G-LAB

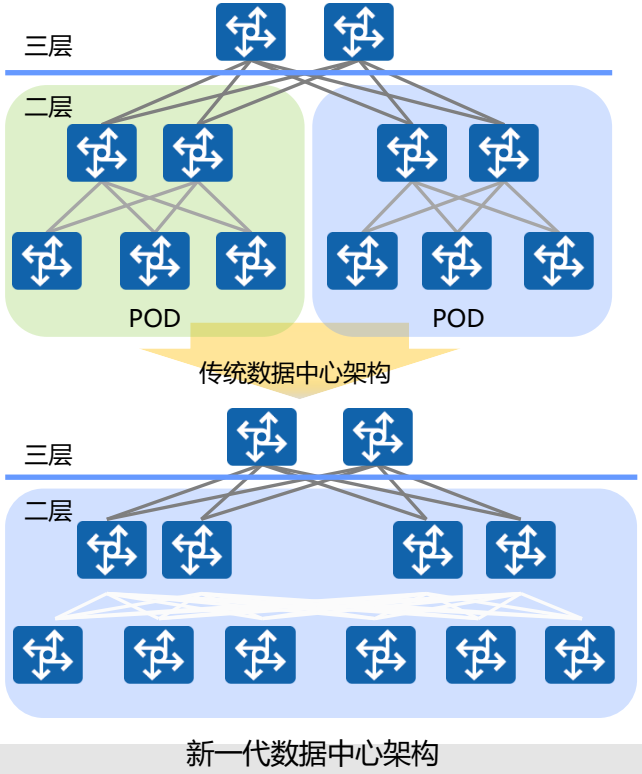
G-CNP v2.0课程

讲师：沈老师

前言

- VxLAN是一个非常重要的overlay技术,在SDN的网络场景中应用交广，比如云网一体化的数据中心场景，又如CloudVPN中的叠加网络。
- 通过Vxlan网络流量的分析，能够端到端理解SDN DCN环境网络中业务的实现。

数据中心发展趋势



新一代数据中心架构

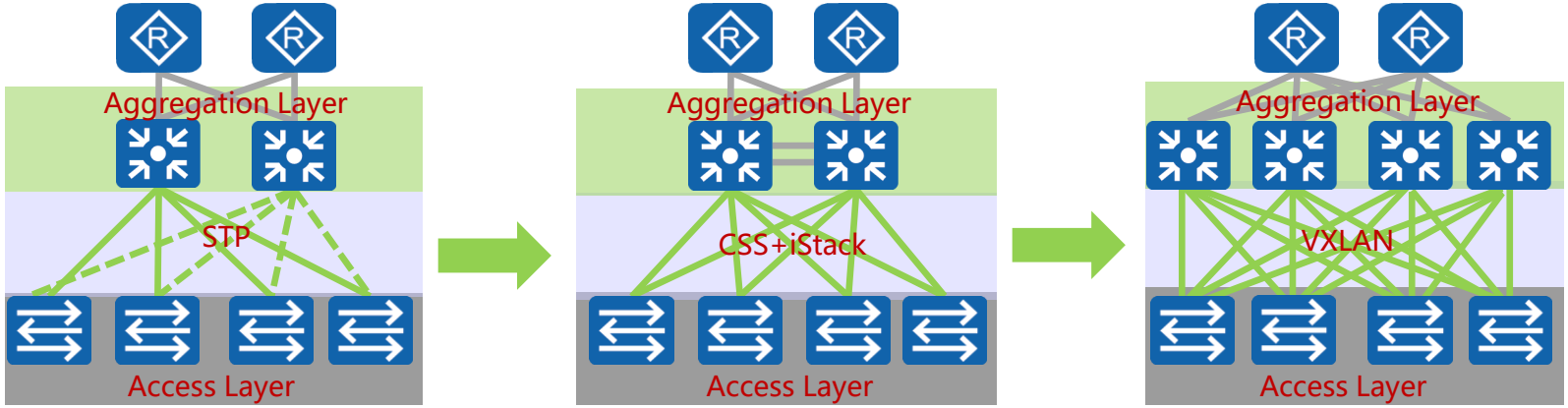
传统数据中心架构

- 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。

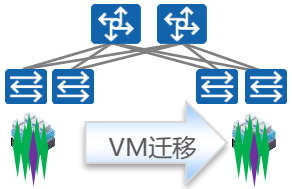
数据中心发展趋势



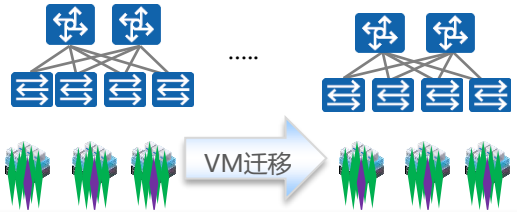
- STP或CSS+iStack传统二层技术不适合构建大规模二层网络， 通过VXLAN可以构建大二层网络，支持扁平化胖树拓扑组网方式，链路带宽利用率高。

云数据中心业务对网络有全新的诉求

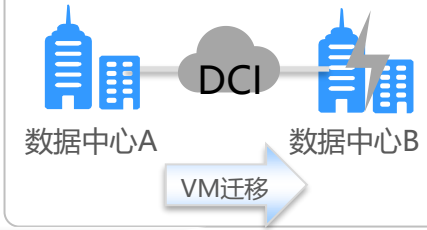
虚拟机POD内自由迁移



虚拟机POD间自由迁移

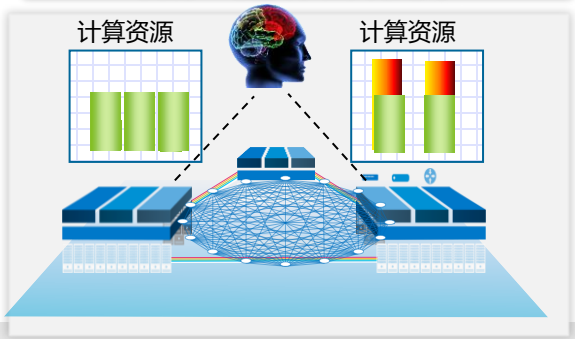


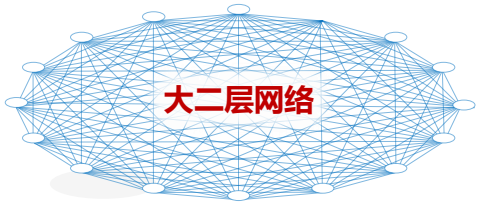
虚拟机数据中心间迁移



提升资源利用率

目标：提升资源利用率

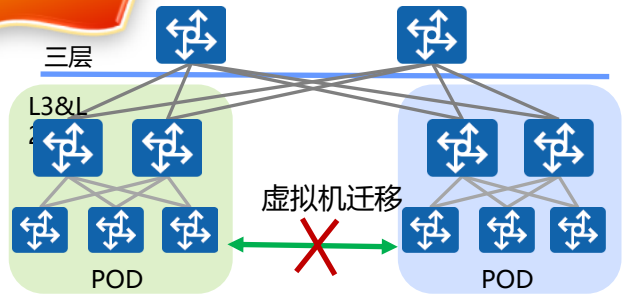




- 虚拟机摆脱地理位置的限制自由迁移，构建跨地理区域的大二层网络。

传统网络为何大不起来

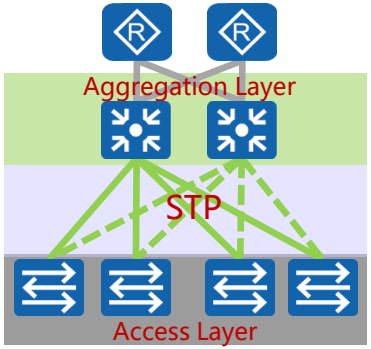
传统数据中心架构



- **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。

?

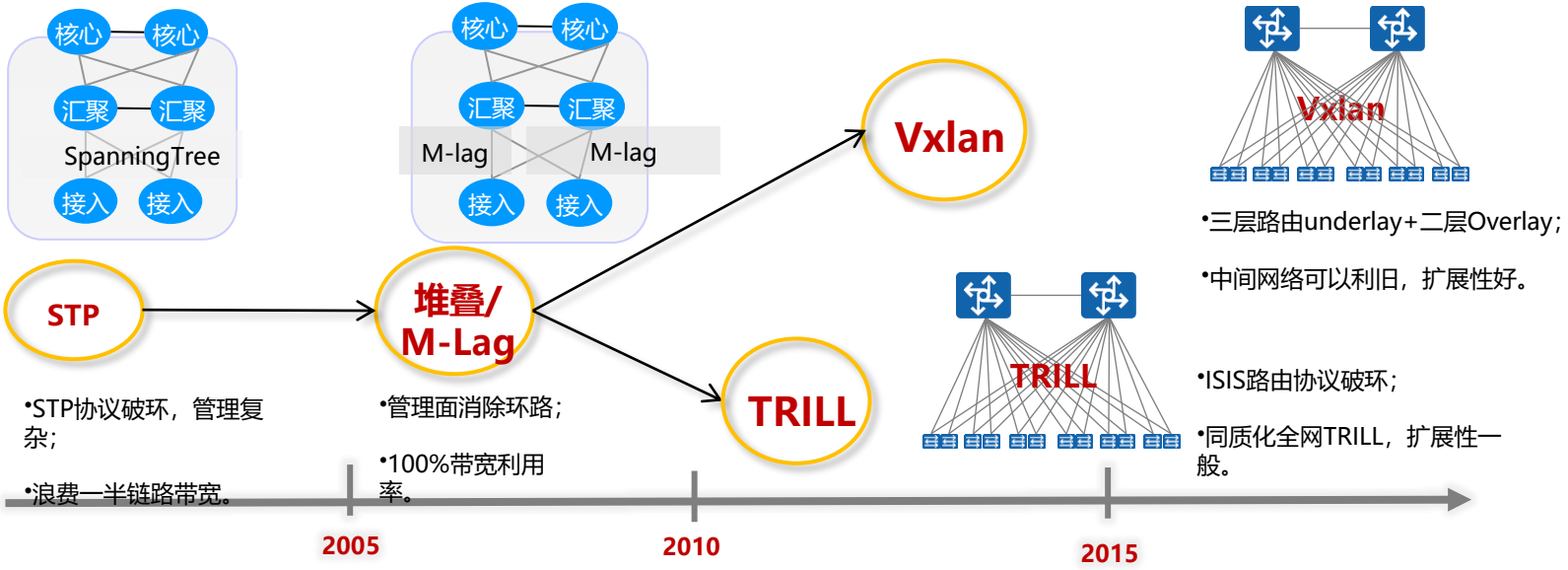
可不可以在现有VLAN技术下，把单个POD做大?



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



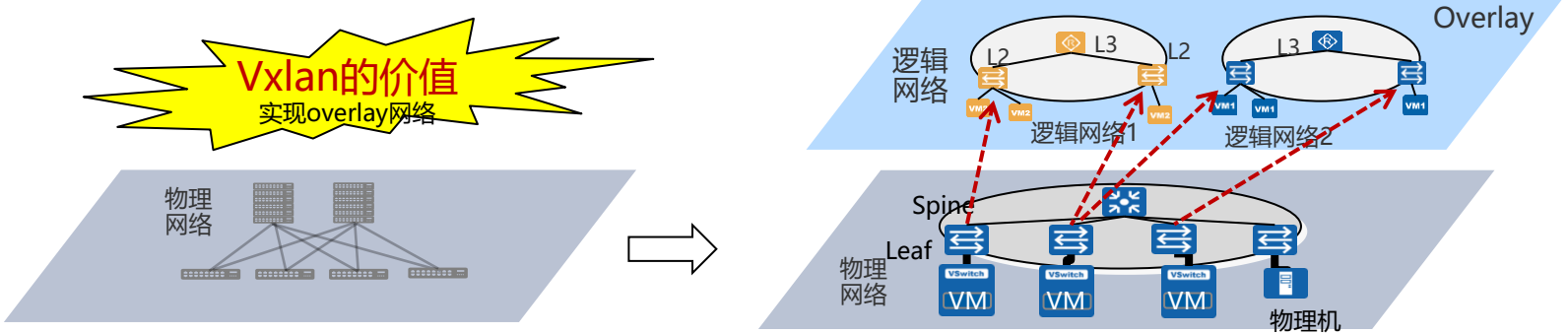
•TRILL技术解决了STP环路组网和规模问题，通过成熟的链路状态路由算法，扩展IS-IS协议，构建无环网络，实现多路径负载分担；

•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。

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：

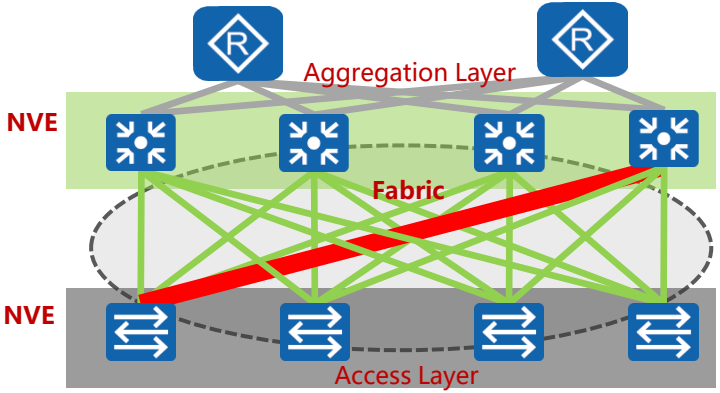
- VXLAN是业界标准的Overlay技术；
- 相比STP的主备路径， VXLAN利用Underlay网络的ECMP能带来更高的网络转发性能；
- VXLAN基于UDP技术构建， 是统一IT和CT两界的overlay技术；
- VXLAN将VLAN的4k子网扩充到16M，支持多租户；
- VXLAN为SDN提供转发面基础。

VXLAN基本概念

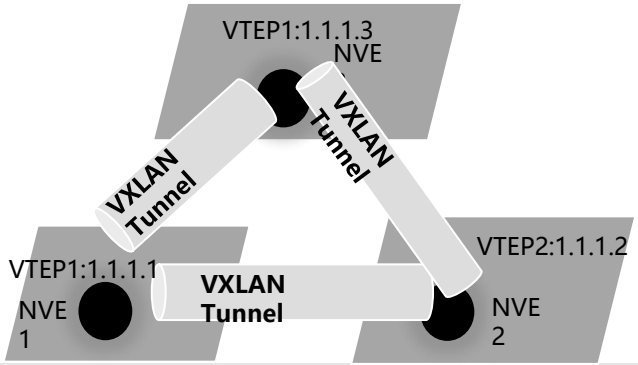
基于NVo3的二层Fabric组网

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。



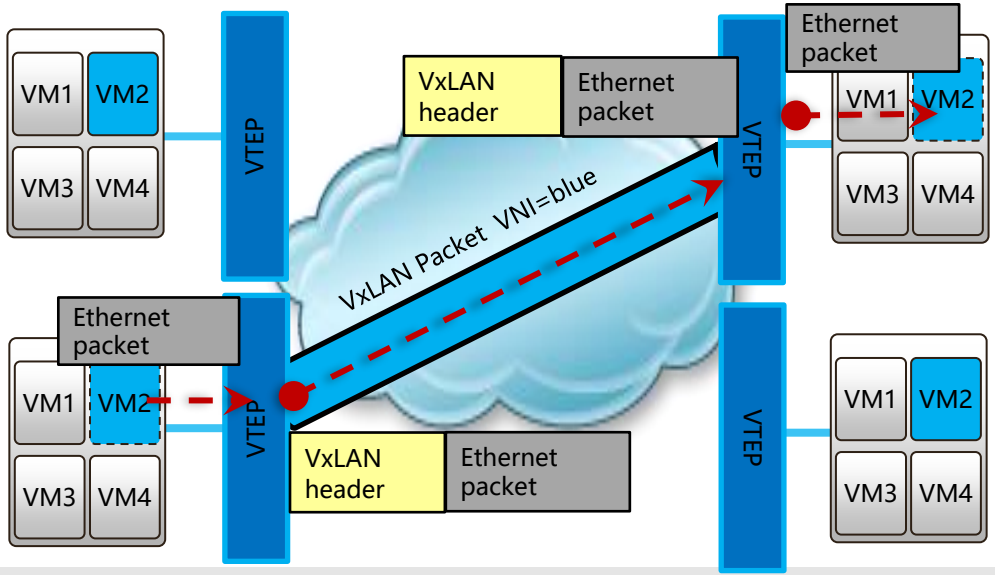
VXLAN(Virtual Extensible LAN，虚拟可扩展局域网)是目前NVO3中影响力最为广泛的一种。它通过LMAC in UDP的报文封装方式，实现基于IP overlay的虚拟局域网。



- 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。

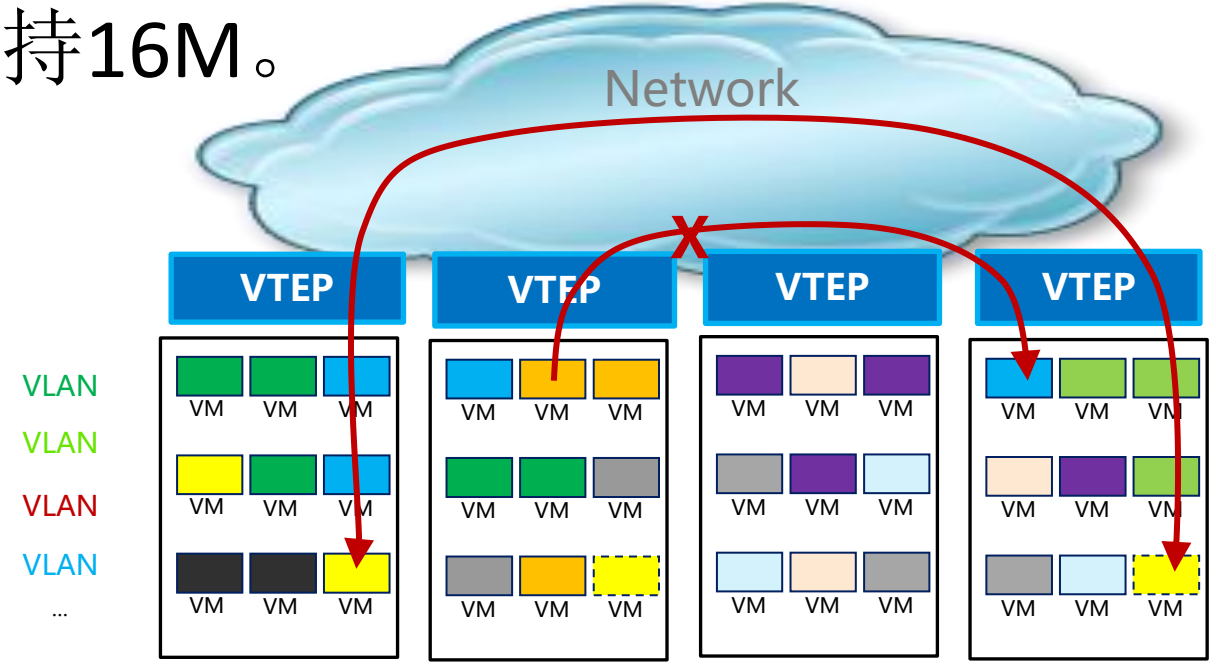
VXLAN 概念 – VTEP

- 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。

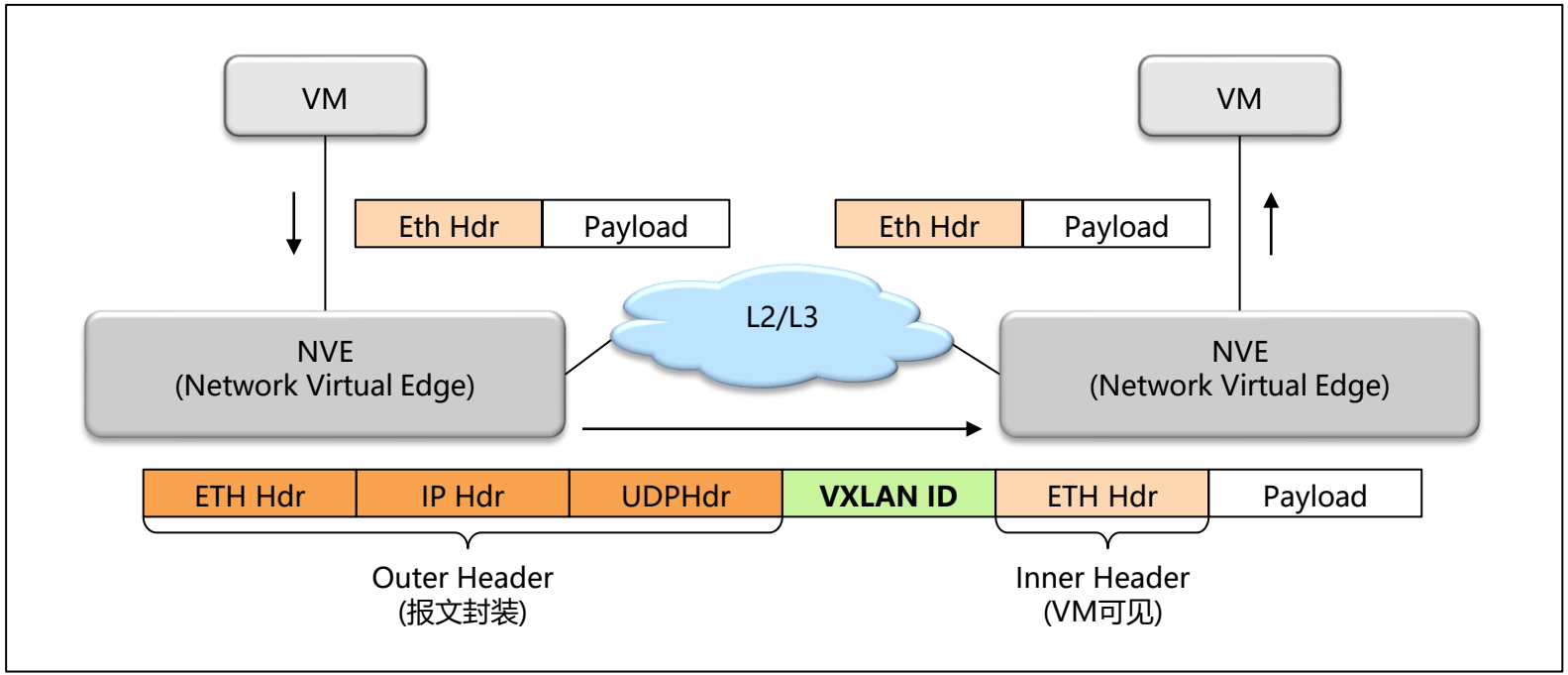


VXLAN – VNI

- VNI-24比特，用于标识虚拟网络，最大支持16M。

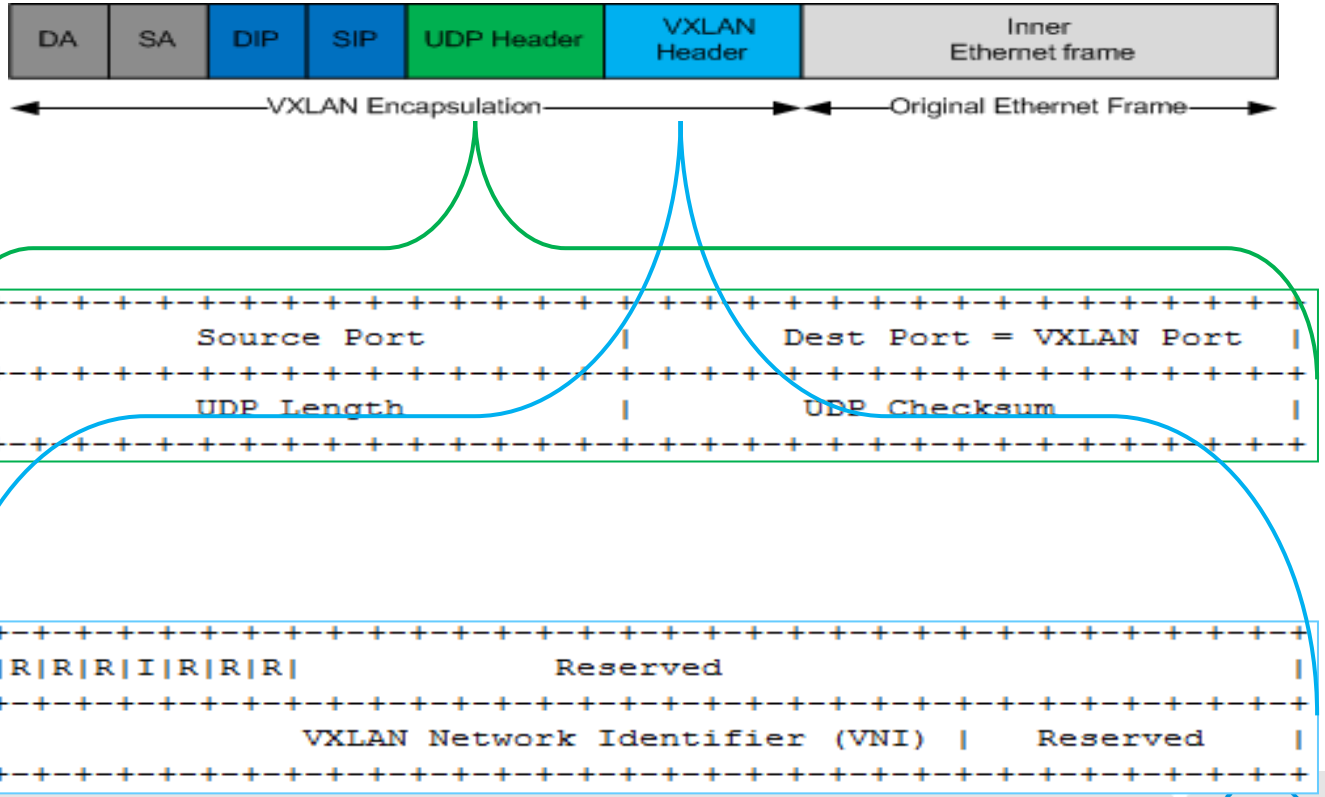


VXLAN 报文格式

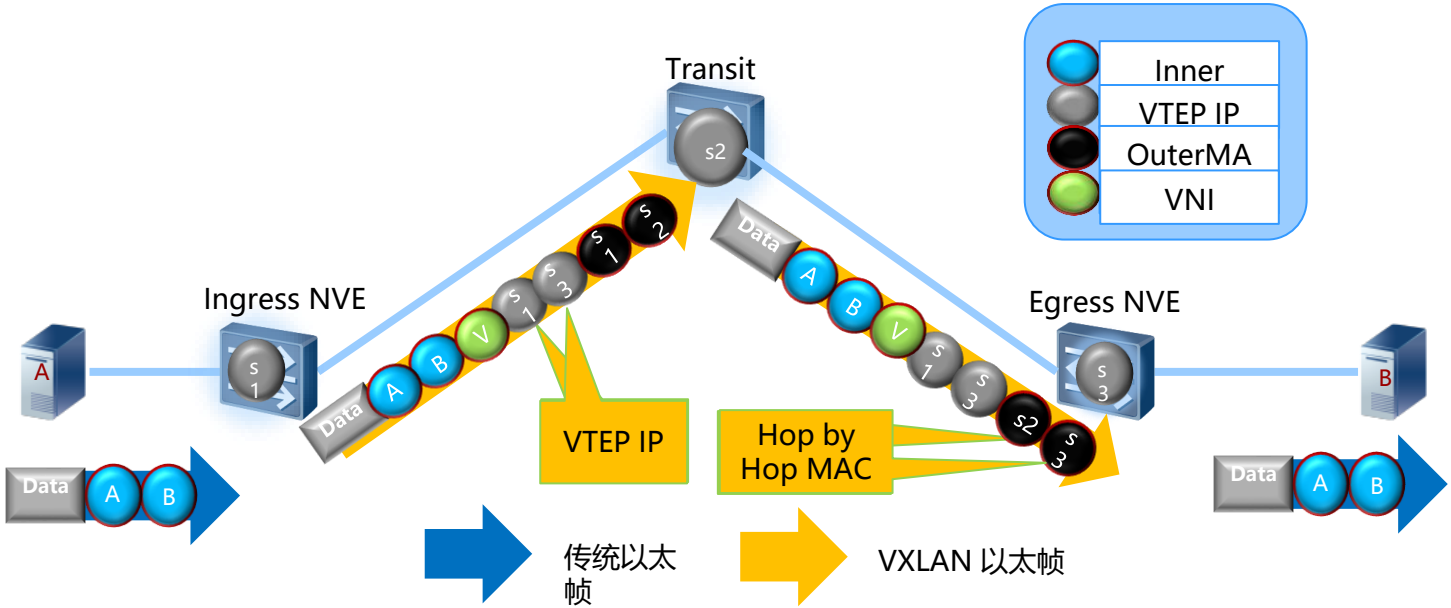


VxLAN 报文封装流程

VXLAN 报文格式

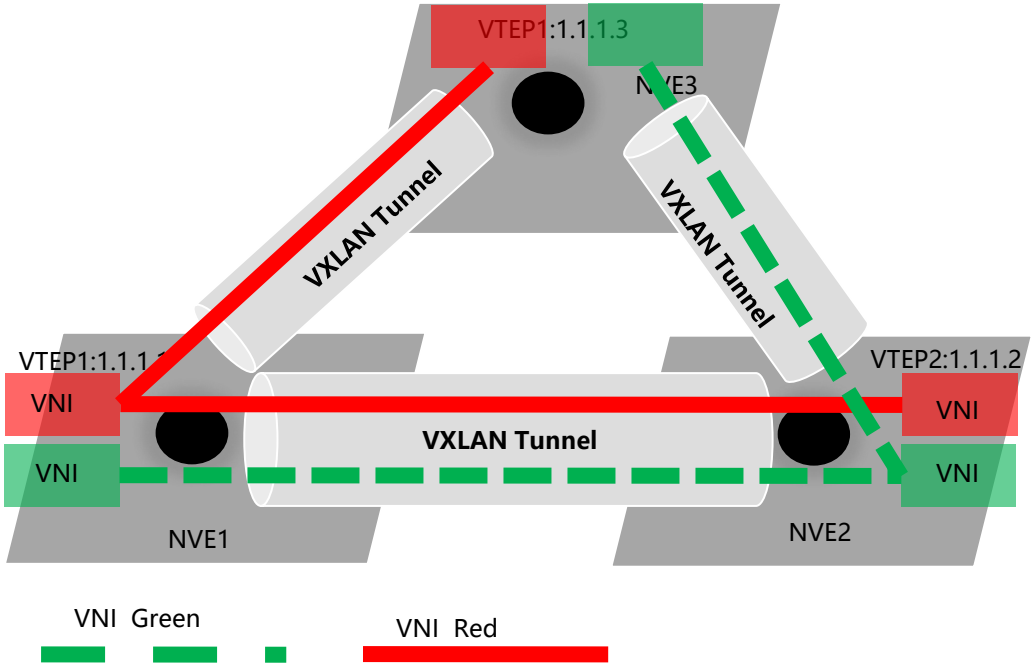


VXLAN 转发数据封装



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。

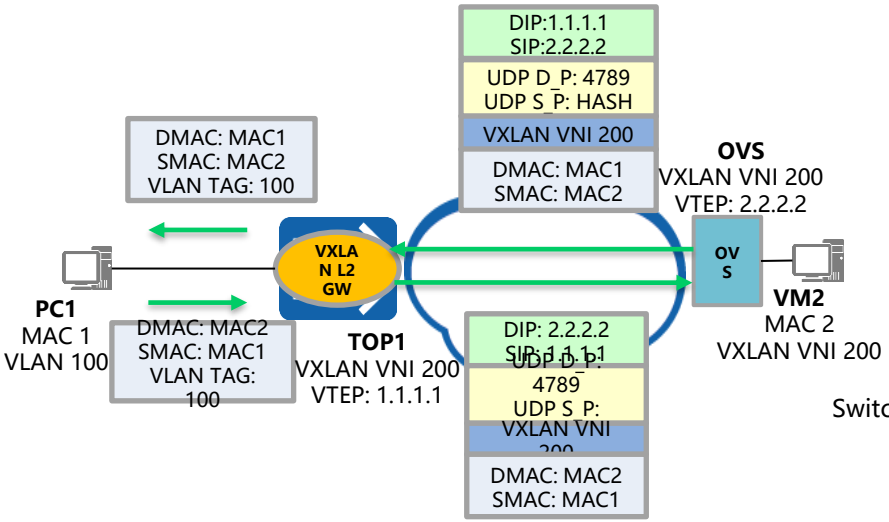
隧道和VNI关系

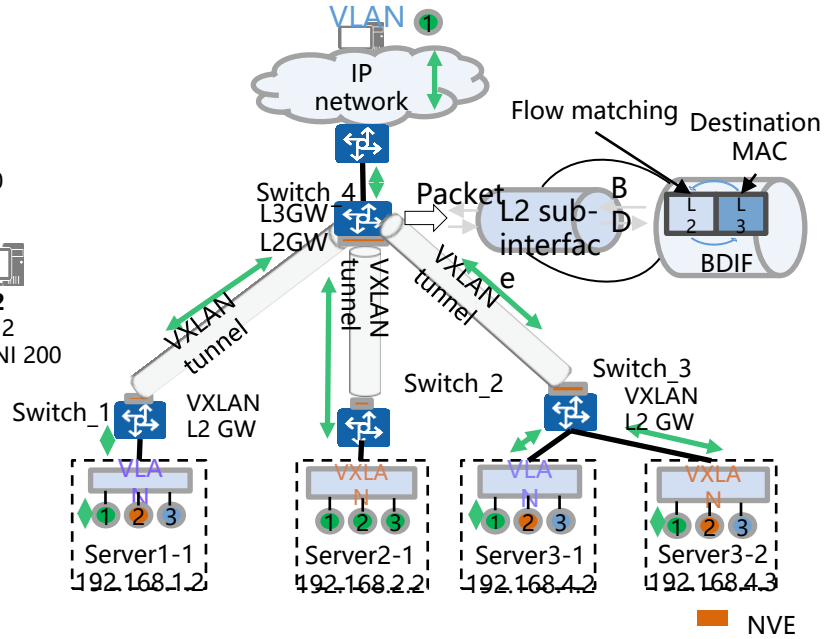


VNI概念

- 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



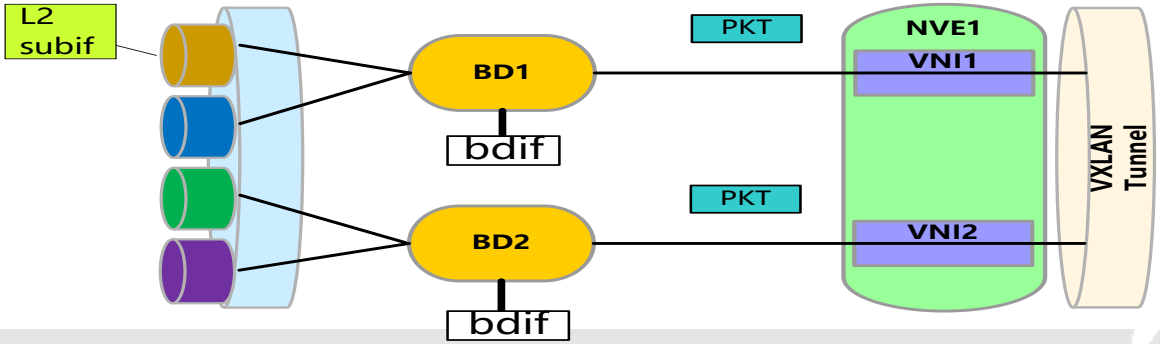


- **VxLAN L2 Gateway:** 允许租户接入VxLAN网络，实现相同VxLAN内部流量互访。

- **VxLAN L3 Gateway:** 实现不同VxLAN直接互访，或者VxLAN与非VxLAN网络互访。

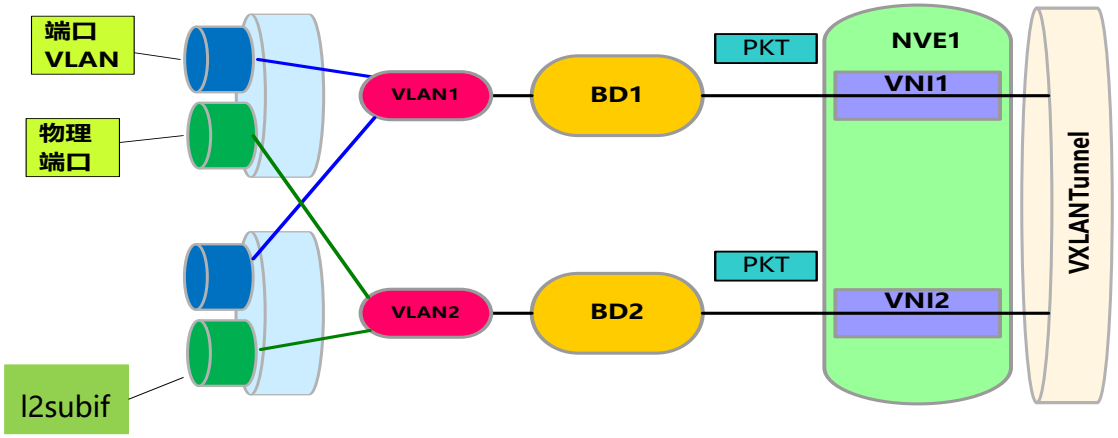
VXLAN接入业务模型（1）

- 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；

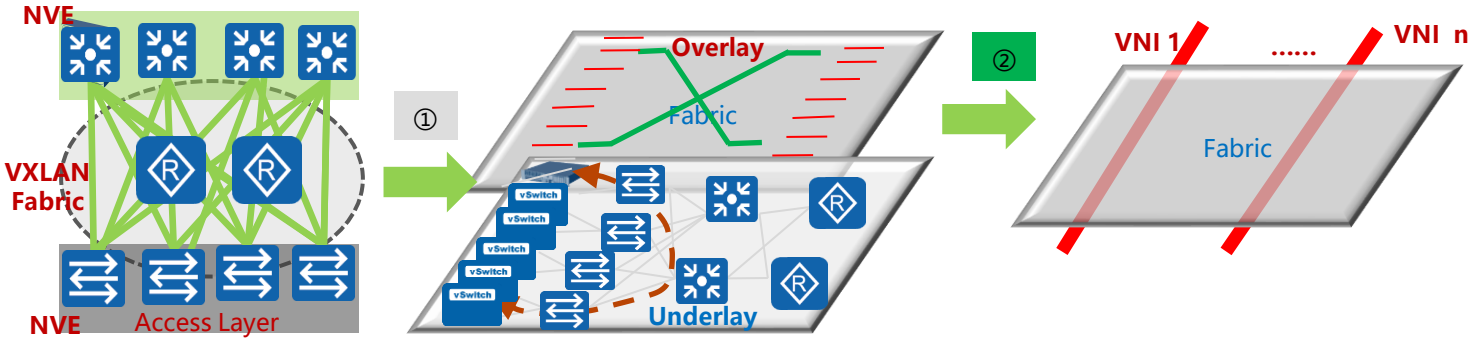


VXLAN接入业务模型（2）

- 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

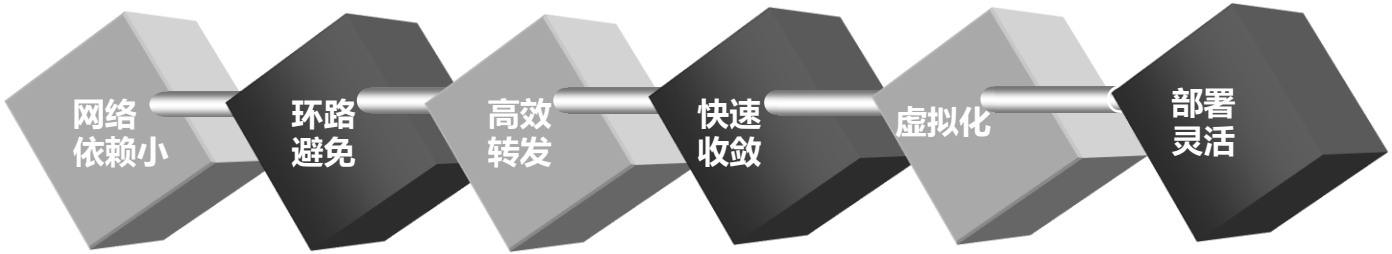


VXLAN的简化理解—---两次虚拟化

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



基于IP的overlay，仅需要边界设备间IP可达。

隧道间水平分割、IP overlay TTL避免环路。

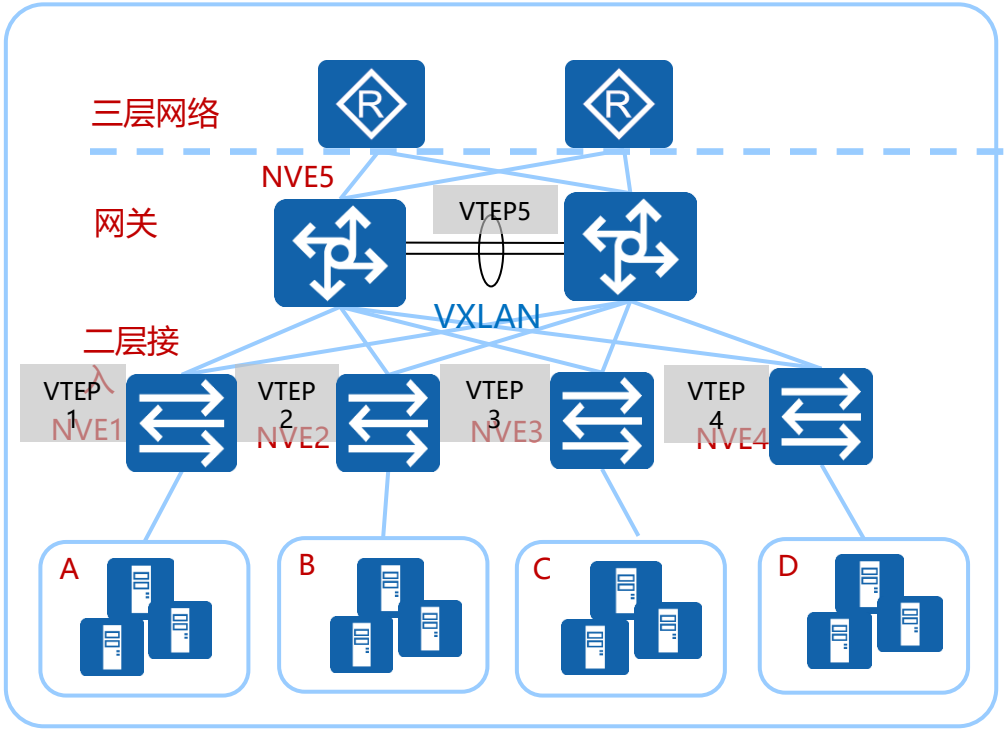
数据流量基于IP路由 SPF及ECMP快速转发。

网络变化实时侦听全网拓扑毫秒收敛。

Overlay+VNI构建虚拟网络，支持多达16M的虚拟网络。

物理设备、vSwitch均能够部署。

VXLAN同子网转发流程



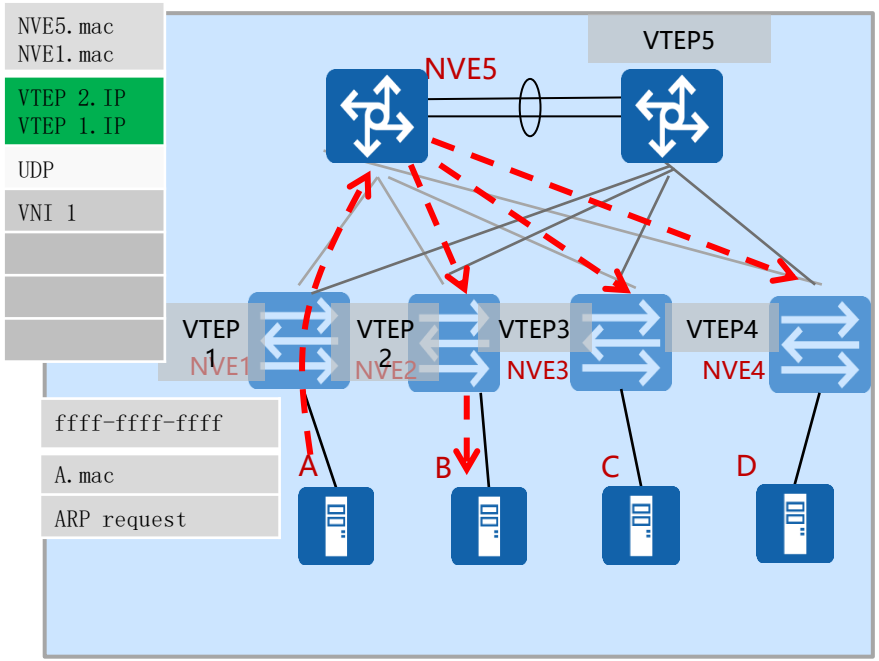
总体流程

- 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）

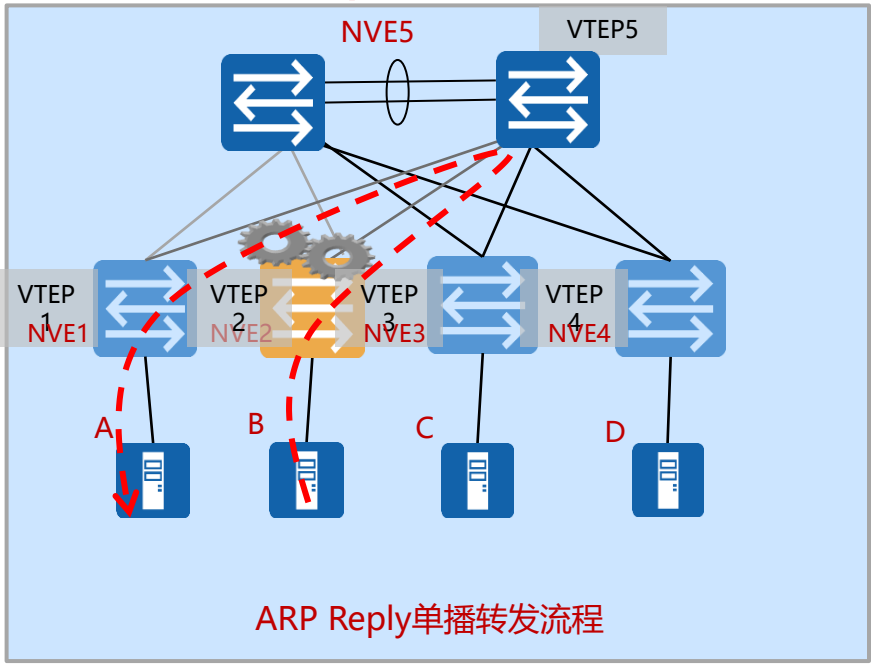
A到B的ARP request 广播



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）

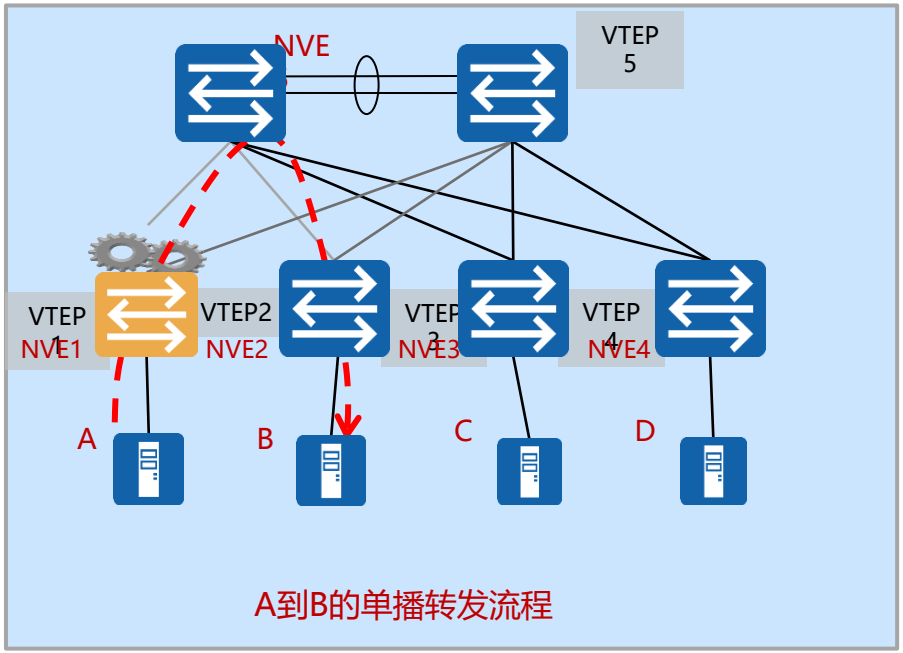
B回复A的ARP reply 单播



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。

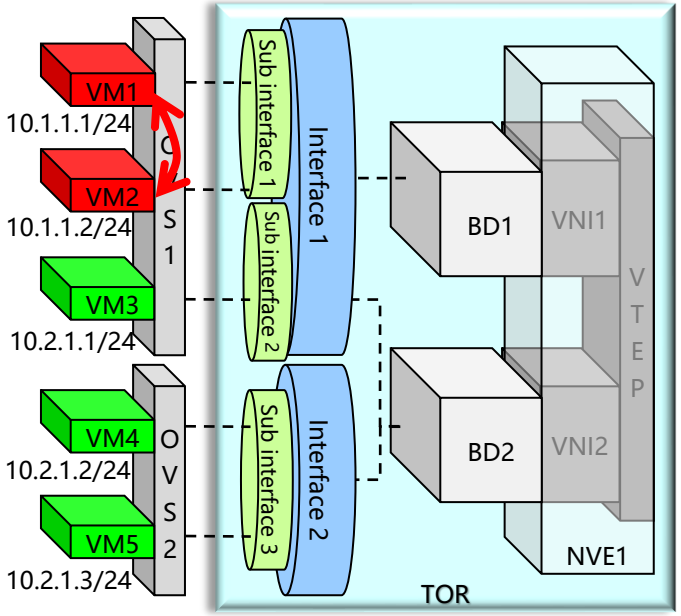
同网段查MAC二层转发（3）

A到B的单播数据报文

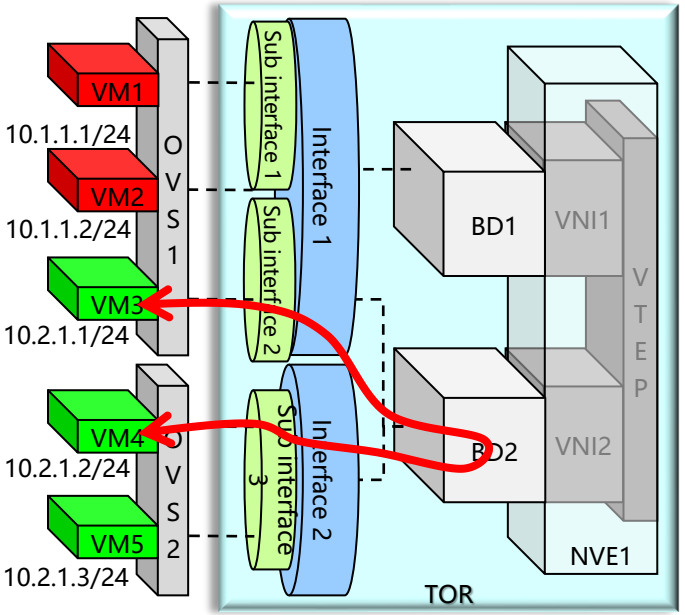


1. NVE1 和 NVE2都学习到了服务器A和B的MAC地址；后续查找MAC则命中；单播流程不同的是外层封装了隧道；underlay是IP转发。
2.

VXLAN转发模型之相同网段VM互访（1）

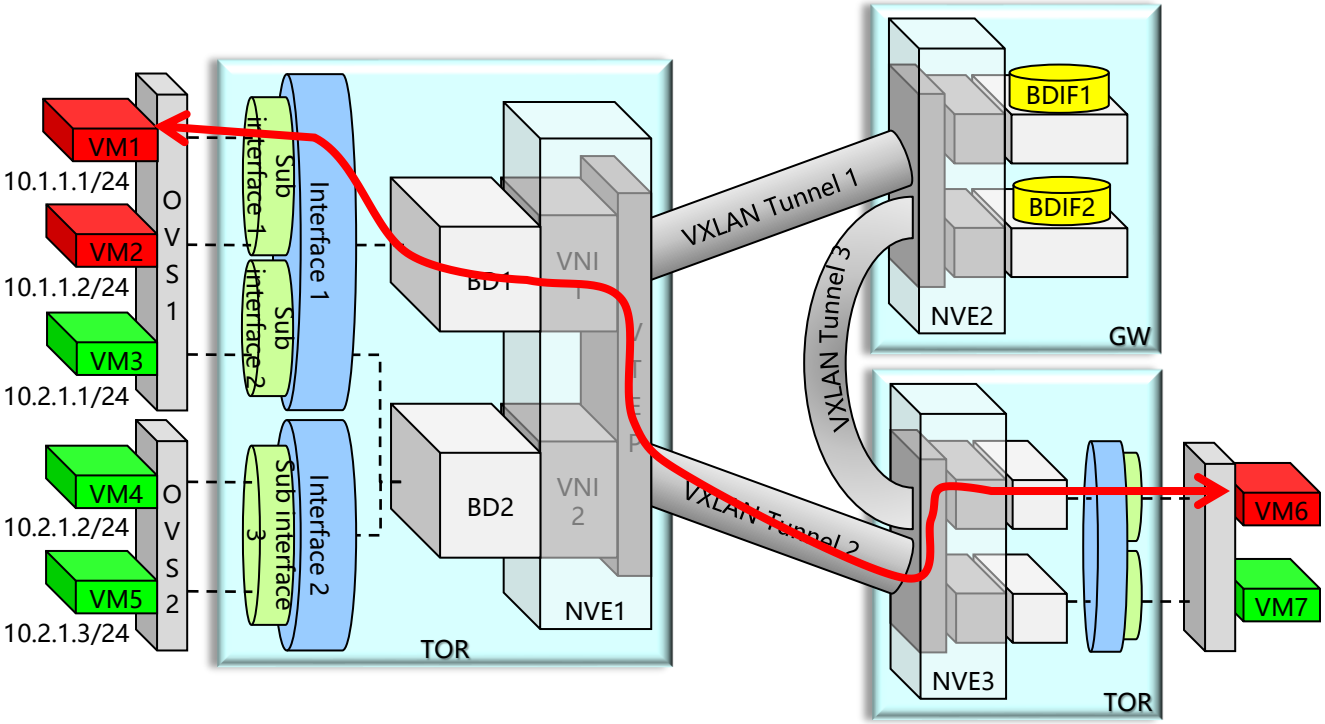


Scenario 1: Both VMs located at the same vSwitches connected to same TOR



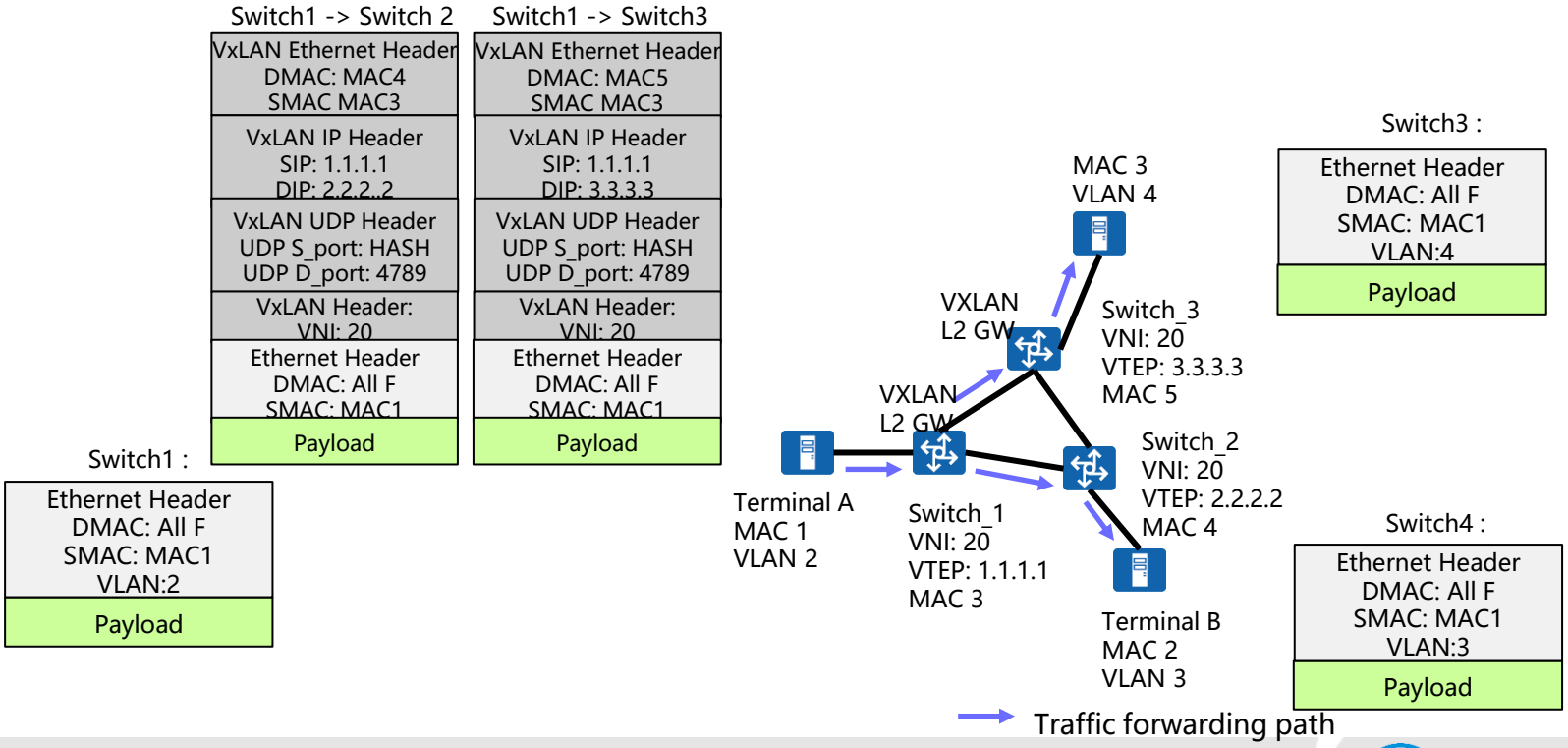
Scenario 2: Both VMs located at different vSwitches connected to same TOR

VXLAN转发模型之相同网段VM互访（2）

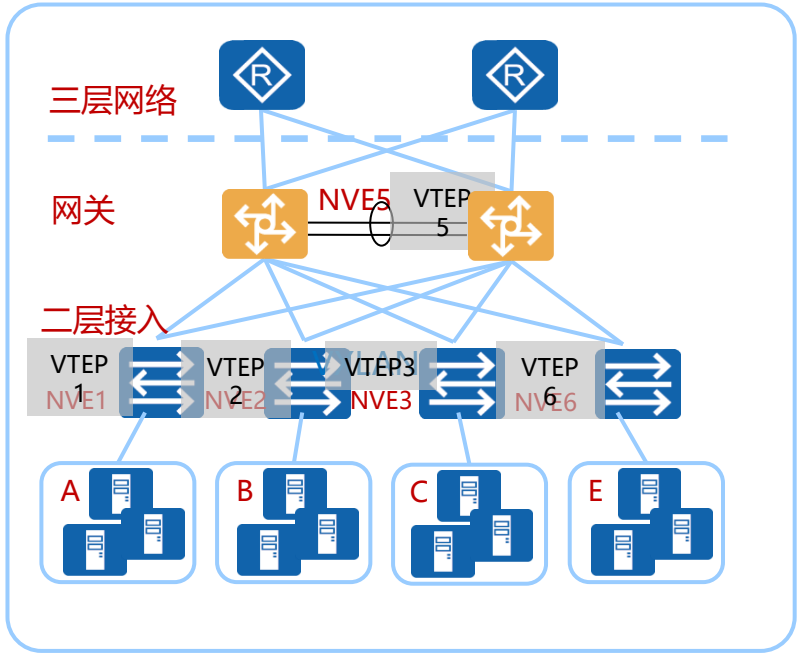


Scenario 3: Both VMs located at different vSwitches connected to different TOR

VXLAN - BUM 报文转发流程



VXLAN数据转发总体流程（跨子网）

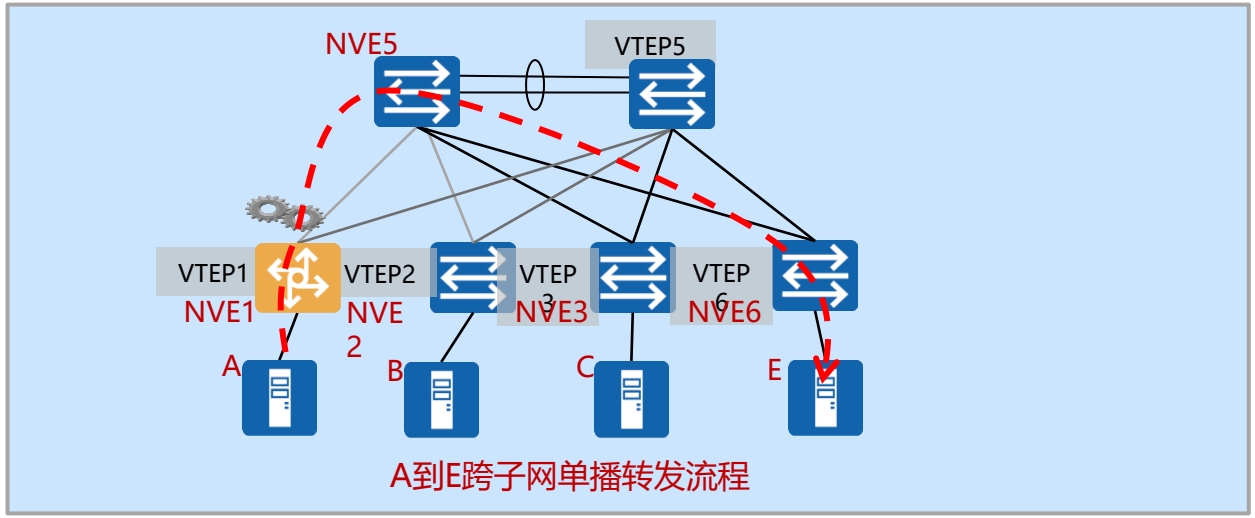


总体流程

- HOST A发送单播数据报文给HOST E。

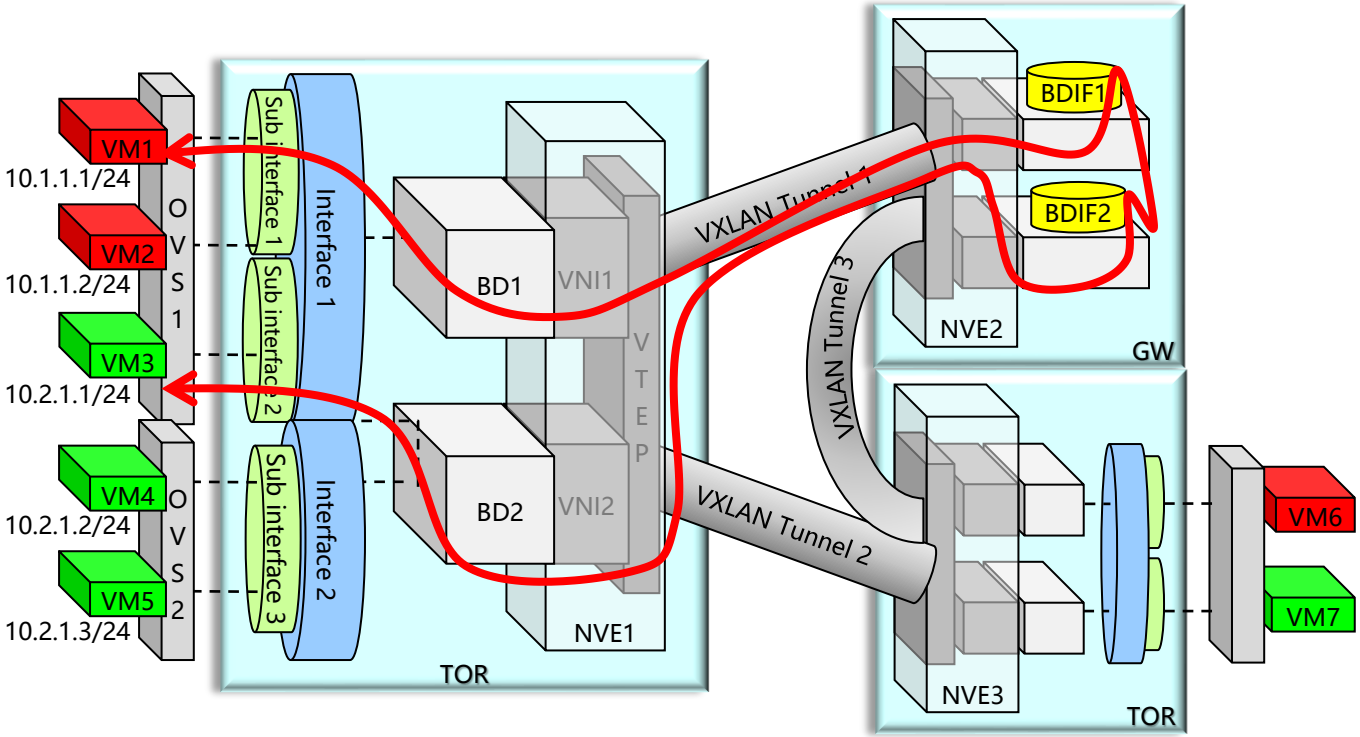
注：NVE5作为三层网关，HOST A属于VNI 1，HOST E属于VNI 2，默认主机与网关都互相学习到ARP表，各个节点MAC都已学习。

A到E单播转发流程



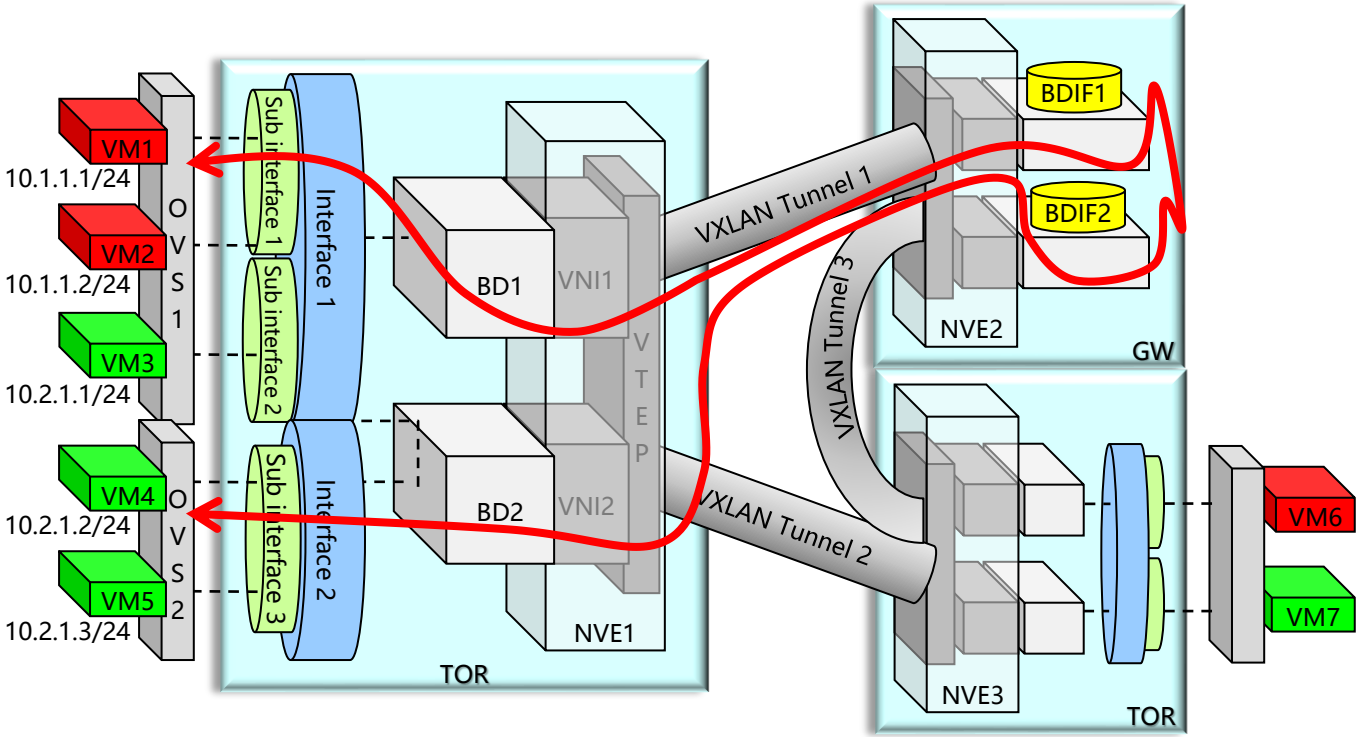
1. NVE1查找网关mac转发表，封装隧道；使用VNI1；
2. 网关解封装报文，根据内层IP头查路由，替换内层以太头，封装VXLAN头部，使用VNI2；
3. NVE6 接收报文并解封装，内层报文根据目的MAC转发。

VXLAN转发模型之不同网段VM互访 (1)



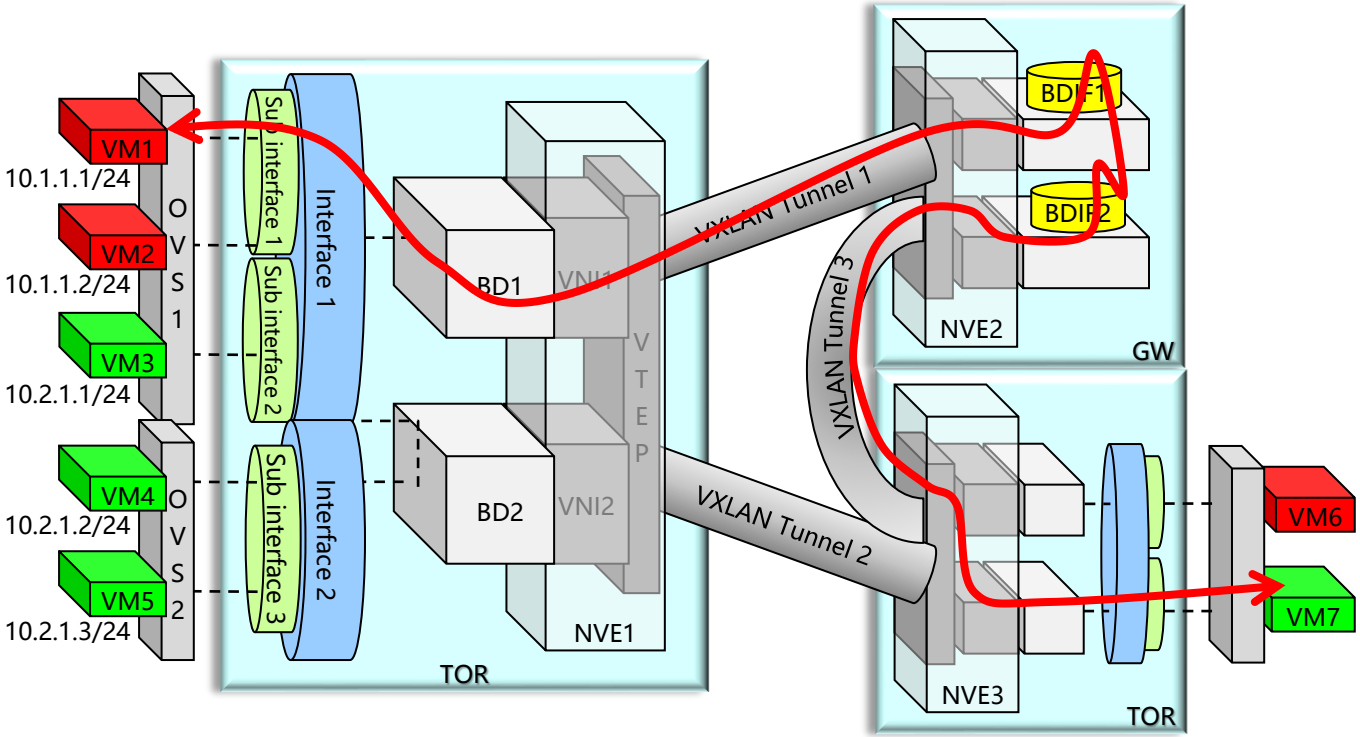
Scenario 1: Both VMs located at the same vSwitches connected to same TOR

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

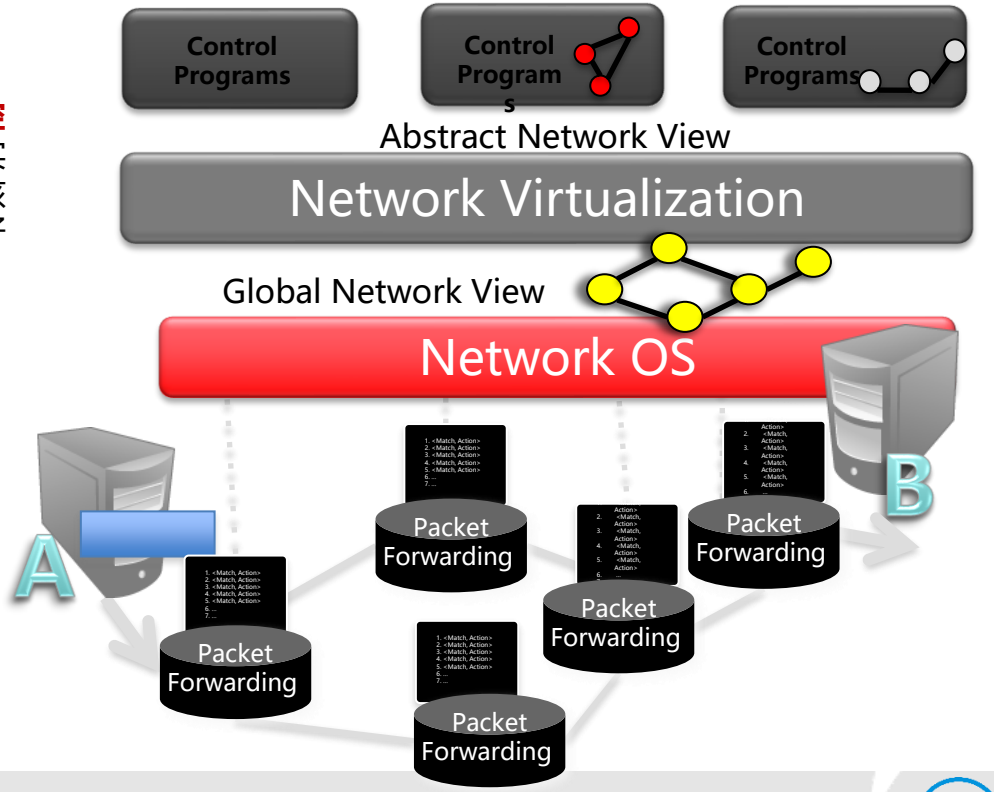
SDN起源

斯坦福大学尼克.麦吉翁教授等人发明**OpenFlow**协议，通过**Controller**集中管控网络转发行为

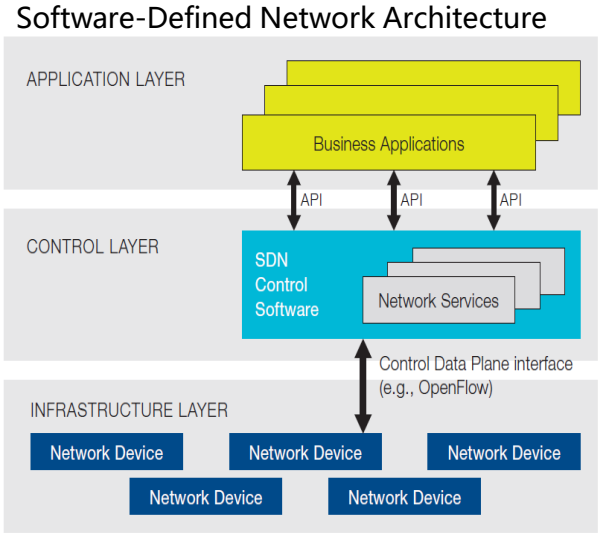
斯坦福大学Nick McKeown教授团队的Clean Slate项目成员

SDN的定义

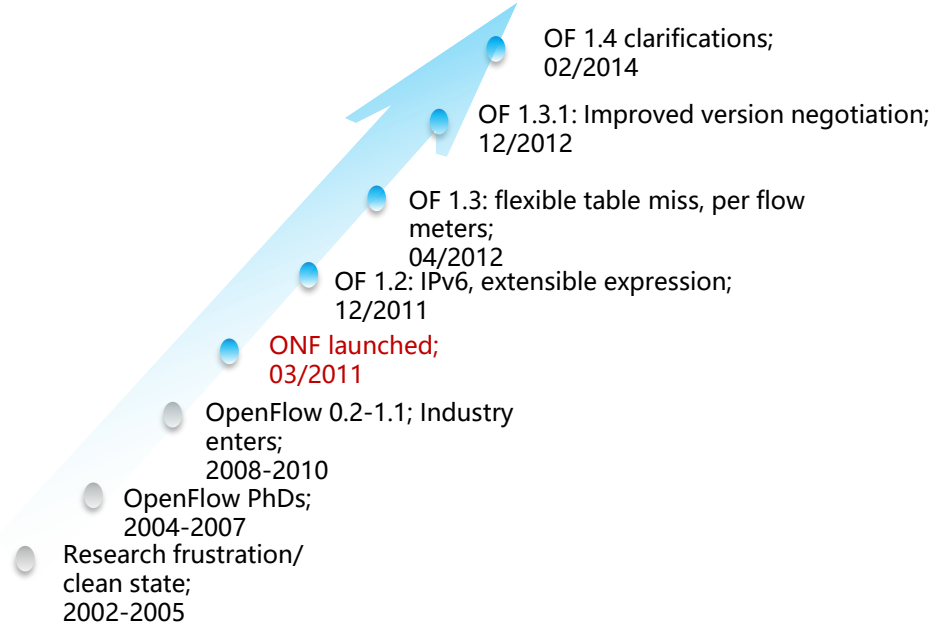
SDN（Software Defined Network），即软件定义网络，其核心技术是通过将网络**设备控制面**与**数据面**分离开来，从而实现了网络流量的灵活控制，为核心网络及应用的创新提供了良好的平台。



ONF定义的SDN基本架构

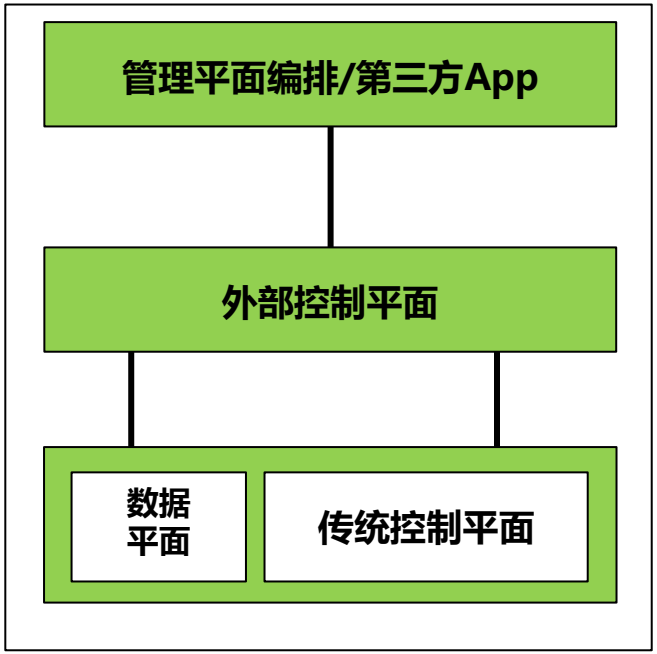


Source: ONF white paper, April 13, 2012



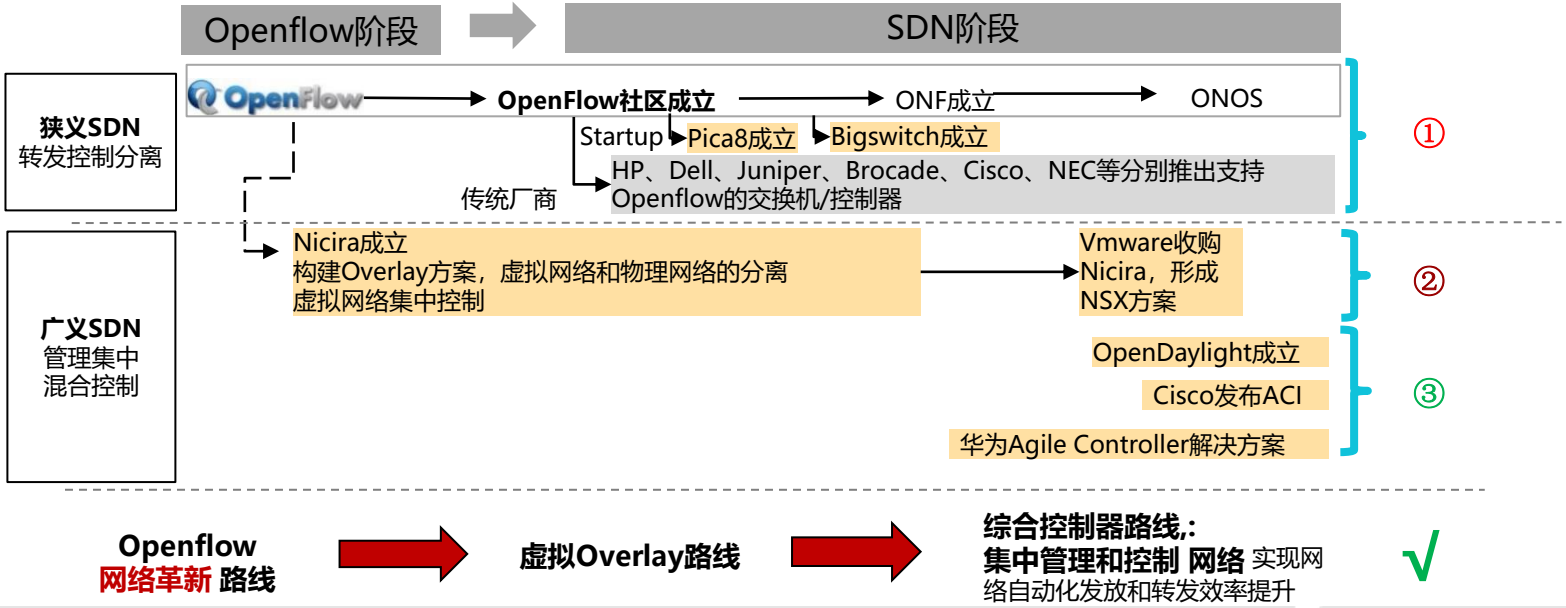
ONF强调OpenFlow-based SDN，强调控制与转发分离以实现转发设备的标准化，重点是OF协议标准化。

传统CT厂商眼中的SDN架构

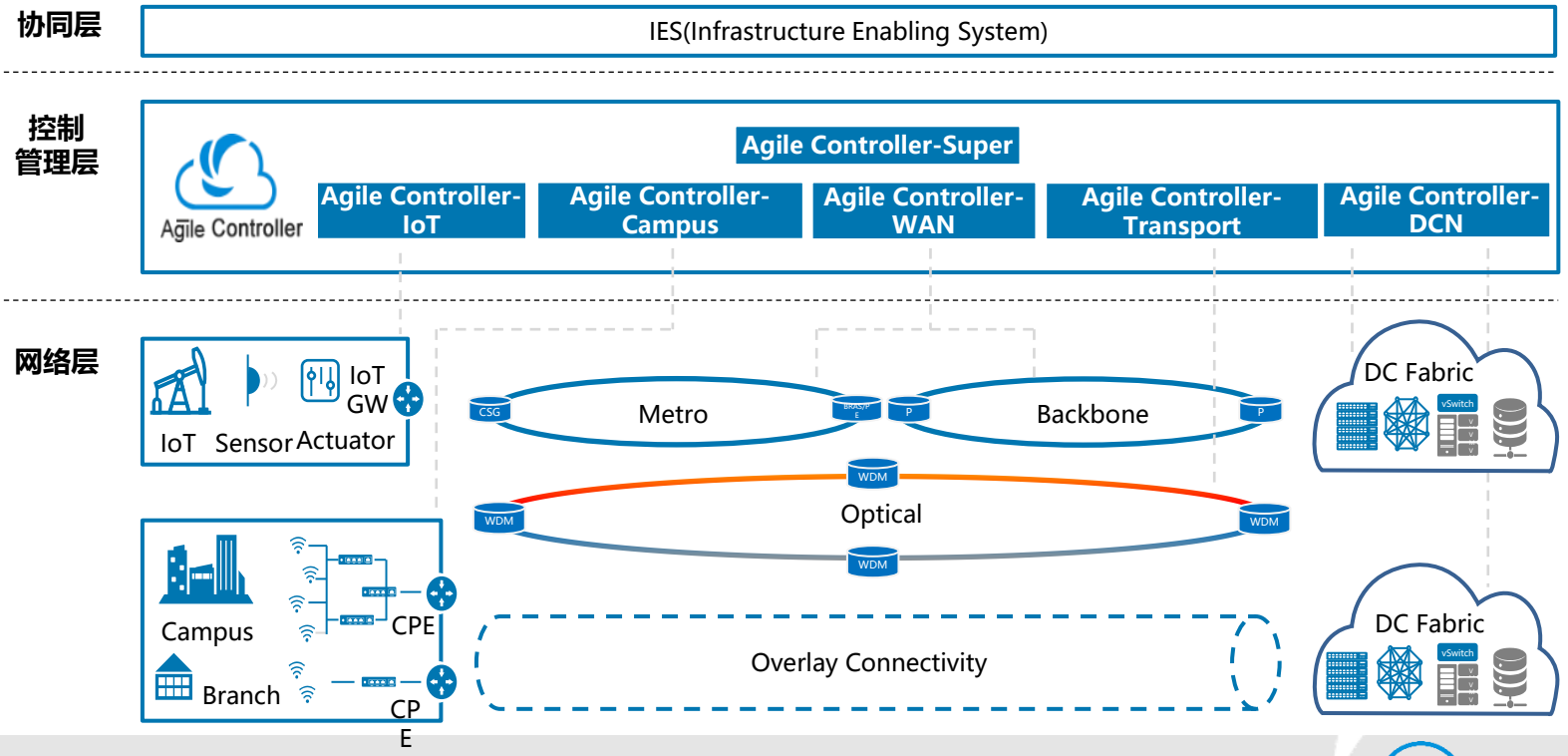


SDN主要技术路线

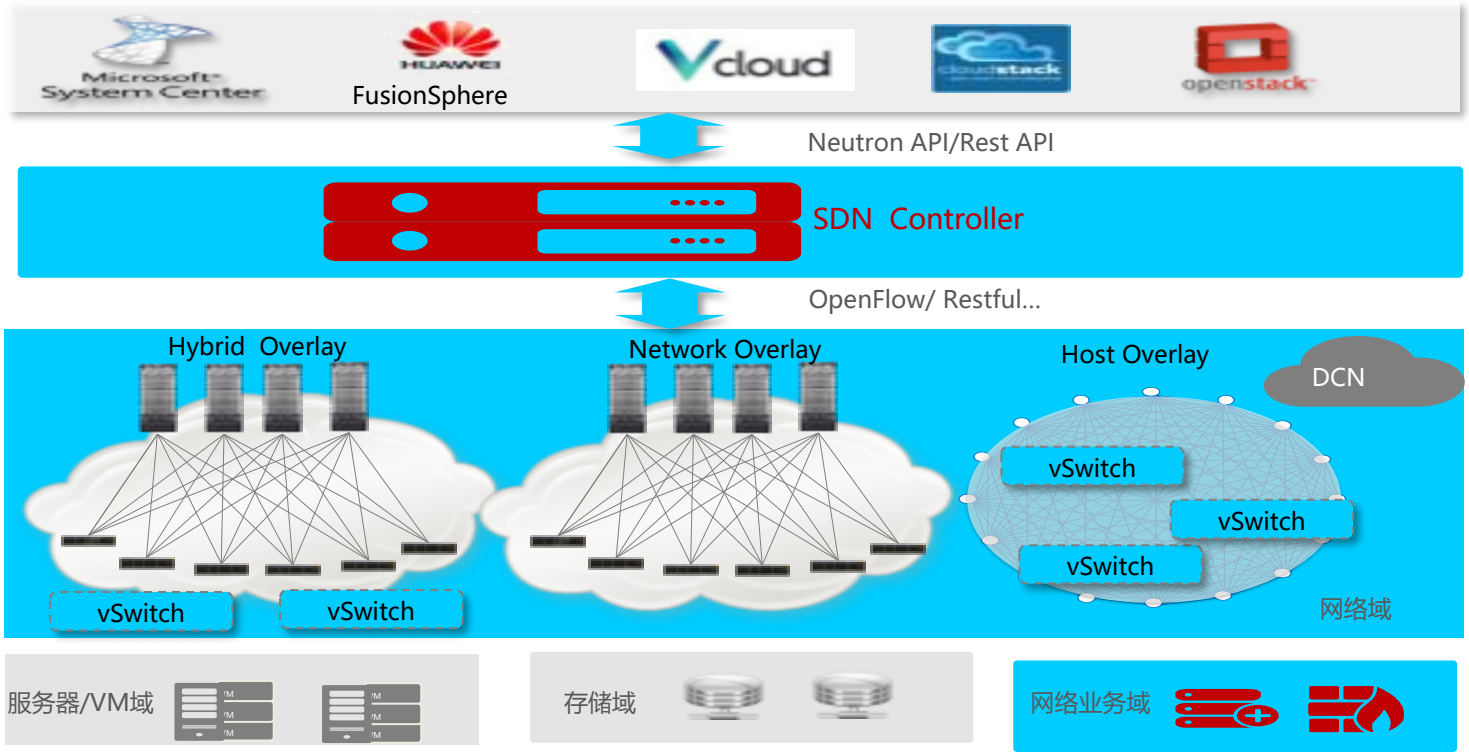
- 2006年斯坦福大学发布 OpenFlow，网络设备转发和控制面分离，通过集中的控制面实现网络流量的灵活控制
- 华为数据中心SDN的核心特征是适度的转控分离结合之外，通过管理与控制分离，实现网络业务自动化发放，助力数据中心业务实现敏捷发放.



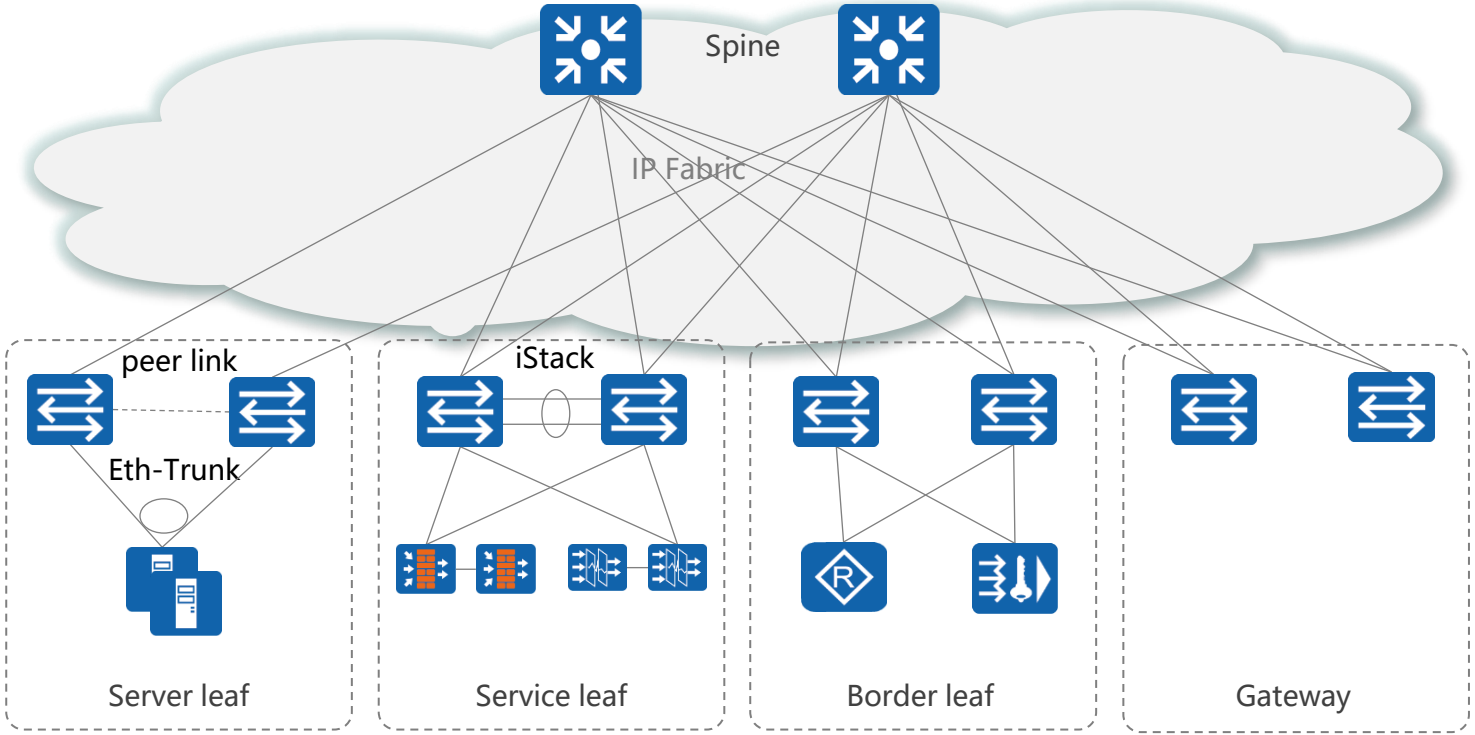
华为SDN解决方案全景图



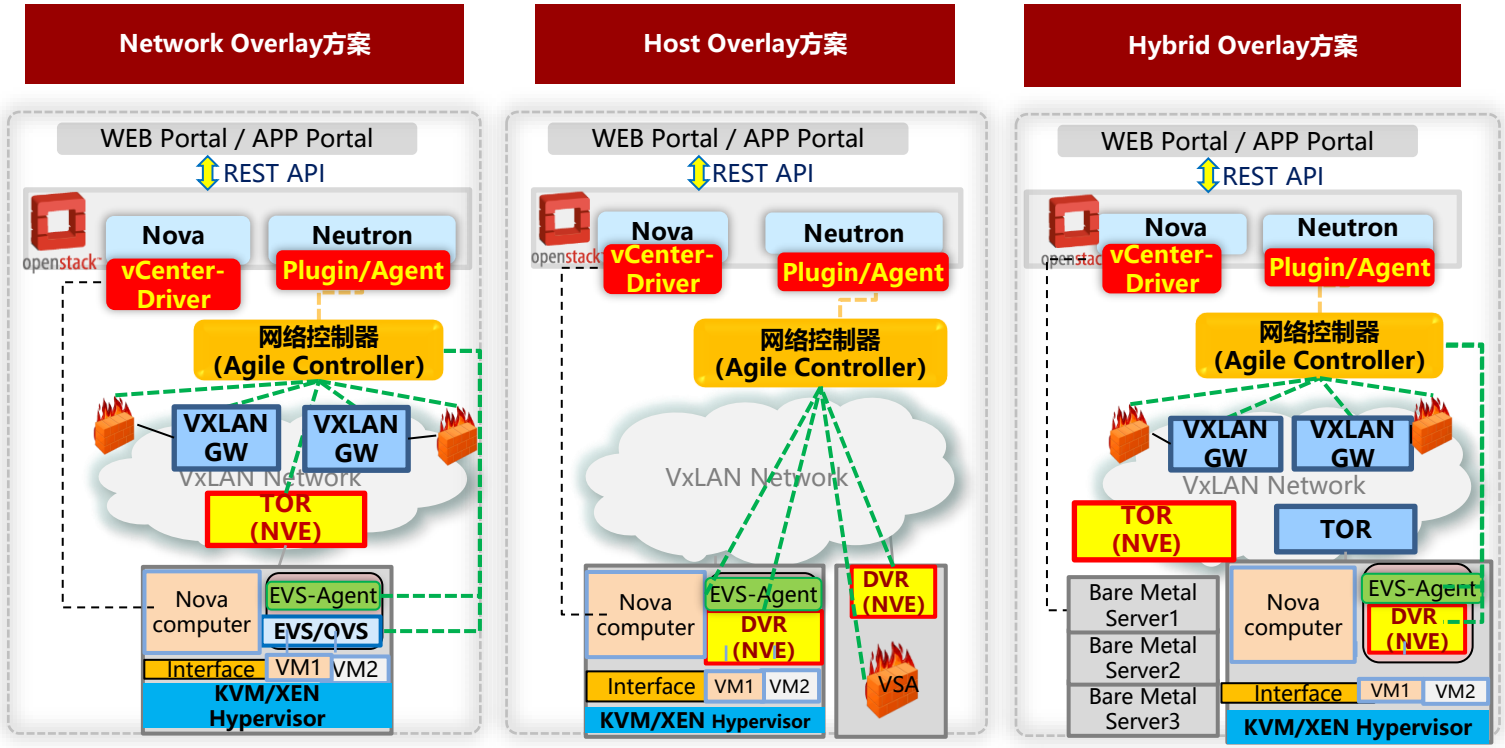
SDN DCN解决方案



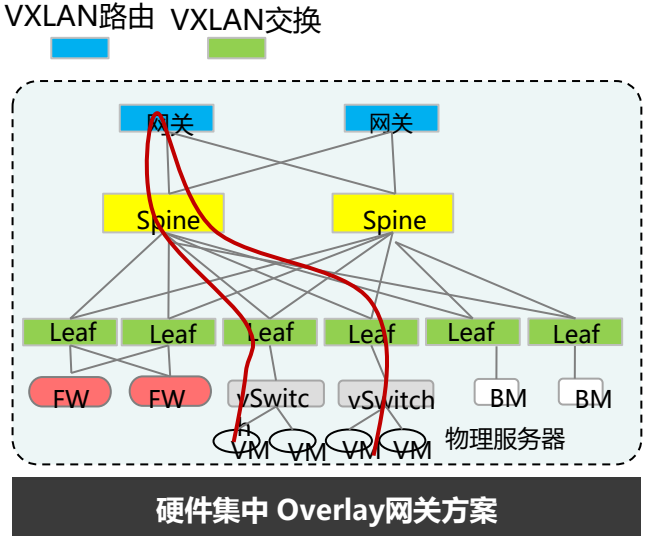
VXLAN 基于Spine-Leaf组网架构



VXLAN三种OverLay组网方案

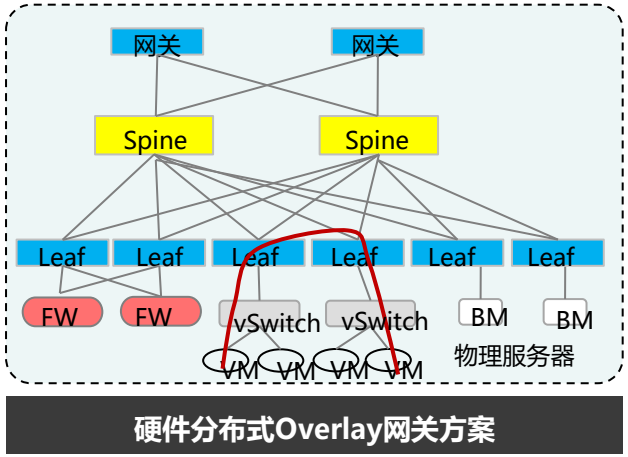


VXLAN OverLay的分布式和集中式



硬件集中 Overlay网关方案

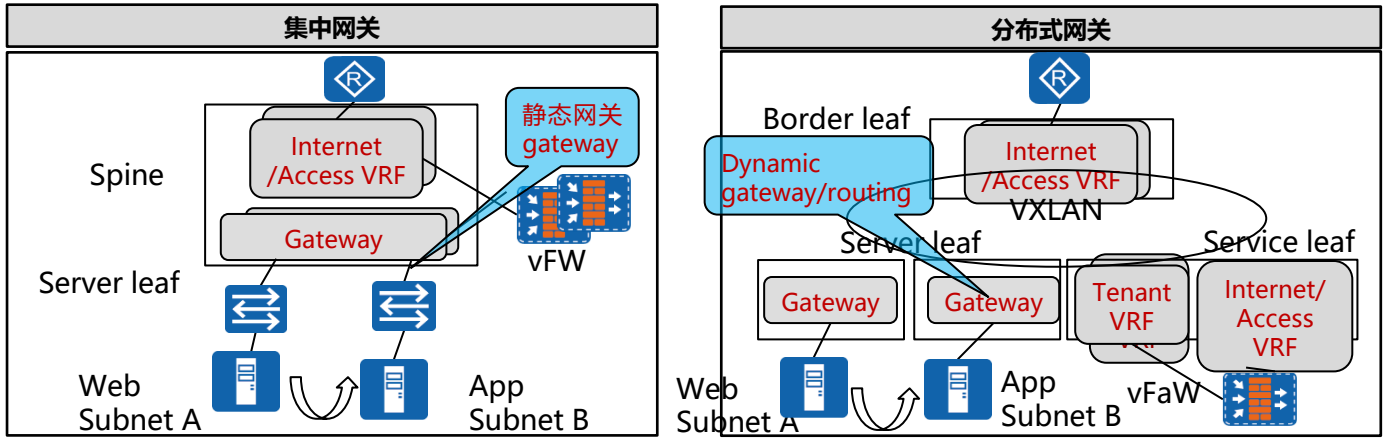
- VXLAN二层VTEP功能：部署在Leaf
- VXLAN三层网关功能：部署在核心层
- Spine：普通路由



硬件分布式Overlay网关方案

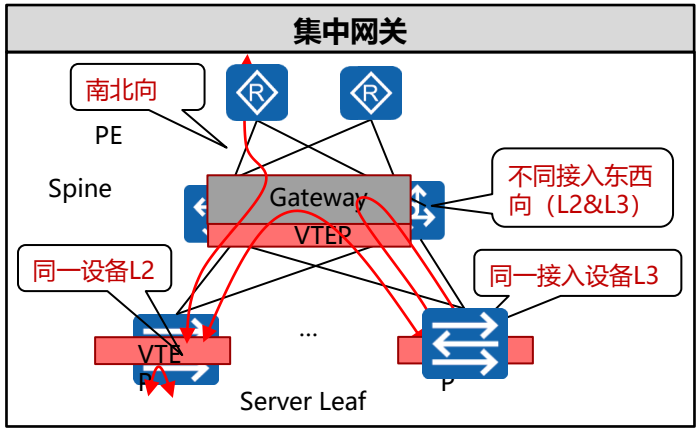
- Leaf 节点既是VxLAN二层VTEP网关，又是东西向流量的三层VxLAN网关
- 南北向流量的网关部署在核心层
- Spine：普通路由

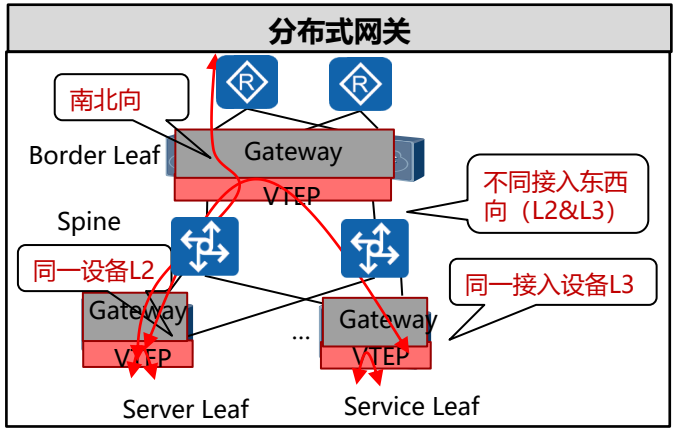
网关部署对比



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

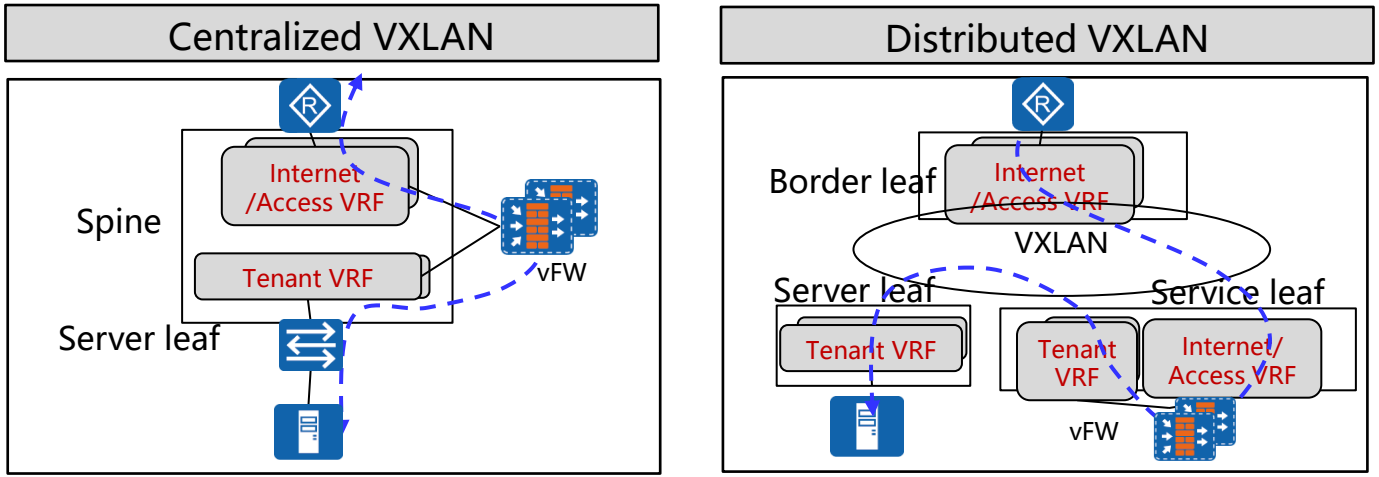
转发路径优化对比





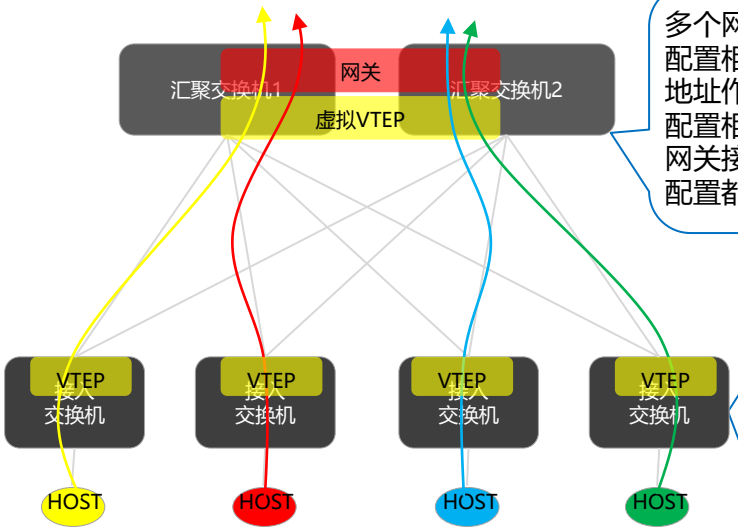
比较项	集中部署	分布式部署
南北向流量(customer to servers)	经过核心	经过核心
不同的接入设备间东西向 流量(L2&L3)	经过核心	经过核心
同一个接入设备下东西向流量(L2)	本地转发	本地转发
同一个接入设备下东西向流量(L3)	经过核心	本地转发

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；

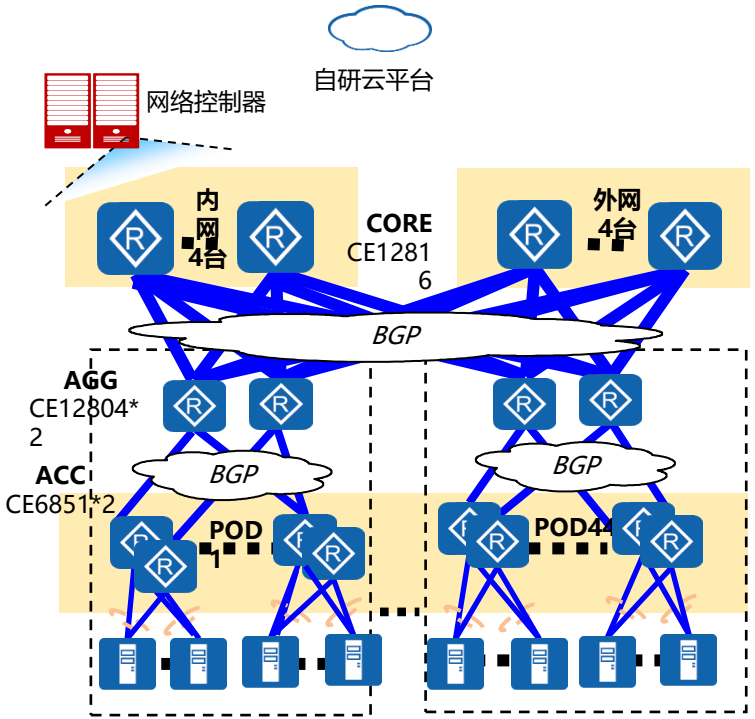
3、网关上行口全故障，依靠OPS联动将VTEP地址发布的优先级降低，保证上行流不发到故障网关。

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。

THANK YOU

Ping 通您的梦想～

腾讯课堂交流群：17942636

ADD：苏州市干将东路666号和基广场401-402；Tel：0512-8188 8288；

课程咨询QQ：2853771087 ； 官网 :www.51glab.com