# Kvm vxlan

来自wiki的名称

## 目录

 [[隐藏](http://192.168.100.158/mw/index.php/Kvm_vxlan)]

* [1 KVM VXLAN网络架构介绍与部署以及问题排查](http://192.168.100.158/mw/index.php/Kvm_vxlan" \l "KVM_VXLAN.E7.BD.91.E7.BB.9C.E6.9E.B6.E6.9E.84.E4.BB.8B.E7.BB.8D.E4.B8.8E.E9.83.A8.E7.BD.B2.E4.BB.A5.E5.8F.8A.E9.97.AE.E9.A2.98.E6.8E.92.E6.9F.A5)
  + [1.1 KVM的VXLAN网络架构介绍](http://192.168.100.158/mw/index.php/Kvm_vxlan" \l "KVM.E7.9A.84VXLAN.E7.BD.91.E7.BB.9C.E6.9E.B6.E6.9E.84.E4.BB.8B.E7.BB.8D)
    - [1.1.1 方案要求](http://192.168.100.158/mw/index.php/Kvm_vxlan" \l ".E6.96.B9.E6.A1.88.E8.A6.81.E6.B1.82)
    - [1.1.2 方案设计](http://192.168.100.158/mw/index.php/Kvm_vxlan" \l ".E6.96.B9.E6.A1.88.E8.AE.BE.E8.AE.A1)
  + [1.2 KVM的VXLAN网络部署](http://192.168.100.158/mw/index.php/Kvm_vxlan" \l "KVM.E7.9A.84VXLAN.E7.BD.91.E7.BB.9C.E9.83.A8.E7.BD.B2)
    - [1.2.1 各节点配置与说明](http://192.168.100.158/mw/index.php/Kvm_vxlan" \l ".E5.90.84.E8.8A.82.E7.82.B9.E9.85.8D.E7.BD.AE.E4.B8.8E.E8.AF.B4.E6.98.8E)
    - [1.2.2 计算节点部署](http://192.168.100.158/mw/index.php/Kvm_vxlan" \l ".E8.AE.A1.E7.AE.97.E8.8A.82.E7.82.B9.E9.83.A8.E7.BD.B2)
    - [1.2.3 内网网络节点部署](http://192.168.100.158/mw/index.php/Kvm_vxlan" \l ".E5.86.85.E7.BD.91.E7.BD.91.E7.BB.9C.E8.8A.82.E7.82.B9.E9.83.A8.E7.BD.B2)
    - [1.2.4 外网网络节点部署](http://192.168.100.158/mw/index.php/Kvm_vxlan" \l ".E5.A4.96.E7.BD.91.E7.BD.91.E7.BB.9C.E8.8A.82.E7.82.B9.E9.83.A8.E7.BD.B2)
  + [1.3 KVM的VXLAN问题排查](http://192.168.100.158/mw/index.php/Kvm_vxlan" \l "KVM.E7.9A.84VXLAN.E9.97.AE.E9.A2.98.E6.8E.92.E6.9F.A5)
    - [1.3.1 各种网络设备的状态](http://192.168.100.158/mw/index.php/Kvm_vxlan" \l ".E5.90.84.E7.A7.8D.E7.BD.91.E7.BB.9C.E8.AE.BE.E5.A4.87.E7.9A.84.E7.8A.B6.E6.80.81)
    - [1.3.2 隧道问题](http://192.168.100.158/mw/index.php/Kvm_vxlan" \l ".E9.9A.A7.E9.81.93.E9.97.AE.E9.A2.98)
    - [1.3.3 MTU设置](http://192.168.100.158/mw/index.php/Kvm_vxlan" \l "MTU.E8.AE.BE.E7.BD.AE)
    - [1.3.4 VLAN设置](http://192.168.100.158/mw/index.php/Kvm_vxlan" \l "VLAN.E8.AE.BE.E7.BD.AE)
    - [1.3.5 其他问题](http://192.168.100.158/mw/index.php/Kvm_vxlan" \l ".E5.85.B6.E4.BB.96.E9.97.AE.E9.A2.98)

## KVM VXLAN网络架构介绍与部署以及问题排查

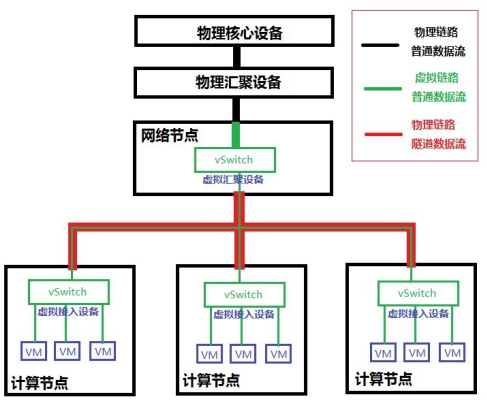
* [KVM的VXLAN网络架构介绍](http://192.168.100.158/mw/index.php/Kvm_vxlan" \l "KVM.E7.9A.84VXLAN.E7.BD.91.E7.BB.9C.E6.9E.B6.E6.9E.84.E4.BB.8B.E7.BB.8D)
* [KVM的VXLAN网络部署](http://192.168.100.158/mw/index.php/Kvm_vxlan" \l "KVM.E7.9A.84VXLAN.E7.BD.91.E7.BB.9C.E9.83.A8.E7.BD.B2)
* [KVM的VXLAN问题排查](http://192.168.100.158/mw/index.php/Kvm_vxlan" \l "KVM.E7.9A.84VXLAN.E9.97.AE.E9.A2.98.E6.8E.92.E6.9F.A5)

### KVM的VXLAN网络架构介绍

#### 方案要求

1. 实现Overlay在三层之上的跨省大二层;
2. 实现基于租户隔离的网络故障域(ARP,IP)
3. 新方案需要考虑原有网络设备的利旧问题,即目前已购买3层汇聚交换机和大量普通2层交换机
4. 尽可能只使用OVS,降低对运维人员的技能要求
5. 网络架构尽可能简单,越简单越稳定
6. 尽可能不使用VXLAN物理设备,主要原因为市面上VXLAN物理设备价格昂贵且技术不成熟

#### 方案设计

[](http://192.168.100.158/mw/index.php/%E6%96%87%E4%BB%B6:Topolgy.png)

说明:

1.计算节点上的虚拟接入设备负责接入VM通过流表(SW\_PORT+VM\_VNI+IP+MAC)限制ARP欺骗以及进行租户隔离

2.网络节点上的虚拟汇聚设备负责转发数据包通过流表(VM\_VNI+IP)实现指定端口转发, 避免广播现象

3.仅仅在计算节点与网络节点之间打隧道,所有计算节点之间不打隧道星形结构,避免环路

4.网络节点上的OVS负责将虚拟链路和物理链路桥接起来

又说明：

1.目前网络节点没有网关功能,因此需要物理汇聚设备提供网关功能

2.后期在网络节点上实现网关功能,去掉物理汇聚设备;网络节点将直接和物理核心设备互联,物理核心设备通过静态路由转发数据包到网络节点

虚拟交换流流表说明:

1.内网虚拟接入交换机(计算节点的br-lan)的流表项

#虚机ARP出包, 防欺骗限制VM\_PORT,VM\_MAC, VM\_IP,

#将ARP数据包通过TABLE1指定的VM列表进行匹配转发

table=0,priority=100,in\_port=VM\_PORT,arp,dl\_src=VM\_MAC,arp\_spa=VM\_IP,actions=mod\_vlan\_vid:VM\_VLAN,set\_field:VM\_VNI->tun\_id,resubmit(,1)

#虚机IP出包, 防欺骗限制VM\_PORT,VM\_MAC, VM\_IP

#将IP数据包通过TABL2指定的VM列表进行匹配转发

table=0,priority=100,in\_port=VM\_PORT,ip,dl\_src=VM\_MAC,nw\_src=VM\_IP,actions=mod\_vlan\_vid:VM\_VLAN,set\_field:VM\_VNI->tun\_id,resubmit(,2)

#将Tunnel-ARP数据包通过TABLE1指定的VM列表进行匹配转发

table=0,priority=100,in\_port=TUNNEL\_PORT,arp,actions=resubmit(,1)

#将Tunnel-IP数据包通过TABLE2指定的VM列表进行匹配转发

table=0,priority=100,in\_port=TUNNEL\_PORT,ip,actions=resubmit(,2)

#根据虚机IP列表接收ARP数据包

#这里要验证arp\_tpa与VM\_VNI是否匹配

table=1,priority=100,tun\_id=0,arp,arp\_tpa=VM\_IP,

actions= strip\_vlan,output:VM\_PORT

table=1,priority=100,tun\_id=VM\_VNI,arp,arp\_tpa=VM\_IP,

actions= strip\_vlan,output:VM\_PORT

#未匹配到该IP,则通过网络节点转发到其他接入交换机

table=1,priority=99,in\_port=TUNNEL\_PORT,actions=drop

table=1,priority=98,actions=output:TUNNEL\_PORT

#根据虚机MAC列表接收IP数据包

#这里要验证nw\_dst与VM\_VNI是否匹配

table=2,priority=100,tun\_id=0,ip,nw\_dst=VM\_IP,dl\_dst=VM\_MAC,actions= strip\_vlan,output:VM\_PORT

table=2,priority=100,tun\_id=VM\_VNI,ip,nw\_dst=VM\_IP,dl\_dst=VM\_MAC,actions=strip\_vlan,output:VM\_PORT

#未匹配到该IP,则通过网络节点转发到其他接入交换机

table=2,priority=99,in\_port=TUNNEL\_PORT,actions=drop

table=2,priority=98,actions=output:TUNNEL\_PORT

2.内网虚拟汇聚设备(内网网络节点的br-lan)的流表项

#根据网关列表将ARP广播包发往上联口

table=0,priority=100,arp,arp\_tpa=GW\_IP actions=output:LINK\_PORT

#根据网关IP列表将数据包发往上联口

table=0,priority=100,ip,nw\_dst=GW\_IP actions=output:LINK\_PORT

#根据汇聚设备MAC列表将回应包发往上联口

table=0,priority=100,dl\_dst=DEV\_MAC actions=output:LINK\_PORT

#根据虚机IP列表将ARP数据包转发到接入交换机

#这里要验证arp\_tpa与VM\_VNI是否匹配

table=0,priority=100,tun\_id=VM\_VNI,arp,arp\_tpa=VM\_IP actions=output:HOST\_PORT

table=0,priority=100,tun\_id=0,arp,arp\_tpa=VM\_IP actions=output:HOST\_PORT

#根据网关MAC列表将3层IP包转发到上联口

#这里要验证nw\_src与VM\_VNI是否匹配

table=0,priority=101,tun\_id=VM\_VNI,ip,nw\_src=VM\_IP,dl\_dst=GW\_MAC actions=output:LINK\_PORT

#根据虚机MAC列表将2层IP包转发到接入交换机

#这里要验证nw\_dst与VM\_VNI是否匹配

table=0,priority=100,tun\_id=VM\_VNI,ip,nw\_dst=VM\_IP actions=output:HOST\_PORT

table=0,priority=100,tun\_id=0,ip,nw\_dst=VM\_IP

actions= output:HOST\_PORT

#根据外部IP列表将数据包发往上联口(云数据库或其他业务使用)

table=0,priority=100,ip,nw\_dst=FORGET\_IP,dl\_dst=GW\_MAC

actions=output:LINK\_PORT

#没有匹配到数据包将被丢弃

#table=0,priority=99 actions=drop

3.外网虚拟接入设备(计算节点br-wan)的流表项

#虚机ARP出包,防欺骗限制VM\_PORT,VM\_MAC,VM\_IP,

#将ARP数据包通过TABLE1指定的VM列表进行匹配转发

table=0,priority=100,in\_port=VM\_PORT,arp,dl\_src=VM\_MAC,arp\_spa=VM\_IP,actions=mod\_vlan\_vid:VM\_VLAN,resubmit(,1)

#虚机IP出包, 防欺骗限制VM\_PORT,VM\_MAC, VM\_IP

#将IP数据包通过TABL2指定的VM列表进行匹配转发

table=0,priority=100,in\_port=VM\_PORT,ip,dl\_src=VM\_MAC,nw\_src=VM\_IP,actions=mod\_vlan\_vid:VM\_VLAN,resubmit(,2)

#将Tunnel-ARP数据包通过TABLE1指定的VM列表进行匹配转发

table=0,priority=100,in\_port=TUNNEL\_PORT,arp,actions=resubmit(,1)

#将Tunnel-IP数据包通过TABLE2指定的VM列表匹配转发

table=0,priority=100,in\_port=TUNNEL\_PORT,ip,actions=resubmit(,2)

#根据虚机IP列表接收ARP数据包

table=1,priority=100,arp,arp\_tpa=VM\_IP,

actions=strip\_vlan,output:VM\_PORT

#未匹配到该IP,则通过网络节点转发到其他接入交换机

table=1,priority=99,in\_port=TUNNEL\_PORT,actions=drop

table=1,priority=98,actions=output:TUNNEL\_PORT

#根据虚机MAC列表接收IP数据包

table=2,priority=100,ip,dl\_dst=VM\_MAC,actions= strip\_vlan,output:VM\_PORT

#未匹配到该IP,则通过网络节点转发到其他接入交换机

table=2,priority=99,in\_port=TUNNEL\_PORT,actions=drop

table=2,priority=98,actions=output:TUNNEL\_PORT

4.外网虚拟汇聚设备(外网网络节点br-wan)的流表项

#根据网关IP列表将ARP广播包发往上联口

table=0,priority=100,arp,arp\_tpa=GW\_IP actions=output:LINK\_PORT

#根据网关MAC列表将数据包发往上联口

table=0,priority=100,dl\_dst=GW\_MAC actions=output:LINK\_PORT

#根据汇聚设备MAC列表将数据包发往上联口

table=0,priority=100,dl\_dst=DEV\_MAC actions=output:LINK\_PORT

#根据VM\_IP列表将ARP数据包转发到接入交换机

table=0,priority=100,arp,arp\_tpa=VM\_IP

actions=output:HOST\_PORT

#根据VM\_MAC列表将IP数据包转发到接入交换机

#不要使用IP, 3层IP包的MAC为GW\_MAC

table=0,priority=100,ip,dl\_dst=VM\_MAC actions=output:HOST\_PORT

#没有匹配到的数据包丢弃

table=0,priority=99,actions=drop

### KVM的VXLAN网络部署

#### 各节点配置与说明

-----------------------------------------------------------------------------------------------

主机名 角色 网卡 IP 说明

-----------------------------------------------------------------------------------------------

kvm-101671101 计算节点 enp3s0f0 10.167.1.101/24 管理,内网隧道vtep

enp3s0f1 100.138.168.101/16 外网隧道vtep

- - (暂无,但需要) 数据网络

-----------------------------------------------------------------------------------------------

kvm-101671102 计算节点 enp3s0f0 10.167.1.102/24 管理,内网隧道vtep

enp3s0f1 100.138.168.102/16 外网隧道vtep

- - (暂无,但需要) 数据网络

-----------------------------------------------------------------------------------------------

node-intranet 内网网络节点 enp7s0f0 10.138.1.251/16 管理,内网隧道vtep

enp1s0f1 - 上联机房内网交换机

-----------------------------------------------------------------------------------------------

node-extranet 外网网络节点 enp1s0f0 10.138.1.117/16 管理

enp7s0f0 100.138.1.251/16 外网隧道vtep

mybond0 - 上联机房外网交换机

------------------------------------------------------------------------------------------------

#### 计算节点部署

以kvm-101671101为例

1.创建内网ovs网桥与外网ovs网桥(并清除默认流表规则)

[root@kvm-101671101 ~]# ovs-vsctl add-br br-lan && ovs-ofctl del-flows br-lan

[root@kvm-101671101 ~]# ovs-vsctl add-br br-wan && ovs-ofctl del-flows br-wan

2.将 br-lan与 br-wan通过创建vtep与对应网络节点相连

[root@kvm-101671101 ~]# ovs-vsctl add-port br-lan vxlan20 \

-- set Interface vxlan20 type=vxlan options:remote\_ip=10.138.1.251 options:key=flow \

-- set Interface vxlan20 ofport\_request=1

[root@kvm-101671101 ~]# ovs-vsctl add-port br-wan vxlan21 \

-- set Interface vxlan21 type=vxlan options:remote\_ip=100.138.1.251 options:key=flow \

-- set Interface vxlan21 ofport\_request=1

3.创建某虚拟机(KdZRlbQfzEI6)对应的内/外网linux bridge,并将它们连接到对应ovs网桥

[root@kvm-101671101 ~]# brctl addbr br1KdZRlbQfzEI6 && ip link set br1KdZRlbQfzEI6 up

[root@kvm-101671101 ~]# ovs-vsctl -- --if-exists del-port lanKdZRlbQfzEI6 \

-- add-port br-lan lanKdZRlbQfzEI6 \

-- set Interface lanKdZRlbQfzEI6 type=internal \

-- set Interface lanKdZRlbQfzEI6 external-ids:iface-status=active \

-- set Interface lanKdZRlbQfzEI6 external-ids:iface-id=$(uuidgen)

[root@kvm-101671101 ~]# ip link set lanKdZRlbQfzEI6 up && brctl addif br1KdZRlbQfzEI6 lanKdZRlbQfzEI6

[root@kvm-101671101 ~]# brctl addbr br0KdZRlbQfzEI6 && ip link set br0KdZRlbQfzEI6 up

[root@kvm-101671101 ~]# ovs-vsctl -- --if-exists del-port wanKdZRlbQfzEI6 \

-- add-port br-wan wanKdZRlbQfzEI6 \

-- set Interface wanKdZRlbQfzEI6 type=internal \

-- set Interface wanKdZRlbQfzEI6 external-ids:iface-status=active \

-- set Interface wanKdZRlbQfzEI6 external-ids:iface-id=$(uuidgen)

[root@kvm-101671101 ~]# ip link set wanKdZRlbQfzEI6 up && brctl addif br0KdZRlbQfzEI6 wanKdZRlbQfzEI6

4.新建虚拟机(KdZRlbQfzEI6)直连对应linux bridge

[root@kvm-101671101 ~]# mkdir -p /data/instance/KdZRlbQfzEI6

[root@kvm-101671101 ~]# cp /data/image/CentOS-7.1-ky-v1.qcow2 /data/instance/KdZRlbQfzEI6/system

[root@kvm-101671101 ~]# virt-install -n KdZRlbQfzEI6 --description KdZRlbQfzEI6 \

--ram 2048 --vcpus 2 --cpu host-model --accelerate --hvm \

--network bridge:br0KdZRlbQfzEI6,model=virtio \

--network bridge:br1KdZRlbQfzEI6,model=virtio \

--disk /data/instance/KdZRlbQfzEI6/system,bus=virtio,cache=writeback,driver\_type=qcow2,size=20 \

--boot hd,cdrom --graphics vnc,listen=0.0.0.0 --noautoconsole \

--input tablet,bus=usb \

--serial file,path=/data/instance/KdZRlbQfzEI6/console.log \

--cdrom /data/iso/welcome.iso

5.使用vnc视图进入虚拟机配置ip

eth0 122.114.158.222/24

eth1 10.122.158.222/17

note: 如何使用VNC视图?

[root@kvm-101671101 ~]# virsh vncdisplay KdZRlbQfzEI6

:39

使用vnc工具vncviewer,输入链接:"<宿主ip>:<vnc端口>"即可;例如此时为10.167.0.101:5939

6.在ovs外网交换机上添加流表规则

计算节点的外网流表规则模板如下:

table=0,priority=100,in\_port=VM\_PORT,arp,dl\_src=VM\_MAC,arp\_spa=VM\_IP,actions=mod\_vlan\_vid:VLAN\_ID,resubmit(,1)

table=0,priority=100,in\_port=VM\_PORT,ip,dl\_src=VM\_MAC,nw\_src=VM\_IP,actions=mod\_vlan\_vid:VLAN\_ID,resubmit(,2)

table=0,priority=100,in\_port=TUNNEL\_PORT,arp,actions=resubmit(,1)

table=0,priority=100,in\_port=TUNNEL\_PORT,ip,actions=resubmit(,2)

table=1,priority=100,arp,arp\_tpa=VM\_IP,actions=strip\_vlan,output:VM\_PORT

table=1,priority=99,in\_port=TUNNEL\_PORT,actions=drop

table=1,priority=98,actions=output:TUNNEL\_PORT

table=2,priority=100,ip,nw\_dst=VM\_IP,dl\_dst=VM\_MAC,actions=strip\_vlan,output:VM\_PORT

table=2,priority=99,in\_port=TUNNEL\_PORT,actions=drop

table=2,priority=98,actions=output:TUNNEL\_PORT

其中的变量的说明与如何取值:

(1) VM\_PORT 虚拟机(KdZRlbQfzEI6)直连的外网网桥(br0KdZRlbQfzEI6)与ovs外网交换机相连的端口(wanKdZRlbQfzEI6)的端口号

[root@kvm-101671101 0912]# ovs-vsctl get Interface wanKdZRlbQfzEI6 ofport

76

(2) VM\_MAC 虚拟机外网网卡的mac地址

[root@kvm-101671101 0912]# virsh domiflist KdZRlbQfzEI6 | grep "br0KdZRlbQfzEI6" | awk '{print $5}'

52:54:00:db:a1:88

(3) VM\_IP 虚拟机外网IP,程序分配(此处为 122.114.158.222)

(4) VLAN\_ID 虚拟机外网IP对应的vlan号,管理员为虚拟机IP划分(此处为 80)

(5) TUNNEL\_PORT 外网隧道vtep端口(vxlan21)的端口号(ofport\_request).即固定值1(人为设定)

所以此时填充流表规则模板,并添加流表规则到br-wan:

[root@kvm-101671101 ~]# cat > KdZRlbQfzEI6-wan-flow << EOF

> table=0,priority=100,in\_port=76,arp,dl\_src=52:54:00:db:a1:88,arp\_spa=122.114.158.222,actions=mod\_vlan\_vid:80,resubmit(,1)

> table=0,priority=100,in\_port=76,ip,dl\_src=52:54:00:db:a1:88,nw\_src=122.114.158.222,actions=mod\_vlan\_vid:80,resubmit(,2)

> table=0,priority=100,in\_port=1,arp,actions=resubmit(,1)

> table=0,priority=100,in\_port=1,ip,actions=resubmit(,2)

> table=1,priority=100,arp,arp\_tpa=122.114.158.222,actions=strip\_vlan,output:76

> table=1,priority=99,in\_port=1,actions=drop

> table=1,priority=98,actions=output:1

> table=2,priority=100,ip,nw\_dst=122.114.158.222,dl\_dst=52:54:00:db:a1:88,actions=strip\_vlan,output:76

> table=2,priority=99,in\_port=1,actions=drop

> table=2,priority=98,actions=output:1

> EOF

[root@kvm-101671101 ~]# ovs-ofctl add-flows br-wan KdZRlbQfzEI6-wan-flow

7.在ovs内网交换机上添加流表规则

计算节点的内网流表规则模板如下:

table=0,priority=100,in\_port=VM\_PORT,arp,dl\_src=VM\_MAC,arp\_spa=VM\_IP,actions=mod\_vlan\_vid:VLAN\_ID,\

set\_field:VM\_VNI->tun\_id,resubmit(,1)

table=0,priority=100,in\_port=VM\_PORT,ip,dl\_src=VM\_MAC,nw\_src=VM\_IP,actions=mod\_vlan\_vid:VLAN\_ID,\

set\_field:VM\_VNI->tun\_id,resubmit(,2)

table=0,priority=100,in\_port=TUNNEL\_PORT,arp,actions=resubmit(,1)

table=0,priority=100,in\_port=TUNNEL\_PORT,ip,actions=resubmit(,2)

table=1,priority=100,tun\_id=0,arp,arp\_tpa=VM\_IP,actions= strip\_vlan,output:VM\_PORT

table=1,priority=100,tun\_id=VM\_VNI,arp,arp\_tpa=VM\_IP,actions=strip\_vlan,output:VM\_PORT

table=1,priority=99,in\_port=TUNNEL\_PORT,actions=drop

table=1,priority=98,actions=output:TUNNEL\_PORT

table=2,priority=100,tun\_id=0,ip,nw\_dst=VM\_IP,actions=strip\_vlan,output:VM\_PORT

table=2,priority=100,tun\_id=VM\_VNI,ip,nw\_dst=VM\_IP,actions=strip\_vlan,output:VM\_PORT

table=2,priority=99,in\_port=TUNNEL\_PORT,actions=drop

table=2,priority=98,actions=output:TUNNEL\_PORT

其中的变量的说明与如何取值:

(1) VM\_PORT 虚拟机(KdZRlbQfzEI6)直连的内网网桥(br1KdZRlbQfzEI6)与ovs内网交换机相连的端口(lanKdZRlbQfzEI6)的端口号

[root@kvm-101671101 0912]# ovs-vsctl get Interface lanKdZRlbQfzEI6 ofport

75

(2) VM\_MAC 虚拟机内网网卡的mac地址

[root@kvm-101671101 0912]# virsh domiflist KdZRlbQfzEI6 | grep "br1KdZRlbQfzEI6" | awk '{print $5}'

52:54:00:54:2a:be

(3) VM\_IP 虚拟机内网IP,程序分配(此处为 10.122.158.222)

(4) VLAN\_ID 虚拟机内网IP对应的vlan号,管理员为虚拟机IP划分(此处为 1001)

(5) TUNNEL\_PORT 内网隧道vtep端口(vxlan20)的端口号(ofport\_request).即固定值1(人为设定)

(6) VM\_VNI 用户分配到的vxlan id,此处为4466757;用以区分租户,同租户下的虚拟机互通.

所以此时填充流表规则模板,并添加流表规则到br-lan:

[root@kvm-101671101 ~]# cat > KdZRlbQfzEI6-lan-flow << EOF

> table=0,priority=100,in\_port=75,arp,dl\_src=52:54:00:54:2a:be,arp\_spa=10.122.158.222,\

actions=mod\_vlan\_vid:1001,set\_field:4466757->tun\_id,resubmit(,1)

> table=0,priority=100,in\_port=75,ip,dl\_src=52:54:00:54:2a:be,nw\_src=10.122.158.222,\

actions=mod\_vlan\_vid:1001,set\_field:4466757->tun\_id,resubmit(,2)

> table=0,priority=100,in\_port=1,arp,actions=resubmit(,1)

> table=0,priority=100,in\_port=1,ip,actions=resubmit(,2)

> table=1,priority=100,tun\_id=0,arp,arp\_tpa=10.122.158.222,actions= strip\_vlan,output:75

> table=1,priority=100,tun\_id=4466757,arp,arp\_tpa=10.122.158.222,actions=strip\_vlan,output:75

> table=1,priority=99,in\_port=1,actions=drop

> table=1,priority=98,actions=output:1

> table=2,priority=100,tun\_id=0,ip,nw\_dst=10.122.158.222,actions=strip\_vlan,output:75

> table=2,priority=100,tun\_id=4466757,ip,nw\_dst=10.122.158.222,actions=strip\_vlan,output:75

> table=2,priority=99,in\_port=1,actions=drop

> table=2,priority=98,actions=output:1

> EOF

[root@kvm-101671101 ~]# ovs-ofctl add-flows br-lan KdZRlbQfzEI6-lan-flow

#### 内网网络节点部署

1.创建内网网络节点的ovs网桥

[root@node-intranet ~]# ovs-vsctl add-br br-lan

2.添加内网网络节点到前述计算节点(10.167.1.101)的隧道vtep

[root@node-intranet ~]# ovs-vsctl add-port br-lan vxlan10.167.1.101 \

-- set Interface vxlan10.167.1.101 type=vxlan options:remote\_ip=10.167.1.101 options:key=flow

3.将br-lan连接到上联(机房内网交换机)网卡设备

[root@node-intranet ~]# ovs-vsctl add-port br-lan enp1s0f1 -- set Port enp1s0f1 trunks=999,1000,1001 \

-- set Interface enp1s0f1 ofport\_request=34

4.为业务(KdZRlbQfzEI6)添加流表规则内网网络节点的流表模版如下:

table=0,priority=100,arp,arp\_tpa=GW\_IP actions=output:LINK\_PORT

table=0,priority=100,ip,nw\_dst=GW\_IP actions=output:LINK\_PORT

table=0,priority=100,dl\_dst=DEV\_MAC actions=output:LINK\_PORT

table=0,priority=100,tun\_id=VM\_VNI,arp,arp\_tpa=VM\_IP actions=output:HOST\_PORT

table=0,priority=100,tun\_id=0,arp,arp\_tpa=VM\_IP actions=output:HOST\_PORT

table=0,priority=101,tun\_id=VM\_VNI,ip,nw\_src=VM\_IP,dl\_dst=GW\_MAC actions=output:LINK\_PORT

table=0,priority=100,tun\_id=VM\_VNI,ip,nw\_dst=VM\_IP actions=output:HOST\_PORT

table=0,priority=100,tun\_id=0,ip,nw\_dst=VM\_IP actions=output:HOST\_PORT

其中的变量的说明与如何取值:

(1) GW\_IP 虚拟机分配的ip(VM\_IP)对应的网关ip(此处 10.122.158.222)

(2) LINK\_PORT br-lan连接上联网卡的端口(人为设置为34)

(3) GW\_MAC 虚拟机内网网关的mac地址(固定值 00:00:0c:07:ac:14),由网络管理员设置

(4) DEV\_MAC 机房汇聚交换机的设备mac(固定值 00:08:e2:86:39:ca),由网络管理员设置

(5) VM\_IP 虚拟机内网网卡的ip地址

(6) HOST\_PORT 宿主与内网网络节点建立的隧道的网络节点端的vtep设备(vxlan10.167.1.101)的端口号

[root@node-intranet 0912]# ovs-vsctl get interface vxlan10.167.1.101 ofpor

57

(7) VM\_VNI 用户分配到的vxlan id,此处为4466757;用以区分租户,同租户下的虚拟机互通

填充流表模板后生成流表,并添加到br-lan中

[root@node-intranet ~]# cat > KdZRlbQfzEI6-net-lan-flow << EOF

> table=0,priority=100,arp,arp\_tpa=10.122.128.1 actions=output:34

> table=0,priority=100,ip,nw\_dst=10.122.128.1 actions=output:34

> table=0,priority=100,dl\_dst=00:08:e2:86:39:ca actions=output:34

> table=0,priority=100,tun\_id=4466757,arp,arp\_tpa=10.122.158.222 actions=output:57

> table=0,priority=100,tun\_id=0,ip,nw\_dst=10.122.158.222 actions=output:57

> table=0,priority=100,tun\_id=0,arp,arp\_tpa=10.122.158.222

> table=0,priority=101,tun\_id=4466757,ip,nw\_src=10.122.158.222,dl\_dst=00:00:0c:07:ac:14 actions=output:34

> table=0,priority=100,tun\_id=4466757,ip,nw\_dst=10.122.158.222 actions=output:57

> table=0,priority=100,tun\_id=0,ip,nw\_dst=10.122.158.222 actions=output:57

> EOF

[root@node-intranet ~]# ovs-ofctl add-flows br-lan KdZRlbQfzEI6-net-lan-flow

#### 外网网络节点部署

1.创建外网网络节点的ovs网桥

[root@node-intranet ~]# ovs-vsctl add-br br-wan

2.添加外网网络节点到前述计算节点(100.138.168.101)的隧道vtep

[root@node-intranet ~]# ovs-vsctl add-port br-wan vxlan100.138.168.101 \

-- set Interface vxlan100.138.168.101 type=vxlan options:remote\_ip=100.138.168.101 options:key=flow

3.将br-wan连接到上联(机房外网交换机)网卡设备

[root@node-intranet ~]# ovs-vsctl add-port br-wan mybond0 -- set Port mybond0 trunks=80,140,210,260,300,440,610 \

-- set Interface mybond0 ofport\_request=32

4.为业务(KdZRlbQfzEI6)添加流表规则

外网网络节点的流表模版如下:

table=0,priority=100,arp,arp\_tpa=GW\_IP actions=output:LINK\_PORT

table=0,priority=100,dl\_dst=GW\_MAC actions=output:LINK\_PORT

table=0,priority=100,dl\_dst=DEV\_MAC actions=output:LINK\_PORT

table=0,priority=100,arp,arp\_tpa=VM\_IP actions=output:HOST\_PORT

table=0,priority=100,ip,nw\_dst=VM\_IP,dl\_dst=VM\_MAC actions=output:HOST\_PORT

table=0,priority=99,actions=drop

参数说明:

(1) GW\_IP 虚拟机分配的ip(VM\_IP)对应的网关ip(此处 122.114.158.222)

(2) LINK\_PORT br-lan连接上联网卡的端口(人为设置为32)

(3) GW\_MAC 虚拟机内网网关的mac地址(固定值 00:00:0c:07:ac:09),由网络管理员设置

(4) DEV\_MAC 机房汇聚交换机的设备mac(固定值 00:0d:bd:c2:e4:00),由网络管理员设置

(5) VM\_IP 虚拟机外网网卡的ip地址

(6) HOST\_PORT 宿主与外网网络节点建立的隧道的网络节点端的vtep设备(vxlan100.138.168.101)的端口号

[root@node-extranet ~]# ovs-vsctl get interface vxlan100.138.168.101 ofport

52

(7) VM\_MAC 虚拟机的外网网卡的mac地址(如前文查询结果 52:54:00:db:a1:88)

填充流表模板后生成流表,并添加到br-wan中

[root@node-extranet ~]# cat > KdZRlbQfzEI6-net-wan-flow << EOF

> table=0,priority=100,arp,arp\_tpa=122.114.158.1 actions=output:32

> table=0,priority=100,dl\_dst=00:00:0c:07:ac:09 actions=output:32

> table=0,priority=100,dl\_dst=00:0d:bd:c2:e4:00 actions=output:32

> table=0,priority=100,arp,arp\_tpa=122.114.158.222 actions=output:52

> table=0,priority=100,ip,nw\_dst=122.114.158.222,dl\_dst=52:54:00:db:a1:88 actions=output:52

> table=0,priority=99,actions=drop

> EOF

[root@node-extranet ~]# ovs-ofctl add-flows br-wan KdZRlbQfzEI6-net-wan-flow

### KVM的VXLAN问题排查

#### 各种网络设备的状态

以排查内网为例

1.检查虚拟机网卡/linux网桥/ovs接口设备是否up;否,则设置up

[root@kvm-101671101 ~]# virsh domiflist KdZRlbQfzEI6

Interface Type Source Model MAC

-------------------------------------------------------

vnet78 bridge br0KdZRlbQfzEI6 virtio 52:54:00:db:a1:88

vnet114 bridge br1KdZRlbQfzEI6 virtio 52:54:00:54:2a:be

[root@kvm-101671101 ~]# ip link show vnet114

[root@kvm-101671101 ~]# ip link show br1KdZRlbQfzEI6

[root@kvm-101671101 ~]# ip link show lanKdZRlbQfzEI6

#### 隧道问题

检查隧道是否完整

[root@kvm-101671101 ~]# ovs-vsctl show

...

Port "vxlan21"

Interface "vxlan21"

type: vxlan

options: {key=flow, remote\_ip="10.138.1.251"}

...

[root@node-intranet ~]# ovs-vsctl show | grep 10.167.1.101

Port "vxlan10.167.1.101"

Interface "vxlan10.167.1.101"

options: {key=flow, remote\_ip="10.167.1.101"}

检查计算节点隧道端口是否是1

[root@kvm-101671101 ~]# ovs-vsctl get Interface vxlan21 ofport

1

#### MTU设置

检查计算节点与网络节点隧道对端vtep物理网卡的MTU是否设置合适(目前为1600)

[root@kvm-101671101 ~]# ip link show enp3s0f0

2: enp3s0f0: <BROADCAST,MULTICAST,UP,LOWER\_UP> mtu 1600 qdisc mq state UP mode DEFAULT qlen 1000

link/ether 0c:c4:7a:84:36:54 brd ff:ff:ff:ff:ff:ff

[root@node-intranet ~]# ip link show enp7s0f0

6: enp7s0f0: <BROADCAST,MULTICAST,UP,LOWER\_UP> mtu 1600 qdisc mq state UP mode DEFAULT qlen 1000

link/ether 6c:92:bf:27:18:b6 brd ff:ff:ff:ff:ff:ff

联系网络管理员确认外/内网交换机是否支持1600的MTU.如果没有则需要设置.

确认连接宿主与网络节点的物理接入交换机是否支持1600MTU.如果没有则需要设置.

#### VLAN设置

检查ip分配的vlan是否在网络节点上联网卡的trunks范围(这个主要是外网会用);如果不在该范围内,需要将其vlan添加到trunks范围

[root@node-intranet ~]# ovs-vsctl get Port enp1s0f1 trunks

[999, 1000, 1001]

使用 ovs-vsctl set Port enp1s0f1 trunks=VLAN1,VLAN2,VLAN3,..VLANx 予以设置