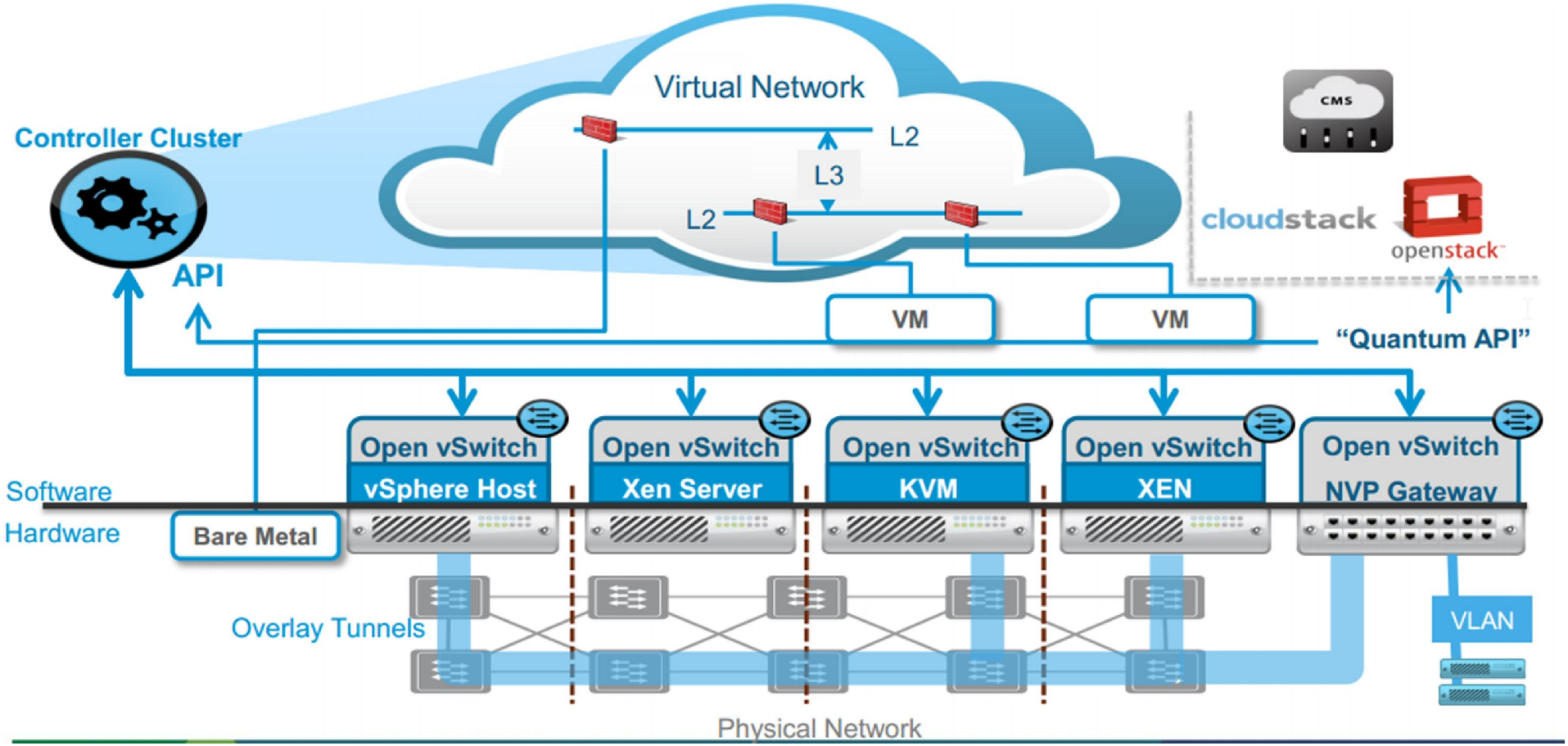
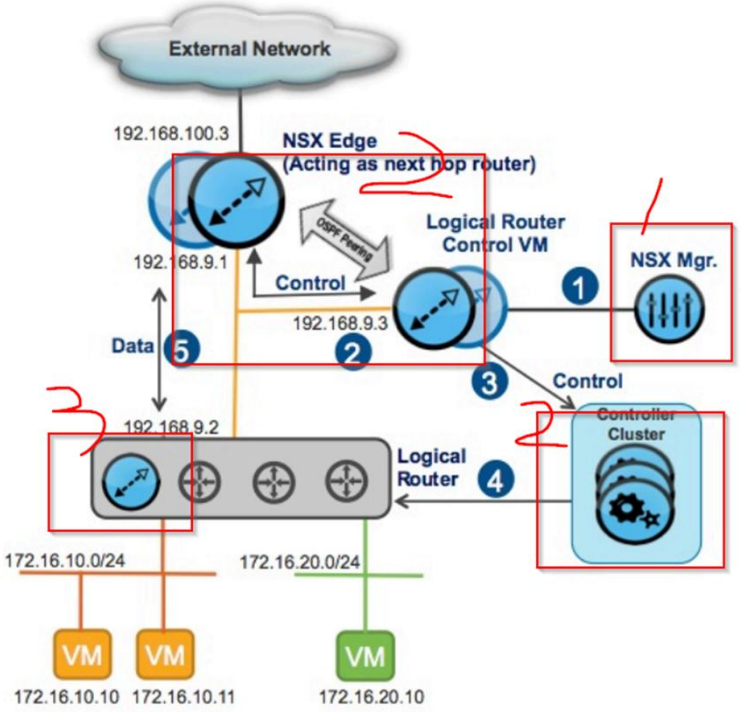
VMware NSX overview

1. 物理L3网络拓扑提供了基于ip的fabric。
2. 虚拟L2网络拓扑完成隧道终结与路由。



NSX由三部分组成

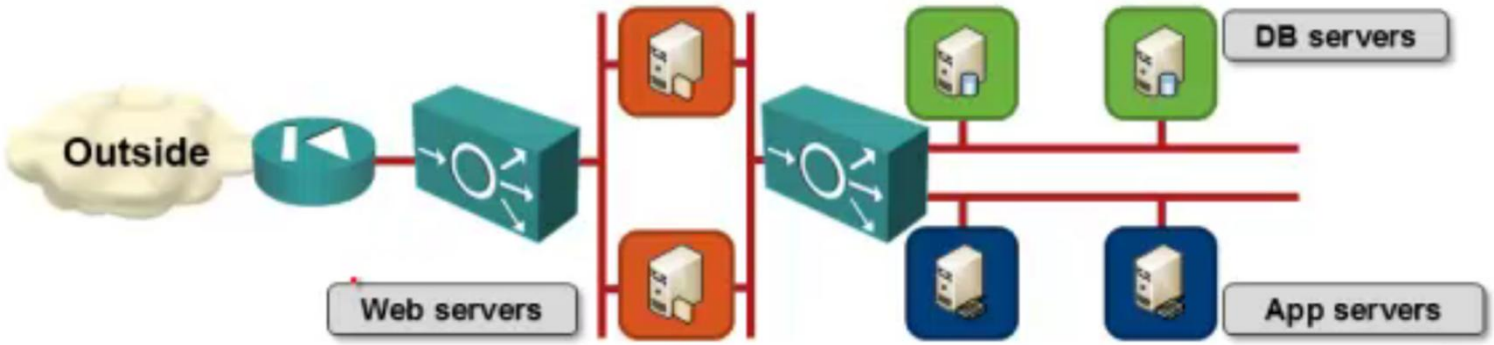
- 管理平面，由NSX manager实现。
- 控制平面，由NSX controller与NSX Edge实现。
- 数据平面，由VDS/OVS、VXLAN plugin、Distributed Logical Router与Distributed Firewall实现。



1. 管理平面
2. 控制平面
3. 数据平面

NSX功能概述

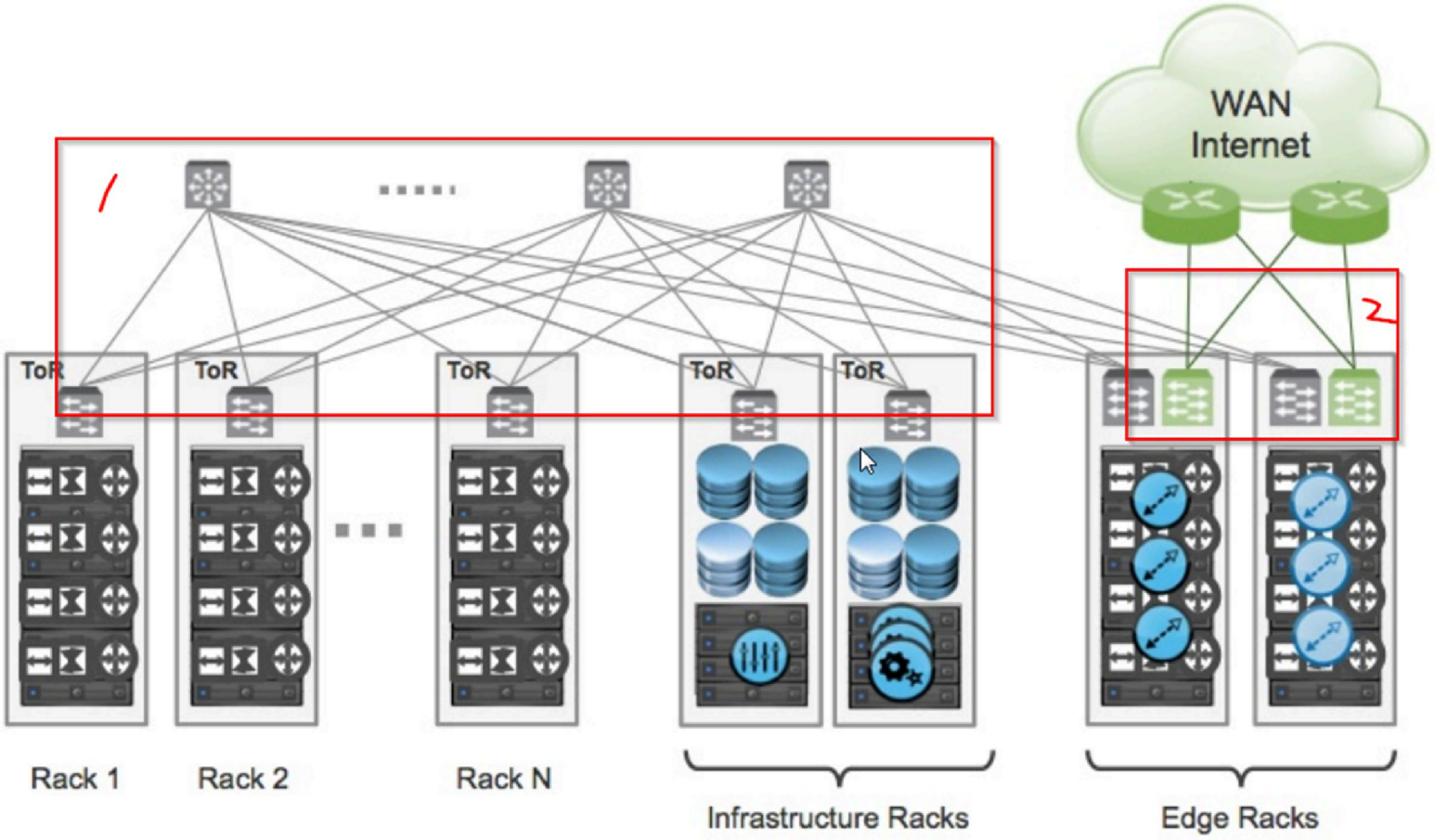
- Logic L2，overlay networking支持。
- Distruted L3， NSX Edge提供路由学习。
- Distributed Firewall，分布式防火墙。
- Logical LB，支持到L7，还可以完成SSL加解密。
- SSL VPN， 实现L2 VPN。



NSX物理网络组成

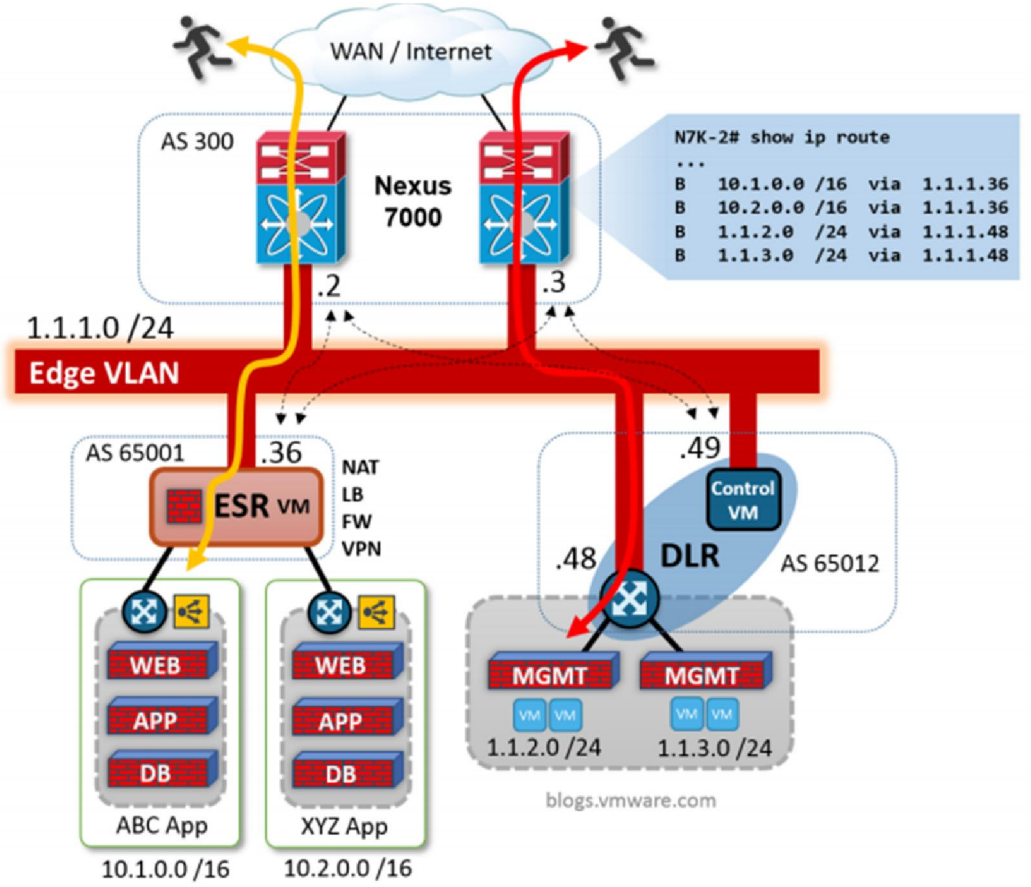
- 网络拓扑。
  - set化部署，管控单独的机架，外网网关单独机架。
  - 图中1框，内网使用全L3 10GbE网络拓扑， 收敛比1：2.5。
  - 图中2框，为外网接入。
  - L3跑openflow。
- ToR交换机配置
  - 网络设备配置文件简化到位置无关。
  - LACP与load-based teaming，提供高带宽以及容错。
  - 使用vlan if，便于ToR与汇聚之间走L3。
  - 多上联， 走ECMP冗余。
  - 支持DSCP或者COS，实现QoS。

NSX物理网络组成 cont.



NSX物理网络VLAN划分

- VLAN 100， 管理
- VLAN 110， vMotion
- VLAN 120， SAN
- VLAN 200， VXLAN
- VLAN 300， Edge

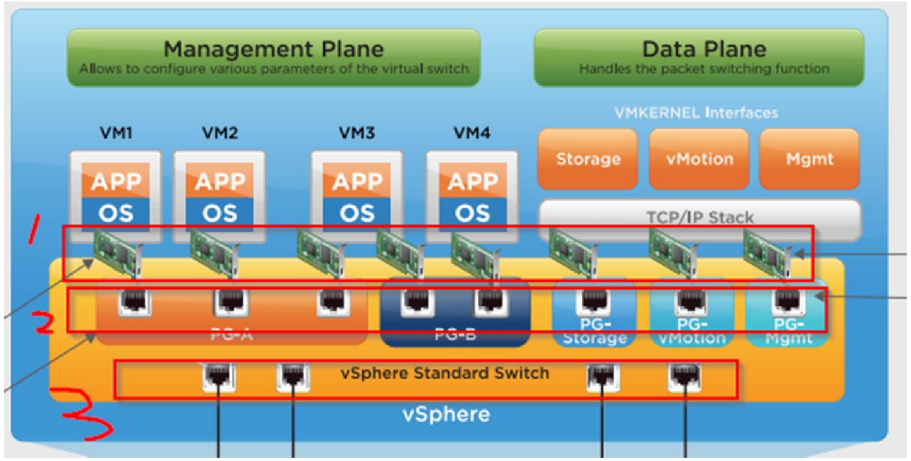


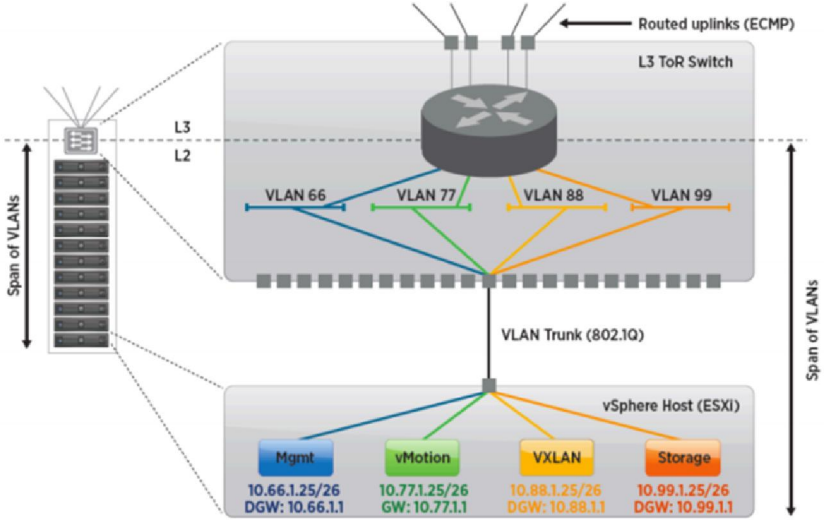
NSX数据平面组成

- 虚拟交换机，完成东西流量的L2转发。
- Distributed Logical Router，完成东西向流量L3转发。
- Firewall，流量隔离，并有TC流量控制的功能。
- Load balancers，
- VPN，为租户登录自定义网络提供接口。

NSX虚拟交换机(VDS/OVS)组成

- 2部分南向接口连接1部分的给guest的虚拟网卡。
- 3部分北向接口通过服务器物理网卡， 连接ToR端口。
- 提供管理、VXLAN租户内网、vMotion迁移、存储访问4类接口。
- VXLAN租户内网接口通过VTEP作为端点，有3中模式
  - LACP， single VTEP使用multiple uplinks
  - Fail Over， single VTEP使用 one uplink
  - Load Balance， multiple VTEPs各自使用one uplink



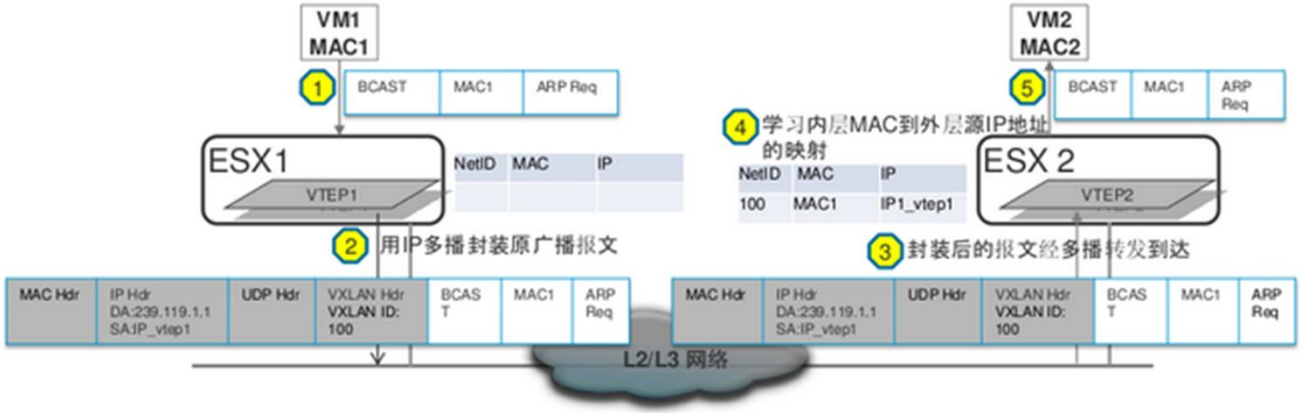


虚拟交换机MAC学习过程

- 上传本机的MAC表给Controller Cluster。
- 从Controller Cluster获取租户全局的MAC表作为cache。
- 如果查询时发生miss，则重新拉取。

VXLAN实现

- VM1以广播的形式发送ARP请求
- VTEP1把ARP请求报文被封装在IP多播报文中，并打上VXLAN标识为100
- VTEP1在IP多播组内里进行多播
- VTEP2接收到ip多播报文后对内层MAC地址到外层IP地址的映射进行学习，解封装后在本地VXLAN标识为100的虚拟局域内广播
- VM2接收到请求自己MAC的ARP请求后做出响应

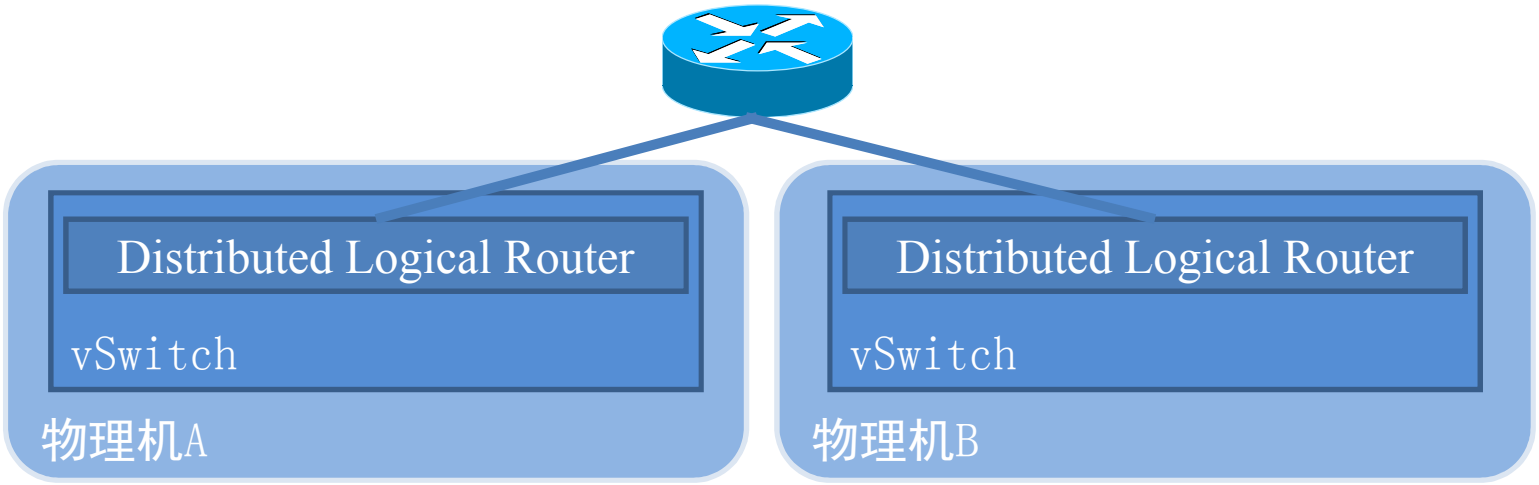


虚拟交换机其它特性

功能点	说明	备注
Port Mirroring	复制port的流量到其它port	debug
NetFlow	网络流量工程建模	通过mib访问
Configuration File	可以按照配置restore	
Network Health Check		
QoS	保证某个流的服务质量	
LACP	实现链路层聚合	上下行都可用

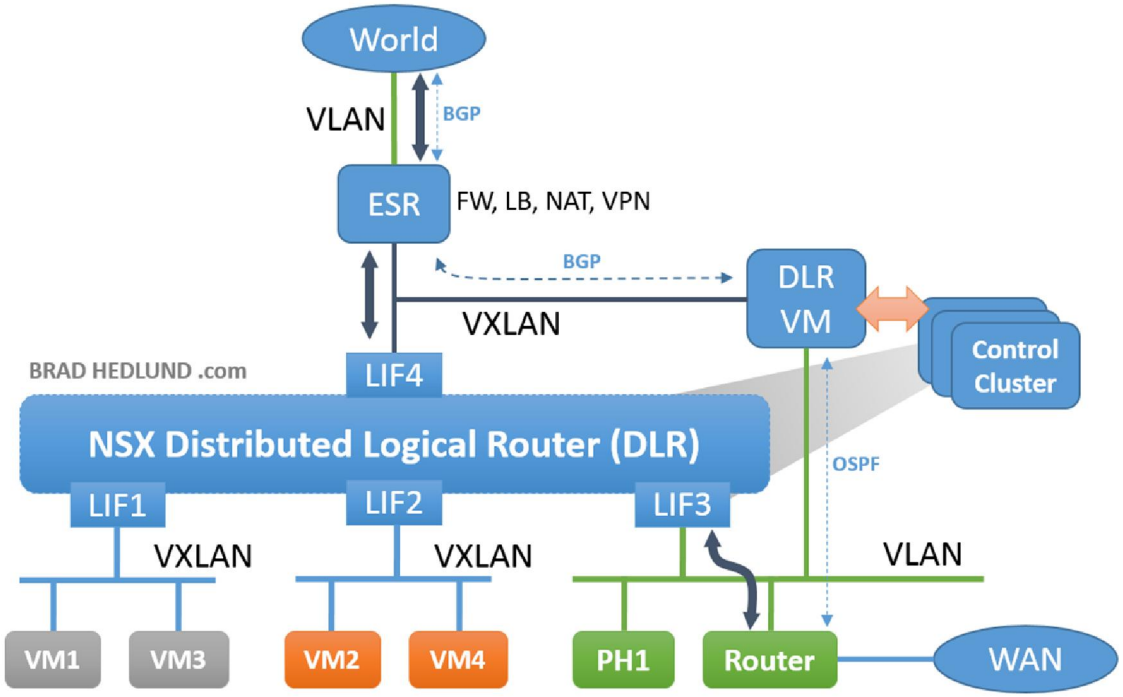
NSX分布式路由器

- 运行在虚拟交换机的接口的ingress上，完成报文的L3转发，可以看作接入交换机的虚拟线卡。
- 分别在各个host的DLR模块，构成统一的整体的DLR，其每个VXLAN接口的IP与vMAC都相同， VLAN接口的IP相同。



分布式路由器拓扑

- DLR的每个接口LIF，都有独立的ARP表。
- DLR可以配置static route，也可以运行dynamic route protocol，比如BGP、OSPF，从DLR VM获得全网路由。



分布式路由器VXLAN转发

- VM1到VM3
  - VM1使用LIF1为default GW。
  - LIF1查询DLR的路由表，在一个VXLAN里，L3转发给LIF3，LIF3通过ARP获得VM3的MAC，L2转发给VM3。
- VM3到VM1
  - VM3使用LIF1为default GW。
  - LIF3查询DLR的路由表，在一个VXLAN里，L3转发给LIF1，LIF1 L2转发给VM1。
- VM1与VM2
  - 其它与VM1到VM3相同，但不在一个VXLAN里面，需要在VETP上添加相应的VXLAN报文头。

分布式路由器VXLAN与VLAN转发

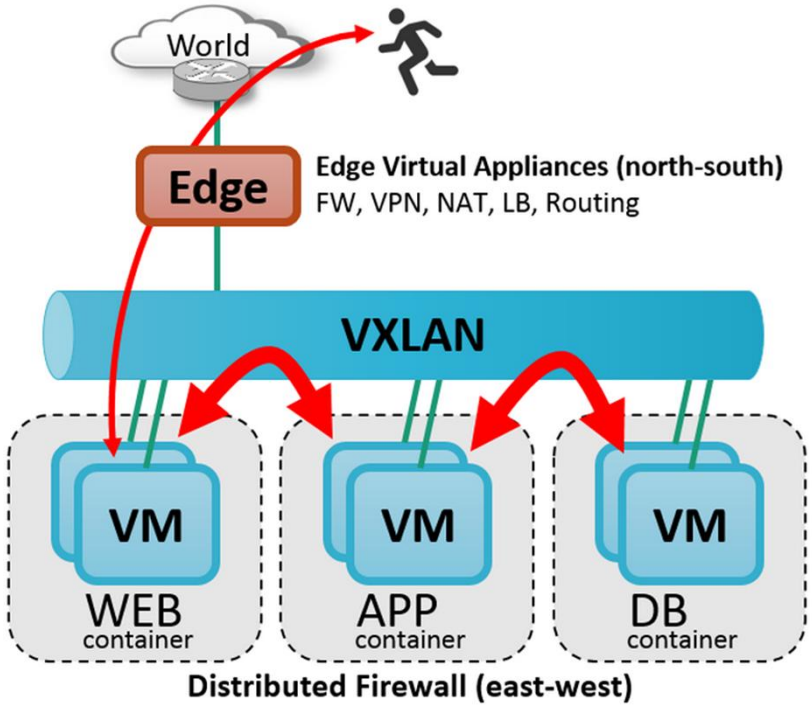
- PH1到VM1
  - PH1使用LIF3为default GW。
  - PH1发LIF3的ARP request， controller cluster收到，选择一个host作为LIF3的Designated Instance（DI）.
  - DI做ARP response，回应自己的pMAC。
  - PH1将数据发送给DI，DI通过自己的DLR完成L3转发到VM1，直接发送或者添加VXLAN头。
- VM1到PH1
  - VM1通过DLR查询路由表， 找到PH1的接口为LIF3。
  - 查询LIF3的ARP table，如果没有命中，发送UDP消息给其的DI，请求发送PH1的ARP request。
  - DI发送ARP request，并将得到的结果发送UDP response给VM1。
  - VM1讲数据发送给DI， DI通过自己的DLR完成L3转发到PH1。

NSX L4-L7流量均衡器

- FTP, HTTP, HTTPS。
- IP hash, URI。
- Source IP, MSRDP, cookie, ssl session-id。
- URL block, URL rewrite, content rewrite。

NSX分布式防火墙

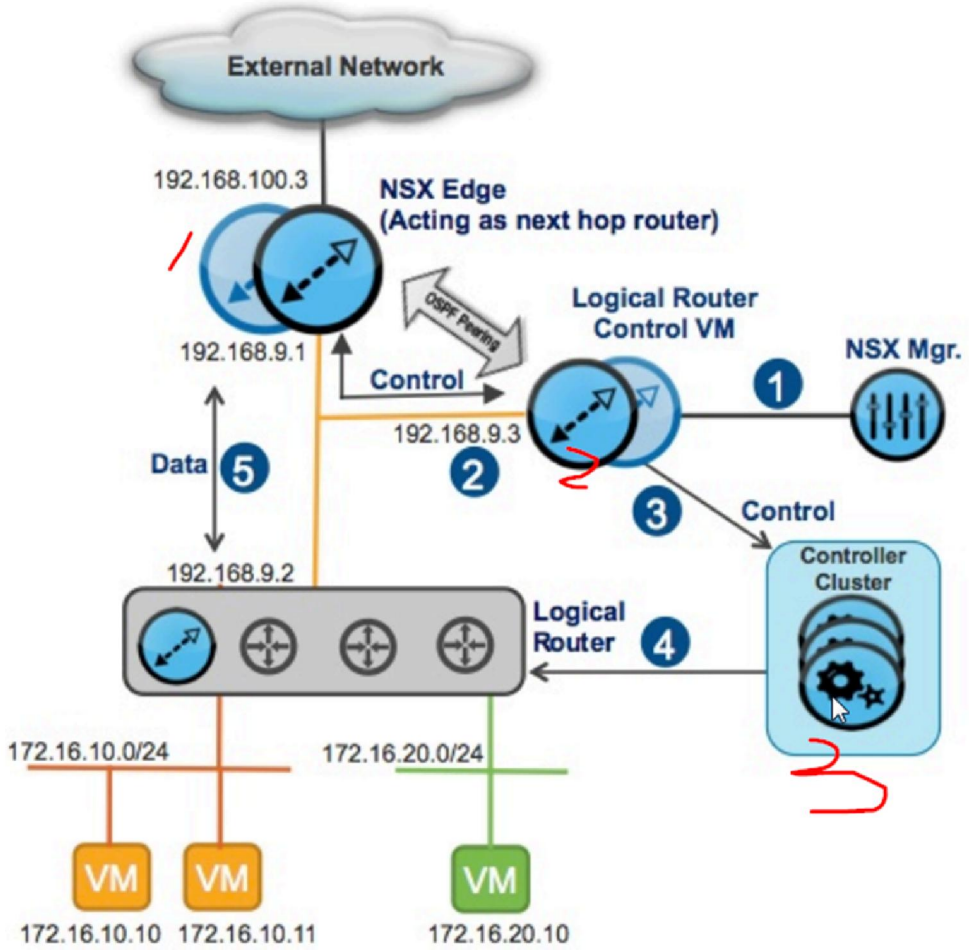
- 每个host，都有stateful的防火墙模块，解决东西流量。
- Edge设备的防火墙模块， 解决南北流量。
- 规则可以基于container， 与IP解耦。



NSX分布式防火墙能力

- stateful L3/4过滤， 可以匹配IP或者containor。
- ARP和L2过滤 。
- Source IP验证。
- 支持IPv4， IPv6。

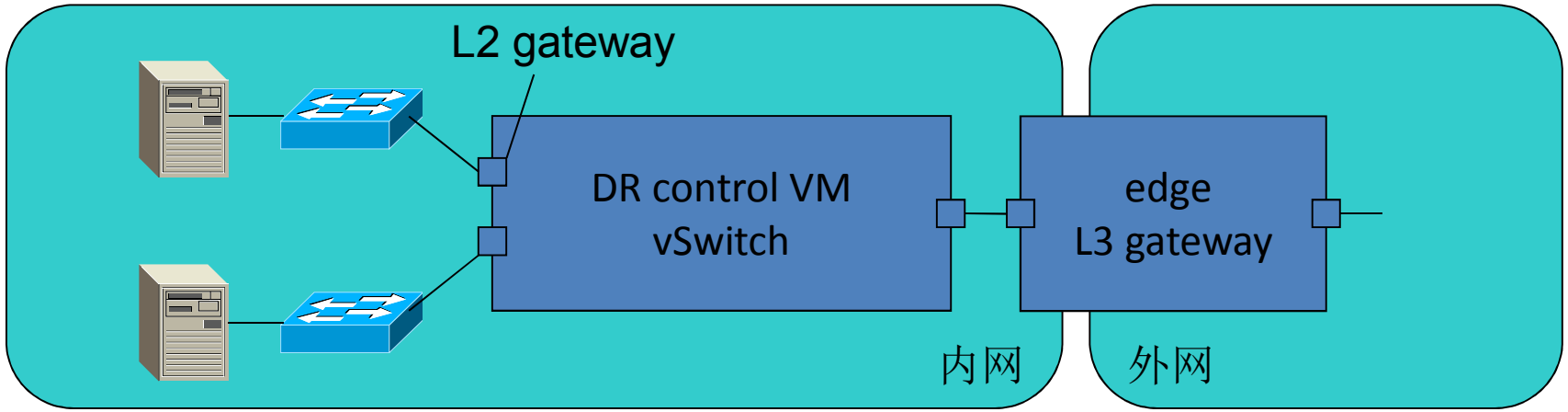
NSX控制平面组成



1. NSX Edge
2. Logical Router Control VM
3. Controller cluster

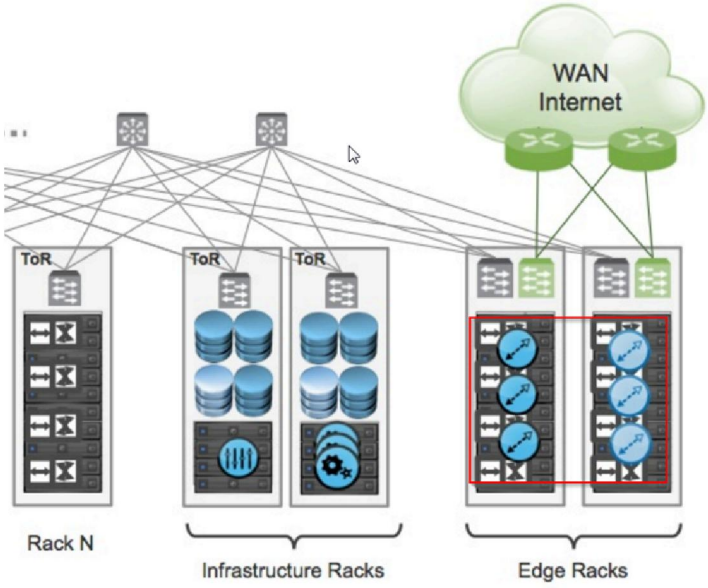
NSX Edge提供外网的L2，L3功能

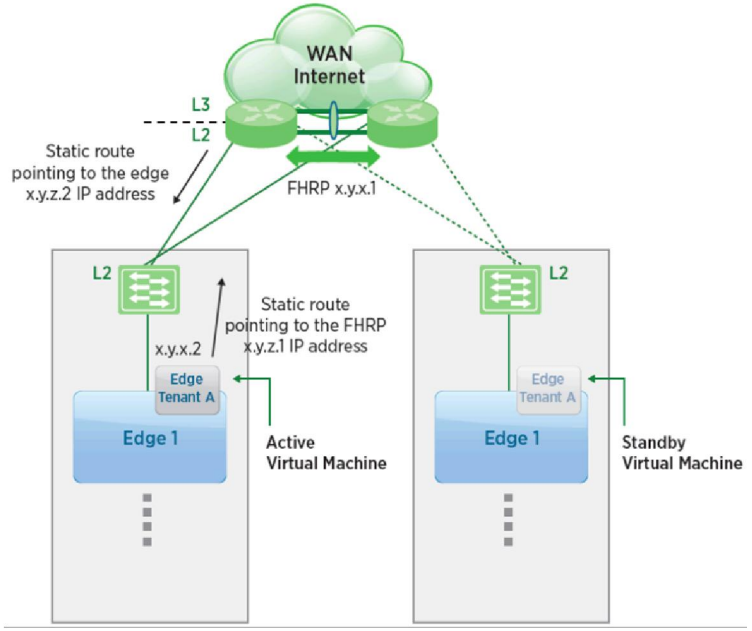
- 作为租户服务器与外网连接的网关，有L2 gateway与L3 gateway两种实现，不能在同一节点部署。
- L2 gateway通常作为外网连接到租户虚拟网络的接口，如下图打通了VXLAN与外网VLAN之间的通信。
  - L2 gateway作为Distributed Router Control VM L3 vSwitch的一个L2 port，controll cluster通过DR来配置L2 gateway。



NSX Edge提供外网的L2，L3功能 cont.

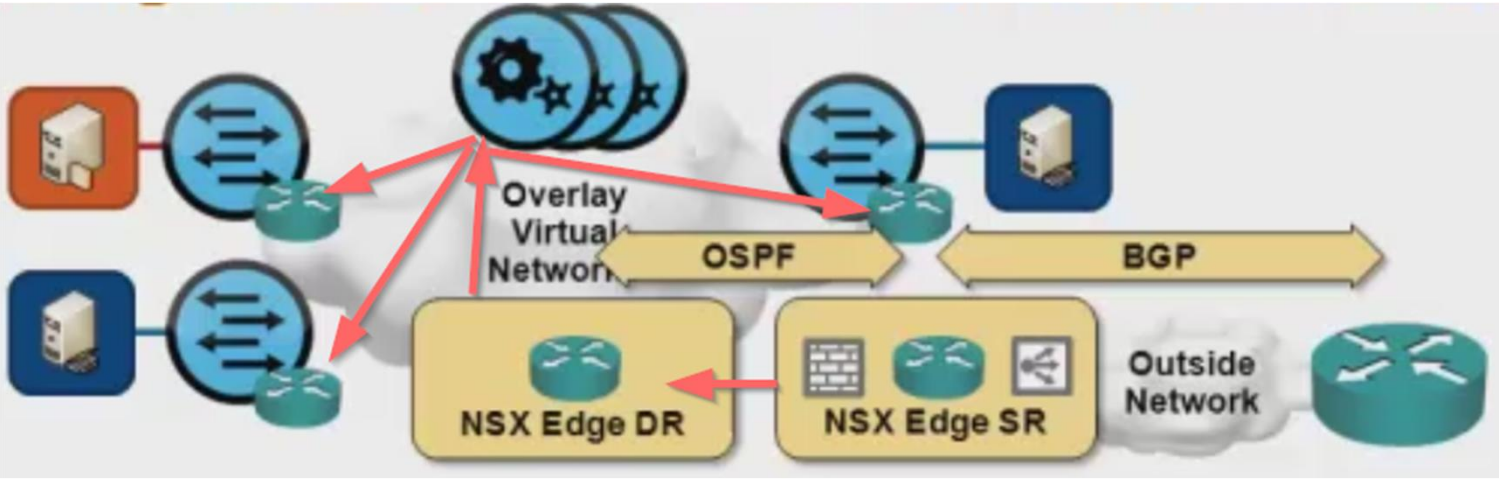
- L3 gateway通常为租户内网VM实例通过float ip与外网通信，提供NAT服务；或者直接配置公网ip的VM做默认路由网关。
  - 支持BGP、OSPF、IS-IS，为Distributed Router Control VM 提供租户网络路由拓扑，最终汇总到control cluster。
  - 每个L3 gateway支持8个北向interface，1000个南向interface。
- 使用 FHRP/VRRP 等协议，多个gateway实现HA。





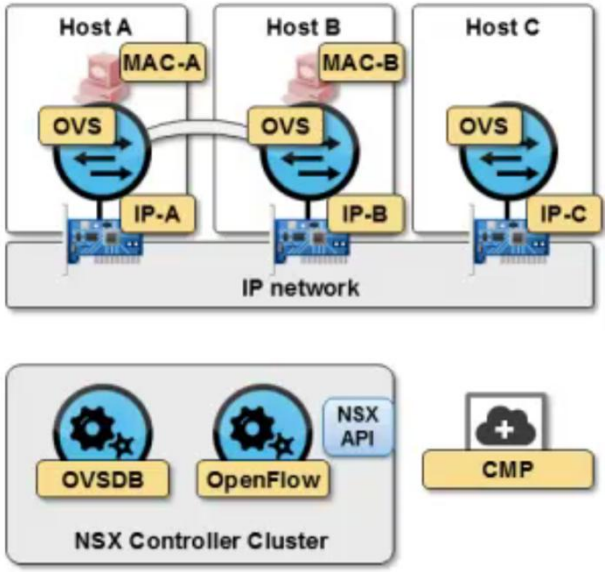
NSX Distributed Router Control VM

- 通过vSwitch的port提高L2 gateway。
- 租户路由汇聚
  - 通过OSPF、BGP等学习到NSX Edge L3 gateway的路由表。
  - 上传给Controller Cluster。
  - Controller Cluster将路由表同步给各个vSwitch的Distributed Logical Router 。



NSX控制平面

- Controller Cluster目标
  - 处理网络拓扑变更，分发配置以及流信息，构建逻辑网络。
- Controller Cluster实现
  - 通过SSL-over-TCP连接配置data plane。



1. 用户通过管理平面启动新VM实例C
2. controller cluster获得C的信息，在ovsdb建立a-c，b-c的条目。
3. controller cluster通知数据平面ovs建立of接口。
4. controller cluster将新的条目推送到数据平面的ovs。

NSX控制平面 cont.

- 通过User World Agent (UWA)，获取到每个虚拟交换机的MAC表，汇总后，将每个租户的全网MAC表同步给响应的虚拟交换机。
- 将从Logical Router Control VM 学到的路由，通过每个host上运行的UWA，分发给host的Distributed Logical Router模块。

NSX管理平面

- NSX manage功能
  - 唯一配置NSX的入口，提供UI， 支持使用REST API。
  - 完成部署controll cluster以及vSwitch的VMM module。
  - 添加与配置hypervisors以及对应的vSwitch。
  - 添加与配置虚拟网络设备（edge、service node）。
  - troubleshooting与信息采集。