OPENSTACK DAYS
CHINA

Topic: VXLAN EVOLVE IN OPENSTACK

Speaker: 王为（wangwei@unitedstack.com）

Agenda

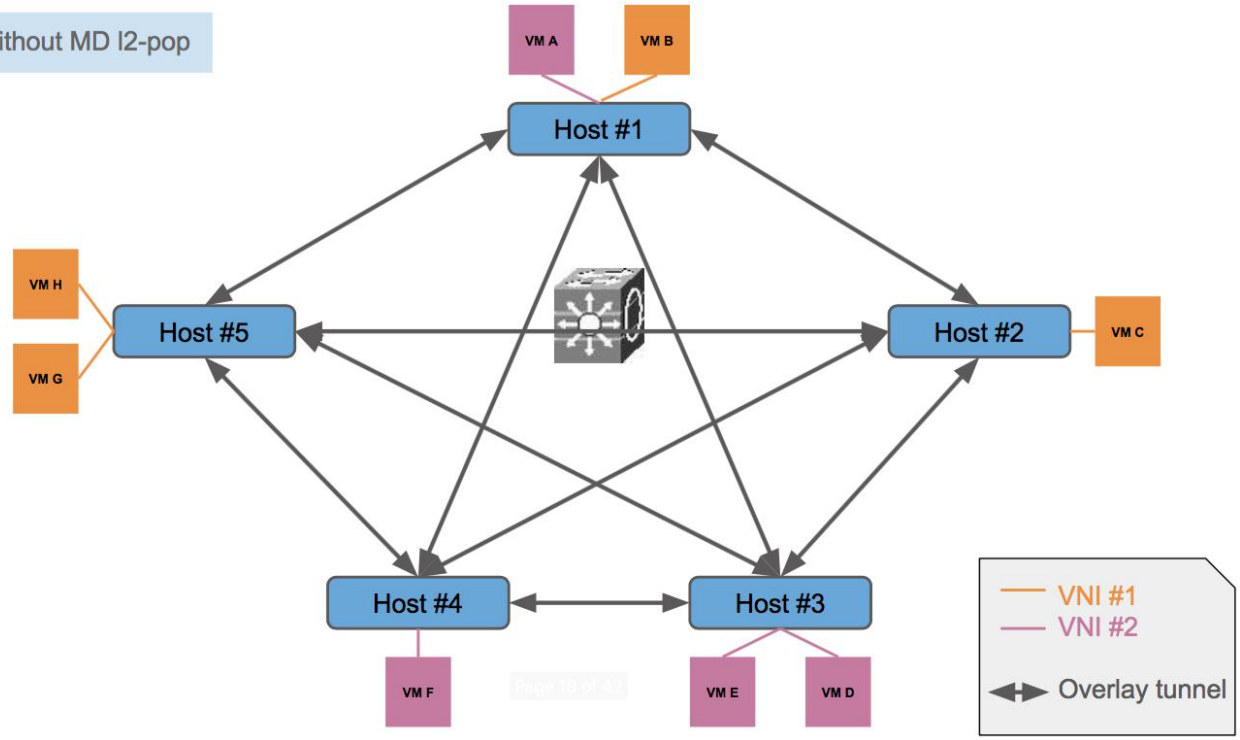
- OVS VXLAN: Flood&Learn;
- BaGPipe: EVPN implemented by software;
- DVR: Neutron as the control plane;
- MP BGP EVPN: Industry VXLAN control plane;
- Networking-bgpvpn: DVR with BGP, Routed Network;
- BGP-EVPN-Advertisement: Bring DirectConnect to OpenStack?

OPENSTACK DAYS CHINA

Agenda

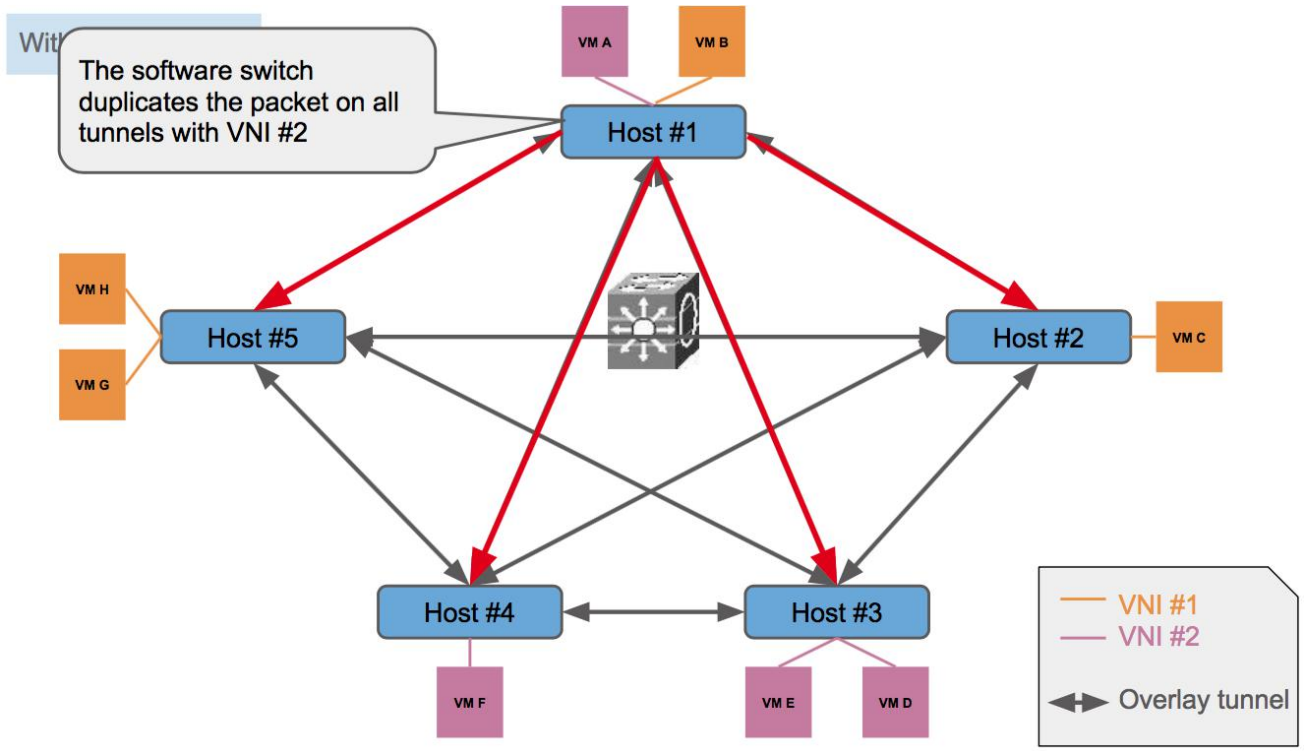
- *OVS VXLAN: Flood&Learn;*
- BaGPipe: EVPN implemented by software;
- DVR: Neutron as the control plane;
- MP BGP EVPN: Industry VXLAN control plane;
- Networking-bgpvpn: DVR with BGP, Routed Network;
- BGP-EVPN-Advertisement: Bring DirectConnect to OpenStack?

OPENSTACK DAYS CHINA



Multicast Based VXLAN

基于组播的 VXLAN 网络其实是没有控制平面的，依赖于数据平面的 flood-and-learn，如果交换机不支持组播的话，将会退化到广播，目前这类的应用已经很少了。



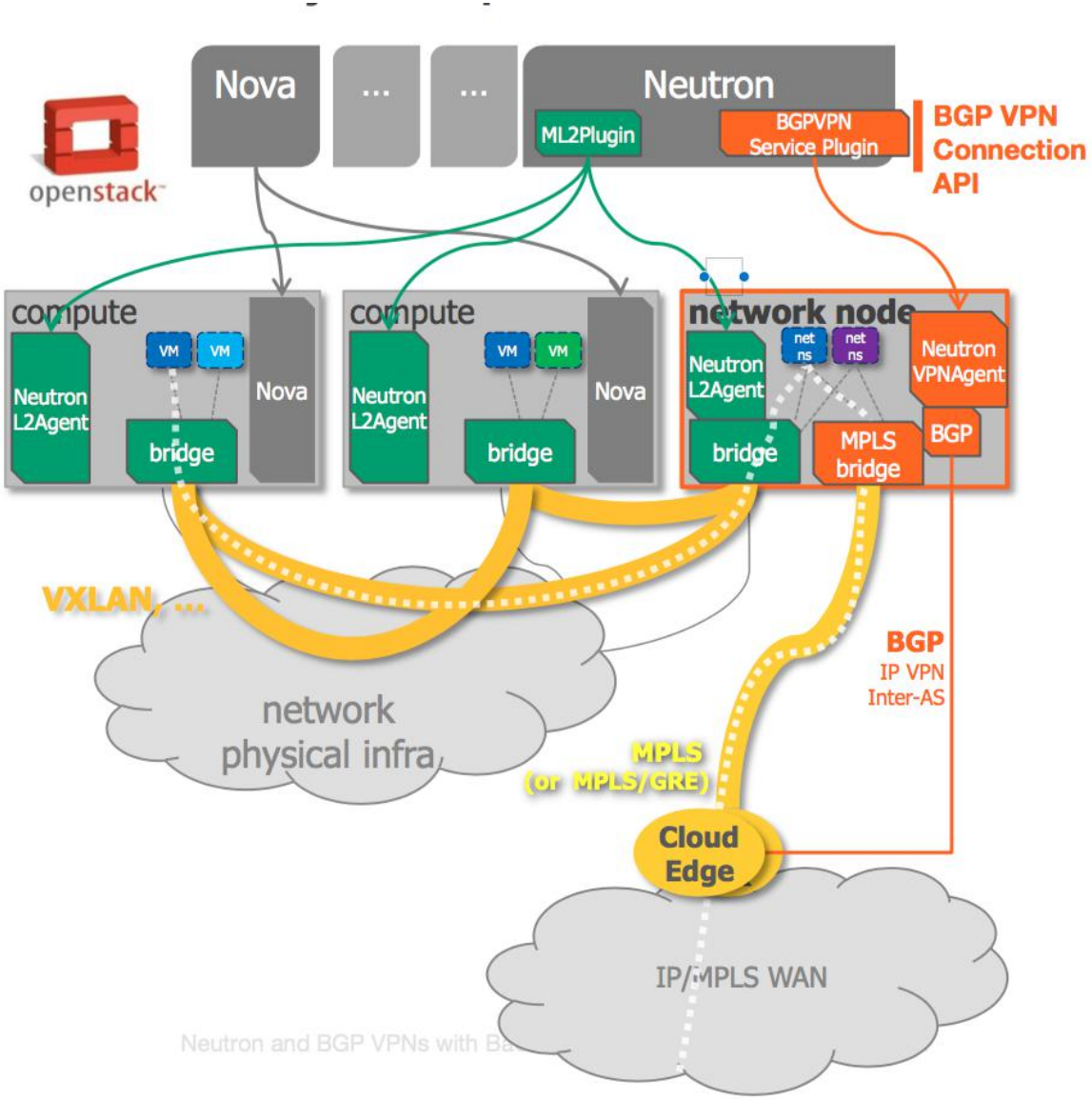
Multicast Based VXLAN

OVS 目前不支持组播 VXLAN，因此会通过 HER 的方式复制到所有隧道上。

Agenda

- OVS VXLAN: Flood&Learn;
- *BaGPipe: EVPN implemented by software;*
- DVR: Neutron as the control plane;
- MP BGP EVPN: Industry VXLAN control plane;
- Networking-bgpvpn: DVR with BGP, Routed Network;
- BGP-EVPN-Advertisement: Bring DirectConnect to OpenStack?

OPENSTACK DAYS CHINA

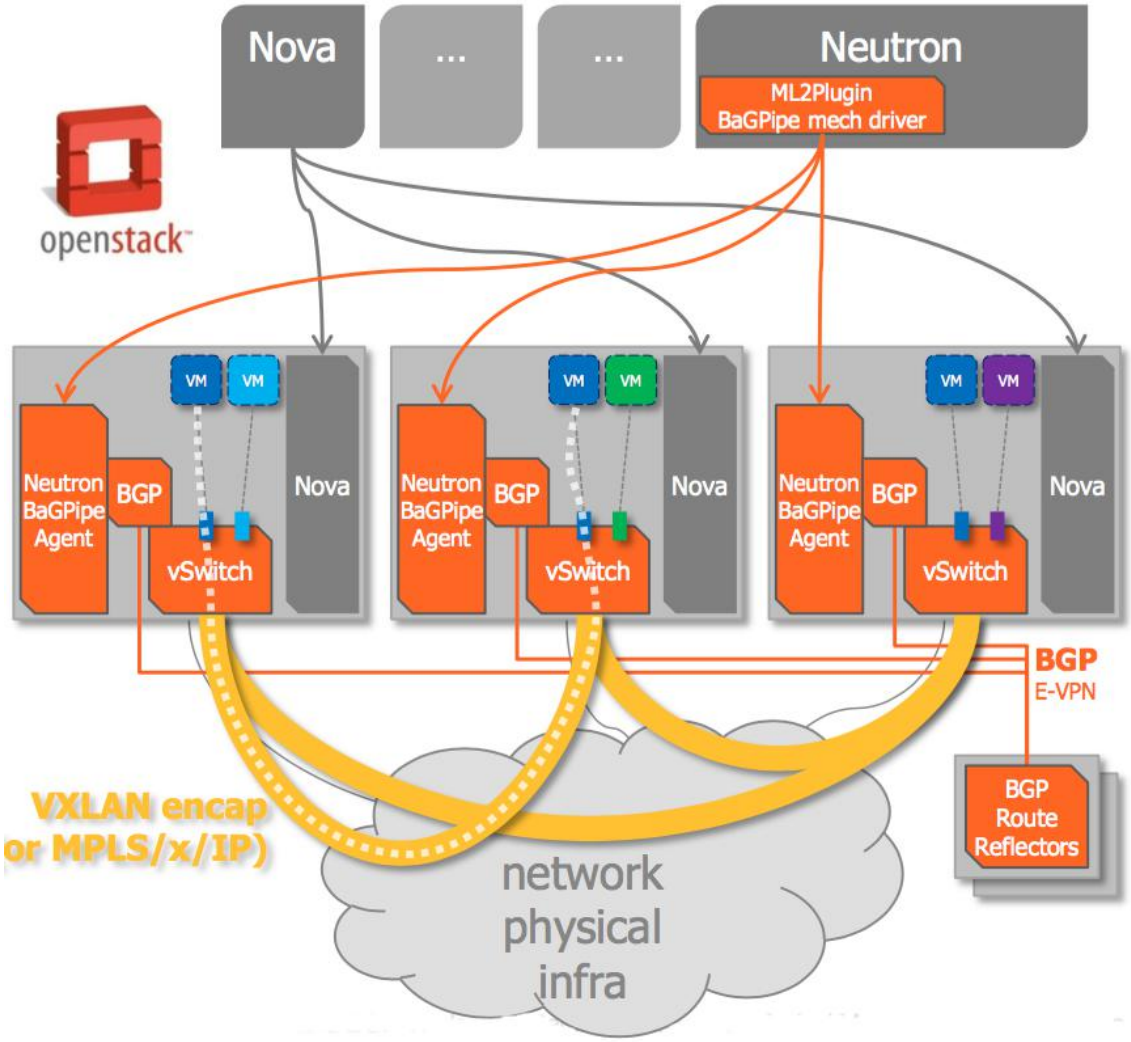


BaGPipe

基于 ExaBGP 实现的轻量级的 EVPN 实现。

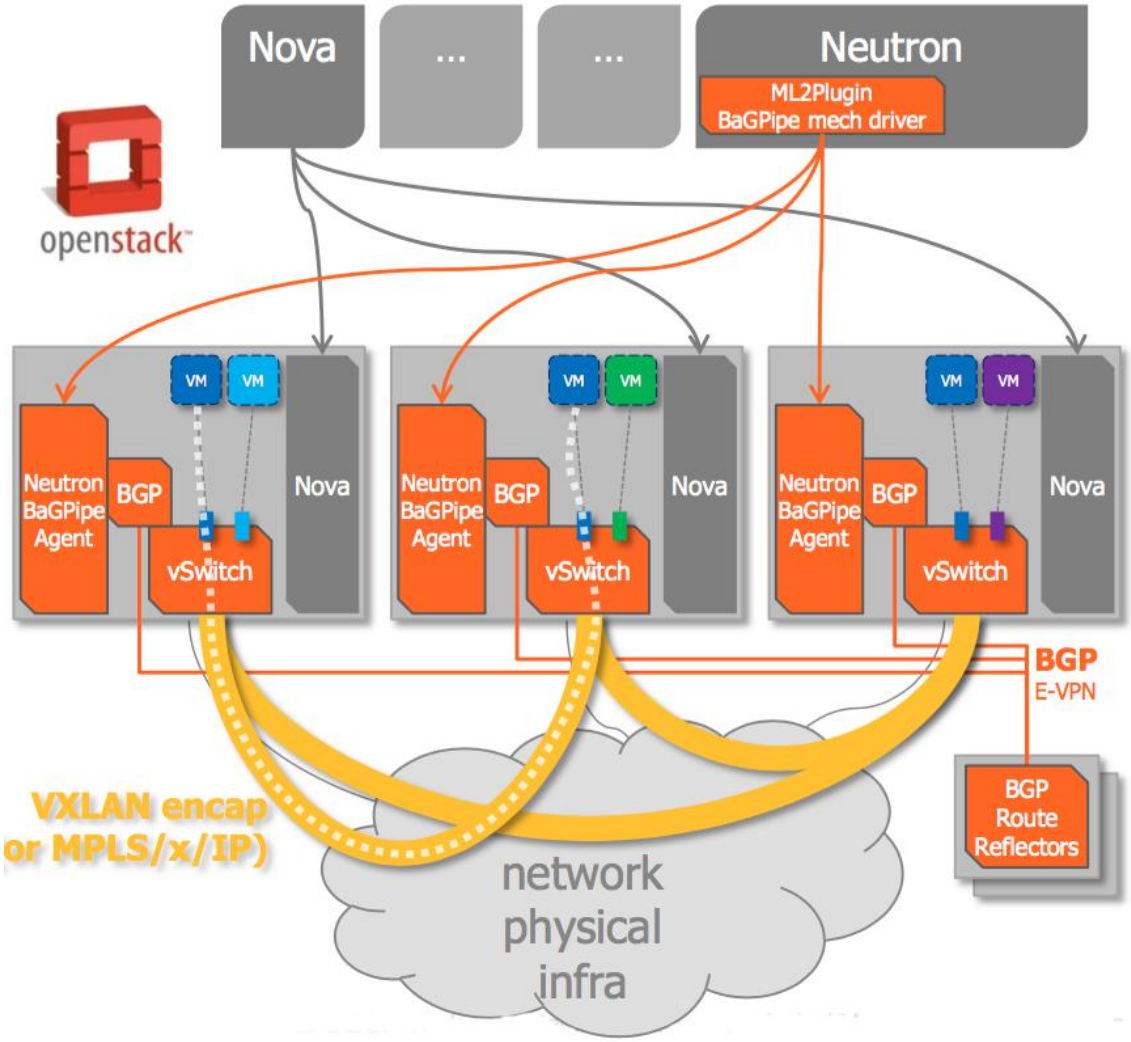
BGP RR 可以使用硬件设备或其他开源的 BGP 实现（Quagga、OpenContrail、GoBGP、BGPd）。

一种 Use case 是作做外网路由宣告。



BaGPipe

另一种更为重要的 Use case 是提供 EVPN 实现的分布式路由，参考 [BGP MPLS-Based Ethernet VPN (RFC7432)]。Scalability 和标准化得到很大的提高，方便接入网元设备和 L2 Gateway，接入其他 VXLAN 网络。



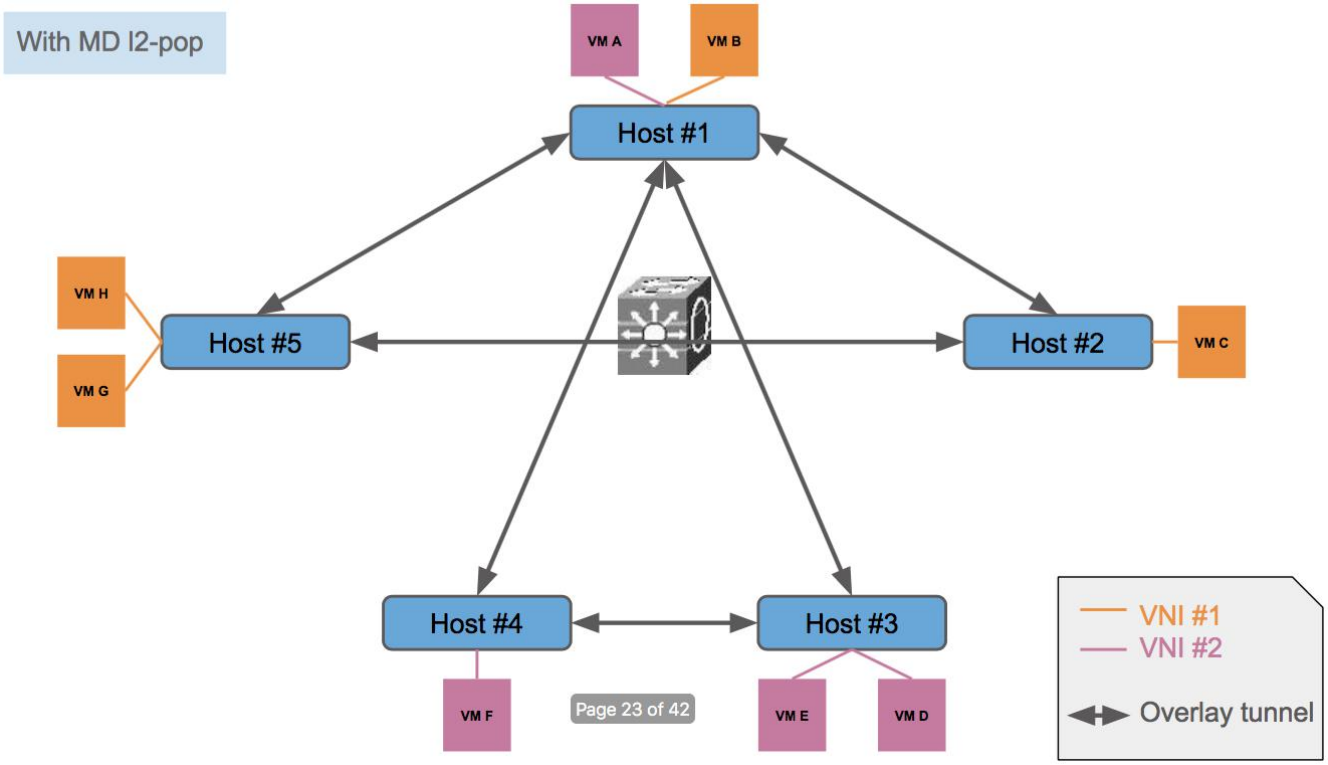
BaGPipe

另一种更为重要的 Use case 是提供 EVPN 实现的分布式路由，参考 [BGP MPLS-Based Ethernet VPN (RFC7432)]。Scalability 和标准化得到很大的提高，方便接入网元设备和 L2 Gateway，接入其他 VXLAN 网络。

Agenda

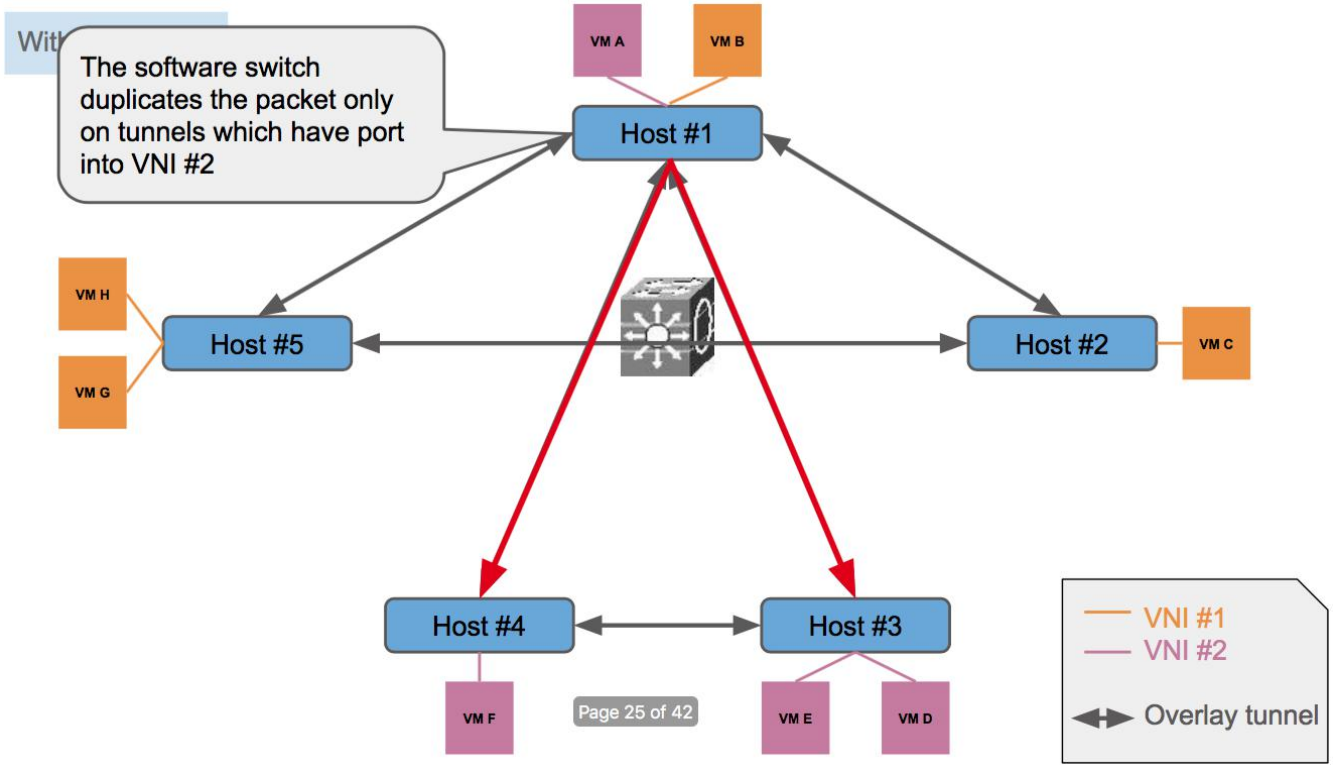
- OVS VXLAN: Flood&Learn;
- BaGPipe: EVPN implemented by software;
- *DVR: Neutron as the control plane;*
- MP BGP EVPN: Industry VXLAN control plane;
- Networking-bgpvpn: DVR with BGP, Routed Network;
- BGP-EVPN-Advertisement: Bring DirectConnect to OpenStack?

OPENSTACK DAYS CHINA



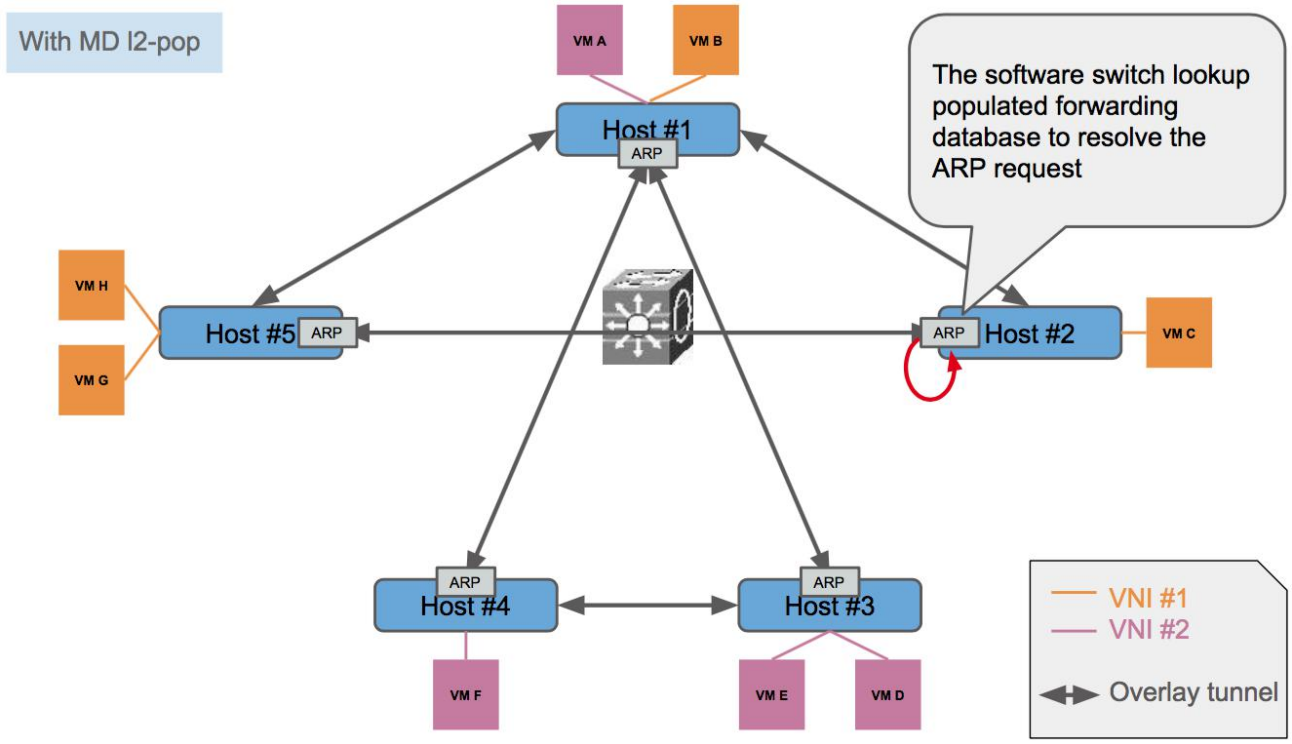
DVR & L2 Population

DVR、L2 Population、ARP Responder 从多个角度优化了网络流量。首先，隧道的建立大幅减少。



DVR & L2 Population

其次，广播只会复制到正确的 VTEPs 上，单播可以直接传输到正确的 VTEP。



DVR & L2 Population

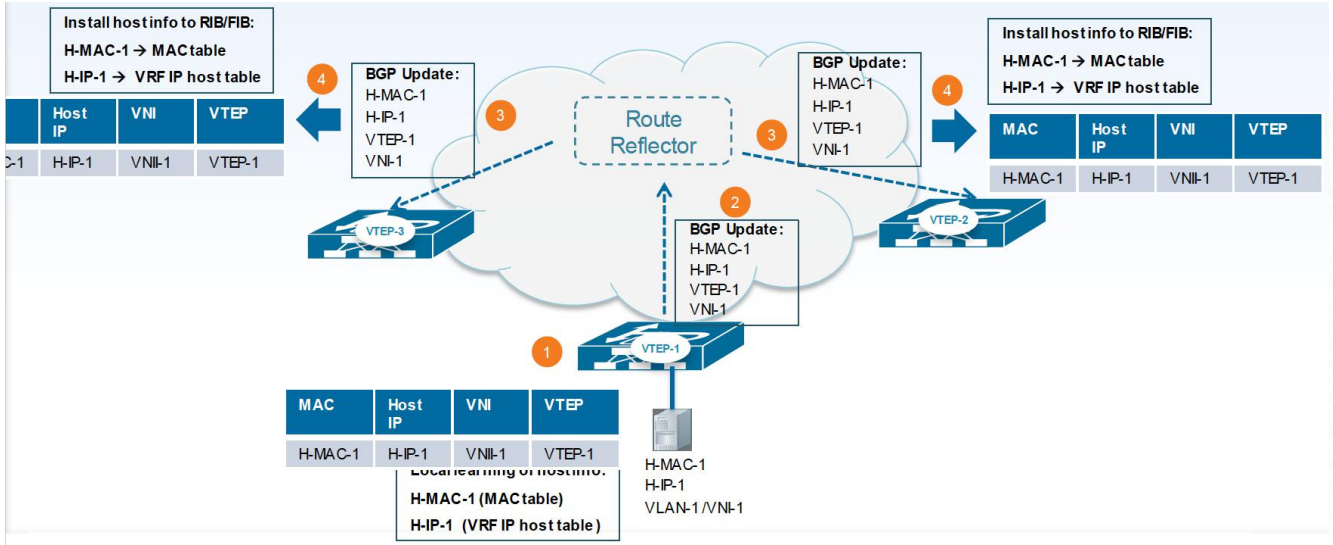
此外，同子网的 ARP 将由 ARP Responder 应答，跨子网的 ARP 缓存在本地的 DVR 虚拟路由器内。

Agenda

- OVS VXLAN: Flood&Learn;
- BaGPipe: EVPN implemented by software;
- DVR: Neutron as the control plane;
- *MP BGP EVPN: Industry VXLAN control plane;*
- Networking-bgpvpn: DVR with BGP, Routed Network;
- BGP-EVPN-Advertisement: Bring DirectConnect to OpenStack?

OPENSTACK DAYS CHINA

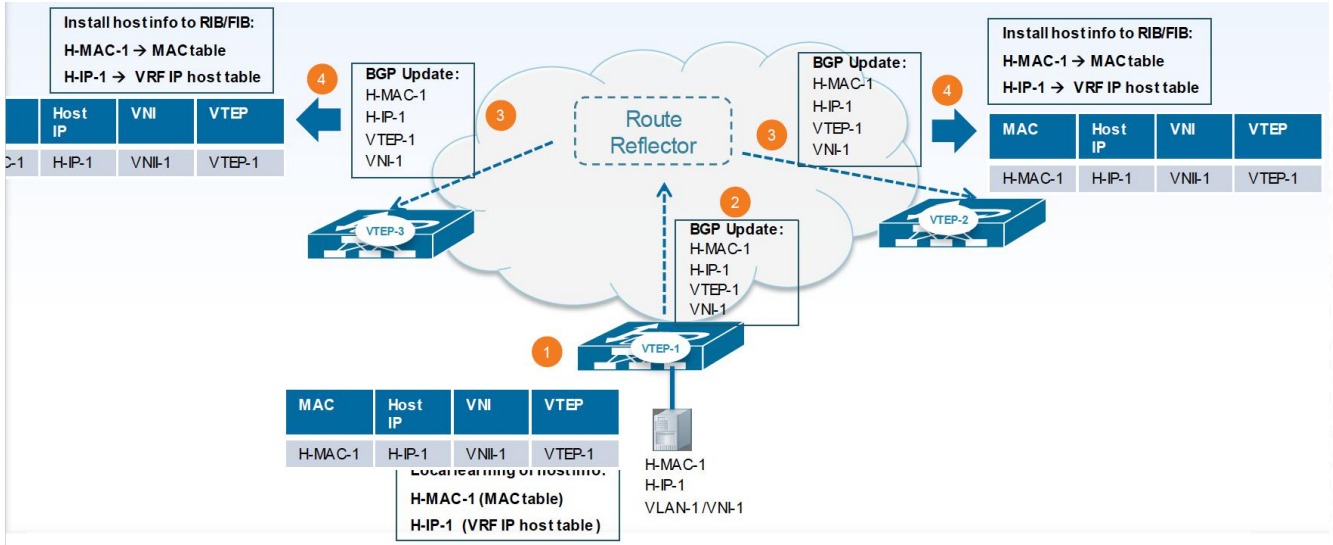
MP BGP EVPN



参考 [BGP MPLS-Based Ethernet VPN (RFC7432)](), [draft-ietf-bess-evpn-inter-subnet-forwarding](), [draft-ietf-rtgwg-bgp-routing-large-dc]()，描述了使用 BGP 作为 VXLAN 的控制平面的参考设计。通过 EVPN，MAC 的学习将类似于三层网络路由的学习，将有助于有效减少泛洪。

OPENSTACK DAYS CHINA

MP BGP EVPN

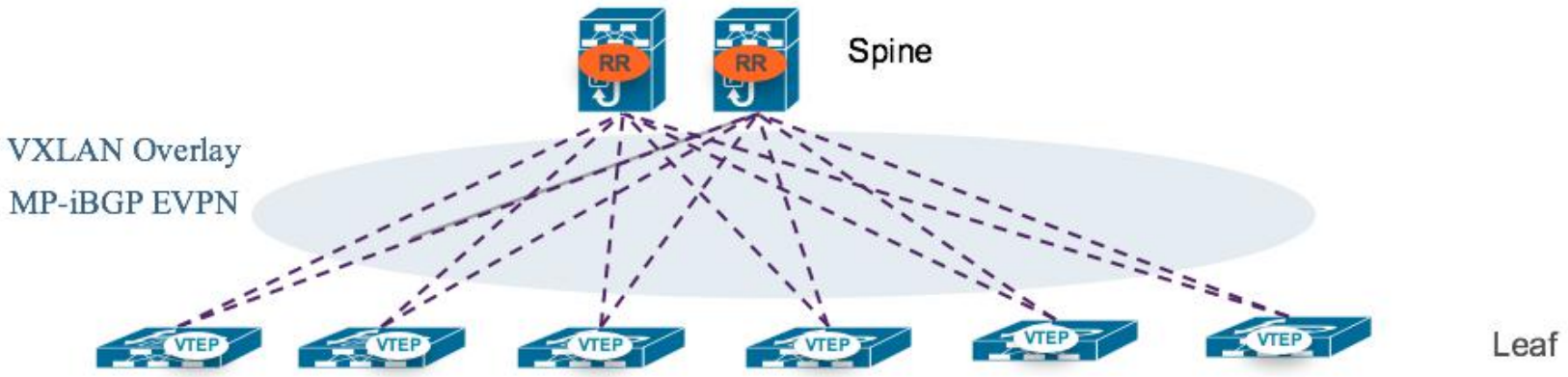


每一个 VTEP 将作为一个 BGP Speaker，向其他 VTEP 通过 EVPN 发送本地的 MAC、IP 信息，BGP RR 可以避免 BGP 的 Full-Mesh，提高通信效率。得益于控制平面，每个 VTEP 将可作为分布式网关、可以抑制 ARP 广播、可以将广播或组播通过单播复制来提升效率、可以对 VTEP 进行认证。

MP BGP EVPN

具体到 BGP 租网上，有几种选择，包括 iBGP、eBGP 的选择和外部网络拓扑。

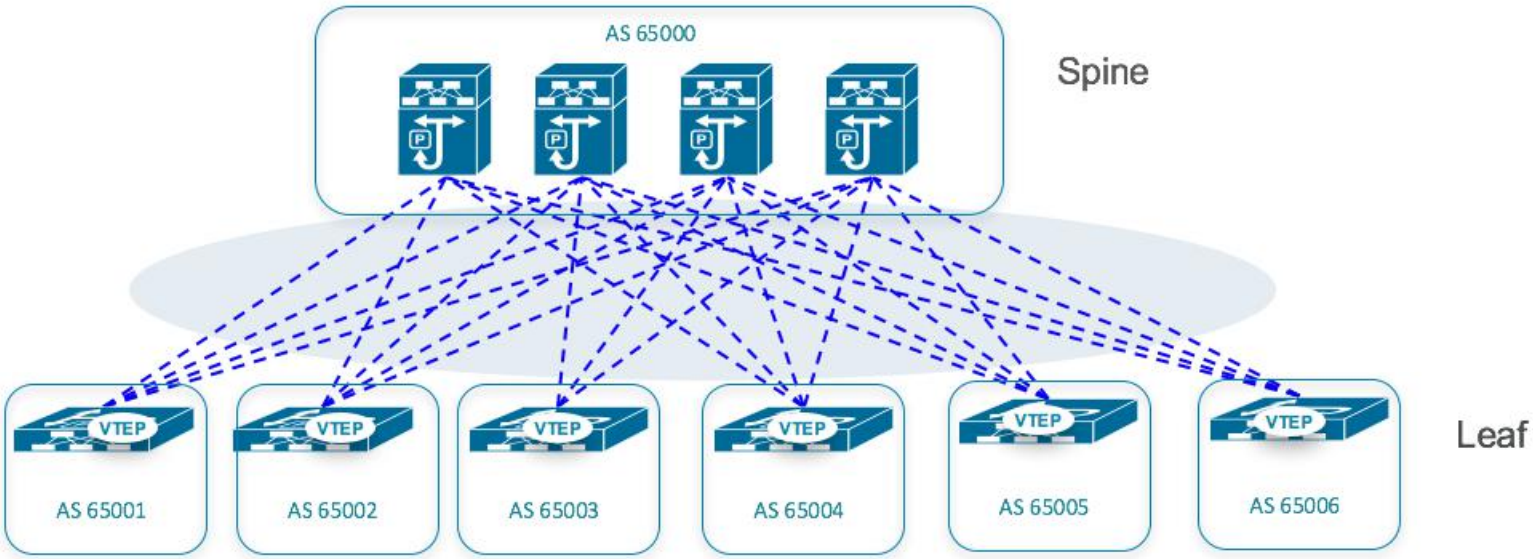
- Spine交换机设置为iBGP的RR，Spine交换机不需要支持VxLAN
- Leaf交换机与Spine交换机建立iBGP邻居关系



MP BGP EVPN

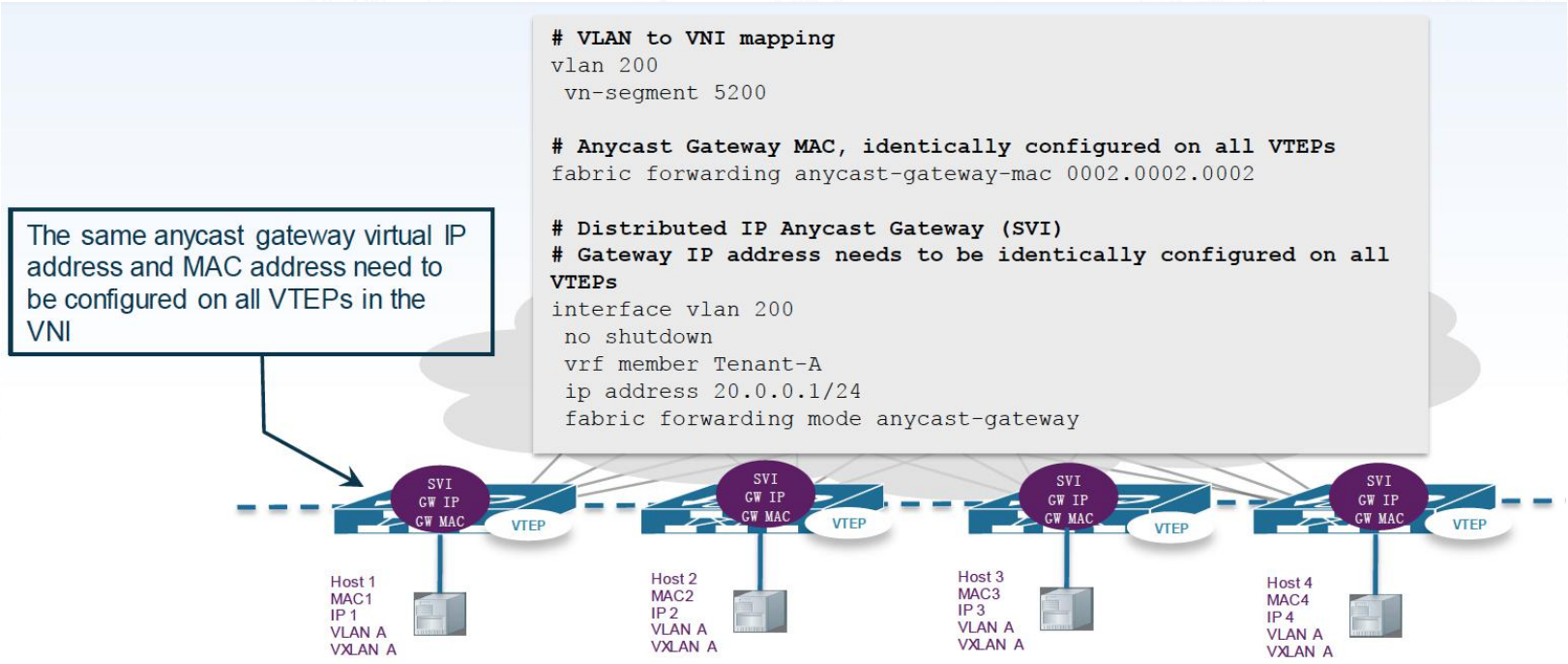
具体到 BGP 租网上，有几种选择，包括 iBGP、eBGP 的选择和外部网络拓扑。

- Spine交换机设置为iBGP的RR，Spine交换机不需要支持VxLAN
- Leaf交换机与Spine交换机建立eBGP邻居关系，Spine交换机间建立iBGP邻居关系



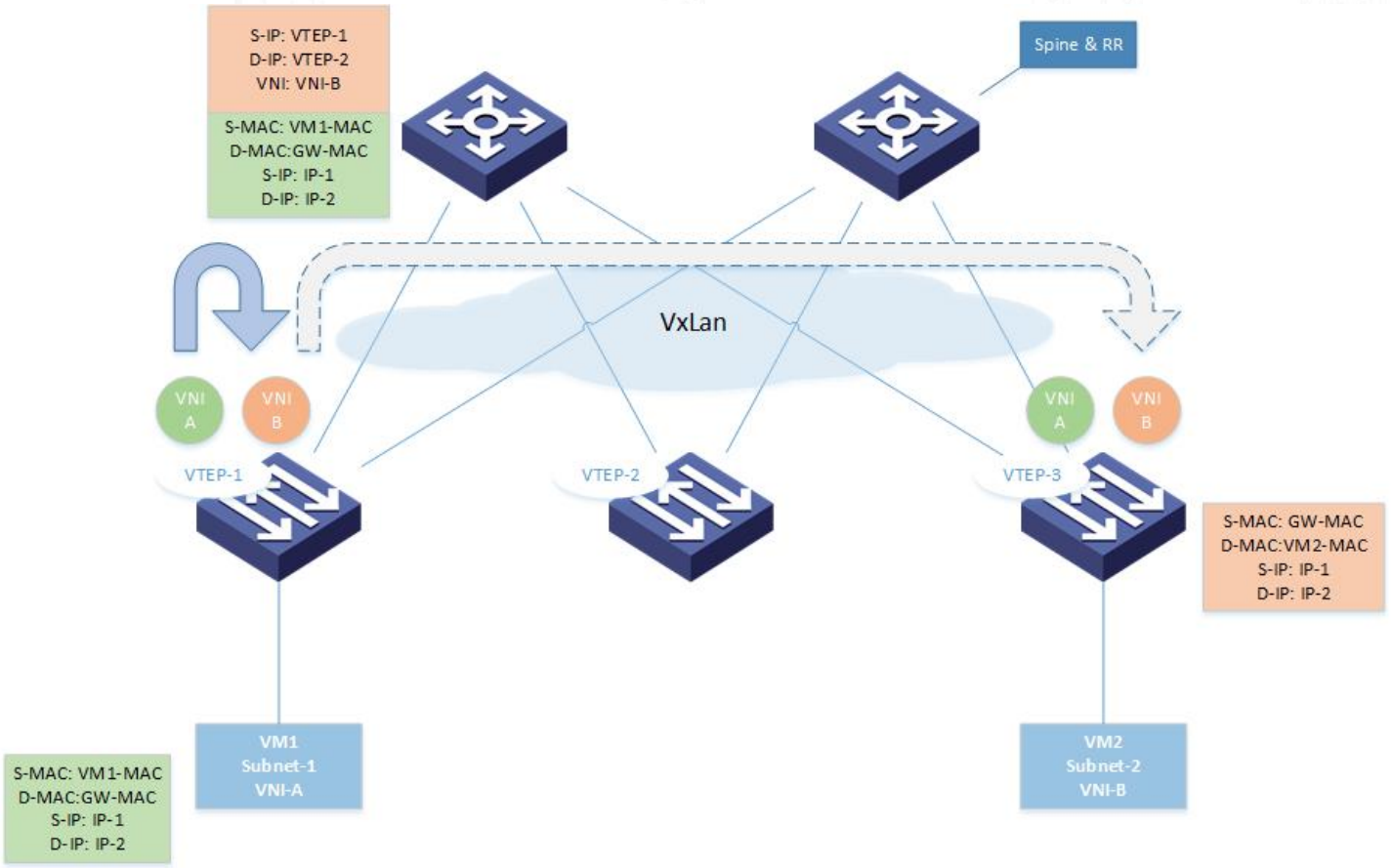
MP BGP EVPN

EVPN VxLan 的实际路由过程可以分成两步来谈，第一部分是虚拟机的 First-hop 的地址，即网关地址，第二部分是如何在不同 VxLan 间路由（IRB）



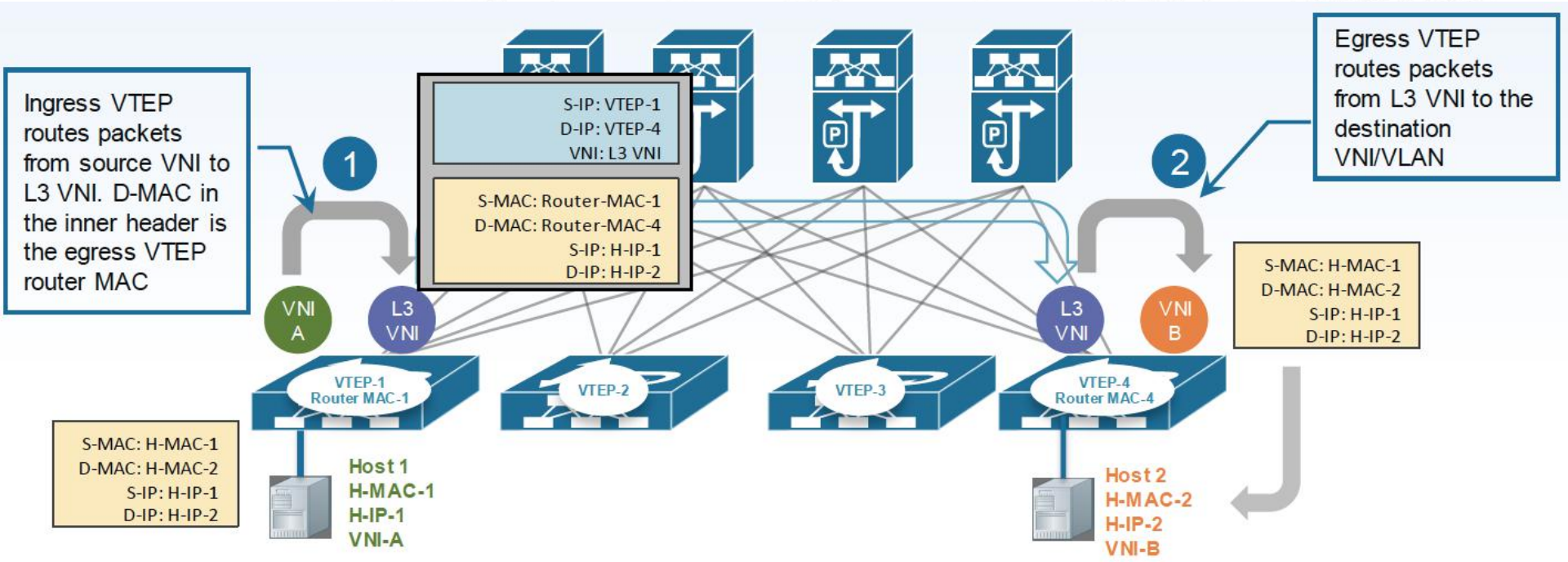
MP BGP EVPN

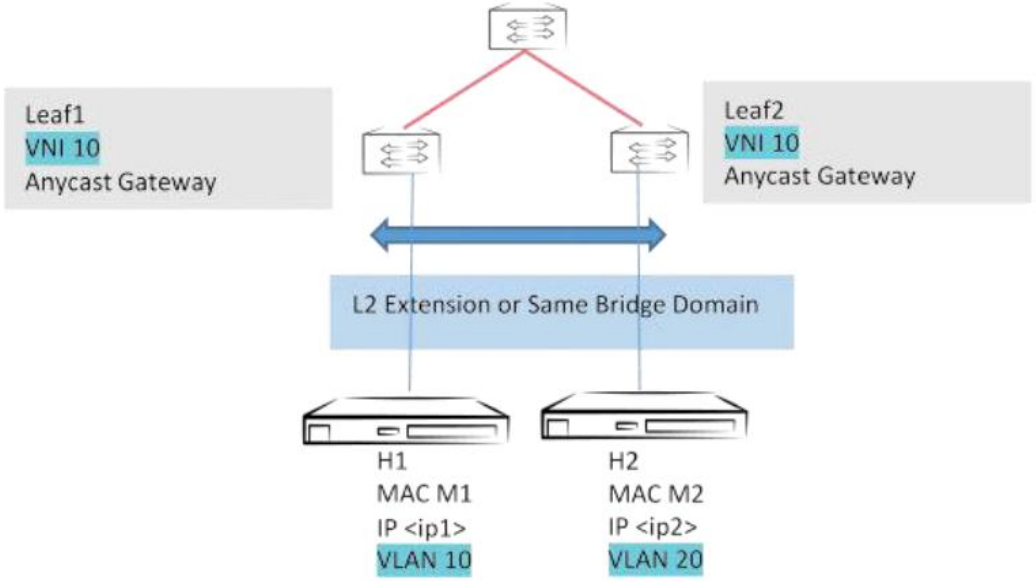
IRB 即 VTEP 提供三层和二层功能，但是对于具体如何路由，目前存在两种方法，分别为 Asymmetric IRB mode 和 Symmetric IRB mode。前者是非对称模式，后者是对称模式

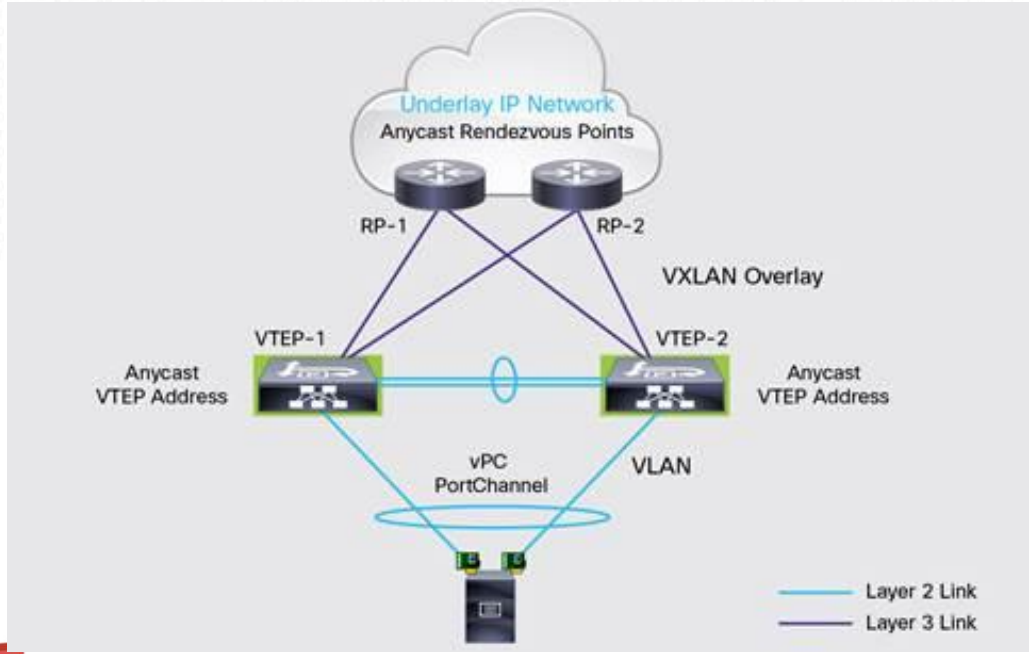


MP BGP EVPN

另一种实现方法是 Symmetric IRB，其实现与 Asymmetric IRB 最显著的不同是源 VTEP 和目标 VTEP 都会承担三层和二层功能，而不像 Symmetric IRB 只在源 VTEP 做路由。这样最终实现是对称的，但前提是必须引入一个新的概念即 L3 VNI。







MP BGP EVPN

其他相关技术，如 VLAN Scoping、vPC/vLAG。
同一个 VXLAN 在不同的 VTEP 下可以对应不同的 VLAN，目前 VXLAN 到 VLAN 映射有 Leaf 层面和 Port 层面两种实现。
为了解决 Leaf 的高可用问题，一对 vPC peer 共享一个 VTEP 地址（anycast VTEP address），组成一个逻辑上的 VTEP，共同分担负载。但是实际上对于 EVPN，每个设备上的 Route ID 都是不同的，独立的广播 BGP 路由。

EVPN 实现	**Open vSwitch Driver VXLAN 实现**
Head-End Replication	Edge Replication
ARP Supression	ARP Responder
MAC advertise	L2 Population
Conversational Learning	Conversational Learning
Asymmetric IRB mode & Anycast Gateway	DVR
vPC/vLAG	Bonding/Teaming

OPENSTACK DAYS CHINA

Agenda

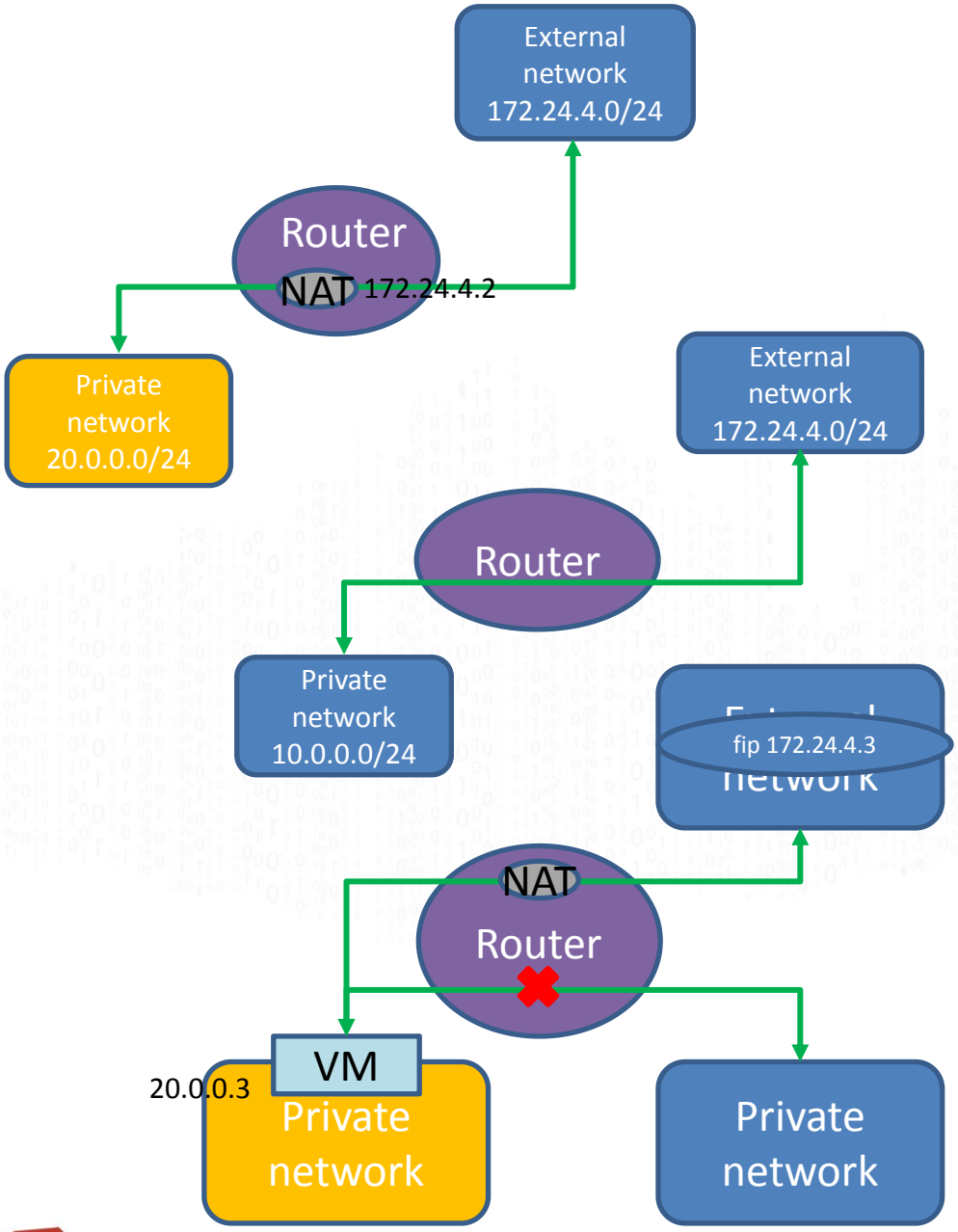
- OVS VXLAN: Flood&Learn;
- BaGPipe: EVPN implemented by software;
- DVR: Neutron as the control plane;
- MP BGP EVPN: Industry VXLAN control plane;
- *Networking-bgpvpn: DVR with BGP, Routed Network;*
- BGP-EVPN-Advertisement: Bring DirectConnect to OpenStack?

OPENSTACK DAYS CHINA

Applications of Neutron BGP Dynamic Routing

- Routed Model for Floating Range spanned in multiple L2 domains
- Directly Routable IPv4/IPv6 Tenant Networks
- DVR with BGP
- Directly Routable IPv4/IPv6 Tenant Networks

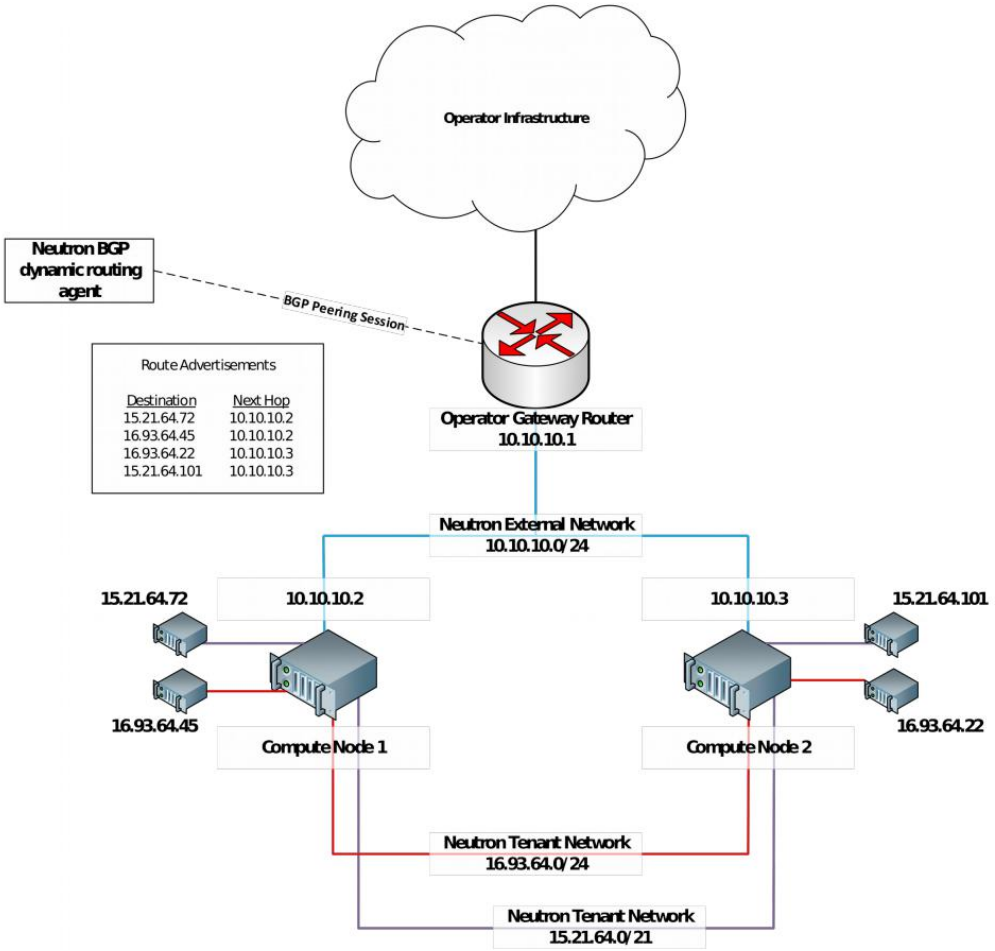
OPENSTACK DAYS CHINA



Networking-bgpvpn with address scope

同一 address scope 可以不做 NAT 直接路由，不同 address scope 需要 NAT 地址转换。

最后一个图为一个结合场景。

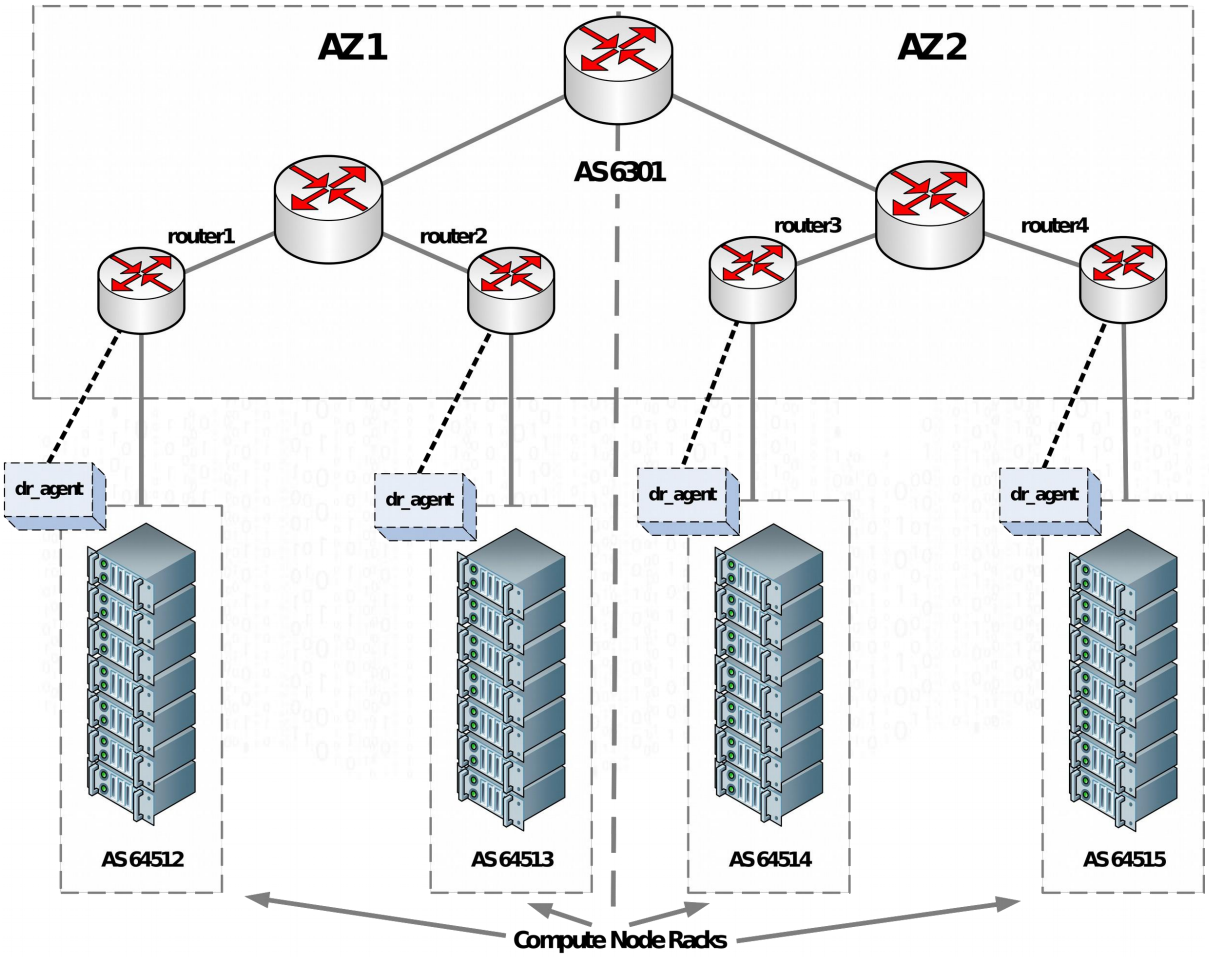


Networking-bgpvpn with DVR

特别对于 DVR 场景，租户网络通过 L2 Pop、DVR 等方式优化了网络效率，但外部网络分散在所有节点上，又造成了一个不可控的广播域。

通过 BGP 可以将二层的泛洪查找尽量转换成三层路由，减少了无效流量。

OPENSTACK DAYS CHINA



Networking-bgpvpn with DVR

针对特别大的网络规模，还可以通过规划和设计优化网络流量

Agenda

- OVS VXLAN: Flood&Learn;
- BaGPipe: EVPN implemented by software;
- DVR: Neutron as the control plane;
- MP BGP EVPN: Industry VXLAN control plane;
- Networking-bgpvpn: DVR with BGP, Routed Network;
- *BGP-EVPN-Advertisement: Bring DirectConnect to OpenStack?*

OPENSTACK DAYS CHINA

BGP-EVPN-Advertisement

- Connectivity to an on premises tenant site using L3VPN connectivity
- Connectivity between OpenStack regions
- Connectivity to a non-OpenStack public data center
- Still in spec review

OPENSTACK DAYS CHINA

Thanks