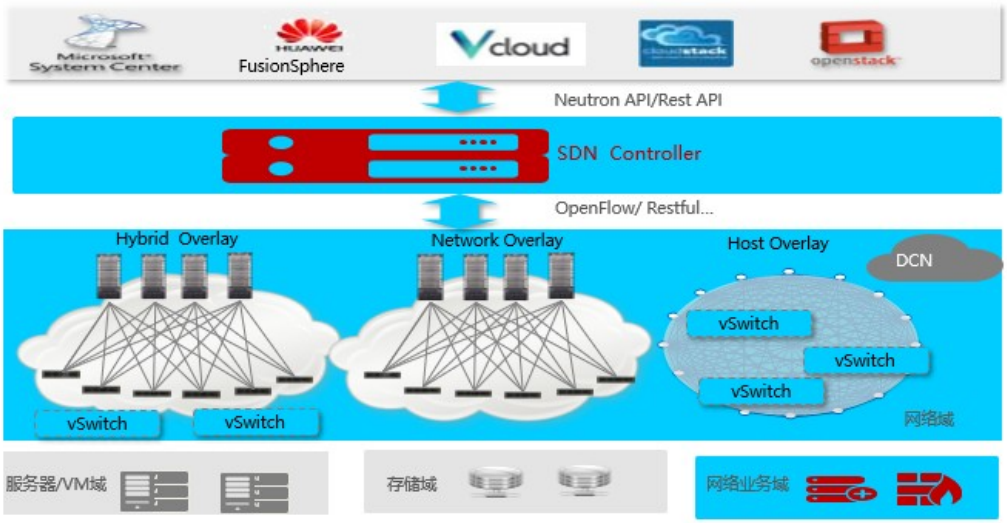
- 2006年斯坦福大学发布 OpenFlow，网络设备转发和控制面分离，通过集中的控制面实现网络流量的灵活控制
- 华为数据中心SDN的核心特征是适度的转控分离结合之外，通过管理与控制分离，实现网络业务自动化发放，助力数据中心业务实现敏捷发放.



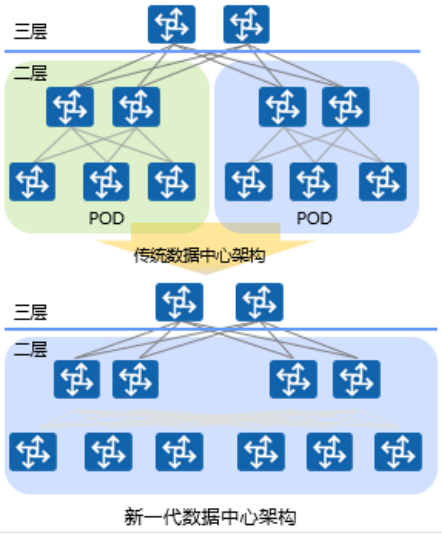
华为SDN解决方案全景图



SDN DCN解决方案



数据中心发展趋势



传统数据中心架构

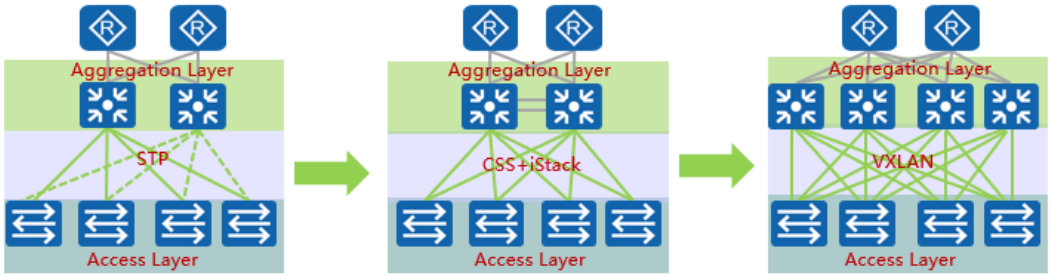
- 传统数据中心组网方式，一般二层只到接入或汇聚交换机，虚拟机的迁移只能局限一个二层区域内。如果需要跨二层区域迁移，需要更改VM的IP地址，应用会中断。

新一代数据中心架构

- 在云计算时代，IDC运营商为了更充分的利用数据中心资源，VM需要更大的迁移范围；
- 由于服务器之间存在大量的横向流量，要求数据报文支持无阻塞转发，网络链路资源得到充分的利用。

- 虚拟机规模受网络规格限制
- 在大二层网络环境下，数据报文是通过查询 MAC 地址表进行二层转发，而 MAC 地址表的容量限制了虚拟机的数量。

数据中心发展趋势

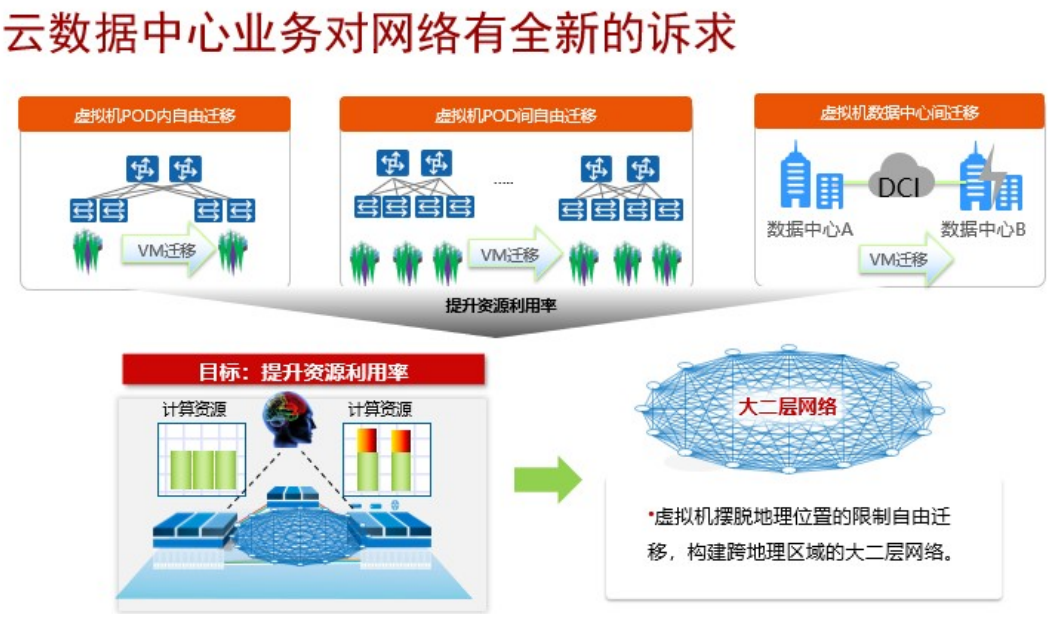


- STP或CSS+iStack传统二层技术不适合构建大规模二层网络， 通过VXLAN可以构建大二层网络，支持扁平化胖树拓扑组网方式，链路带宽利用率高。

- 网络隔离能力限制
- 当前主流的网络隔离技术是 VLAN 或 VPN（Virtual Private Network），在大规模的虚拟化网络中部署存在如下限制：由于 IEEE 802.1Q 中定义的 VLAN Tag 域只有 12 比特，仅能

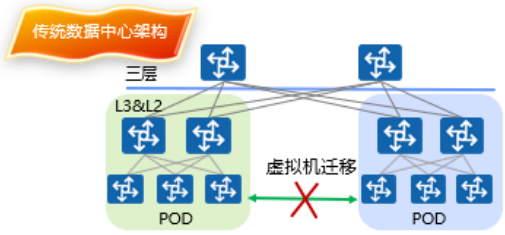
表示 4096 个 VLAN，无法满足大二层网络中标识大量用户群的需求。

- 传统二层网络中的 VLAN/VPN 无法满足网络动态调整的需求。

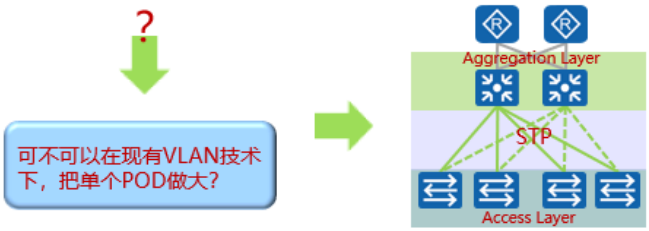


- Google 是大规模服务器集群的实践者，服务器间的大量通信要求网络是无阻塞的；
- Google 服务器单集群的规模已经达到了 1 万台左右；
- 国内的互连网厂家也在考虑规划 2 万台服务器的集群；
- 网络接口数量和容量是决定集群规模的核心因素；
- 构建跨地理区域的服务器集群，提高系统容灾能力。
- 大二层网络:
- 大规模的二层网络；
- 要求网络横向流量提供无阻塞能力。

传统网络为何大不起来



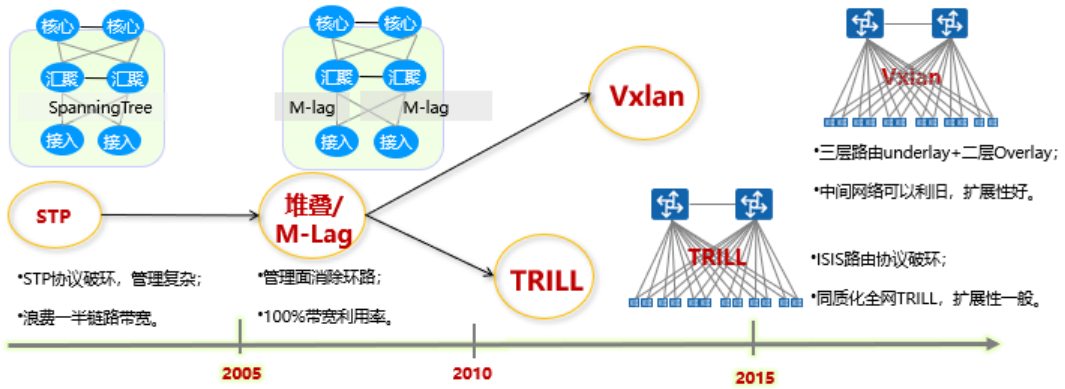
- VLAN无法跨越三层边界：这样传统数据中心组网方式通常是网关部署在汇聚交换机，汇聚交换机间通过三层核心互通。虚拟机的迁移只能局限于POD内。如果需要跨POD二层区域迁移，需要更改VM的IP地址，应用会中断。



STP技术在解决网络环路问题的同时，存在以下主要缺陷：

- STP收敛时间长，通常不超过50个网络节点，不适宜云数据中心大规模组网。
- STP构建无环网络时，需要阻断一半的链路；带宽利用率低。

数据中心网络架构发展趋势



•TRILL技术解决了STP环路组网和规模问题，通过成熟的链路状态路由算法，扩展IS-IS协议，构建无环网络，实现多路径负载分担；

•VXLAN技术具有更好的可用性和扩展性，更易运维，已经成为IT&CT厂商力推的技术。

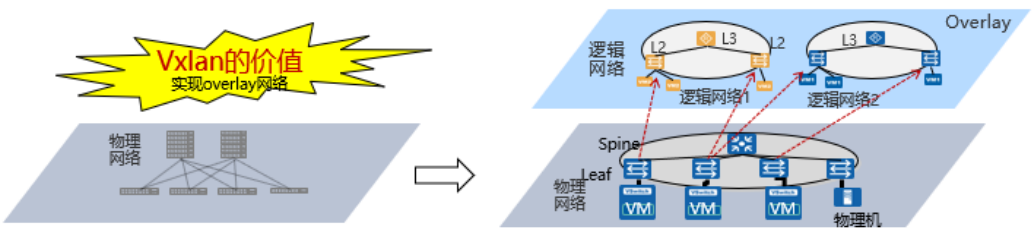
- IT+CT 形成合力；
- TRILL 是革命性的技术，Vxlan 是改良的技术。

VXLAN 是业界 Overlay技术的事实标准

云数据中心高端网络诉求		VXLAN	CSS/SVF	TRILL
>4K租户	出租型的数据中心，需要支持海量租户	**16M**	4K	4K（最新标准可升级到16M）
保护现有网络投资	可在现有网络的基础上构建新的Fabric	**对现有网络无要求，只要支持普通L3即可**	全网新建	全网新建
SDN支持能力	可平滑升级到SDN网络	**Overlay是SDN的重要路线之一**	不支持	不支持
标准协议	不同厂商可实现互通	**标准**	各厂商私有	标准
跨DC能力	可跨越IP WAN构建大二层	**支持**	不支持	不支持

- VXLAN在支持SDN，多租户等方面能力更强，因此成为业界的技术热点。

VXLAN的价值



VXLAN的概念：

- Overlay 网络定义将一个计算网络构建在另一个网络之上；
- 核心是实现封装，将网络业务与底层设施解耦；
- 封装技术使用的VXLAN；
- VXLAN隧道封装的端点就叫NVE（Network Virtualization Edge），负责原始以太报文的VXLAN封装和解封装。

VXLAN的优势：

- VXLAN是业界标准的Overlay技术；
- 相比STP的主备路径，VXLAN利用Underlay网络的ECMP能带来更高的网络转发性能；
- VXLAN基于UDP技术构建，是统一IT和CT两界的overlay技术；
- VXLAN将VLAN的4k子网扩充到16M，支持多租户；
- VXLAN为SDN提供转发面基础。

- 物理网路
- 物理网络，带宽高，容量大；
- 大二层网络需要 STP 解决环路问题；
- 二层网络隔离受限，仅 4k VLAN；
- 虚机迁移不够灵活，需要改变物理网络配置。

- Overlay 网络：
- Overlay 实现了某种程度的 ID 和位置信息的分离, 有更好的移动性, 满足二层网络弹性需求；

- Overlay 按需部署业务网络, 业务变化的时候 Underlay 网络不需要改变；
- 兼容性好，通过 Overlay 实现与物理网络解耦。

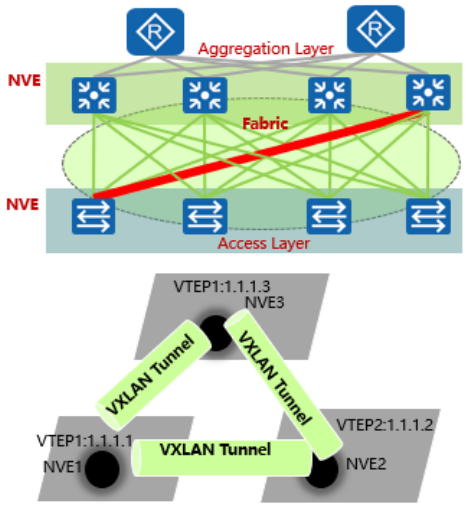
VXLAN基本概念

基于NVo3的二层Fabric组网

NVO3(Network Virtualization Over Layer 3)，基于三层IP overlay网络构建虚拟网络技术统称为NVO3，目前比较有代表性的有：VXLAN、NVGRE、STT。
运行NVO3的设备叫做NVE（Network Virtualization Edge），它位于overlay网络的边界，实现二、三层的虚拟化功能。

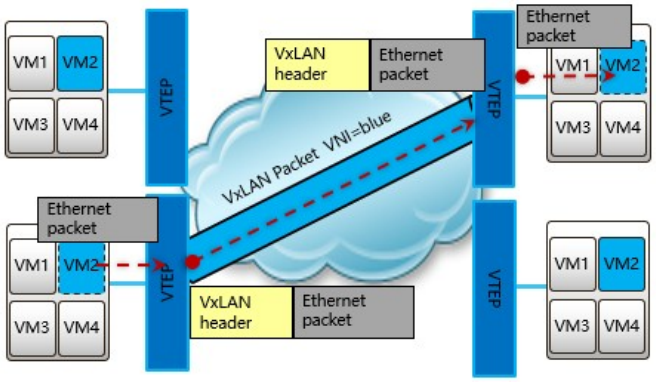
VXLAN(Virtual Extensible LAN，虚拟可扩展局域网)是目前NVO3中影响力最为广泛的一种。它通过LMAC in UDP的报文封装方式，实现基于IP overlay的虚拟局域网。

- VXLAN网络中的NVE以VTEP进行标识，VTEP（VXLAN Tunnel EndPoint，VXLAN隧道端点）；
- 每一个NVE至少有一个VTEP，VTEP使用NVE的IP地址表示；
- 两个VTEP可以确定一条VXLAN隧道。



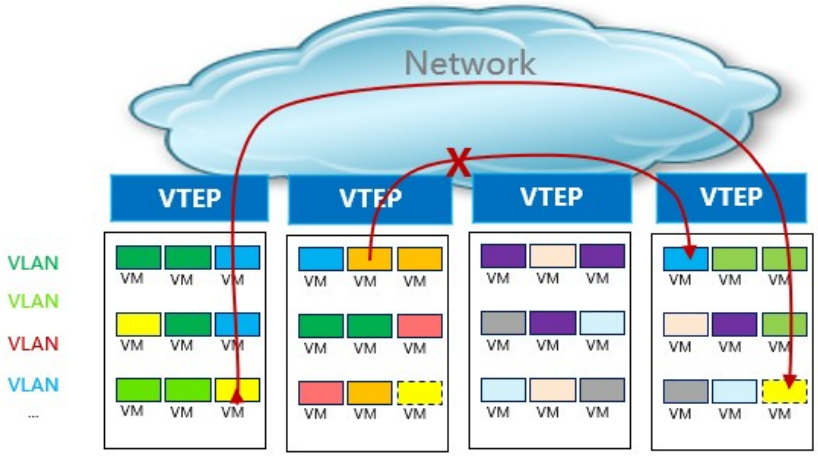
VXLAN 概念 - VTEP

- VXLAN网络中的NVE以VTEP进行标识，VTEP（VXLAN Tunnel EndPoint，VXLAN隧道端点）；
- 每一个NVE至少有一个VTEP，VTEP使用NVE的IP地址表示；
- 两个VTEP可以确定一条VXLAN隧道，VTEP间的这条VXLAN隧道将被两个NVE间的所有VNI所公用。

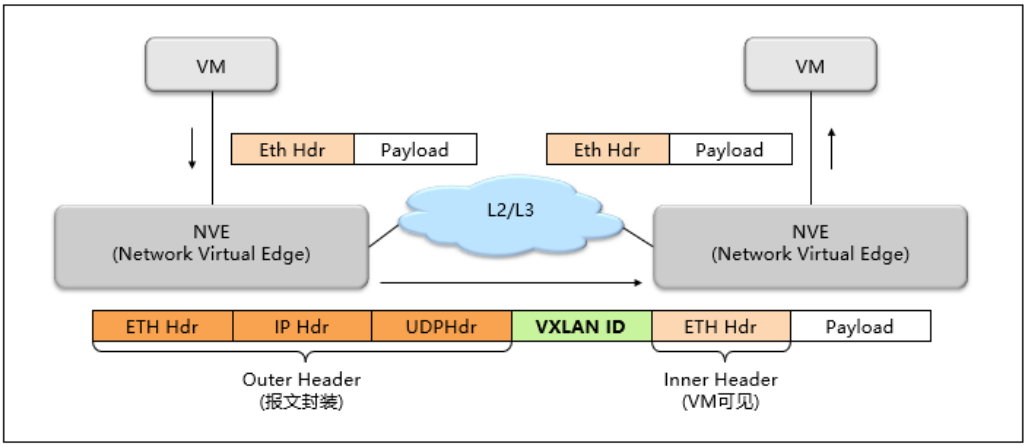


VXLAN - VNI

- VNI-24比特，用于标识虚拟网络，最大支持16M。

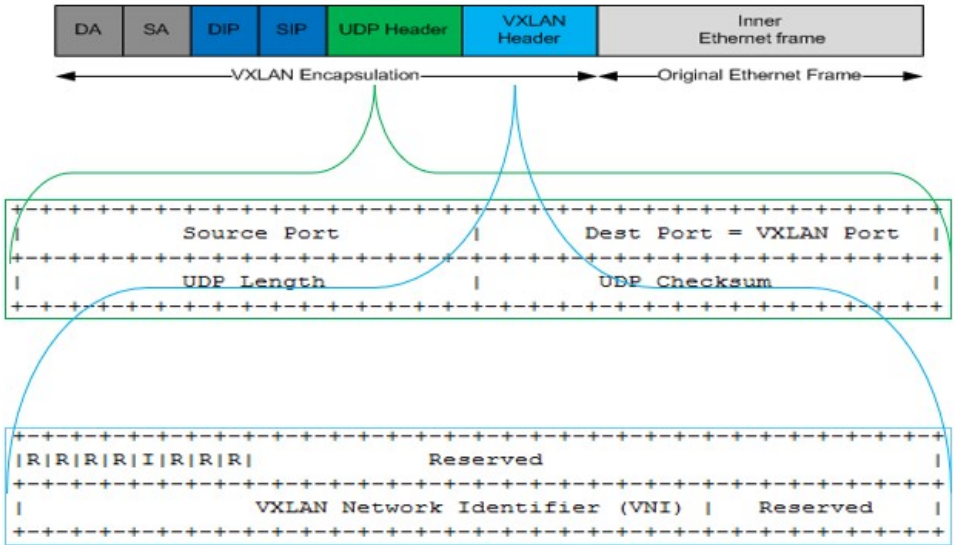


VXLAN 报文格式

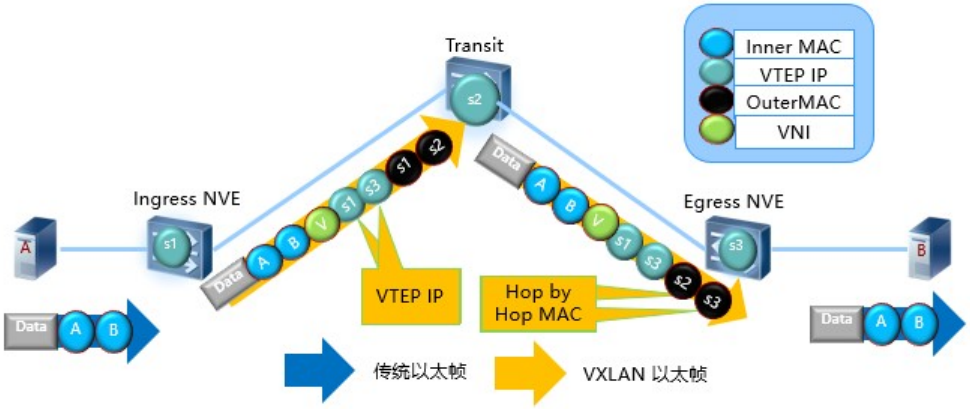


VxLAN 报文封装流程

VXLAN 报文格式

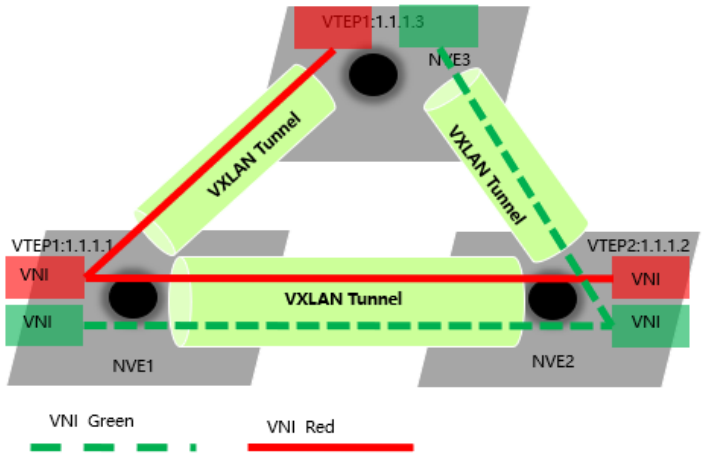


VXLAN 转发数据封装



源终端的二层报文能够穿越IP网络到达目的终端，VXLAN网络对于主机来说相当于是Bridge Fabric。

隧道和VNI关系

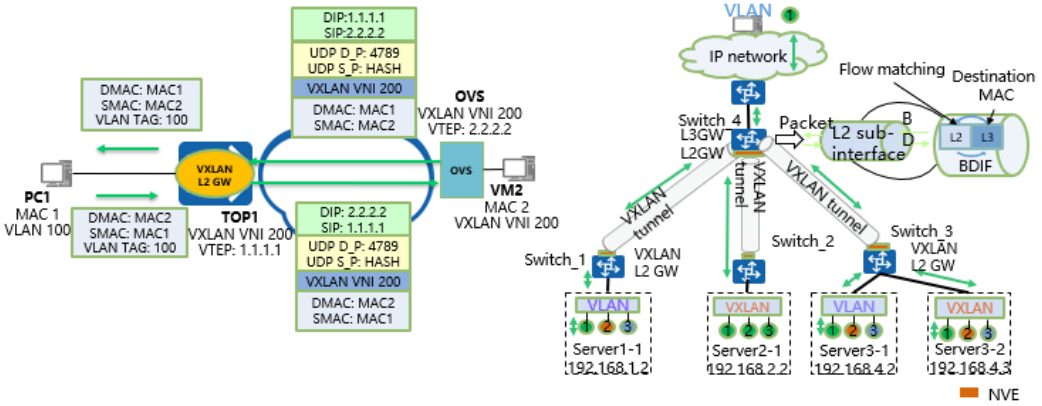


- VNI概念

■ 标识VXLAN网络中的二层域。

■ 两个VTEP可以确定一条VXLAN隧道，VTEP间的这条VXLAN隧道将被两个NVE间的所有VNI所公用。

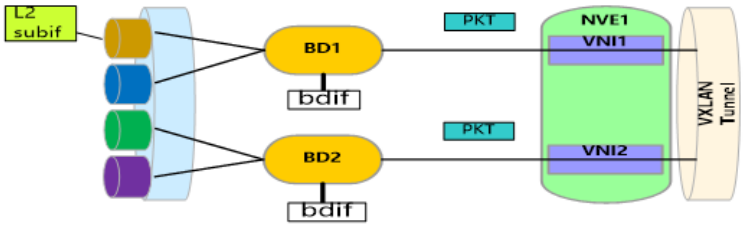
VXLAN 网关



- **VxLAN L2 Gateway:** 允许租户接入VxLAN网络，实现相同VxLAN内部流量互访。
- **VxLAN L3 Gateway:** 实现不同VxLAN直接互访，或者VxLAN与非VxLAN网络互访。

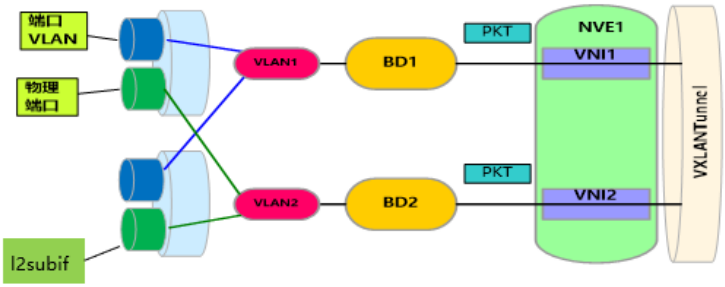
VXLAN接入业务模型 (1)

- VXLAN网关使用EVC的业务模型，模型构件主要包含：BD（Bridge-Domain）、VNI（Virtual Net Instance）、NVE（Network Virtualization Edge）、二层子接口（l2 subif）、VXLAN隧道。
  - L2-Subif：用于用户接入，子接口上可以配置一层tag接入或者不配置tag接入；
  - BD（Bridge-Domain）：标识一个二层广播域，BD和VNI 1:1映射。所有广播域功能基于BD支持，如MAC学习、二层查表、广播复制等；
  - NVE（Network Virtualization Edge）：主要用于本地VTEP地址管理，VXLAN隧道管理，头端复制列表管理；
  - VXLAN隧道：VXLAN隧道用于VXLAN报文的转发，用本地VTEP地址+远端VTEP地址标识；
  - BDIF：BD域的三层路由接口，用于二层流量进入三层进行路由转发；

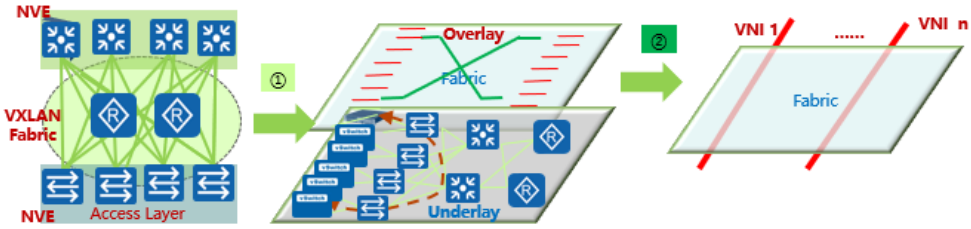


VXLAN接入业务模型 (2)

- 全局VLAN接入模型：主要应用在L2VPN服务场景，VLAN绑定bd，提供将传统port+vlan接口接入VXLAN网络的能力；二层子接口绑定BD。



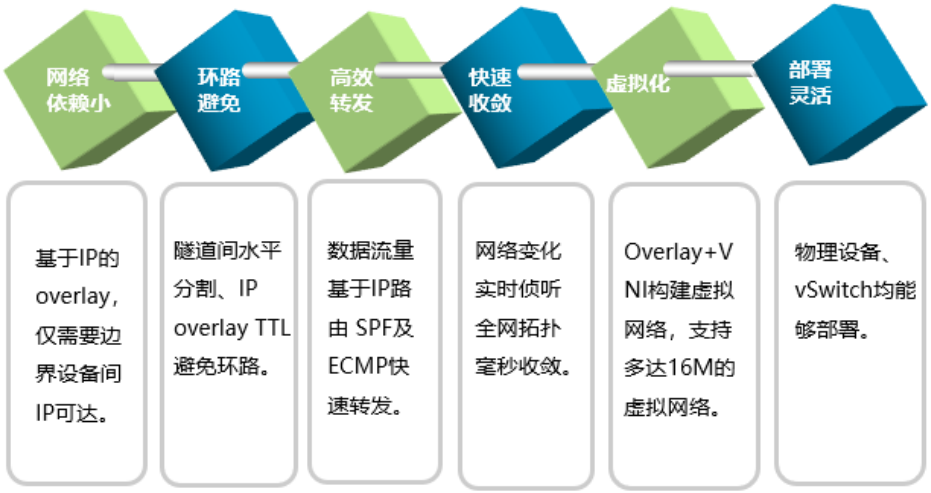
VXLAN逻辑抽象



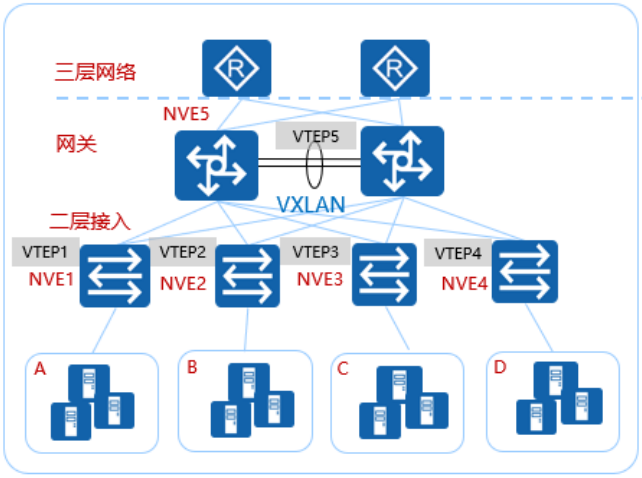
- VXLAN的简化理解----两次虚拟化

1、第一次虚拟化：利用隧道技术将边缘设备互连透传二层报文；整网抽象理解成一台端口数目扩展的超大LAN switch。
2、第二次虚拟化利用VNI将这台超大的交换机虚拟出多个二层的广播域，和VLAN本质是一样的，VNI类比VLANID. 并通过定义VXLAN header中的VNI字段，将子网范围由4K扩展至16M。

VXLAN的主要优点



VXLAN同子网转发流程



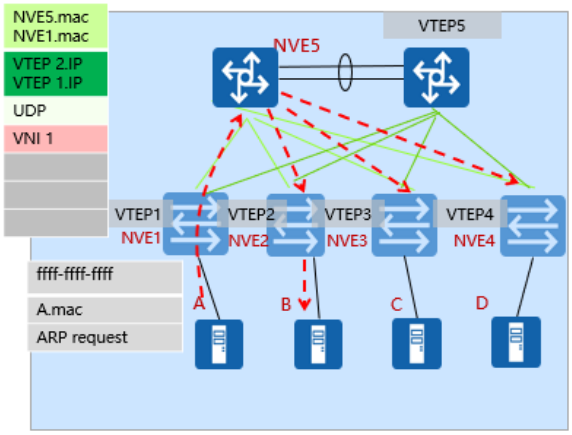
总体流程

- HOST A发送ARP Request报文到HOST B。
- HOST B回应ARP Reply报文到HOST A。
- HOST A发送单播数据报文到HOST B。

注：A、B、C、D都属于同一VNI 1。不考虑ARP广播优化使能。

同网段查MAC二层转发 (1)

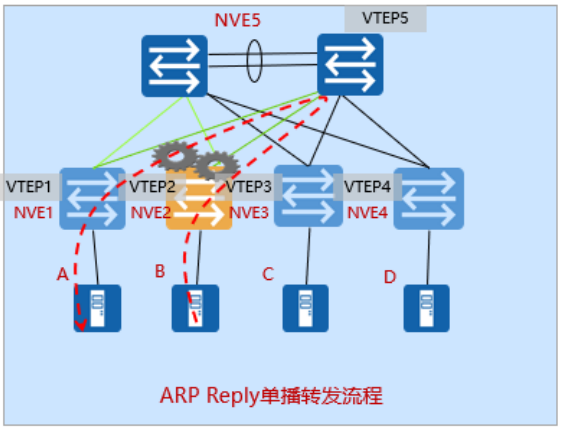
A到B的ARP request 广播



1. NVE1发现是广播报文，在VNI1内广播ARP request报文；报文做隧道封装；
2. 中间节点，IP透传overlay报文；
3. NVE2/3/4/5 接收报文,解隧道封装，原始报文本地VNI1内广播；学习到服务器A mac。

同网段查MAC二层转发 (2)

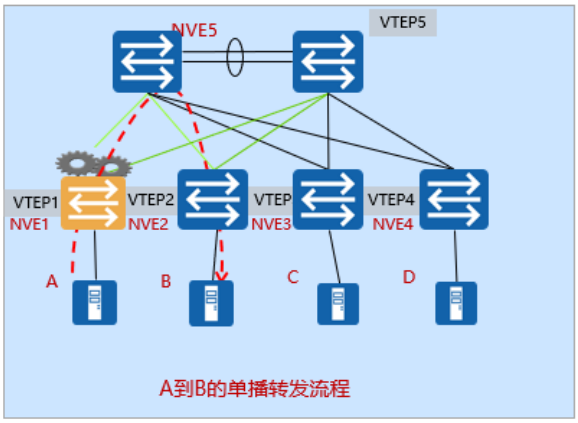
B回复A的ARP reply 单播



1. NVE2查找服务器A mac转发表，命中出接口为隧道(NVE2至NVE1隧道）；报文封装后三层转发；
2. 中间节点，IP透传overlay报文；
3. NVE1接收报文并解封装，在本地转发；NVE1学习到服务器B的MAC地址。

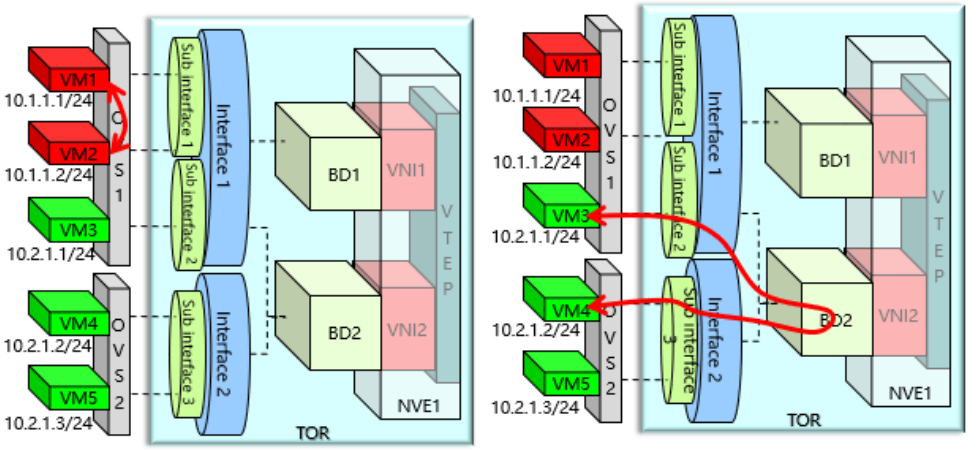
同网段查MAC二层转发 (3)

A到B的单播数据报文



1. NVE1 和 NVE2都学习到了服务器A和B的MAC地址；后续查找MAC则命中；单播流程，和VLAN一样的；
2. 唯一不同的是外层封装了隧道；underlay是IP转发。

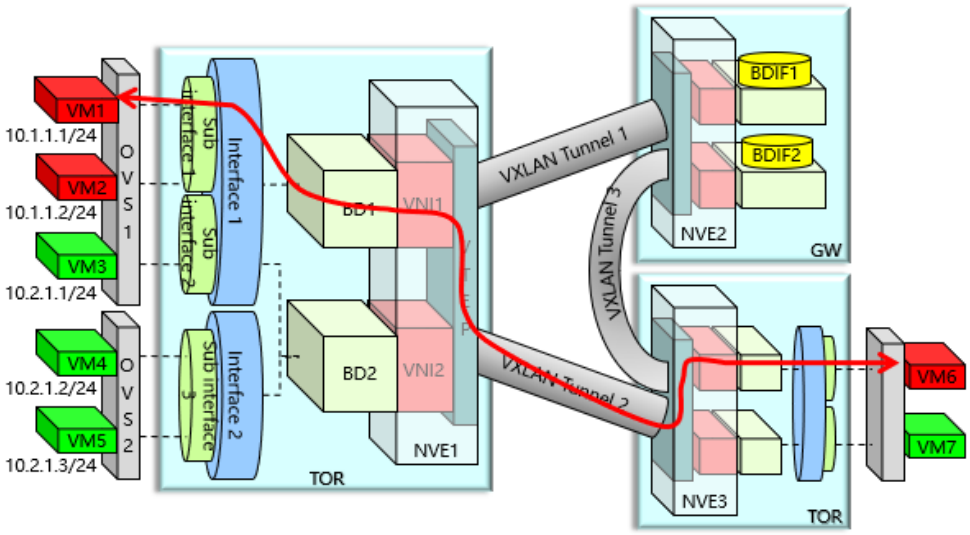
VXLAN转发模型之相同网段VM互访 (1)



Scenario 1: Both VMs located at the same vSwitches connected to same TOR

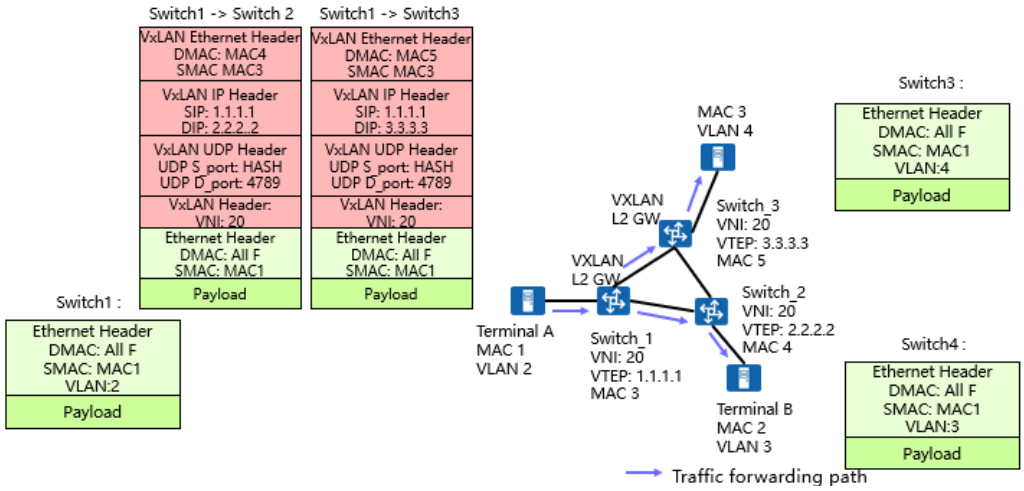
Scenario 2: Both VMs located at different vSwitches connected to same TOR

VXLAN转发模型之相同网段VM互访 (2)

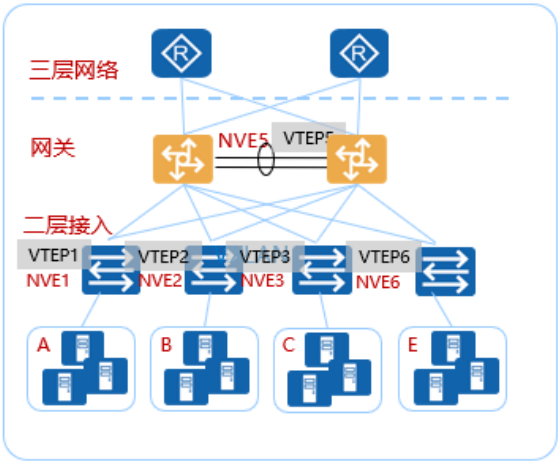


Scenario 3: Both VMs located at different vSwitches connected to different TOR

VXLAN - BUM 报文转发流程



VXLAN数据转发总体流程 (跨子网)



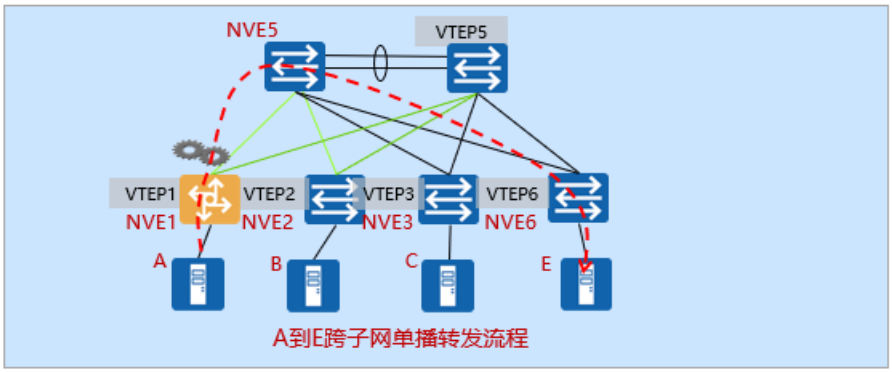
总体流程

■HOST A发送单播数据报文给HOST E。

注：NVE5作为三层网关，HOST A属于VNI 1，HOST E属于VNI 2，默认主机与网关都互相学习到ARP表，各个节点MAC都已学习。

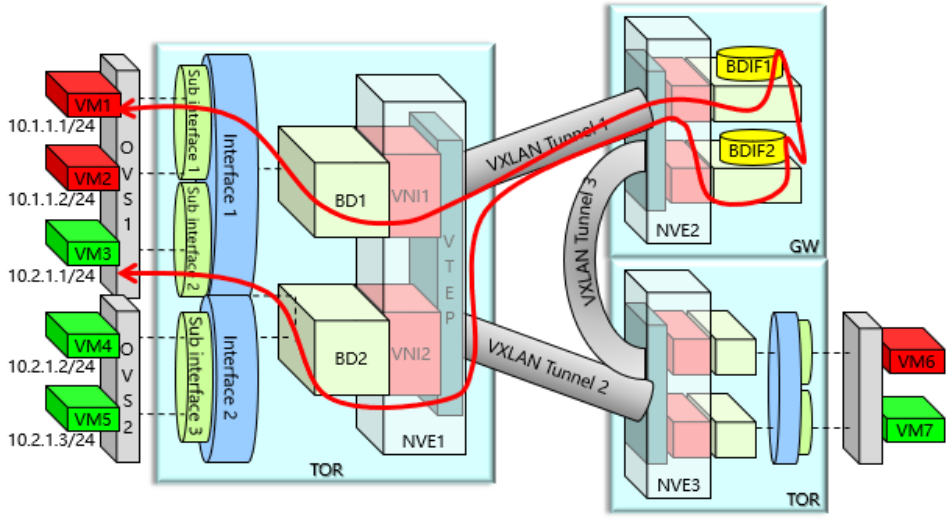
- HOST A 发送单播数据报文给 HOST E。
- 注：NVE5 作为三层网关，HOST A 属于 VNI 1，HOST E 属于 VNI 2，默认主机与网关都互相学习到 ARP 表，各个节点 MAC 都已学习。

A到E单播转发流程



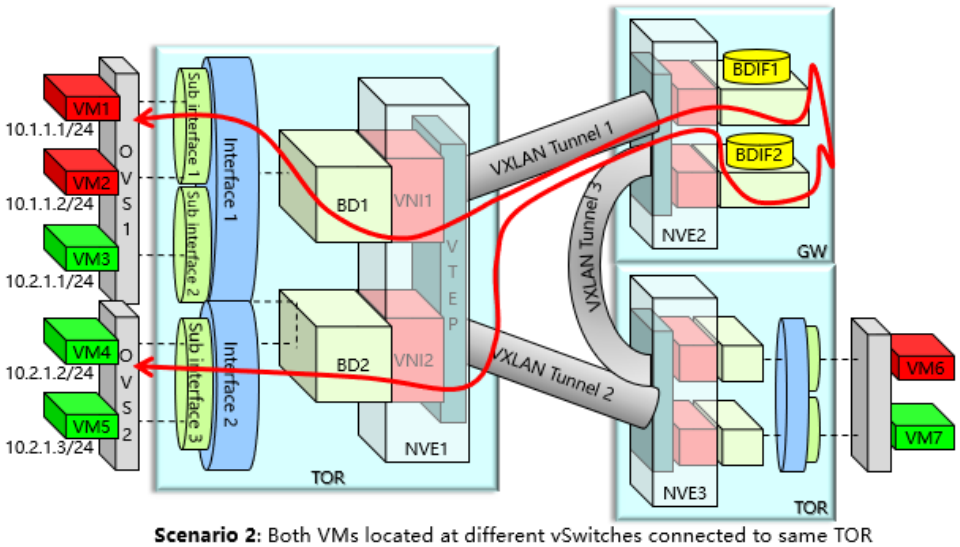
1. NVE1查找网关mac转发表，封装隧道；使用VNI1；
2. 网关解封装报文，根据内层IP头查路由，替换内层以太头，封装VXLAN头部，使用VNI2；
3. NVE6 接收报文并解封装，内层报文根据目的MAC转发。

VXLAN转发模型之不同网段VM互访 (1)



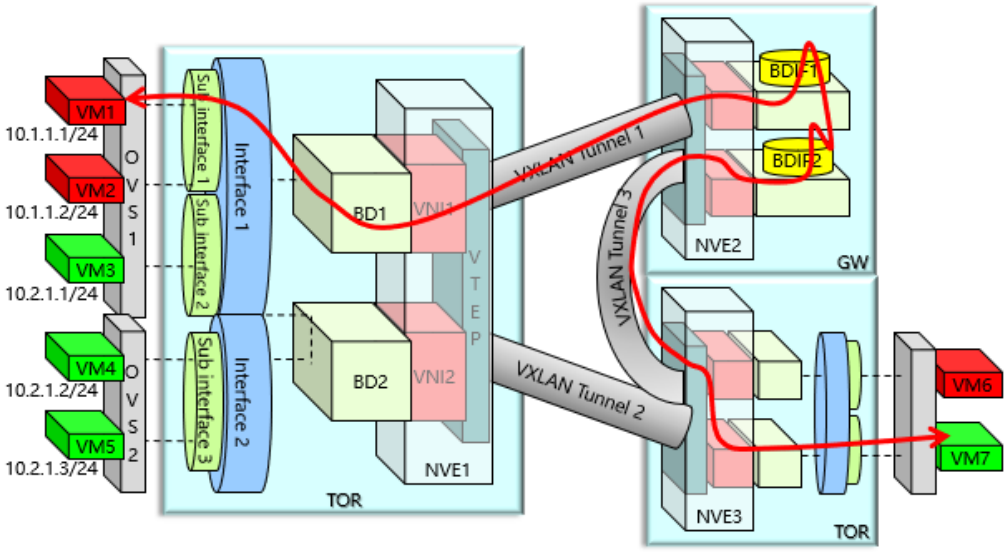
Scenario 1: Both VMs located at the same vSwitches connected to same TOR

VXLAN转发模型之不同网段VM互访 (2)



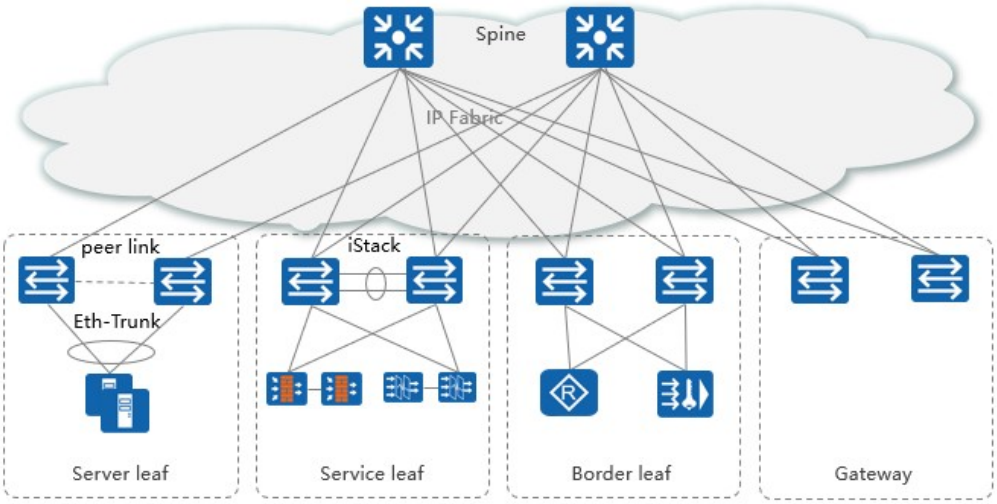
Scenario 2: Both VMs located at different vSwitches connected to same TOR

VXLAN转发模型之不同网段VM互访 (3)

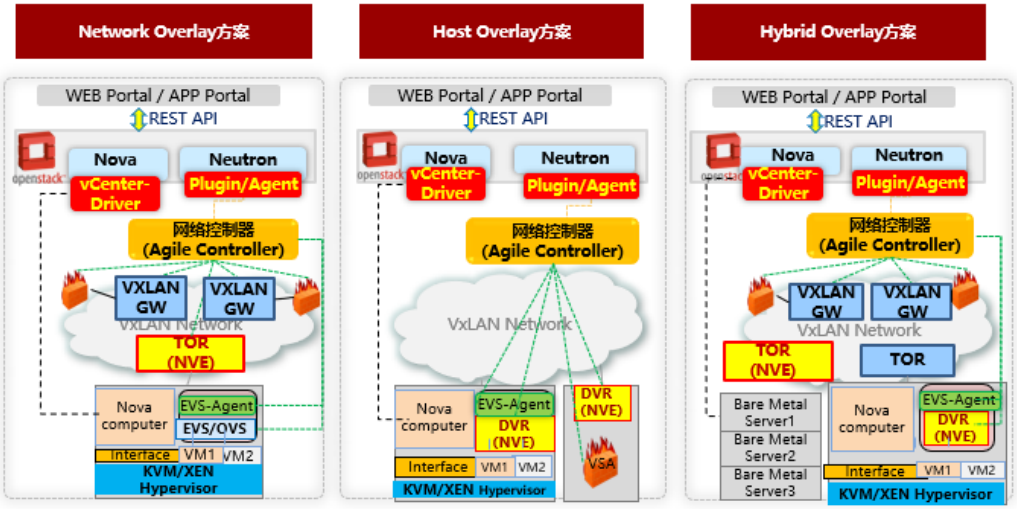


Scenario 3: Both VMs located at different vSwitches connected to different TOR

VXLAN 基于Spine-Leaf组网架构



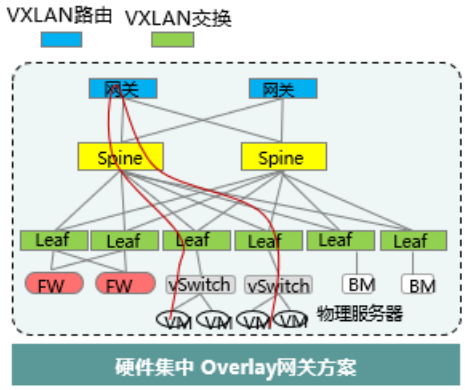
VXLAN三种OverLay组网方案



- 软件方案：
- 不改造现有物理设备，与具体厂商硬件设备解耦，无需配置物理网络，实现大规模逻辑二层网络的自动创建。
- 硬件方案：
- 新建物理网络，通过 VXLAN Overlay 网络，实现自动化业务发放。
- 混合方案：
- 通过 SDN 实现对虚拟网络及物理网络（交换机、防火墙

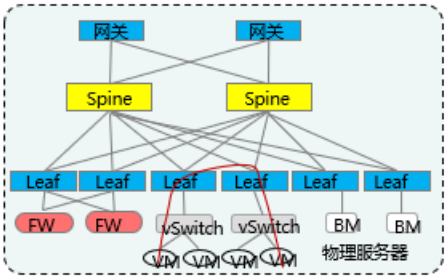
及 F5）的配置管理和自动化业务发放。

VXLAN OverLay的分布式和集中式



硬件集中 Overlay网关方案

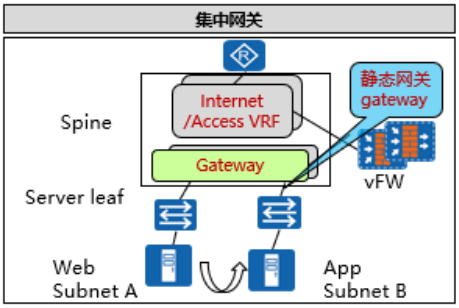
- VXLAN二层VTEP功能：部署在Leaf
- VXLAN三层网关功能：部署在核心层
- Spine：普通路由

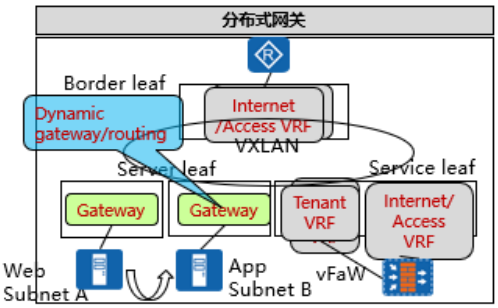


硬件分布式Overlay网关方案

- Leaf 节点既是VxLAN二层VTEP网关，又是东西向流量的三层VxLAN网关
- 南北向流量的网关部署在核心层
- Spine：普通路由

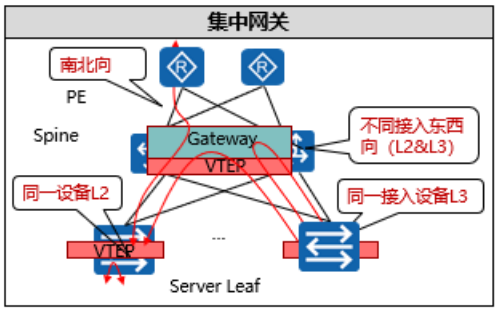
网关部署对比

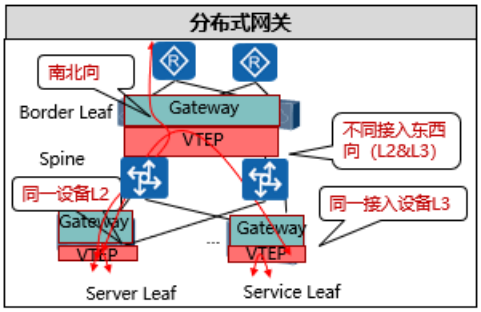




比较项	集中部署	分布式部署
VM迁移网关部署变化	VM迁移，网关部署不变	VM迁移，网关动态迁移 从源接入设备删除网关，在目的接入设备创建网关
VM迁移影响的表项	集中网关刷新ARP 接入设备刷新MAC	所有设备刷新主机路由 接入设备刷新MAC

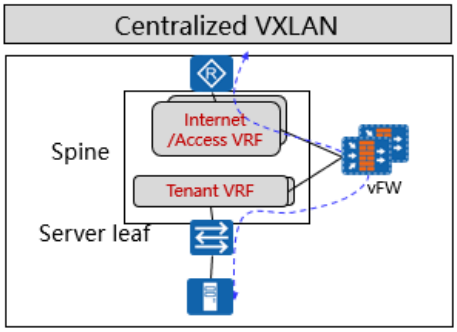
转发路径优化对比

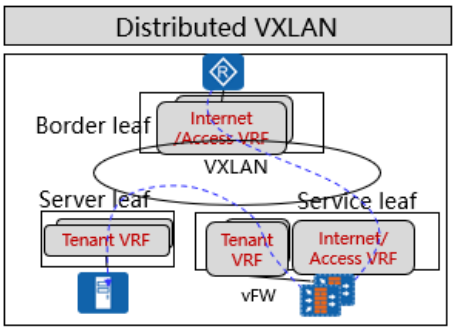




比较项	集中部署	分布式部署
南北向流量(customer to servers)	经过核心	经过核心
不同的接入设备间东西向 流量(L2&L3)	经过核心	经过核心
同一个接入设备下东西向流量(L2)	本地转发	本地转发
同一个接入设备下东西向流量(L3)	经过核心	本地转发

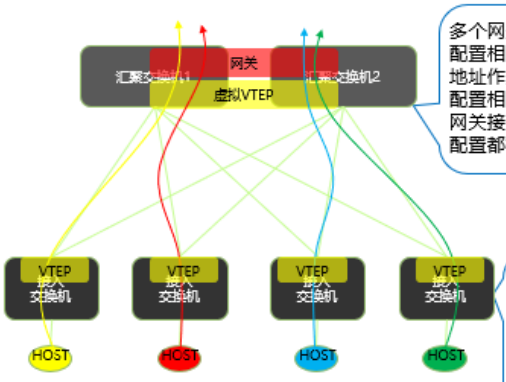
防火墙流量过滤对比





比较项	集中部署	分布式部署
防火墙引流方案	防火墙旁挂集中网关 集中网关单点部署策略将流量引到防火墙	防火墙与网关非直接 多点网关通过隧道连到防火墙 分布网关多点部署策略

集中网关高可靠



多个网关设备上：
配置相同的loopback地址作为VTEP；
配置相同的网关接口，网关接口的IP、MAC等配置都相同。

1、网关故障，在恢复时，让恢复设备发布低优先的路由（通过控制器调整），保证上下行流量发给其他网关，待网关ARP恢复后，再恢复路由优先级，让此网关参与负载分担；

2、网关下行口故障，underlay路由收敛；

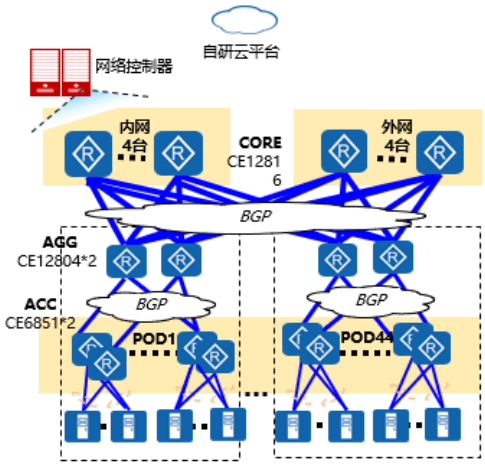
3、网关上行口全故障，依靠OPS联动将VTEP地址发布的优先级降低，保证上行流不发到故障网关。

接入设备上MAC和IP路由转发表

目的MAC	远端VTEP
网关虚MAC	vVTEP
目的VTEP路由	下一跳
vVTEP IP	AGG1&AGG2

与传统的堆叠、VRRP网关比，有如下一些优势：
网关之间不需要运行类似VRRP、GLBP的基于子网粒度的心跳协议，网关信令处理压力小。
相比VRRP三层网关，物理网关之间流量能够实现Flow-based Loadbalancing，网关能够扩展到多台。
可以通过路由协议控制器vVTEP的路由发布，实现流量无损网关扩容或升级。

中国xx互联网A公司超大规模公有云



客户诉求

- 大规模、高性能、低收敛比；
- 全DC迁移；
- 基于SDN控制器的自动化精细运维。

解决方案

- 基础网络采用Spine-Leaf架构；Overlay网络使用VXLAN构建大二层，VTEP部署在Leaf交换机；
- Vxlan三层网关集中部署；部署多活网关，负载分担提高吞吐；
- 网络控制器：定制北向API，对接私有云平台实现网络业务的自动化部署；
- 通过控制器实现路径可视、流量可视，帮助运维。

客户价值

- 多活硬件集中式网关，保证了规模、性能的同时，符合传统数据中心网络管理人员的运维习惯。

思考题

1. 关于VXLAN的描述，以下错误的是？（ ）
   A. VXLAN采用MAC in MAC封装方式。
   B. 类似于MPLS的标签转发，VXLAN报文通过VNI进行报文转发。
   C. VXLAN是NVO3（Network Virtualization over Layer 3）中的一种网络虚拟化技术。
   D. VXLAN报文外层UDP Header的源端口号一般填内层报文头通过哈希算法计算后的值。
2. VXLAN网络构建方案主要有以下哪几种？（ ）
   A. Network Overlay
   B. Host Overlay
   C. Hybrid Overlay

- 参考答案：
- AB
- ABC

本章总结

- 云数据中心的需求与挑战
- VXLAN基本原理
- VXLAN报文转发原理
- VXLAN常见组网
- 应用案例