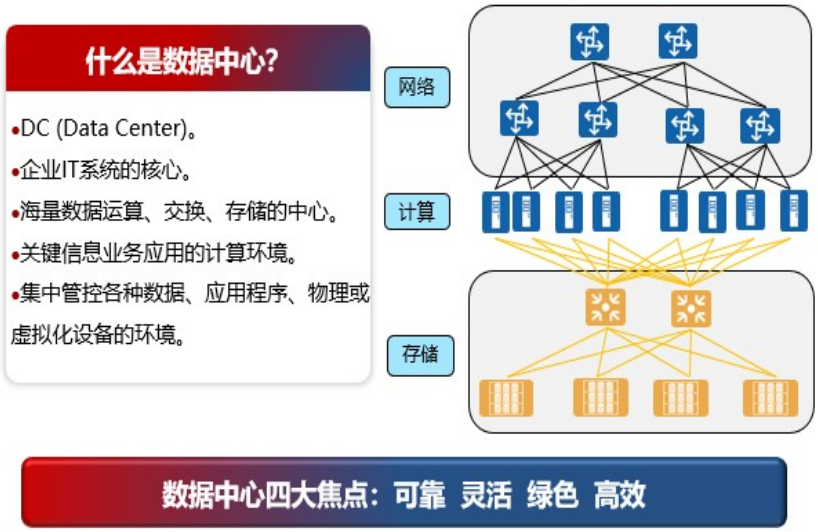
VXLAN：Virtual eXtensible Local Area Network 虚拟可扩展局域网

前言

- 服务器虚拟化能够大幅降低IT建设运维成本，提高业务部署灵活性。
- 虚拟机在传统数据中心网络中只能在二层网络中进行无缝迁移，一旦在跨三层网络中进行迁移，就会造成业务中断。
- 于是VXLAN技术应运而生，大大提高了虚拟机迁移的灵活性，使海量租户不受网络IP地址变更和广播域限制的影响，同时也大大降低了网络管理的难度。

数据中心的基本概念及特点



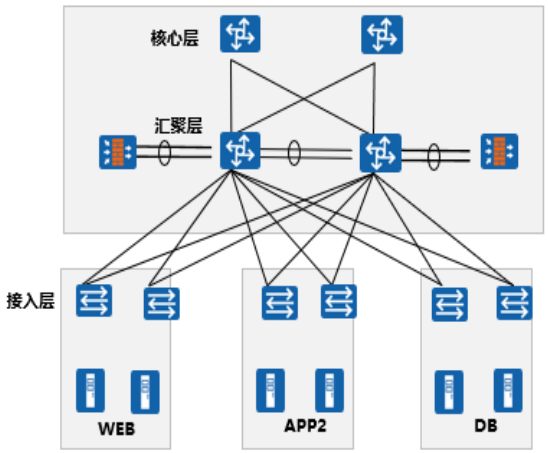
- 数据中心（Data Center）是一套完整、复杂的集合系统，它不仅包括计算机系统和其它与之配套的设备（例如通信和存

储系统），还包含数据通信系统、环境控制设备、监控设备以及各种安全装置。

- 数据中心通常是指在一个物理空间内实现信息集中处理、存储、传输、交换、管理的场所。
- 服务器、网络设备、存储设备等通常都是数据中心的关键设备。
- 设备运行所需要的环境因素，如供电系统、制冷系统、机柜系统、消防系统、监控系统等通常都被认为是关键物理基础设施。
- 互联网数据中心（Internet Data Center，IDC）是互联网中数据存储和处理的中心，是互联网中数据交互最为集中的地方。

传统数据中心网络结构

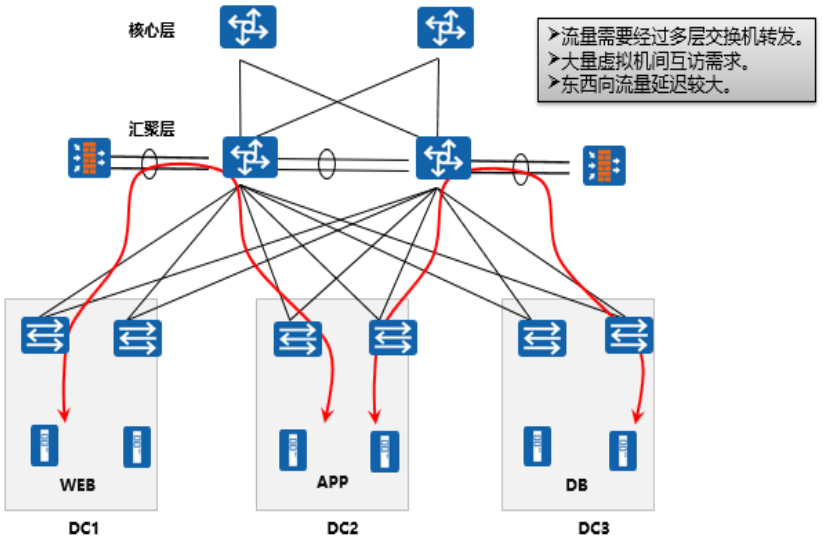
- 数据中心按不同业务功能进行分区。
- 传统网络是标准的三层网络架构。



- 传统网络模型在很长一段时间内，支撑了各种类型的数据中心。
- 按照功能模块划分，传统数据中心可分为核心区、外网服务器区、内网服务器区、互联网服务器区、数据中心管理区、数据交换&测试服务器区、数据存储功能区、数据容灾功能区等。

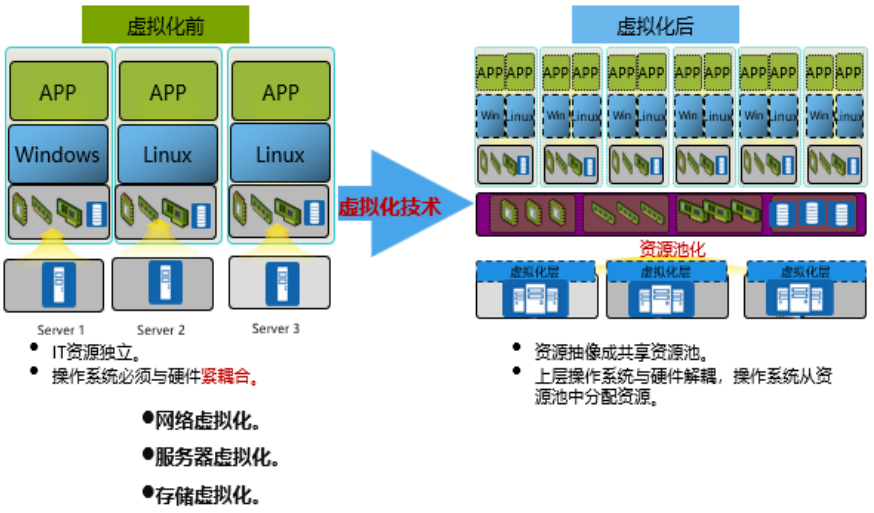
- 在服务器区，再根据不同的应用类型划分不同的层次，例如，数据库层、应用服务器层、WEB 服务器层等。
- 传统数据中心的网络结构是按照经典的三层架构（接入、汇聚、核心）进行部署的。

挑战一：计算节点低延迟需求



- 同一物理服务器部署大量虚拟机，造成流量并发量大增。
- 数据流量模型也从传统的南北向流量转变为东西向流量。
- 网络中存在大量多对一、多对多的东西向流量。
- 对接入层和汇聚层设备的处理能力提出了更高的要求。

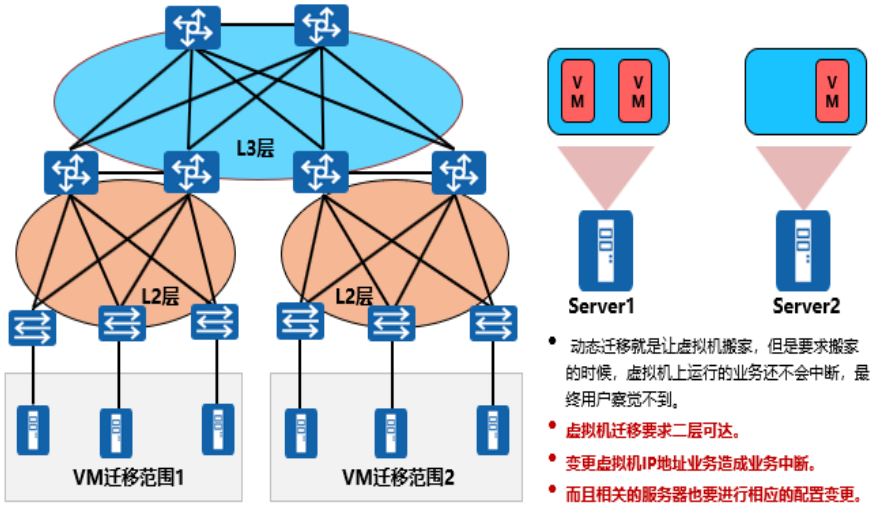
挑战二：虚拟化应用大量部署



- 传统的数据中心内，服务器主要用于对外提供服务，不同业务区域之间可通过划分为不同的安全分区或 VLAN 进行隔离。
- 一个分区通常集中了该业务所需的计算、网络及存储资源，不同的分区之间或者禁止互访，或者经由三层网络进行互访，数据中心的网络流量大部分集中于南北向。在这种设计下，不同分区间计算资源无法共享，资源利用率低下的问题越来越突出。
- 通过虚拟化技术、云计算管理技术等，将各个分区间的资源进行池化，实现数据中心资源的有效利用。
- 随着这些新技术的兴起和应用，新的业务需求如虚拟机迁移、数据同步、数据备份、协同计算等在数据中心内开始实现部署，数据中心内部东西向流量开始大幅度增加。
- 虚拟机动态迁移技术在实际应用中很常见，比如需要对一台服务器进行升级和维护时，可以通过 VM 迁移技术将这台服务器上的 VM 先迁移到另一台服务器上，其间所提供的服务不中断，然后等服务器升级和维护结束后再将 VM 迁移回来即可。

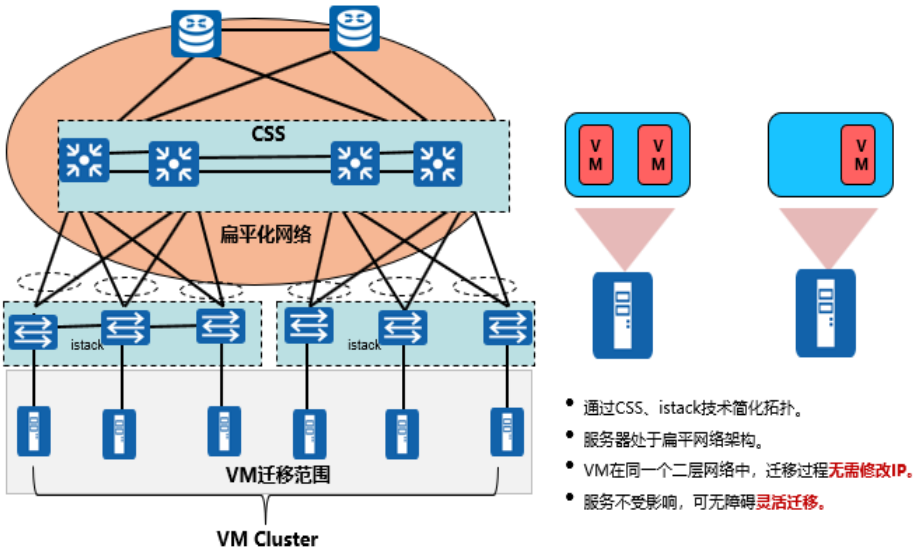
- 虚拟机动态迁移技术还可以充分利用计算资源，比如某公司的网购平台在某段时间内在某片区域提供促销活动，其间业务量大大增加，这样可以将其他业务量小的区域内的 VM 动态迁移过来，这样不会中断其他区域服务的情况下，集中利用资源，活动结束后再将 VM 调整回原先的区域。

传统三层网络架构下虚拟机动态迁移带来的问题



- 虚拟机动态迁移，就是在保证虚拟机上服务正常运行的同时，将一个虚拟机系统从一个物理服务器移动到另一个物理服务器的过程。
- 该过程对于最终用户来说是无感知的，从而使得管理员能够在不影响用户正常使用的情况下，灵活调配服务器资源，或者对物理服务器进行维修和升级。
- 一旦服务器跨二层网络迁移，就需要变更 IP 地址，那么原来这台服务器所承载的业务就会中断，而且牵一发动全身，其他相关的服务器也要变更相应的配置，影响巨大。

当前的一些解决方案：拓扑简化思想



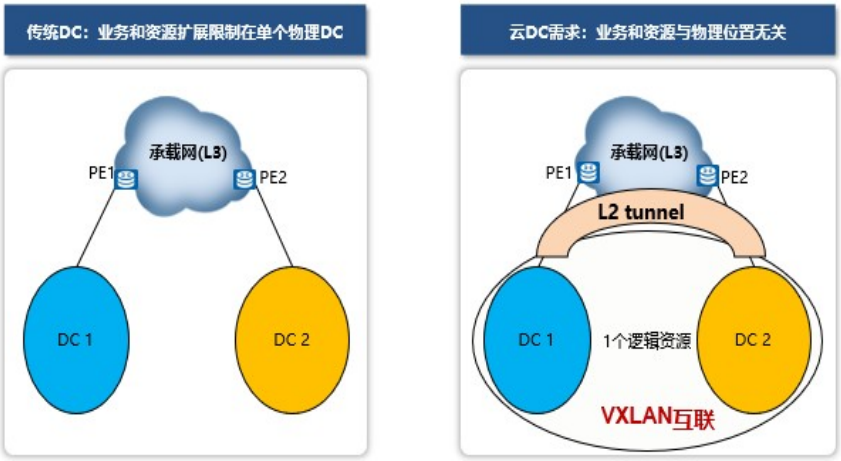
- 为了打破这种跨三层网络限制，实现虚拟机的大范围甚至跨地域的动态迁移，就要求把 VM 迁移可能涉及的所有服务器都纳入到同一个二层网络中，这样才能实现 VM 大范围的无障碍迁移。
- 在汇聚核心层部署 CSS，在接入层部署 istack，可实现简化拓扑结构的目的。
- 设备无需使能 STP 等二层环路保护机制，更有效地提高了链路资源利用率。
- 但设备性能问题并没有得到根本解决。
- 该解决方案比较适于在一个数据中心内部进行 VM 迁移操作 。

该解决方案存在的问题

- MAC地址数量陡增，接入设备压力较大。
- 多租户隔离环境中设备VLAN资源紧张。
- 二层网络范围过大，影响网络通信效率。
- 传统解决方案较适用于DC内部大二层互联应用。

- 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，链路带宽利用率高。

多数据中心大二层互联 - VXLAN

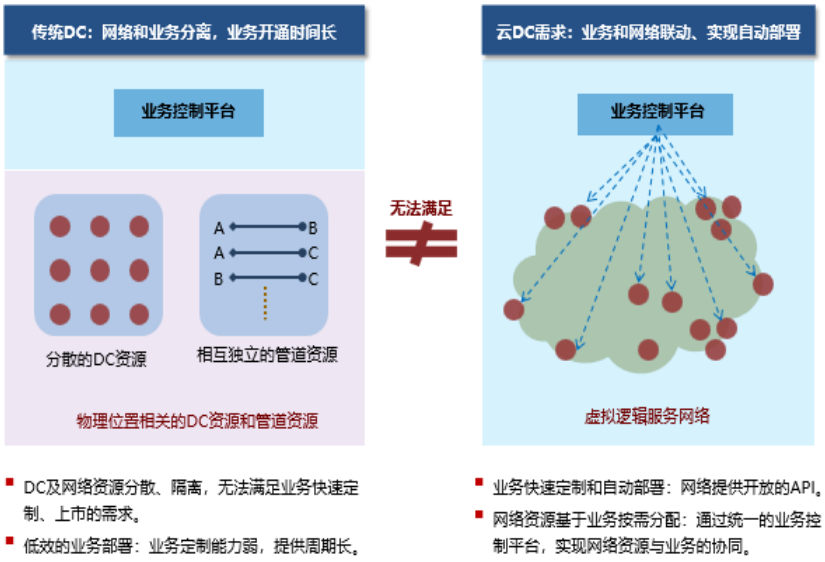


- 早期的虚拟机管理及迁移依附于物理网络，因此数据中

心内部东西向流量主要是二层流量。

- 为扩大二层物理网络的规模，提高链路利用率，出现了 TRILL、SPB 等大二层网络技术。
- 随着虚拟化数据中心规模的不断扩大，以及云化管理的不断深入，物理网络的种种限制越来越不能满足虚拟化的要求，由此提出了 VXLAN、NVGRE 等 Overlay 技术。
- 在 Overlay 方案中，物理网络的东西向流量类型逐渐由二层向三层转变，通过增加封装，将网络拓扑由物理二层变为逻辑二层，同时提供了逻辑二层的划分管理，更好地满足了多租户的需求。
- VXLAN、NVGRE 等 Overlay 技术都是通过将 MAC 封装在 IP 之上，实现对物理网络的屏蔽，解决了物理网络 VLAN 数量限制、接入交换机 MAC 表资源有限等问题，同时通过提供统一的逻辑网络管理工具，更方便地实现了虚拟机在进行迁移时网络策略跟随的问题，大大降低了虚拟化对网络的依赖，成为了目前网络虚拟化的主要发展方向。

挑战三：业务快速创新、自动发放需求



- DC及网络资源分散、隔离，无法满足业务快速定制、上市的需求。
- 低效的业务部署：业务定制能力弱，提供周期长。
- 业务快速定制和自动部署：网络提供开放的API。
- 网络资源基于业务按需分配：通过统一的业务控制平台，实现网络资源与业务的协同。

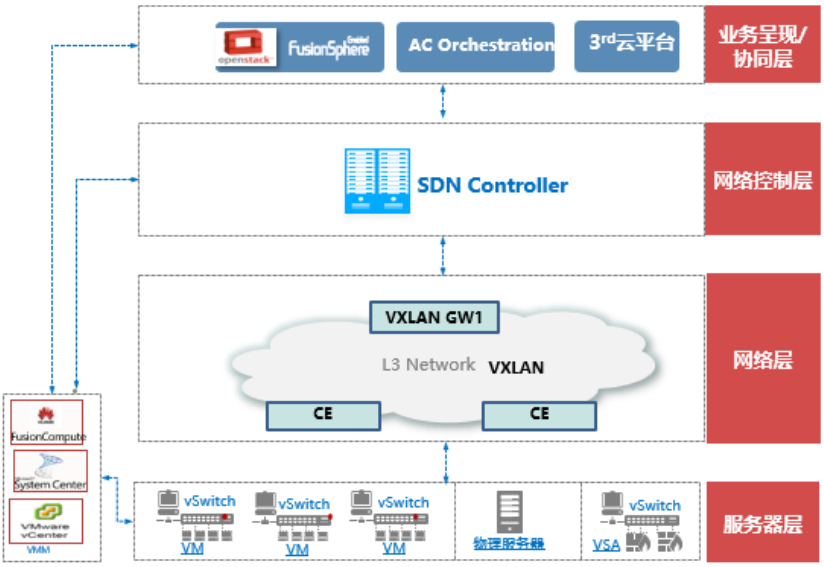
云数据中心的网络解决方案

- VXLAN 技术主要解决多租户环境下的二层互联问题。
- VXLAN 通过隧道技术在不改变三层网络拓扑的前提下构建跨数据中心的逻辑二层网络拓扑。
- VXLAN 技术有效解决了 vlan 数量的限制问题。
- VXLAN 技术对二层网络做了优化不会造成广播风暴等问

题。

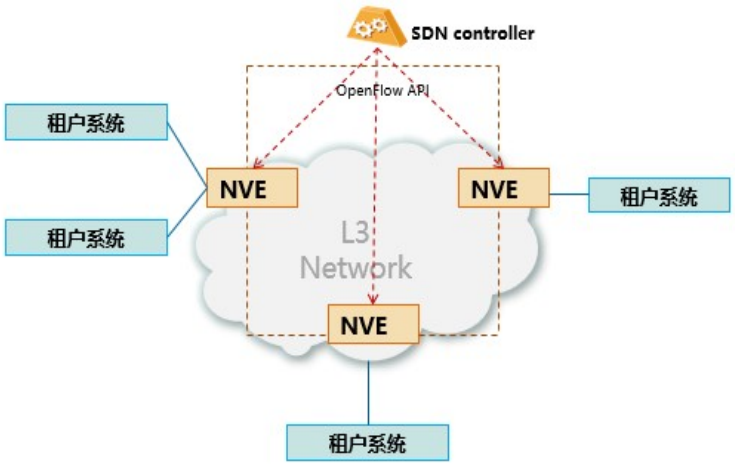
- SDN 技术主要是简化网络的部署、运维、调整等。

VXLAN部署的典型网络架构



- VXLAN/NVGRE/STT 是三种典型的 NVO3 技术。
- 是通过 MAC In IP 技术在 IP 网络之上构建逻辑二层网络。
- 同一租户的 VM 彼此可以二层通信、跨三层物理网络进行迁移。
- 相比传统 L2 VPN 等 Overlay 技术，NVO3 的 CE 侧是虚拟或物理主机，而不是网络站点。
- 此外主机具有可移动性。
- 目前，一般是 IT 厂商主导，通过服务器的 Hypervisor 来构建 Overlay 网络。

VXLAN组网逻辑架构

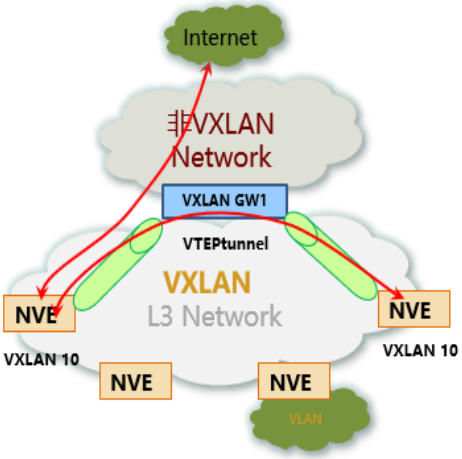


- VXLAN（Virtual eXtensible Local Area Network，RFC7348）是 IETF NVO3（Network Virtualization over Layer 3）定义的 NVO3 标准技术之一，采用 MAC in UDP 封装方式，将二层报文用三层协议进行封装，可对二层网络在三层范围进行扩展，同时支持 24bits 的 VNI ID（16M 租户能力），满足数据中心大二层 VM 迁移和多租户的需求。
- 在 VXLAN NVO3 网络模型中，部署在 VXLAN 网络边缘的设备称为 VXLAN NVE（Network Virtualization Edge，网络虚拟边缘），主要负责 VLAN 网络与 VXLAN 网络间的封装和解封装。经过 NVE 封装转换后，NVE 间就可基于 L3 基础网络建立 Overlay 二层虚拟化网络。
- VXLAN 技术特点：
- 位置无关性：业务可在任意位置灵活部署，缓解了服务器虚拟化后相关的网络扩展问题。
- 可扩展性：在传统网络架构上规划新的 Overlay 网络，部署方便，同时避免了大二层的广播风暴问题，可扩展性极强。
- 部署简单：由高可靠 SDN Controller 完成控制面的配置和管理，避免了大规模的分布式部署，同时集中部署模式可加

速网络和安全基础架构的配置，提高可扩展性。

- 适合云业务：可实现千万级别的租户间隔离，有力地支持了云业务的大规模部署。
- 技术优势：VXLAN 利用了现有通用的 UDP 进行传输，成熟性极高。

VXLAN网关



- **NVE:**
- 目前有软件NVE（一般安装在服务器上；例如OVS）和硬件NVE（一般集成在交换机上，例如CE6850）。由于软件NVE是在原设备中安装一个软件包，硬件NVE是在原设备中增加一个硬件模块，而原设备多数是VLAN的二层设备，所以，NVE又是VXLAN的二层网关，主要实现VXLAN与VLAN、MAC等的二层映射。
- **VXLAN网关:**
- 与VXLAN NVE类似，但是地位更高一些的是另一个VXLAN角色，即VXLAN三层网关，简称VXLAN GW，主要实现VXLAN报文头与IP报文头的映射。
- 不管二层VXLAN网关还是三层VXLAN网关，都是主要实现了VXLAN网络和非VXLAN网络之间的连接。

- NVE 是服务器虚拟化层的一个功能模块，虚拟机通过虚拟化软件直接建立 VTEP 隧道。
- NVE 也可以是一台支持 VXLAN 的接入交换机集中为多租户提供 VXLAN 网关服务。
- VXLAN 网关可以实现不同 VXLAN 下租户间通信，也能实现 VXLAN 用户与非 VXLAN 用户间通信，这和 VLANIF 接口的功能是类似的。

NVO3标准术语

VN	Virtual Network虚拟网络
VNI	虚拟网络实例（Virtual Network Instance），可以为L2或L3网络，一个租户可以对应一个或多个VNI。
VNID	虚拟网络ID（Virtual Network Identifier），标识一个虚拟网络。
NVE	虚拟网络边缘，可以位于物理网络边缘设备，也可以位于Hypervisor，可以是二层转发或三层转发。
VN Context	该字段位于Overlay封装头部，用于Egress NVE设备确定VNI。
Hypervisor	运行在物理服务器内的虚拟化软件，为服务的VM提供共享计算资源、内存、存储，而且Hypervisor内经常内嵌Virtual Switch。
Tenant End System	租户终端系统，可以是物理服务器也可以是VM。

业界其他技术实现 - NVO3技术背景

类别	简单描述	主导厂商
NVGRE	通过Mac In GRE封装实现虚拟二层网络的方法。	Microsoft&Intel 、Arista、Broadcom、Dell、Emulex、HP。
VXLAN	Virtual eXtensible Local Area Network 通过Mac In UDP封装实现虚拟二层网络的方法。	Vmware、Cisco、Arista、DELL、Broadcom、Citrix、Red Hat。
STT	Stateless Transport Tunneling 通过Mac In TCP封装实现虚拟二层网络的方法。	Nicira、RackSpace、Ebay、Intel。

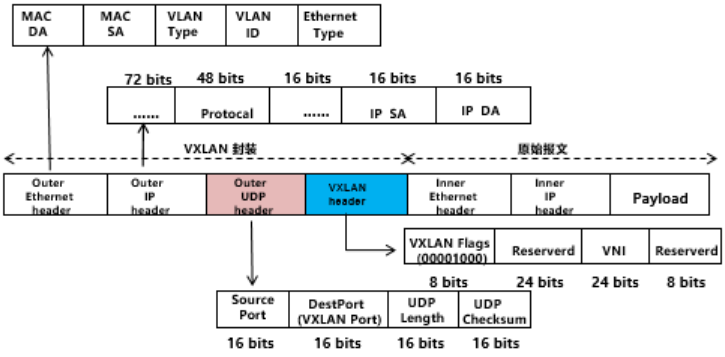
- VXLAN/NVGRE/STT 是 NVO3 三种典型技术，总体是通过 MAC In IP 技术来构建 Over 在 IP 网络之上的二层网络，使同一个租户的 VM 彼此可以二层通信、跨三层物理网络进行迁移。相比传统 L2 VPN 等 Overlay 技术，NVO3 的 CE 侧是虚拟或物理 Host，而不是网络 Site，另外 Host 具有移动性。目

前，一般是 IT 厂商主导，通过服务器内 Hypervisor 来构建 Overlay 网络。

- NVGRE 主要支持者是 Microsoft。与 VXLAN 不同的是，NVGRE 没有采用标准传输协议（TCP/IP），而是借助通用路由封装协议（GRE）。NVGRE 使用 GRE 头部的低 24bit 位作为租户网络标识符。
- VXLAN（Virtual Extensible LAN）虚拟可扩展局域网，是一种 Overlay 的网络技术，使用 MAC in UDP 的方法进行封装。后文会详细介绍 VXLAN 技术。
- STT（Stateless Transport Tunneling）是一种 MAC Over Ip 的协议，和 VXLAN、NVGRE 类似，都是把二层的帧封装在一个 Ip 报文的 payload 中，在 Ip 报文的 payload 中，除了虚拟网络的二层包以外，还要把构造的一个 TCP 头和一个 STT 头加在最前面。

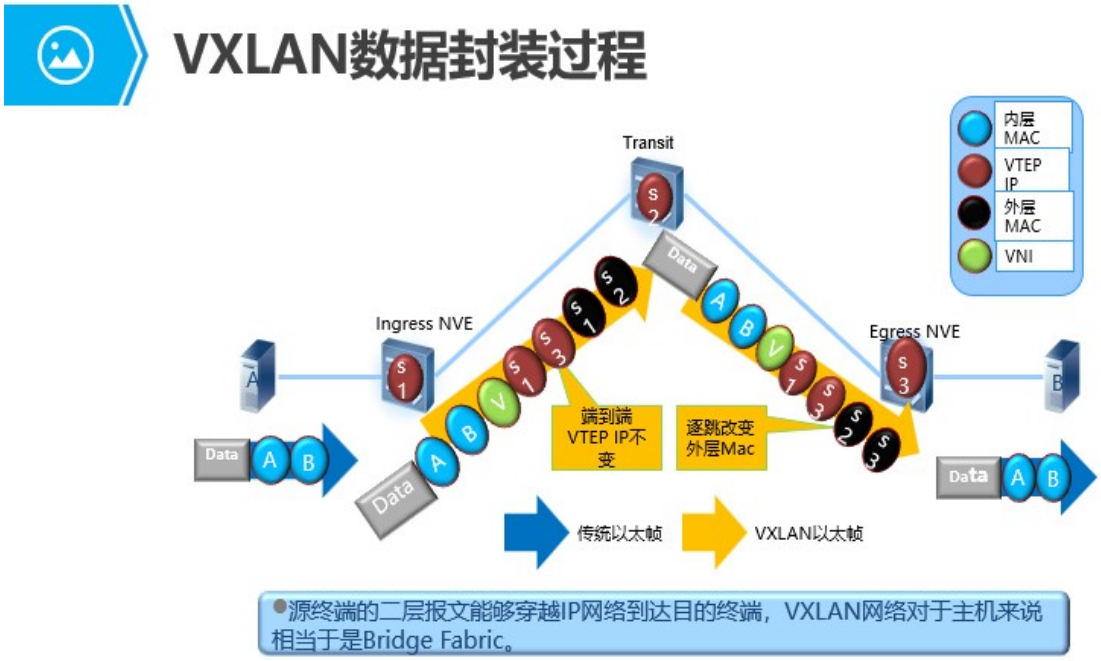
VXLAN报文封装

- VXLAN是IETF定义的NVO3（Network Virtualization over Layer3）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。
- 支持24bits的VNI ID，满足数据中心大二层VM迁移和多租户的需求。



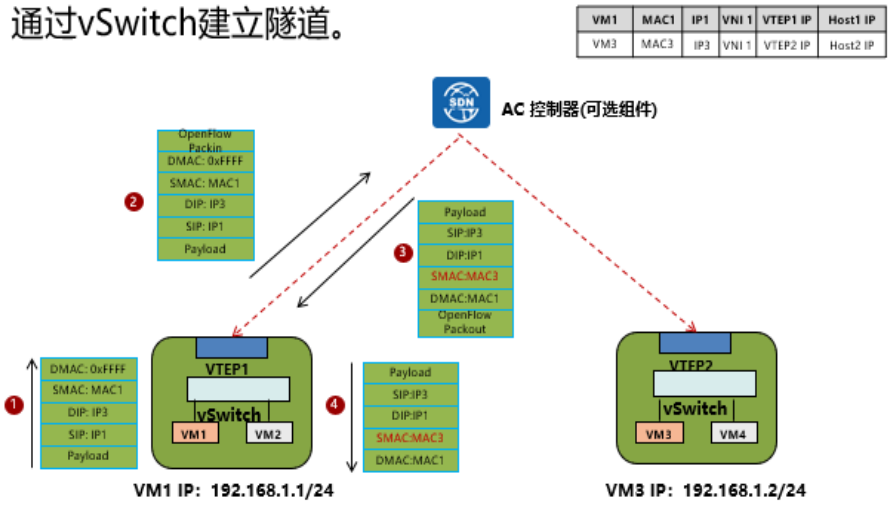
- VXLAN 头封装：
- VNI：VXLAN 网络标识，24 比特，用于区分 VXLAN 段。

- Reserved：24 比特和 8 比特，必须设置为 0。
- 外层 UDP 头封装：
- 目的 UDP 端口号是 4789。源端口号是内层以太报文头通过哈希算法计算后的值。
- 外层 IP 头封装：
- 源 IP 地址为发送报文的虚拟机所属 VTEP 的 IP 地址；目的 IP 地址是目的虚拟机所属 VTEP 的 IP 地址。
- 外层 Ethernet 头封装：
- SA：发送报文的虚拟机所属 VTEP 的 MAC 地址。
- DA：目的虚拟机所属 VTEP 上路由表中直连的下一跳 MAC 地址。
- VLAN Type：可选字段，当报文中携带 VLAN Tag 时，该字段取值为 0x8100。
- Ethernet Type：以太报文类型，IP 协议报文该字段取值为 0x0800。



VXLAN通信流程 (1)

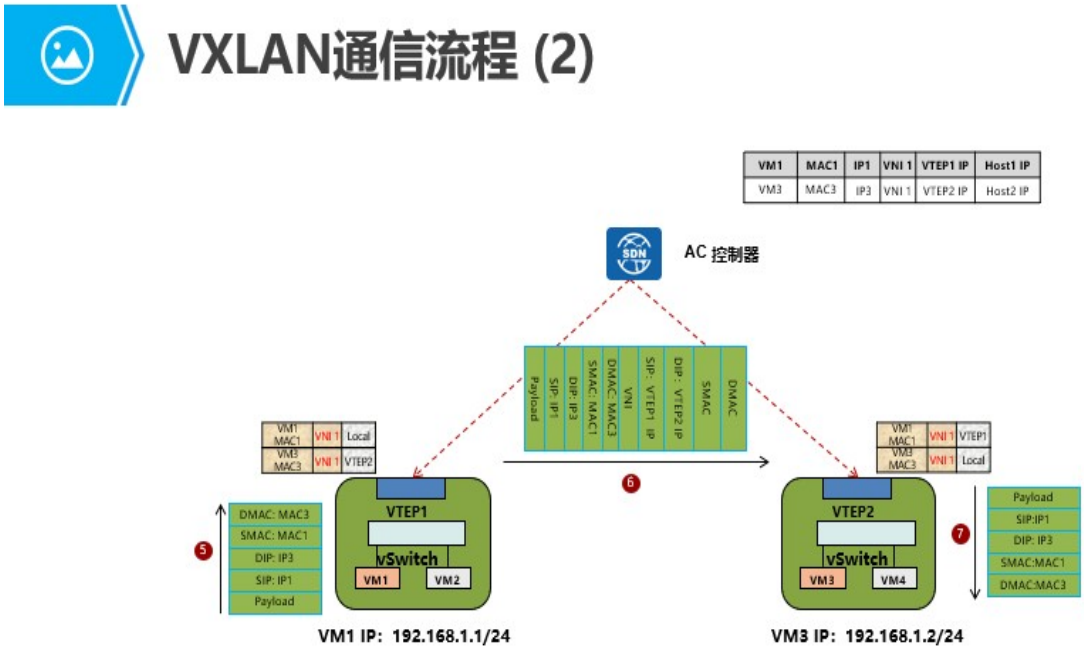
- 通过vSwitch建立隧道。



- vSwitch 为虚拟机工具 Hypervisor 层中集成的虚拟交换机。
- VTEP 是在虚拟机所属服务器的 Hypervisor 中的 vSwitch 间进行建立的。
- 控制器为可选组件。
- VTEP—Virtual Tunnel End Point 虚拟隧道端点，即 VXLAN 隧道的入口和出口。在这里 VM 流量经过 vSwtich 交换后会导入到 VXLAN 隧道里，入口是 VTEP。
- VNI— Virtual Network Instance 虚拟网络实例，一个 VNI 就是一个虚拟网络，一个 VNI 用一个 VNI ID 标识。例子里 VM1 和 VM2 在同一个虚拟网络里，VNI ID 为 1。
- VM1 地址为 IP1，VM3 地址为 IP3，IP1 与 IP3 为同一子网。
- ARP 协议交互过程：
- VM1 先发送 ARP 报文，请求 VM3 的 MAC 地址。
- ARP 报文通过 OpenFlow 报文封装发给 AC，AC 上使能 ARP 代理。
- AC 通过 IP3 找到 MAC3，响应 ARP 报文，SIP 为 IP3，

SMAC 为 MAC3。

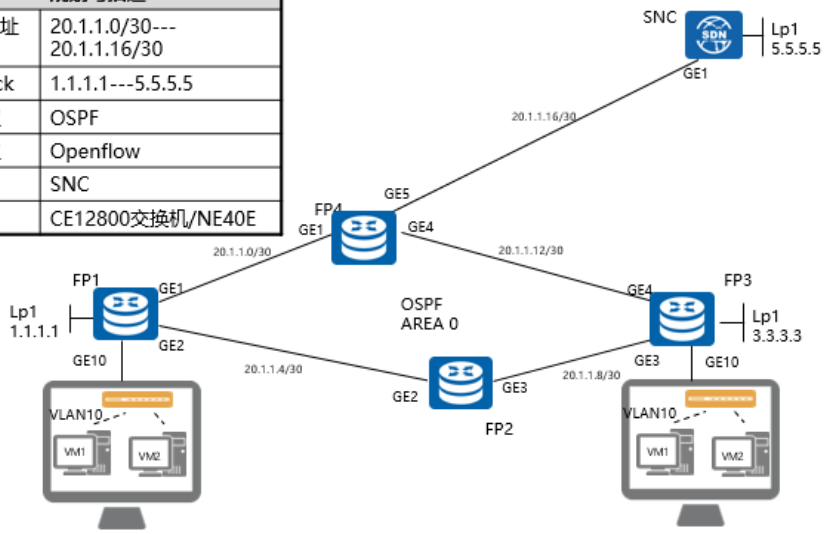
- vSwtich 将 ARP 响应报文发给 VM1。



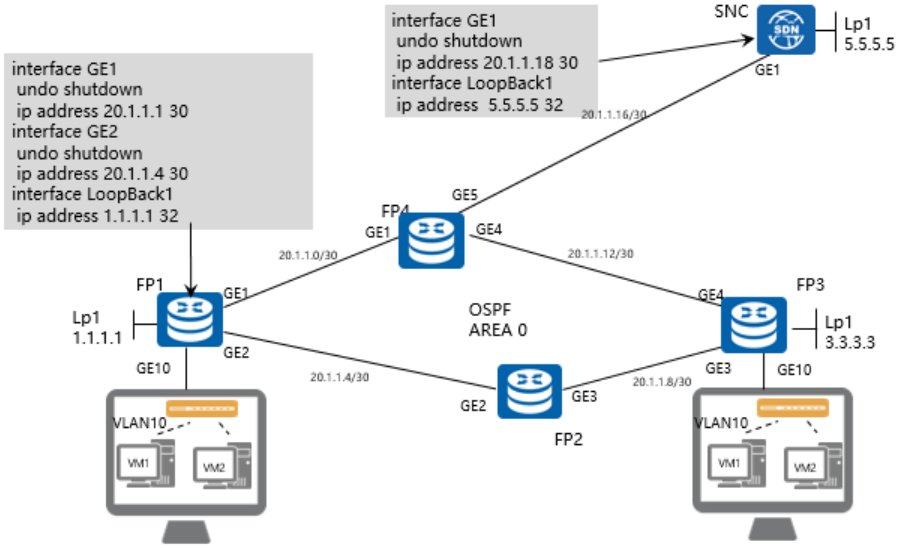
- 报文转发过程：
- VM1 发送数据报文，SIP 为 IP1，DIP 为 IP3，SMAC 为 MAC1，DMAC 为 MAC3。
- VXLAN 封装，内层 SIP 为 IP1，DIP 为 IP3，外层 SIP 为 VTEP1 IP，外层 DIP 为 VTEP2 IP。
- 解封装 VXLAN 头部，发往 VM3。

基于SDN的VXLAN基本组网

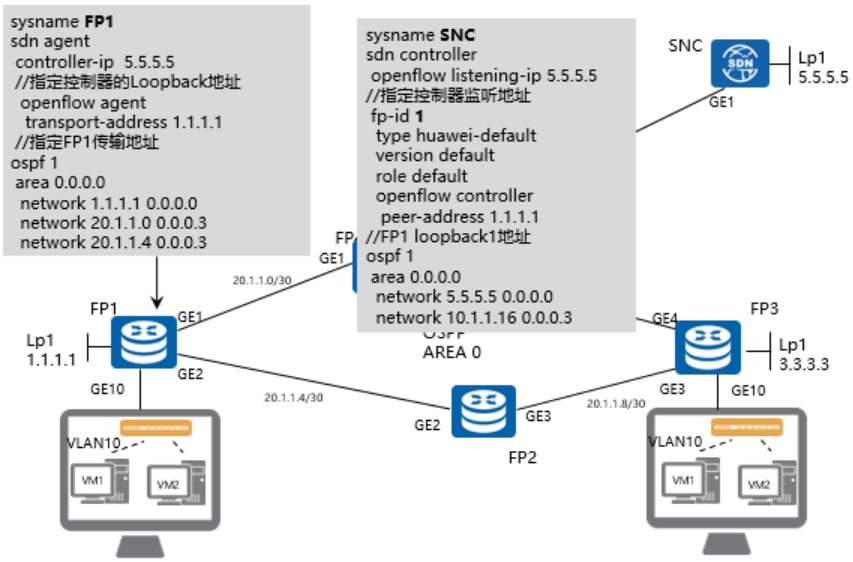
规划与描述	
互联IP地址	20.1.1.0/30--- 20.1.1.16/30
Loopback	1.1.1.1---5.5.5.5
路由协议	OSPF
南向协议	Openflow
控制器	SNC
转发器	CE12800交换机/NE40E



基于SDN的VXLAN基本组网 - 接口配置



基于SDN的VXLAN基本组网 - 协议配置

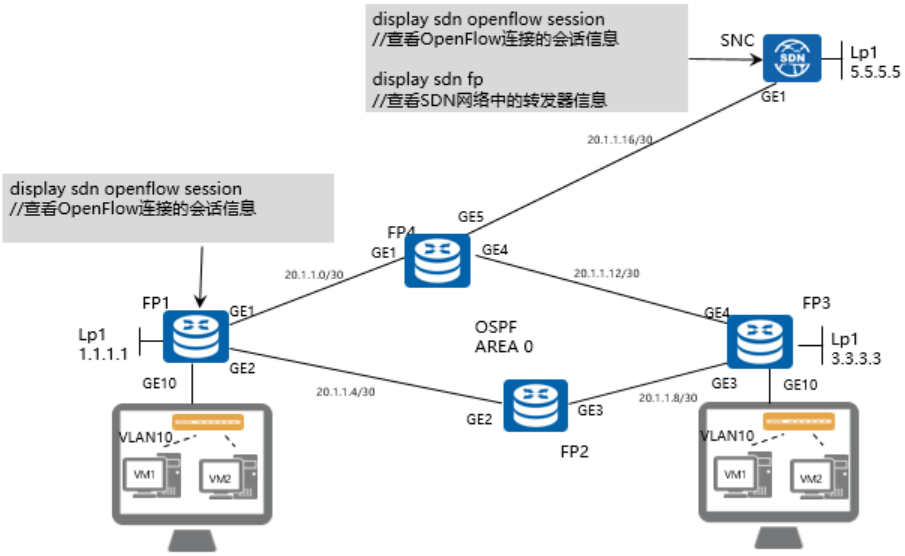


- 1.配置控制器和转发器建立 OpenFlow 通信通道
- SNC 上配置：
- openflow listening-ip 5.5.5.5 //配置 SDN 控制器侦听地址
- fp-id 1 //配置转发器 ID
- type huawei-default //配置转发器的设备类，其中 huawei-default 表示转发器是华为设备；ovs-default 表示转发器是 OVS（openvswitch）设备
- version default //配置转发器的版本是 default
- role default //配置转发器的角色是 default
- openflow controller //配置控制器与转发器之间的通信通道，采用 OpenFlow 连接并进入 OpenFlow 视图
- peer-address 1.1.1.1 //指定转发器 FP1 的 Loopback1 地址
- FP1 配置
- controller-ip 5.5.5.5 //指定控制器的 Loopback 地址，并进入 Controller 视图
- openflow agent //配置转发器与控制器之间的

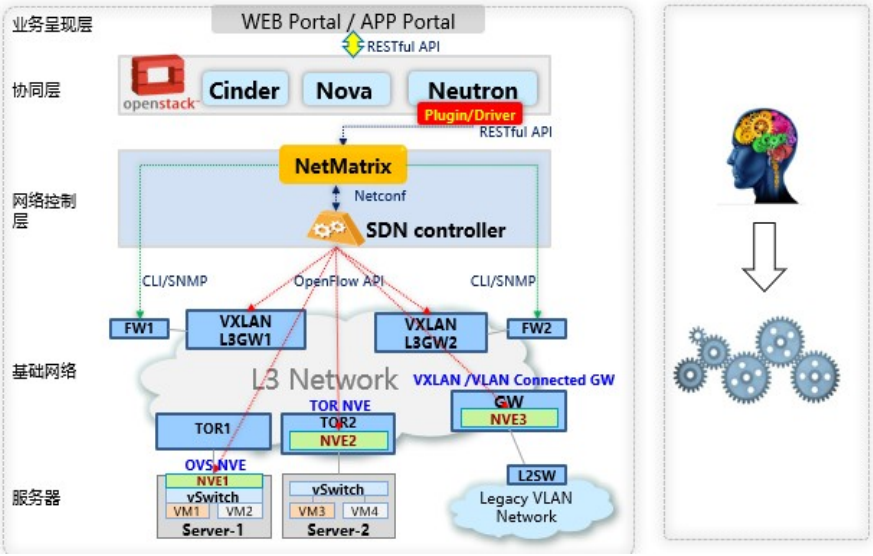
通信通道采用 OpenFlow 连接并进入 OpenFlow Agent 视图

- transport-address 1.1.1.1 //配置 OpenFlow 连接的本端地址

基于SDN的VXLAN基本组网 - 查看配置



VXLAN组网应用（UNICA、联通公有云、上海电信公有云）



- 业务呈现层：
- 面向运营商、企业、租户、RSP 的 Portal。

- 提供业务灵活定制化界面。
- 协同层：
- 标准、开放的 OpenStack 架构，并兼容多厂商。
- 实现存储、计算和网络资源的协同。
- 网络控制层：
- 网络控制平台由 NetMatrix 和 SNC 组成，完成网络建模和网络实例化。
- 北向支持开放 API 接口，实现业务快速定制和自动发放。
- 南向支持 OpenFlow/Netconf 等接口，实现统一管理和控制物理和虚拟网络。
- 基础网络层：
- 物理网络和虚拟网络统一规划和设计的 Overlay 网络。
- 基于硬件的 VXLAN 网关提高业务性能。
- 支持对传统 VLAN 网络的兼容。

思考题

1. 下面哪项为配置SDN控制器侦听地址的命令?
   A. openflow listening-ip 1.1.1.1
   B. sdn controller souce-address 1.1.1.1
   C. controller-ip 1.1.1.1
   D. sdn listening-ip 1.1.1.1
2. VXLAN支持哪几种常用的配置方式?
   A. 通过虚拟化软件配置
   B. 通过SDN控制器配置
   C. 通过SNMP协议配置

- 1、答案：A。

- 2、答案：AB。