OPENSTACK DAYS
CHINA

Topic：VXLAN EVOLVE IN OPENSTACK

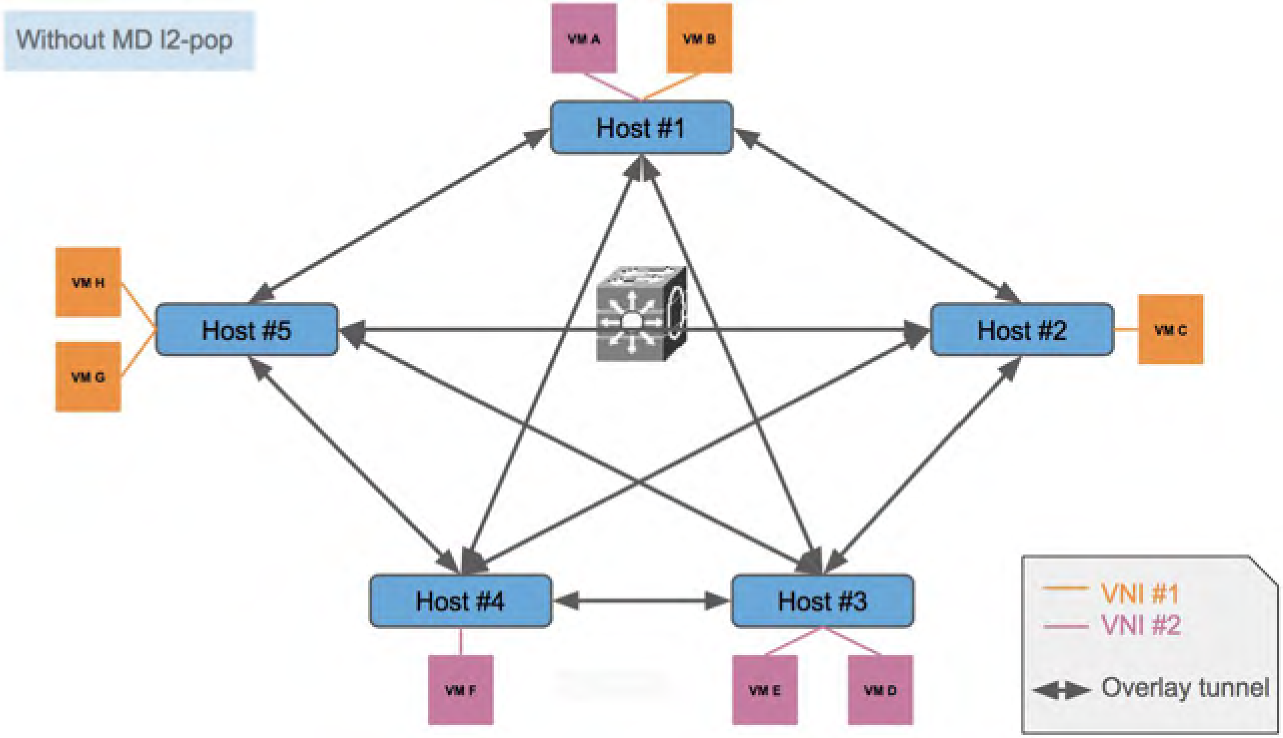
Speaker：王为（wangwei@unitedstack.com）

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?

Agenda

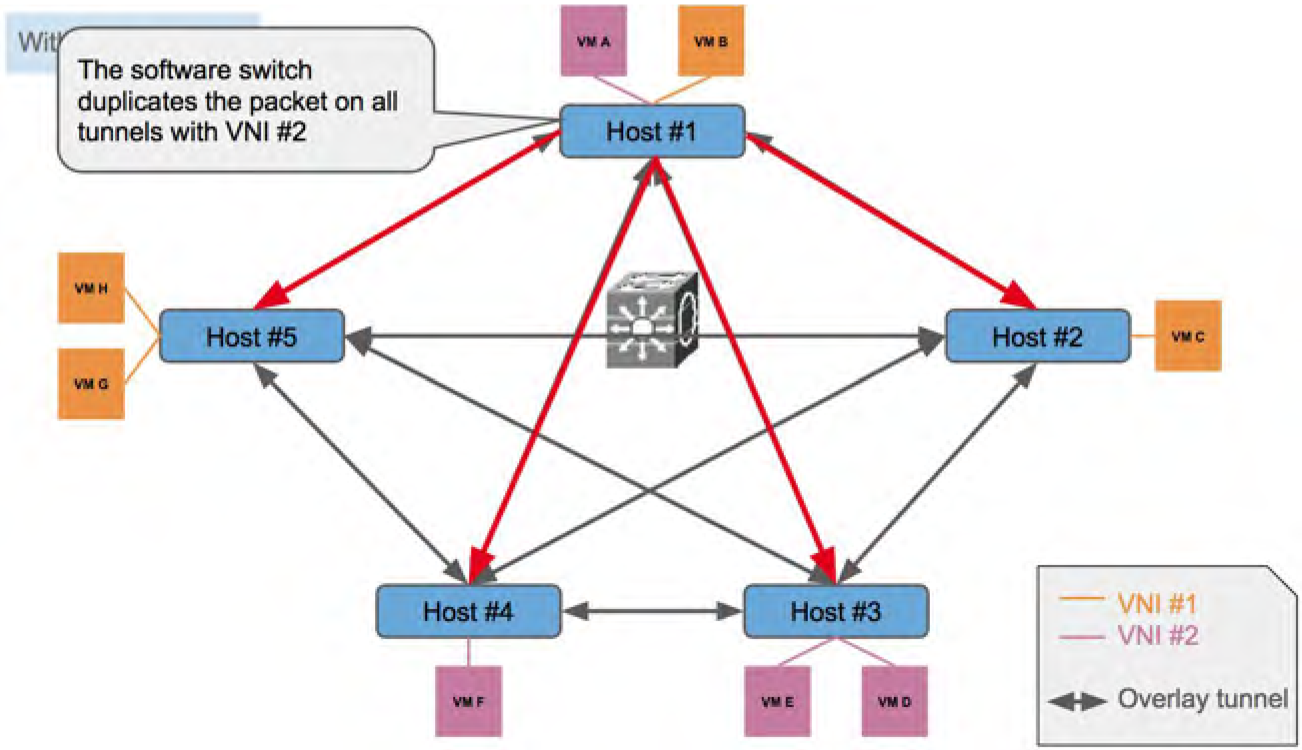
- *OVS VXLAN: Flood&Learn;*
- BaGPipe: EVPN implemented by software;
- DVR: Neutron as the control plane;
- MP BGP EVPN: Industry VXLAN control plane;
- Networking-bgpvpn: DVR with BGP, Routed Network;
- BGP-EVPN-Advertisement: Bring DirectConnect to OpenStack?


Without MD l2-pop
VM A
VM B
Host #1
VM H
VM G
Host #5
Host #2
VM C
Host #4
Host #3
VM F
VM E
VM D
VNI #1
VNI #2
Overlay tunnel

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



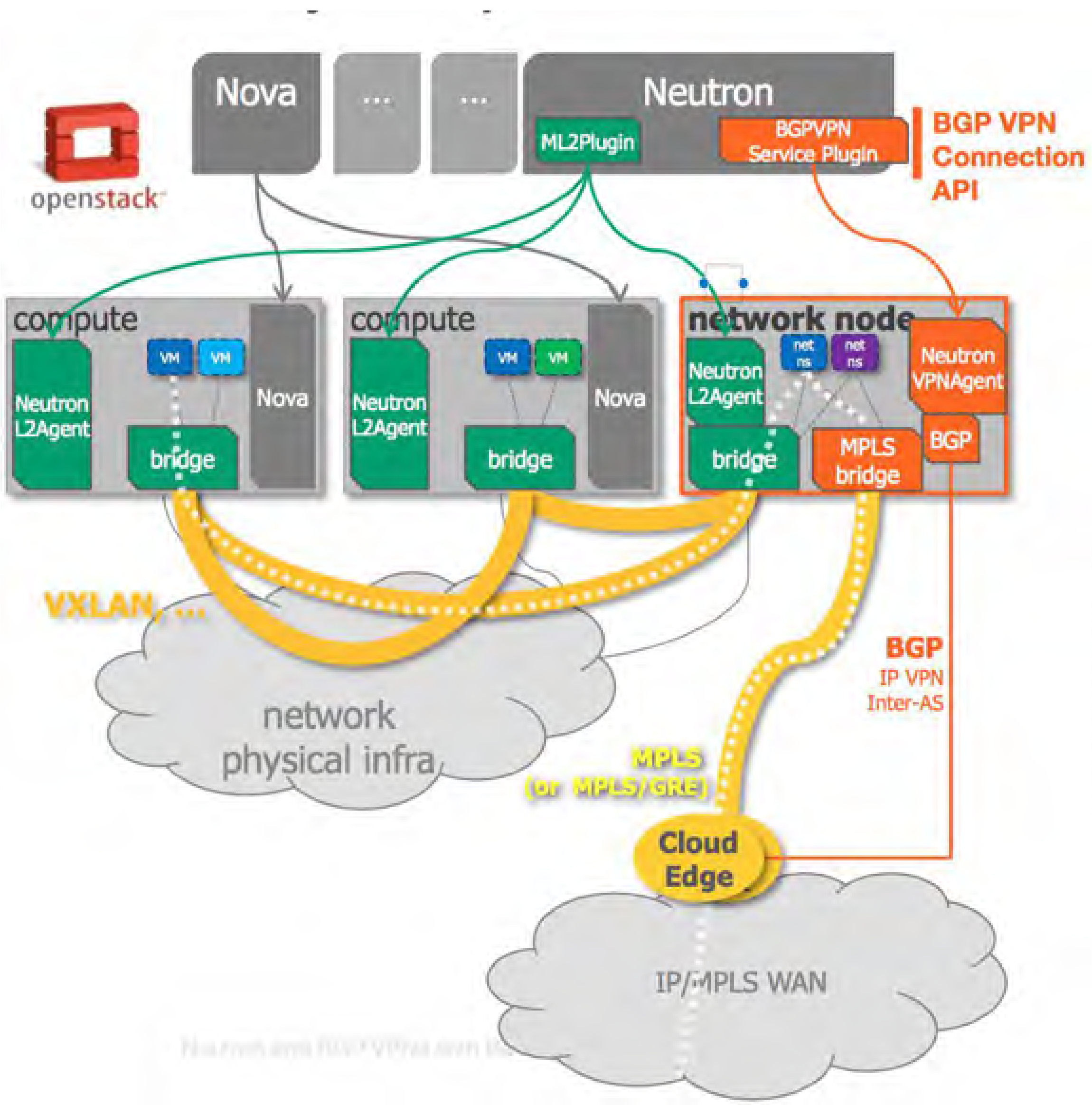
The software switch duplicates the packet on all tunnels with VNI #2

VM A
VM B
Host #1
VM H
VM G
Host #5
Host #2
VM C
Host #4
Host #3
VM F
VM E
VM D
VNI #1
VNI #2
Overlay tunnel

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。

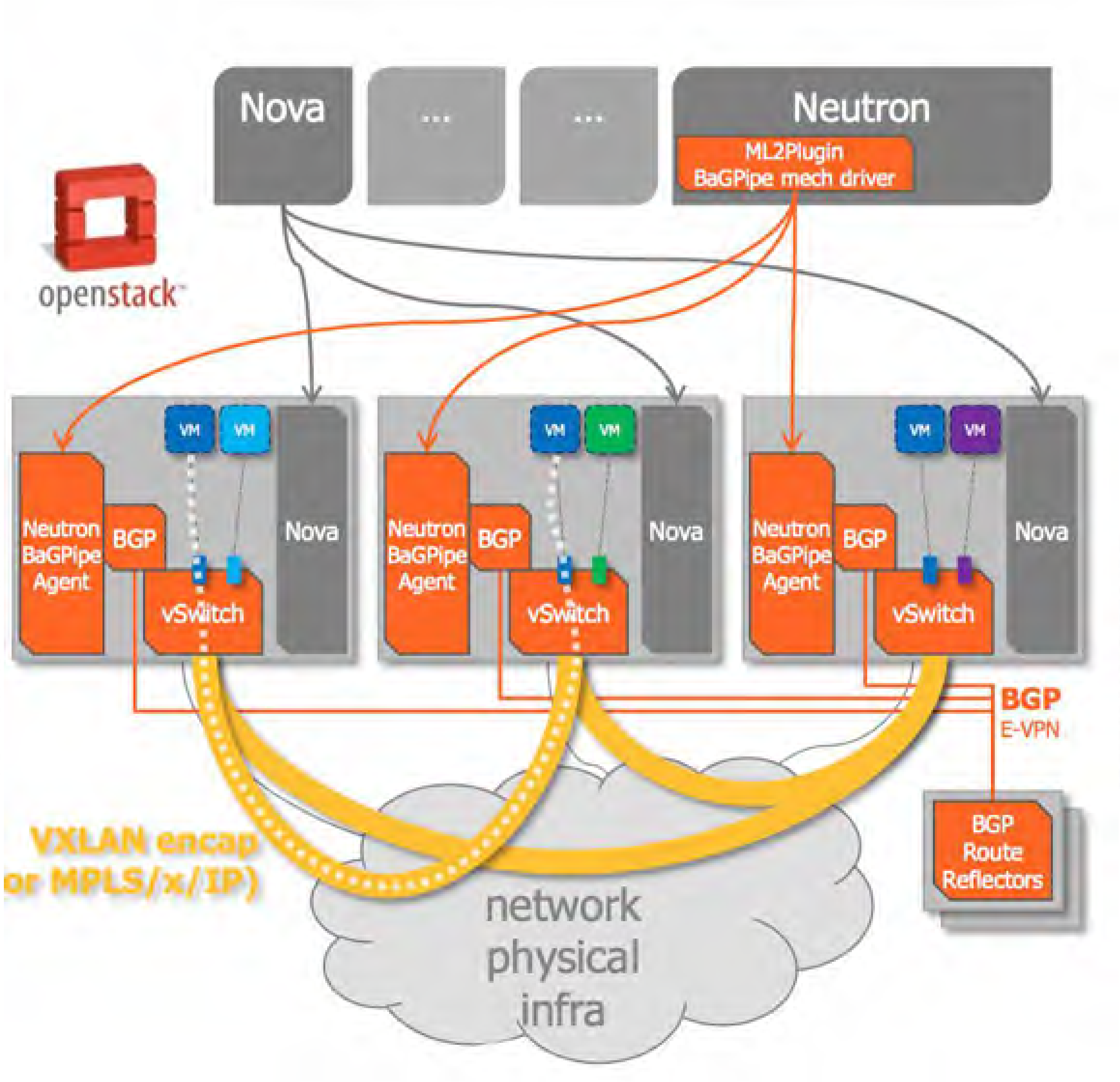
Agenda

- 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?


Nova
...
...
Neutron
ML2Plugin
BGPVPN Service Plugin
BGP VPN Connection API
openstack
compute
VM
VM
Neutron L2Agent
bridge
Nova
compute
VM
VM
Neutron L2Agent
bridge
Nova
network node
net ns
net ns
Neutron L2Agent
Neutron VPNAgent
bridge
MPLS bridge
BGP
VXLAN, ...
network physical infra
MPLS (or MPLS/GRE)
BGP
IP VPN
Inter-AS
Cloud Edge
IP/MPLS WAN

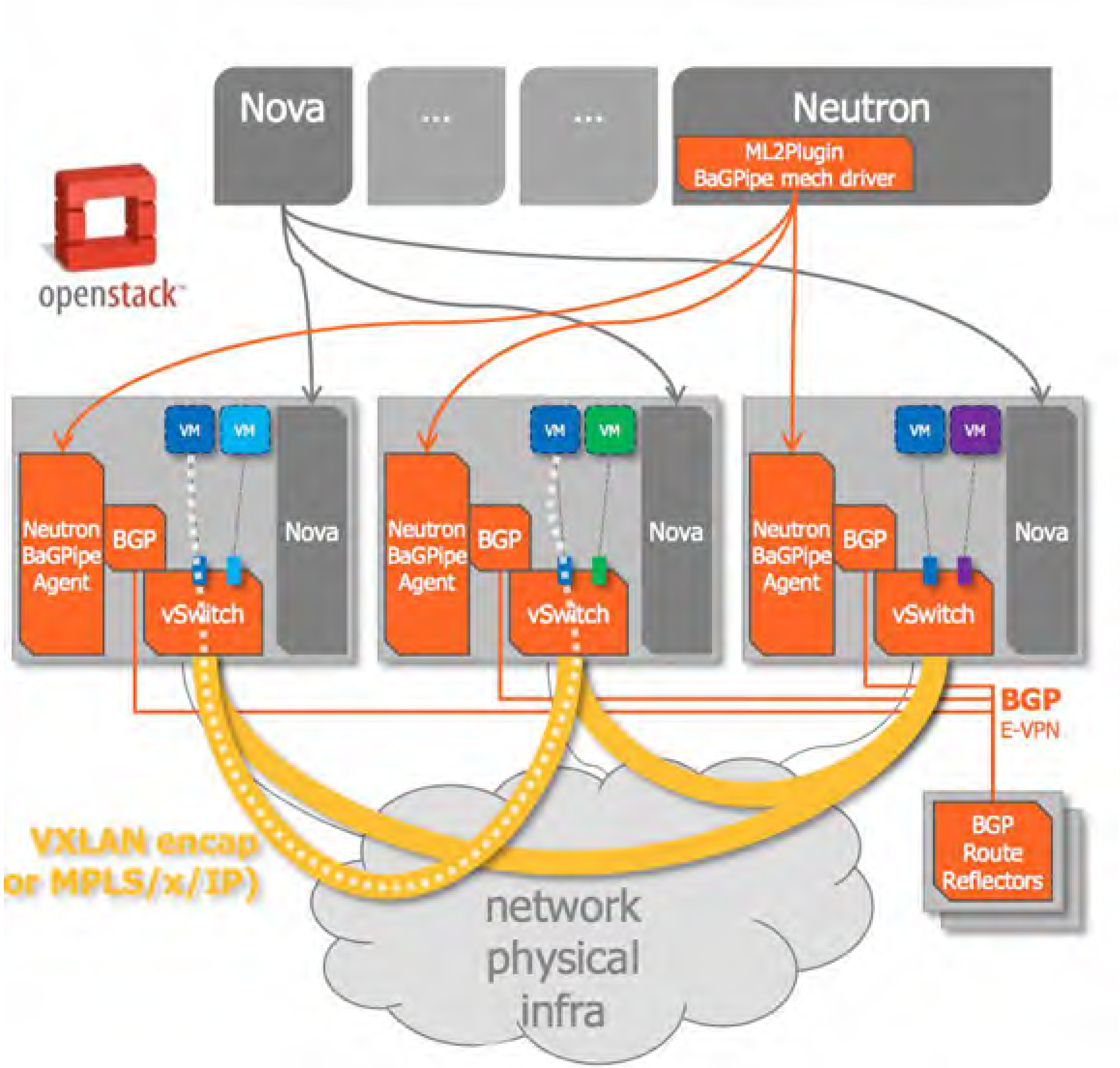
BaGPipe

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。


openstack
Nova
...
...
Neutron
ML2Plugin
BaGPipe mech driver
VM
VM
Neutron BaGPipe Agent
BGP
vSwitch
Nova
VM
VM
Neutron BaGPipe Agent
BGP
vSwitch
Nova
VM
VM
Neutron BaGPipe Agent
BGP
vSwitch
Nova
BGP
E-VPN
VXLAN encap
or MPLS/x/IP)
network
physical
infra
BGP
Route
Reflectors

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。

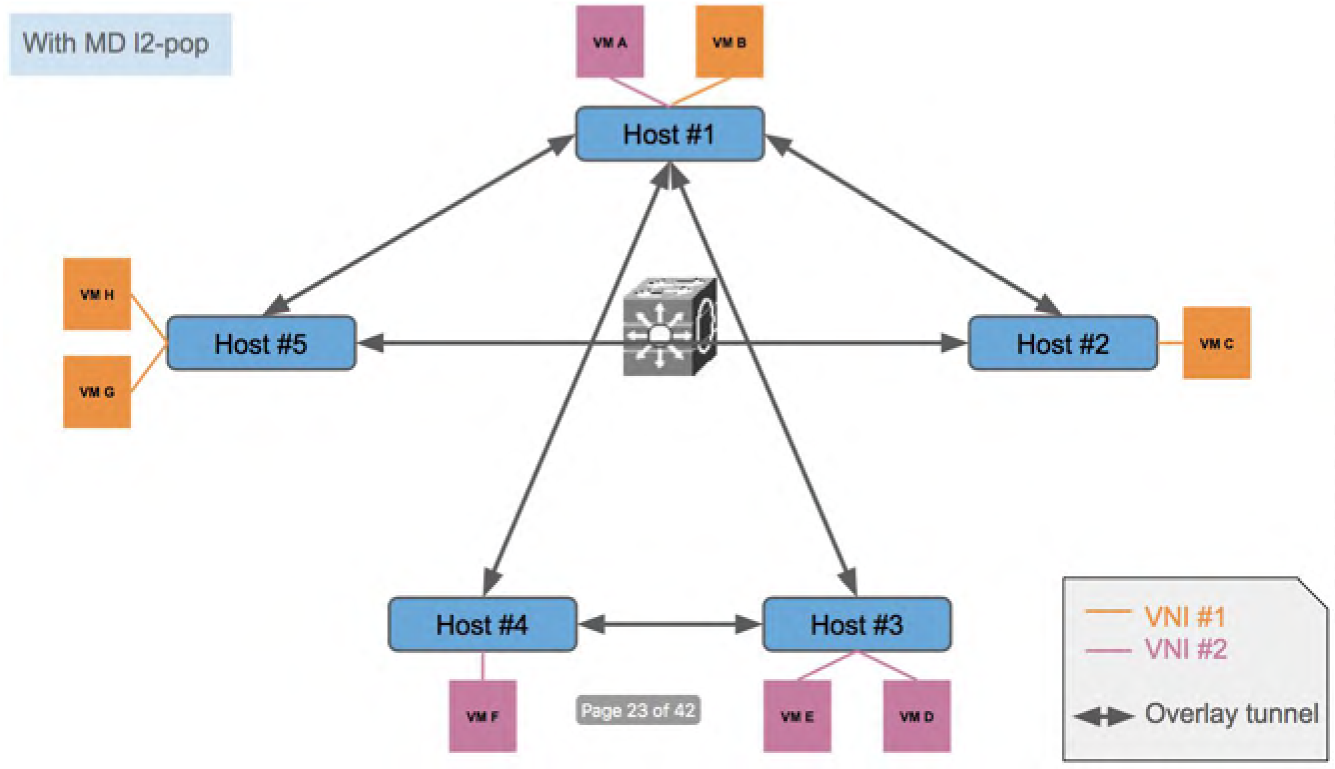

Nova
...
...
Neutron
ML2Plugin
BaGPipe mech driver
openstack
VM
VM
Neutron BaGPipe Agent
BGP
vSwitch
Nova
VM
VM
Neutron BaGPipe Agent
BGP
vSwitch
Nova
VM
VM
Neutron BaGPipe Agent
BGP
vSwitch
Nova
BGP
E-VPN
BGP Route Reflectors
VXLAN encap
or MPLS/x/IP)
network physical infra

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

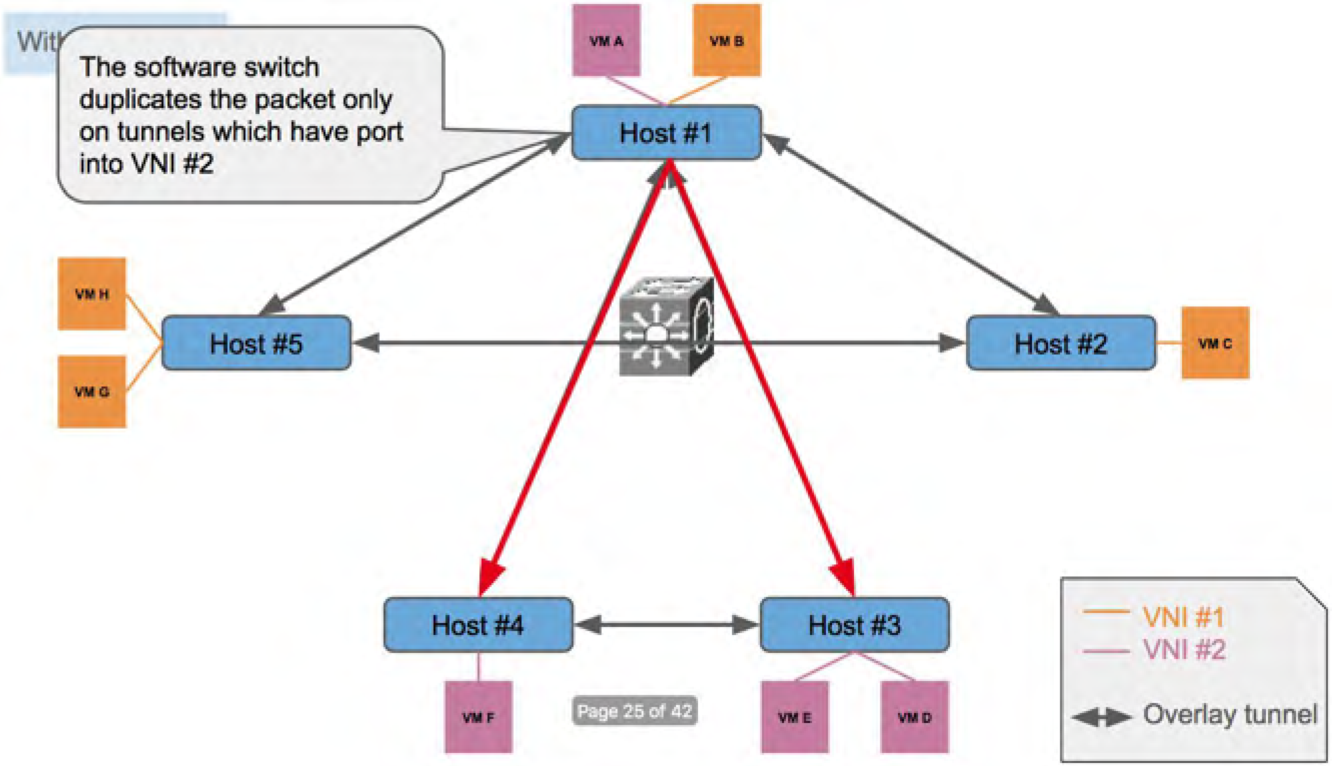
- OVS VXLAN: Flood&Learn;
- BaGPipe: EVPN implemented by software;
- *DVR: Neutron as the control plane;*
- MP BGP EVPN: Industry VXLAN control plane;
- Networking-bgpvpn: DVR with BGP, Routed Network;
- BGP-EVPN-Advertisement: Bring DirectConnect to OpenStack?


With MD l2-pop
VM A
VM B
Host #1
VM H
VM G
Host #5
Host #2
VM C
Host #4
Host #3
VM F
Page 23 of 42
VM E
VM D
VNI #1
VNI #2
Overlay tunnel

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。

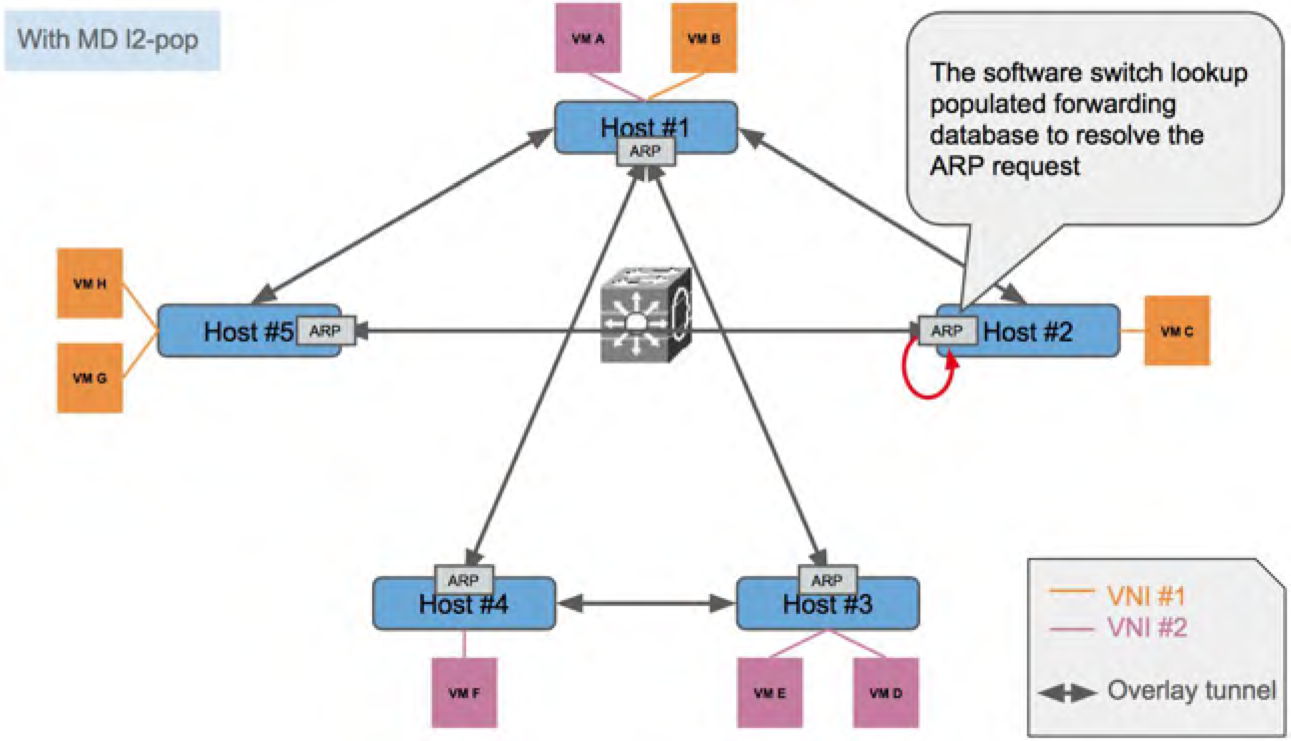
DVR & L2 Population



Wit
The software switch duplicates the packet only on tunnels which have port into VNI #2
VM A
VM B
Host #1
VM H
VM G
Host #5
Host #2
VM C
Host #4
Host #3
VM F
Page 25 of 42
VM E
VM D
VNI #1
VNI #2
Overlay tunnel

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


With MD l2-pop
VM A
VM B
Host #1
ARP
VM H
VM G
Host #5
ARP
ARP
Host #2
VM C
ARP
Host #4
ARP
Host #3
VM F
VM E
VM D
VNI #1
VNI #2
Overlay tunnel

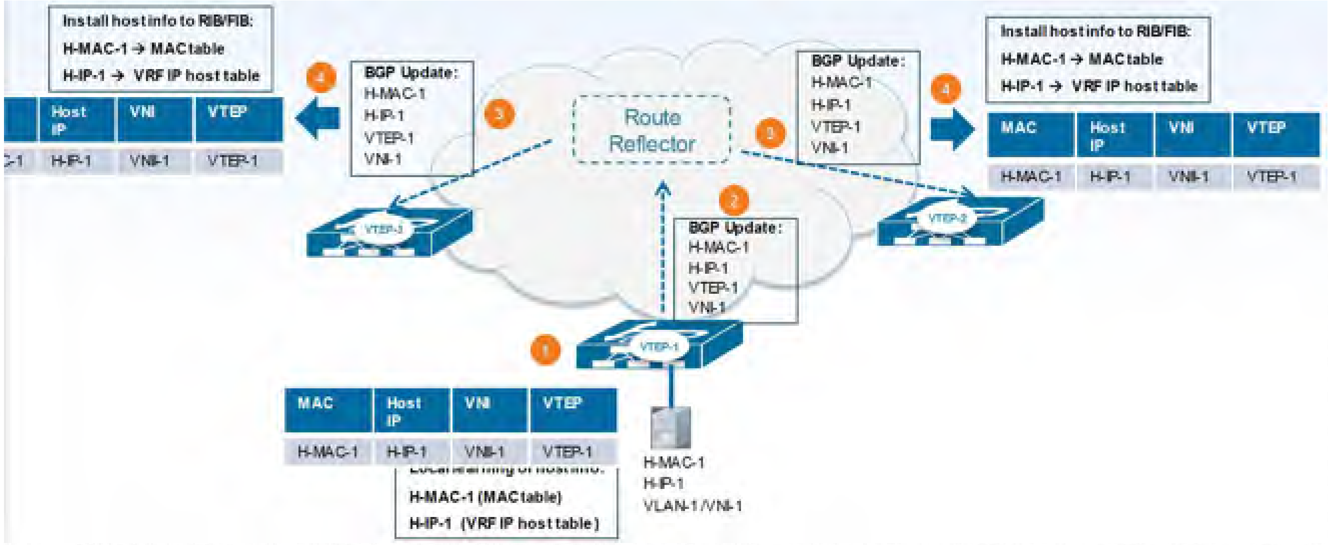
The software switch lookup populated forwarding database to resolve the ARP request

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?

MP BGP EVPN



Install host info to RIB/FIB:
H-MAC-1 → MAC table
H-IP-1 → VRF IP host table

	Host IP	VNI	VTEP
C-1	H-IP-1	VNI-1	VTEP-1

BGP Update:
H-MAC-1
H-IP-1
VTEP-1
VNI-1

BGP Update:
H-MAC-1
H-IP-1
VTEP-1
VNI-1

Install host info to RIB/FIB:
H-MAC-1 → MAC table
H-IP-1 → VRF IP host table

MAC	Host IP	VNI	VTEP
H-MAC-1	H-IP-1	VNI-1	VTEP-1

BGP Update:
H-MAC-1
H-IP-1
VTEP-1
VNI-1

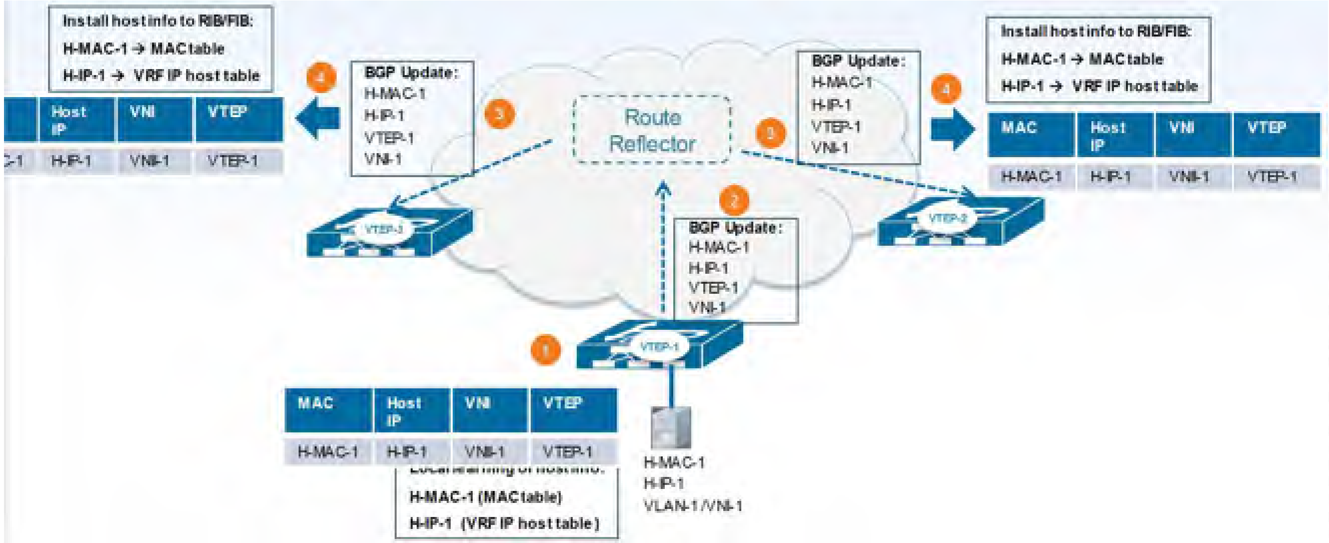
MAC	Host IP	VNI	VTEP
H-MAC-1	H-IP-1	VNI-1	VTEP-1

Local learning of host info:
H-MAC-1 (MAC table)
H-IP-1 (VRF IP host table)

H-MAC-1
H-IP-1
VLAN-1/VNI-1

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。

MP BGP EVPN

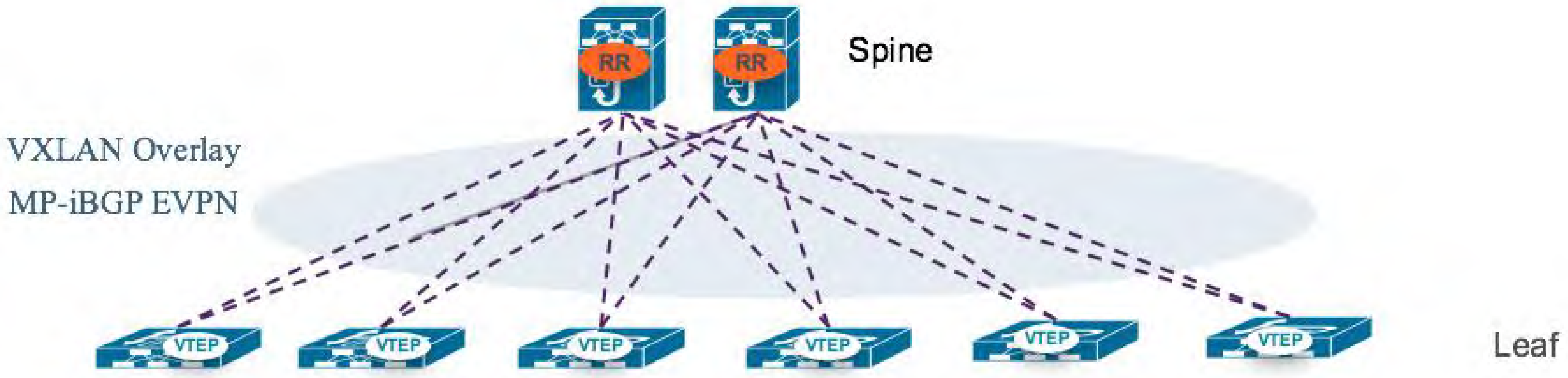

Install host info to RIB/FIB:
H-MAC-1 → MAC table
H-IP-1 → VRF IP host table
4
Host IP
VNI
VTEP
H-IP-1
VNI-1
VTEP-1
BGP Update:
H-MAC-1
H-IP-1
VTEP-1
VNI-1
3
Route Reflector
3
BGP Update:
H-MAC-1
H-IP-1
VTEP-1
VNI-1
4
Install host info to RIB/FIB:
H-MAC-1 → MAC table
H-IP-1 → VRF IP host table
MAC
Host IP
VNI
VTEP
H-MAC-1
H-IP-1
VNI-1
VTEP-1
VTEP-3
2
BGP Update:
H-MAC-1
H-IP-1
VTEP-1
VNI-1
VTEP-2
1
VTEP-1
MAC
Host IP
VNI
VTEP
H-MAC-1
H-IP-1
VNI-1
VTEP-1
H-MAC-1 (MAC table)
H-IP-1 (VRF IP host table)
H-MAC-1
H-IP-1
VLAN-1/VNI-1

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。

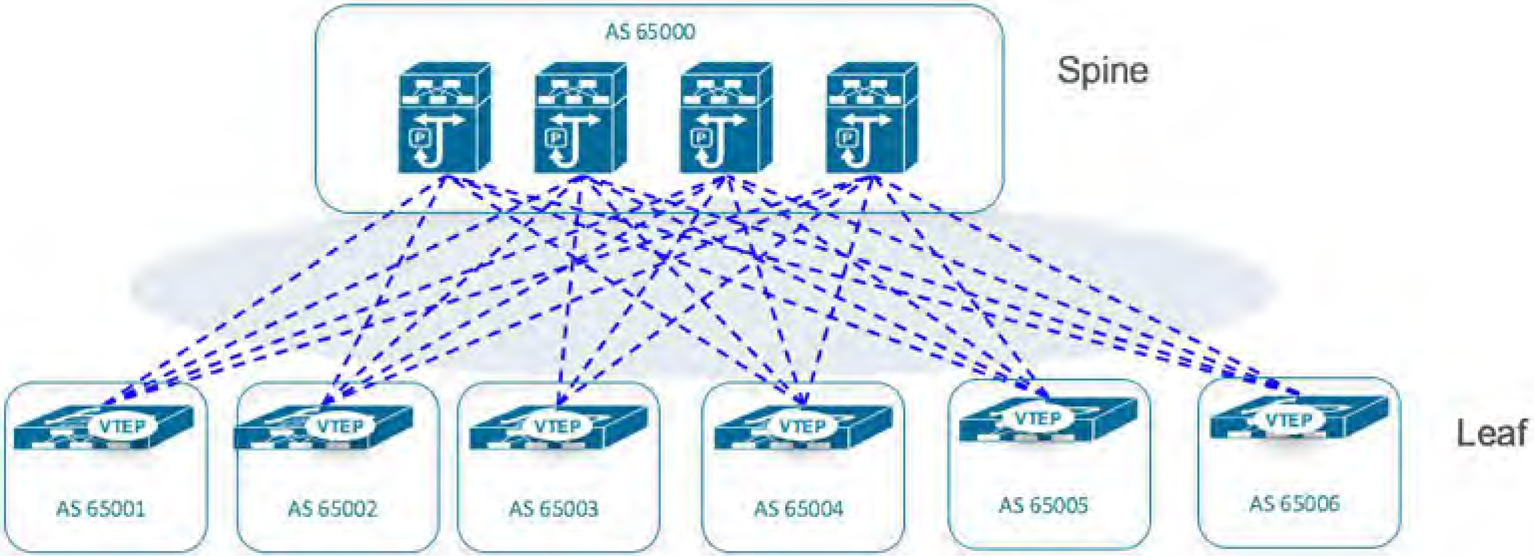
- Spine交换机设置为iBGP的RR，Spine交换机不需要支持VxLAN
- Leaf交换机与Spine交换机建立iBGP邻居关系


RR
RR
Spine
VXLAN Overlay
MP-iBGP EVPN
VTEP
VTEP
VTEP
VTEP
VTEP
VTEP
Leaf

MP BGP EVPN

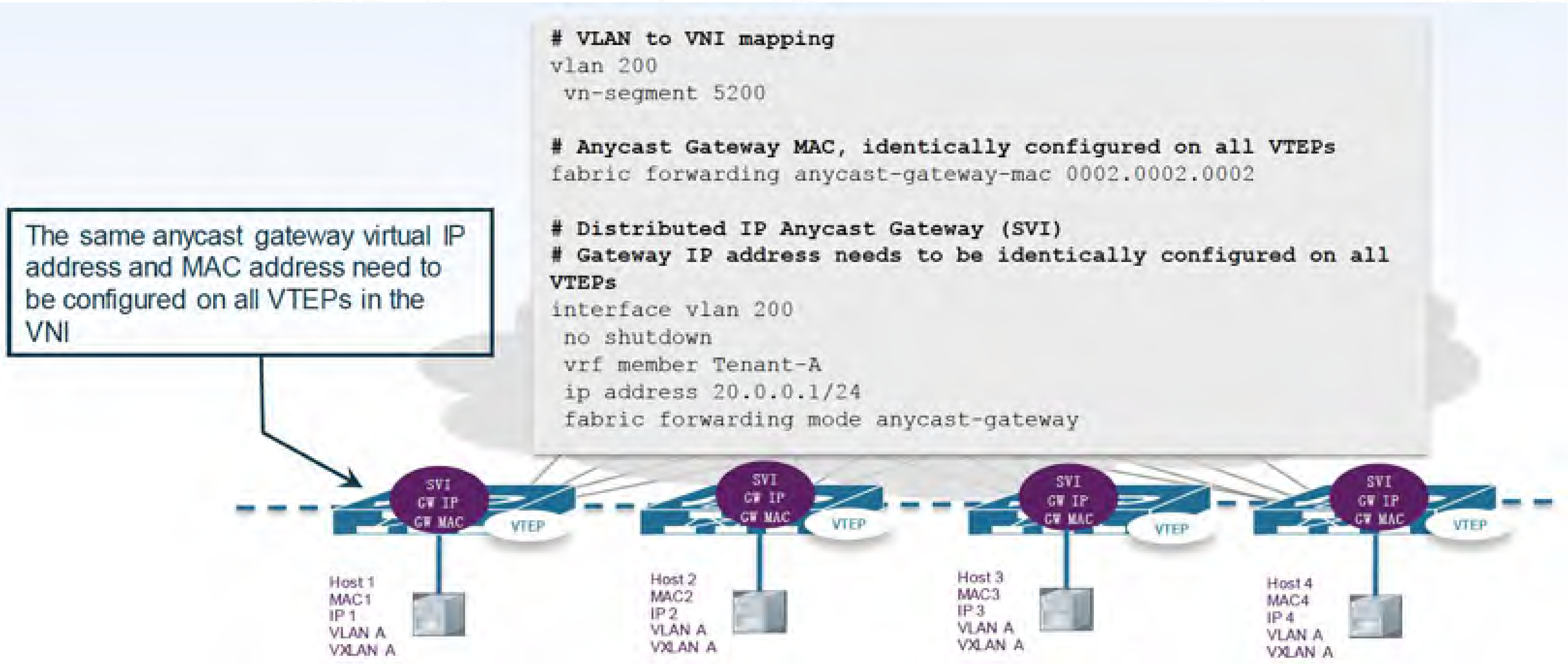
具体到 BGP 租网上，有几种选择，包括 iBGP、eBGP 的选择和外部网络拓扑。

- 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


AS 65000
Spine
VTEP
VTEP
VTEP
VTEP
VTEP
VTEP
AS 65001
AS 65002
AS 65003
AS 65004
AS 65005
AS 65006
Leaf

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）



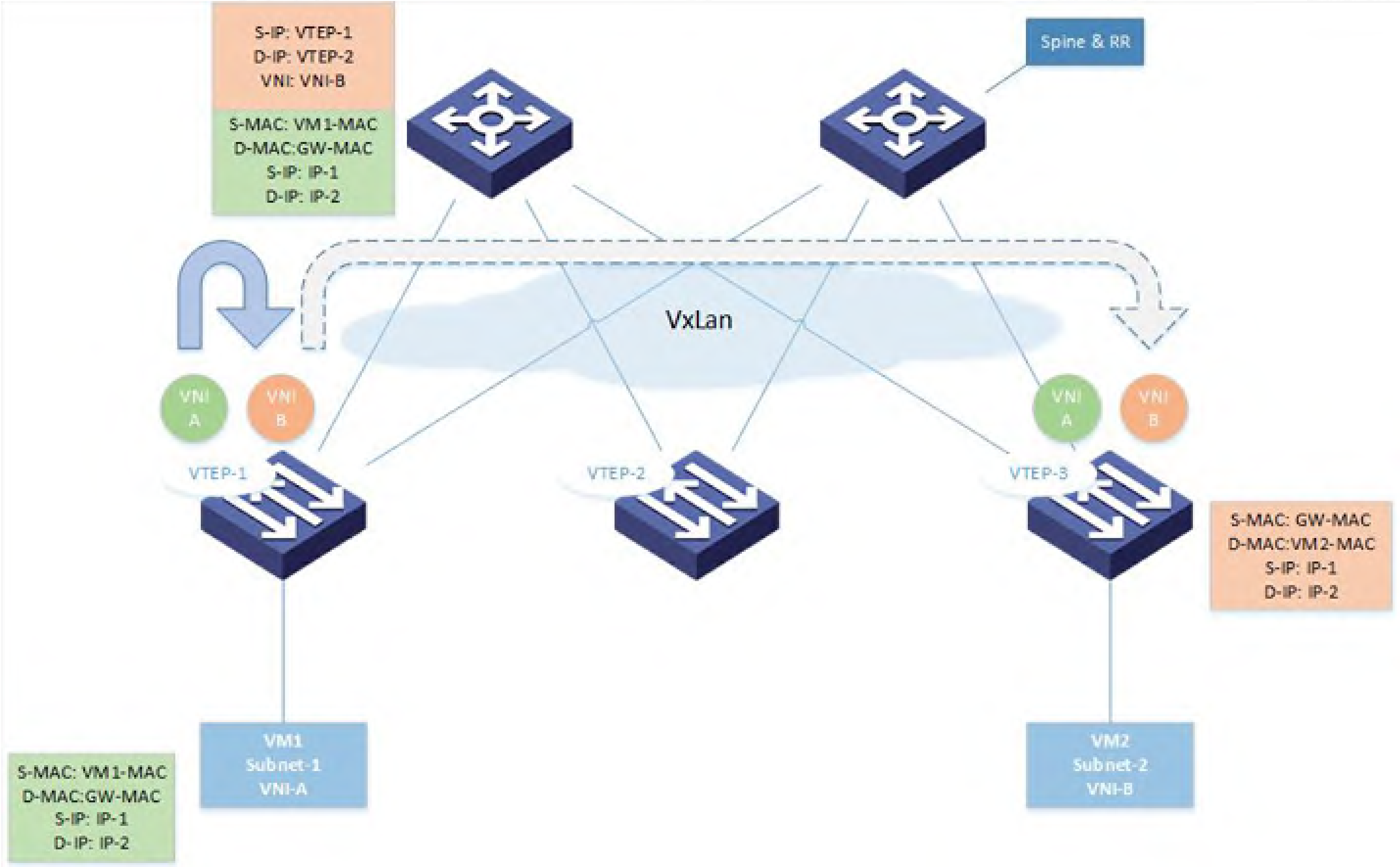
```
# VLAN to VNI mapping
vlan 200
 vn-segment 5200

# Anycast Gateway MAC, identically configured on all VTEPs
fabric forwarding anycast-gateway-mac 0002.0002.0002

# Distributed IP Anycast Gateway (SVI)
# Gateway IP address needs to be identically configured on all
VTEPs
interface vlan 200
 no shutdown
 vrf member Tenant-A
 ip address 20.0.0.1/24
 fabric forwarding mode anycast-gateway
```

The same anycast gateway virtual IP address and MAC address need to be configured on all VTEPs in the VNI

SVI
GW IP
GW MAC
VTEP
Host 1
MAC1
IP 1
VLAN A
VXLAN A
Host 2
MAC2
IP 2
VLAN A
VXLAN A
Host 3
MAC3
IP 3
VLAN A
VXLAN A
Host 4
MAC4
IP 4
VLAN A
VXLAN A

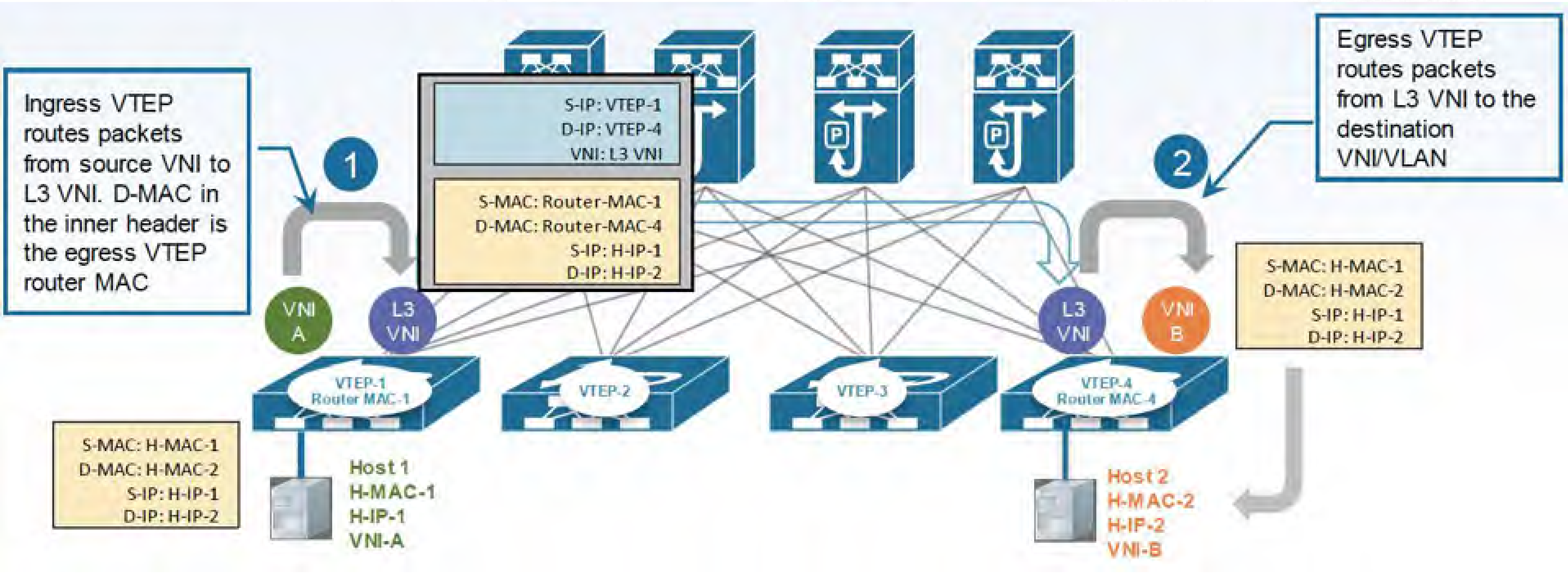
MP BGP EVPN

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


S-IP: VTEP-1
D-IP: VTEP-2
VNI: VNI-B
S-MAC: VM1-MAC
D-MAC:GW-MAC
S-IP: IP-1
D-IP: IP-2
Spine & RR
VxLan
VNI A
VNI B
VNI A
VNI B
VTEP-1
VTEP-2
VTEP-3
S-MAC: GW-MAC
D-MAC:VM2-MAC
S-IP: IP-1
D-IP: IP-2
VM1
Subnet-1
VNI-A
VM2
Subnet-2
VNI-B
S-MAC: VM1-MAC
D-MAC:GW-MAC
S-IP: IP-1
D-IP: IP-2

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。



Ingress VTEP routes packets from source VNI to L3 VNI. D-MAC in the inner header is the egress VTEP router MAC

Egress VTEP routes packets from L3 VNI to the destination VNI/VLAN

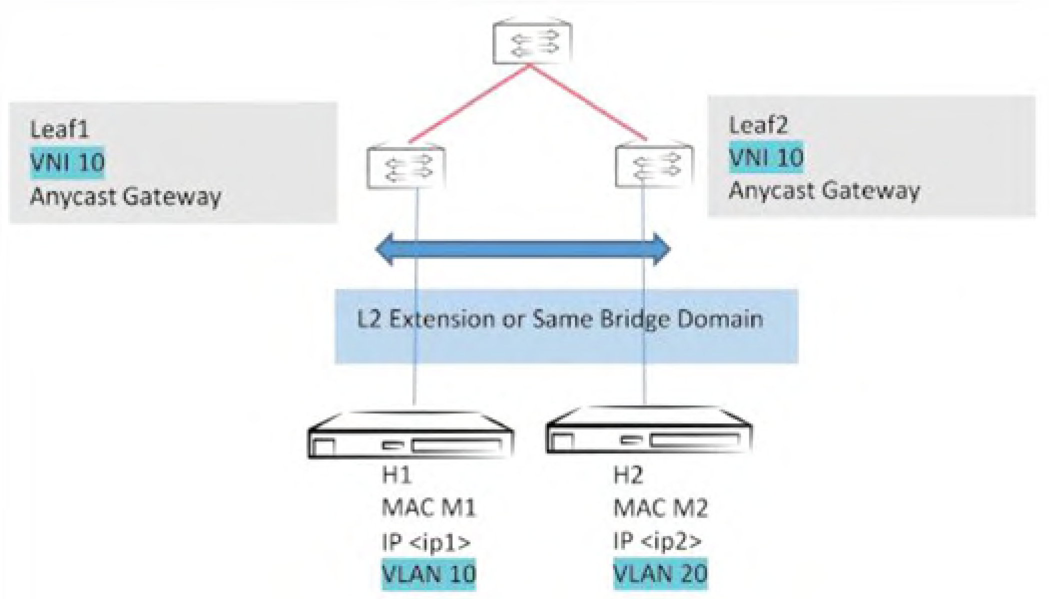
S-IP: VTEP-1
D-IP: VTEP-4
VNI: L3 VNI

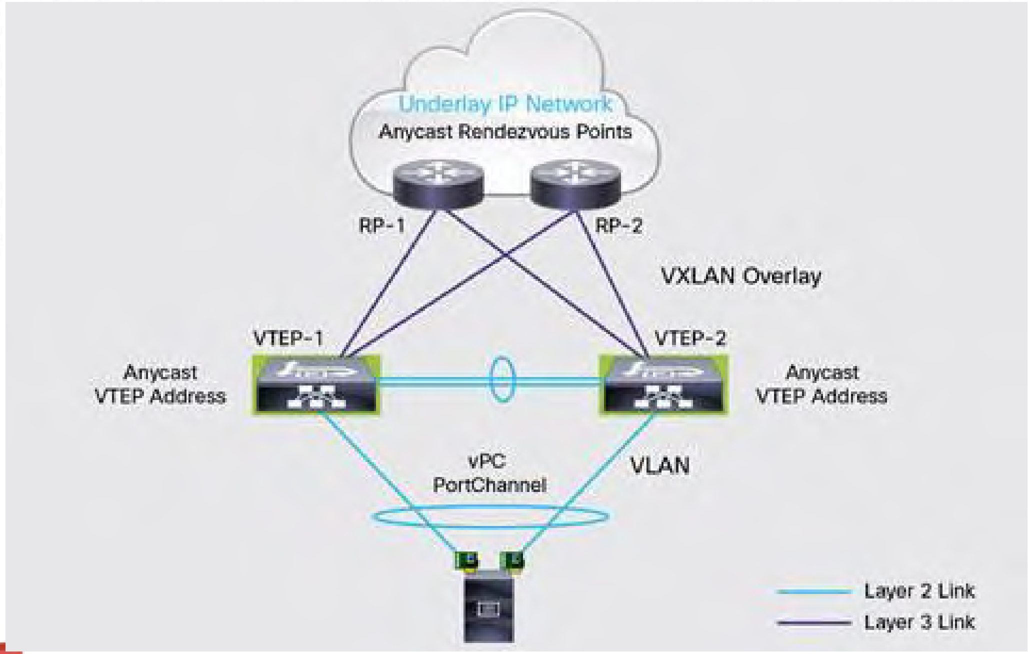
S-MAC: Router-MAC-1
D-MAC: Router-MAC-4
S-IP: H-IP-1
D-IP: H-IP-2

S-MAC: H-MAC-1
D-MAC: H-MAC-2
S-IP: H-IP-1
D-IP: H-IP-2

S-MAC: H-MAC-1
D-MAC: H-MAC-2
S-IP: H-IP-1
D-IP: H-IP-2

1
2
VNI A
L3 VNI
L3 VNI
VNI B
VTEP-1 Router MAC-1
VTEP-2
VTEP-3
VTEP-4 Router MAC-4
Host 1
H-MAC-1
H-IP-1
VNI-A
Host 2
H-MAC-2
H-IP-2
VNI-B


Leaf1
VNI 10
Anycast Gateway
Leaf2
VNI 10
Anycast Gateway
L2 Extension or Same Bridge Domain
H1
MAC M1
IP <ip1>
VLAN 10
H2
MAC M2
IP <ip2>
VLAN 20


Underlay IP Network
Anycast Rendezvous Points
RP-1
RP-2
VXLAN Overlay
VTEP-1
VTEP-2
Anycast
VTEP Address
Anycast
VTEP Address
vPC
PortChannel
VLAN
Layer 2 Link
Layer 3 Link

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。

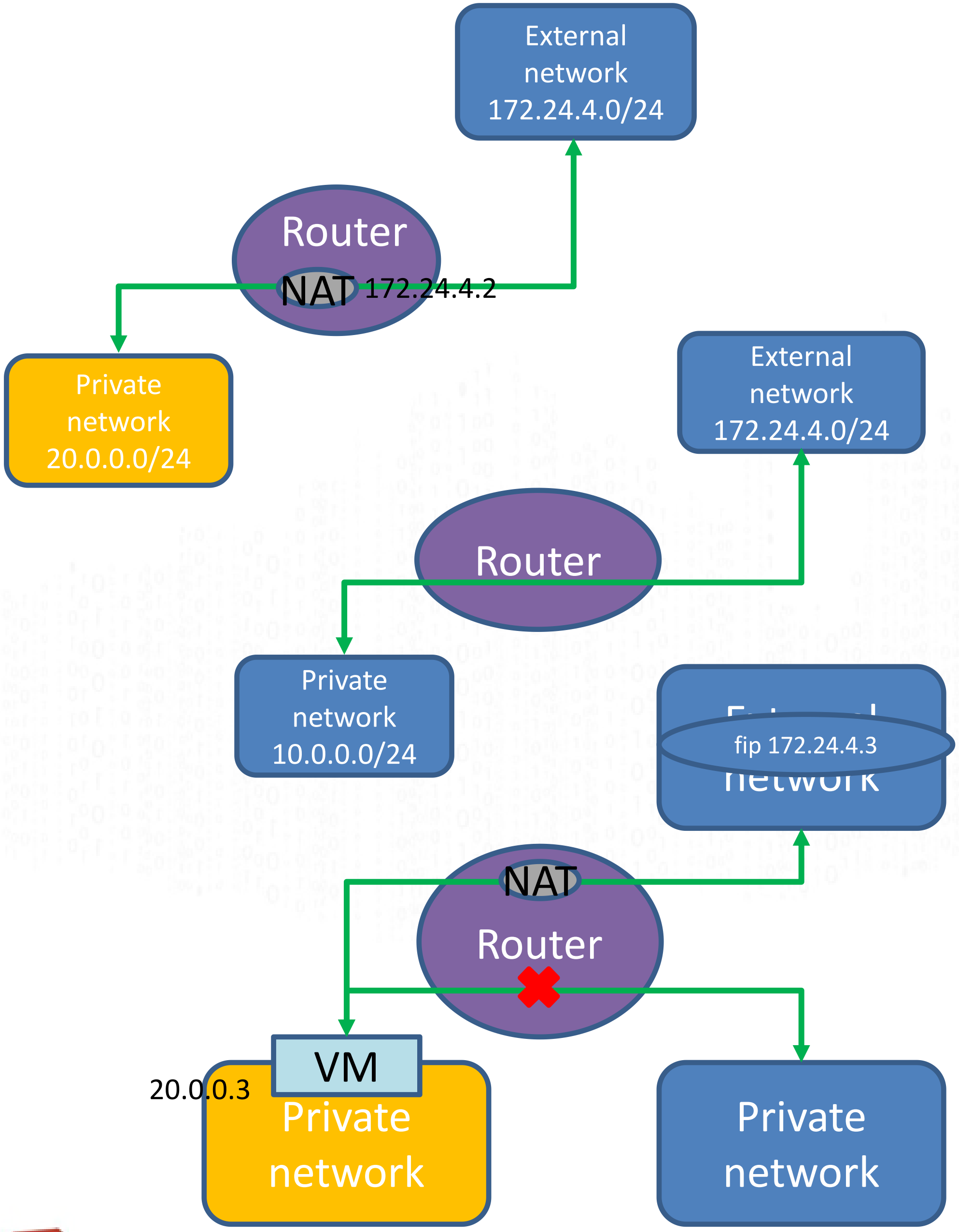
EVPN 实现	**Open vSwitch Driver VXLAN 实现**
Head-End Replication	Edge Replication
ARP Supression	ARP Responder
MAC advertise	L2 Population
Conversational Learning	Conversational Learning
Asymmetric IRB mode & Anycast Gateway	DVR
vPC/vLAG	Bonding/Teaming

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?

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

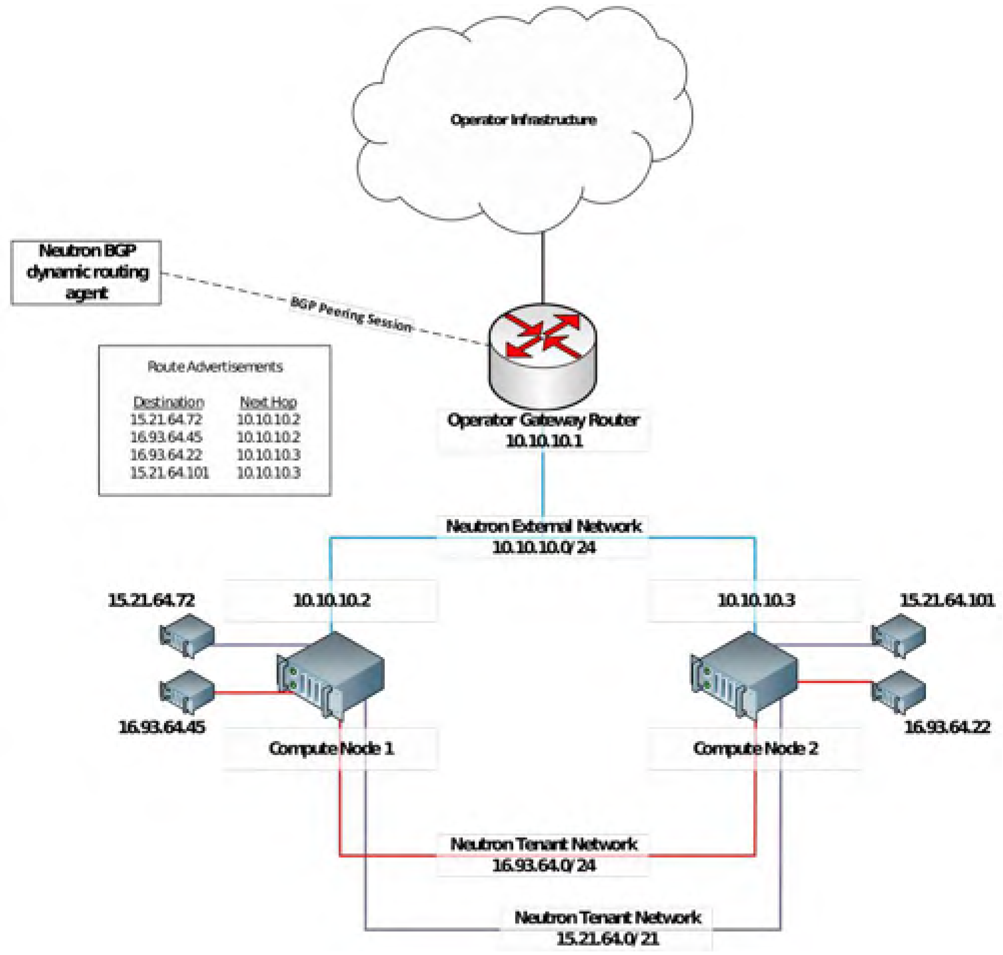

External network 172.24.4.0/24
Router
NAT
172.24.4.2
Private network 20.0.0.0/24
External network 172.24.4.0/24
Router
Private network 10.0.0.0/24
fip 172.24.4.3
NAT
Router
VM
20.0.0.3
Private network
Private network

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

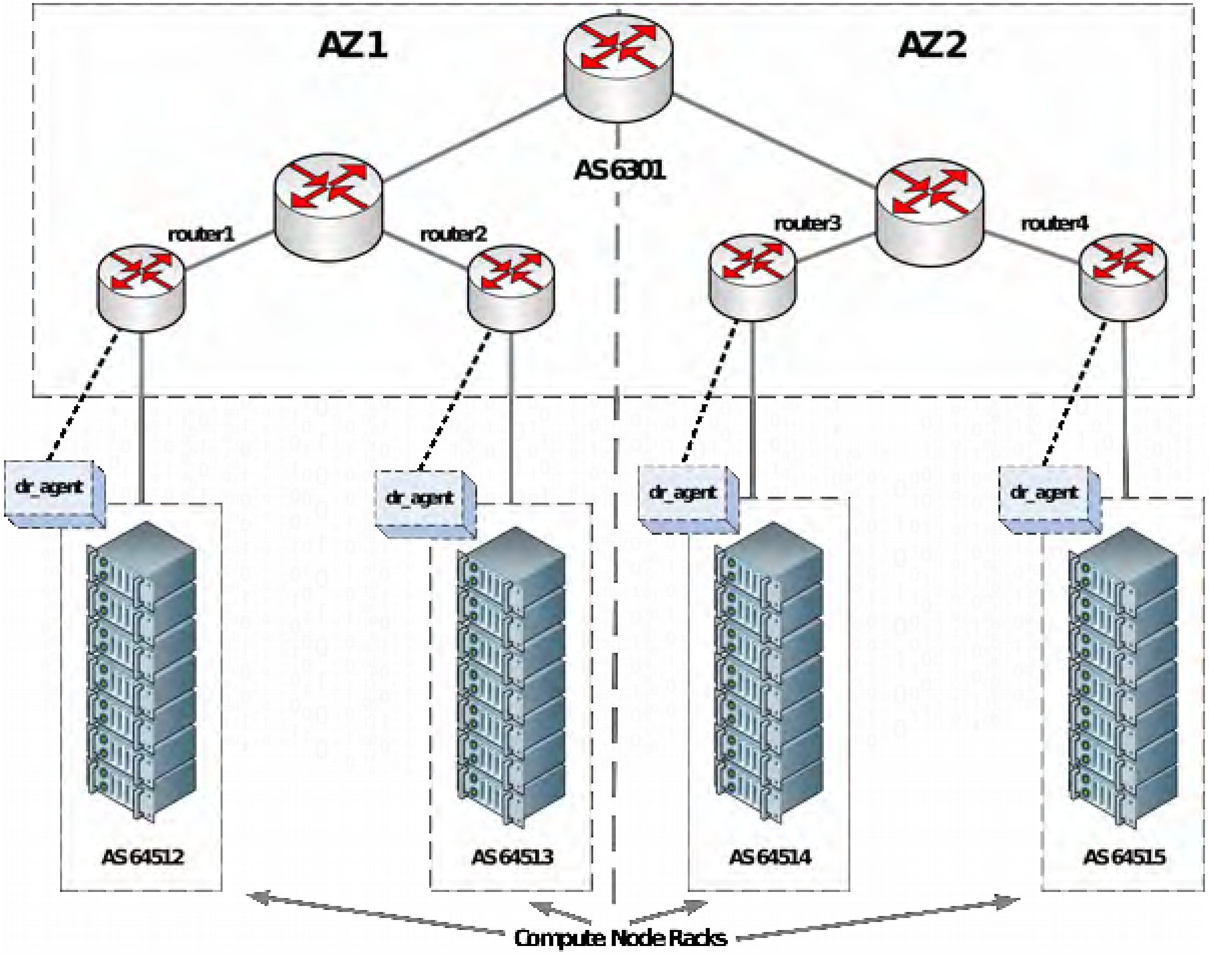

Operator Infrastructure
Neutron BGP dynamic routing agent
BGP Peering Session
Operator Gateway Router
10.10.10.1
Neutron External Network
10.10.10.0/24
15.21.64.72
10.10.10.2
10.10.10.3
15.21.64.101
16.93.64.45
16.93.64.22
Compute Node 1
Compute Node 2
Neutron Tenant Network
16.93.64.0/24
Neutron Tenant Network
15.21.64.0/21

Route Advertisements

Destination	Next Hop
15.21.64.72	10.10.10.2
16.93.64.45	10.10.10.2
16.93.64.22	10.10.10.3
15.21.64.101	10.10.10.3

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。


AZ1
AZ2
AS6301
router1
router2
router3
router4
dr_agent
dr_agent
dr_agent
dr_agent
AS64512
AS64513
AS64514
AS64515
Compute Node Racks

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?*

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

Thanks