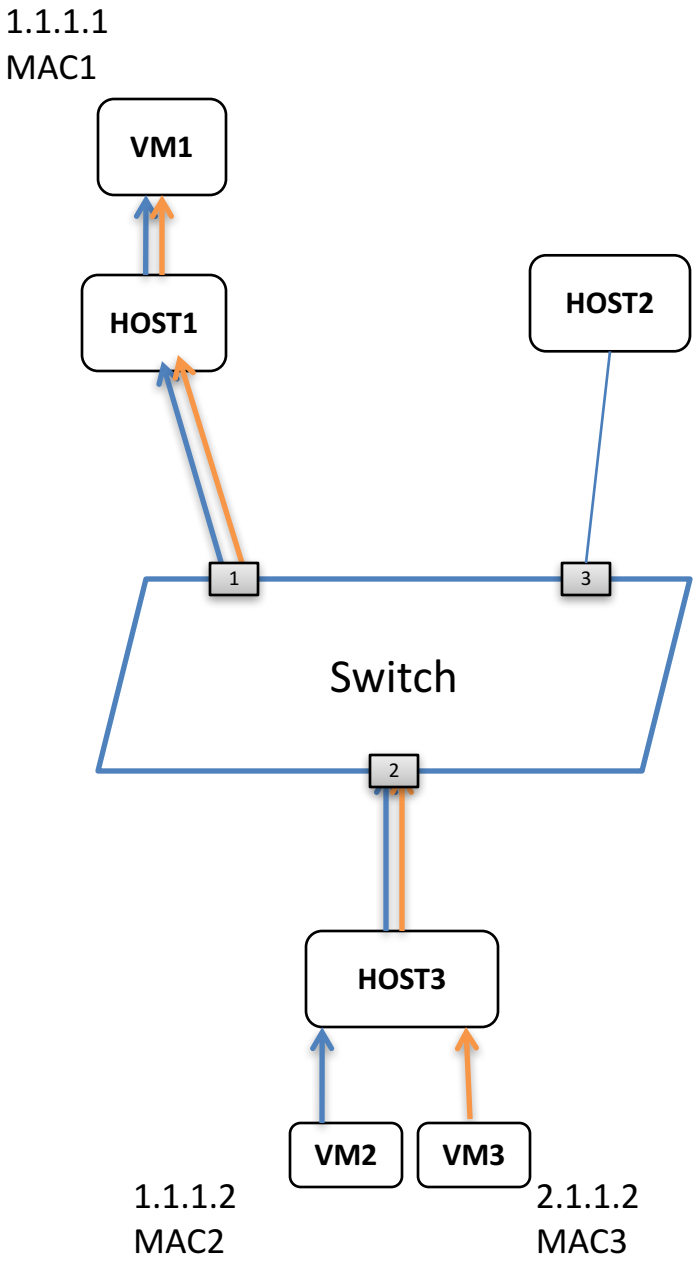
热迁移引起的虚拟机网络中断问题分析

中移（苏州）软件技术有限公司

2016年12月

场景1: 热迁移 - 同子网通，跨子网不通



OpenStack环境概要：

三台计算节点，两个租户网络(VLAN类型）
net1: 1.1.1.0/24 VLAN ID 100 GW 1.1.1.254
net2: 2.1.1.0/24 VLAN ID 200 GW 2.1.1.254

迁移前：同子网/跨子网访问VM1都正常。

物理交换机上的配置：

VLAN 100: port 1、2、3
VLAN 200: port 1、2、3
interface VLAN100: 1.1.1.254
Interface VLAN200: 2.1.1.254

交换机上的转发表项:

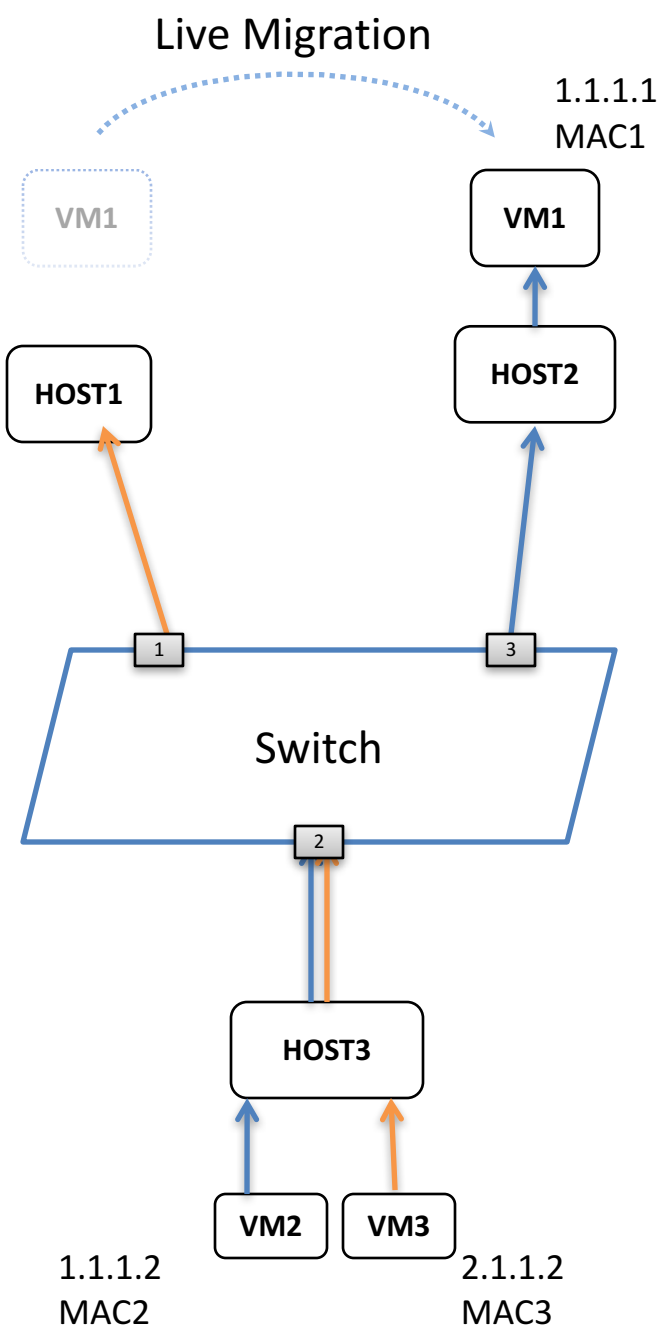
FDB:

VLAN	MAC Address	Type	age	Ports
* 100	MAC1	dynamic	100	1
* 100	MAC2	dynamic	100	2
* 200	MAC3	dynamic	100	2

ARP Cache

Address	Age	MAC Address	Interface
1.1.1.1	00:16:02	MAC1	VLAN100
1.1.1.2	00:16:02	MAC2	VLAN100
2.1.1.2	00:16:01	MAC3	VLAN200

: Traffic from VM2 to VM1
: Traffic from VM3 to VM1

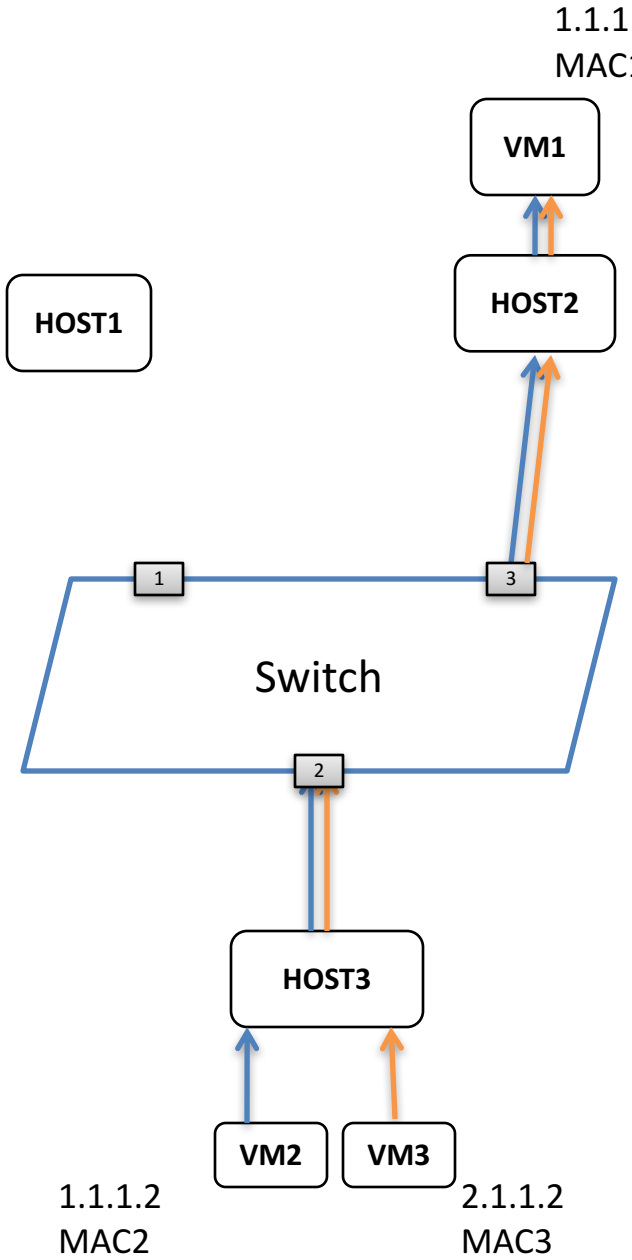


VM1从HOST1热迁移到HOST2之后

同子网访问VM1都正常。
跨子网访问VM1不通。

: Traffic from VM2 to VM1

: Traffic from VM3 to VM1



问题出现后，等待一段时间。

跨子网访问VM1恢复正常。

问题：迁移后的虚拟机，为什么跨子网访问不通？

: Traffic from VM2 to VM1

: Traffic from VM3 to VM1

分别从虚拟网络，物理网络两个层面进行排查。

虚拟网络层面: Linux bridge(安全组， veth pair), ovs bridge(流表、端口属性等)，物理网卡配置等。由于此场景下，非虚拟网络的问题，不再展开描述。

物理网络层面: VLAN,路由,arp, ACL配置等。

不同Vendor的交换机arp cache表更新行为*表现不同。

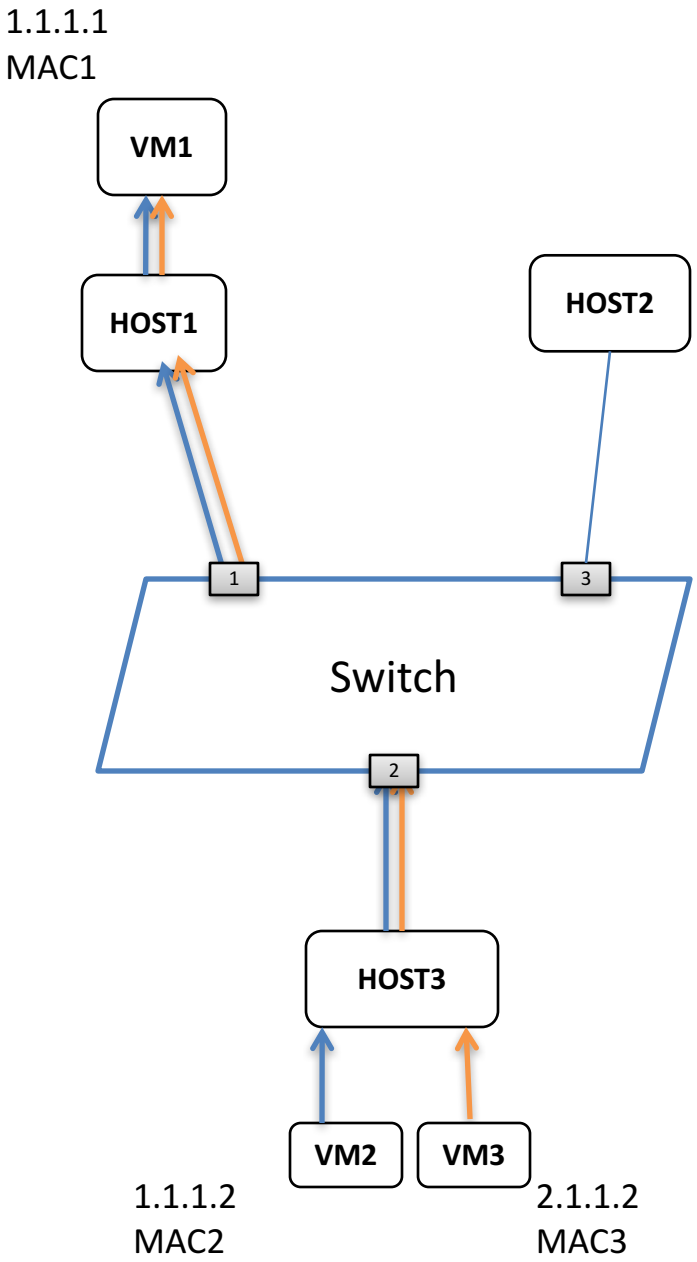
一种是：rarp报文触发FDB表的更新，并由此触发arp cache表的更新。

另一种是: rarp报文触发FDB表的更新，但不会触发arp cache表的更新。

此种场景下,只有arp/FDB表项正常更新，这就确保了虚机迁移后，交换机能使用转发规则将报文发给虚拟机。

* 仅表明从黑盒角度观察，不同类型的交换机或不同配置下的交换机对于同样的场景有不同的行为，并不说明交换机不支持此类功能。

场景2: 热迁移-同子网/跨子网都不通



OpenStack环境概要：

三台计算节点，两个租户网络(VLAN类型）
net1: 1.1.1.0/24 VLAN ID 100 GW 1.1.1.254
net2: 2.1.1.0/24 VLAN ID 200 GW 2.1.1.254

迁移前：同子网/跨子网访问VM1都正常。

物理交换机上的配置：

VLAN 100: port 1、2、3
VLAN 200: port 1、2、3
interface VLAN100: 1.1.1.254
Interface VLAN200: 2.1.1.254

交换机上的转发表项:

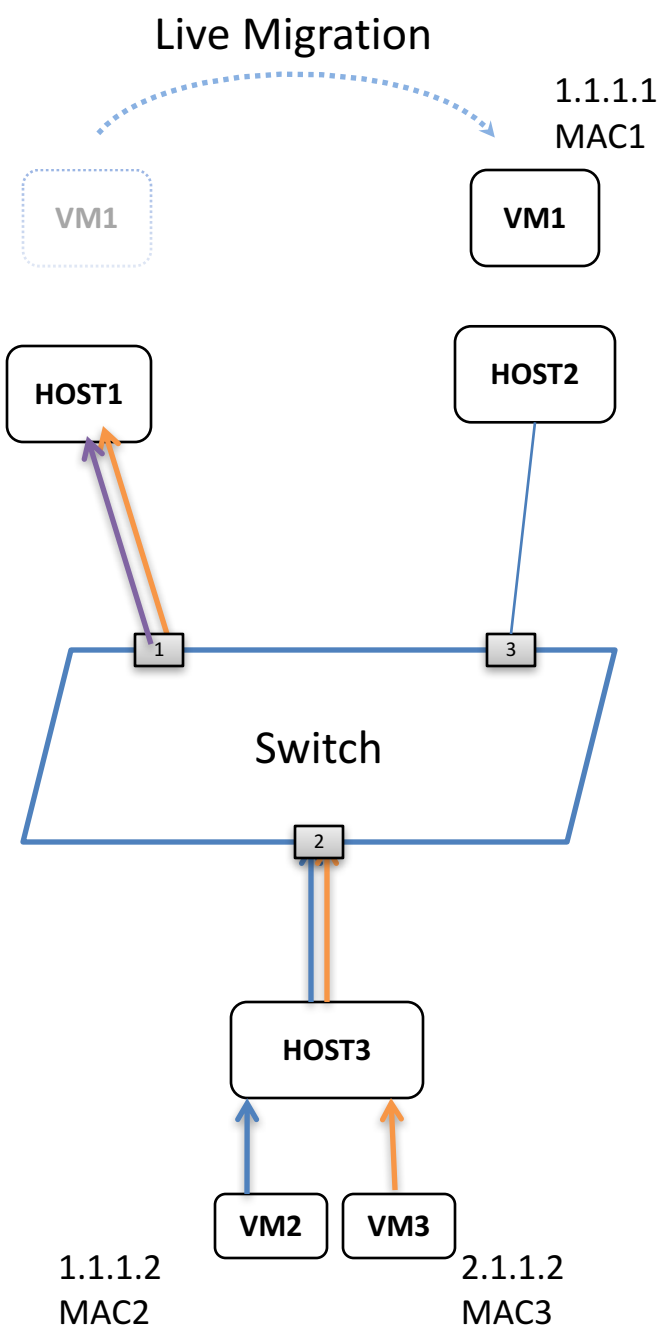
FDB:

VLAN	MAC Address	Type	age	Ports
* 100	MAC1	dynamic	100	1
* 100	MAC2	dynamic	100	2
* 200	MAC3	dynamic	100	2

ARP Cache

Address	Age	MAC Address	Interface
1.1.1.1	00:16:02	MAC1	VLAN100
1.1.1.2	00:16:02	MAC2	VLAN100
2.1.1.2	00:16:01	MAC3	VLAN200

: Traffic from VM2 to VM1
: Traffic from VM3 to VM1



VM1从HOST1热迁移到HOST2之后

同子网访问VM1不通。
跨子网访问VM1不通。

交换机上的转发表项:

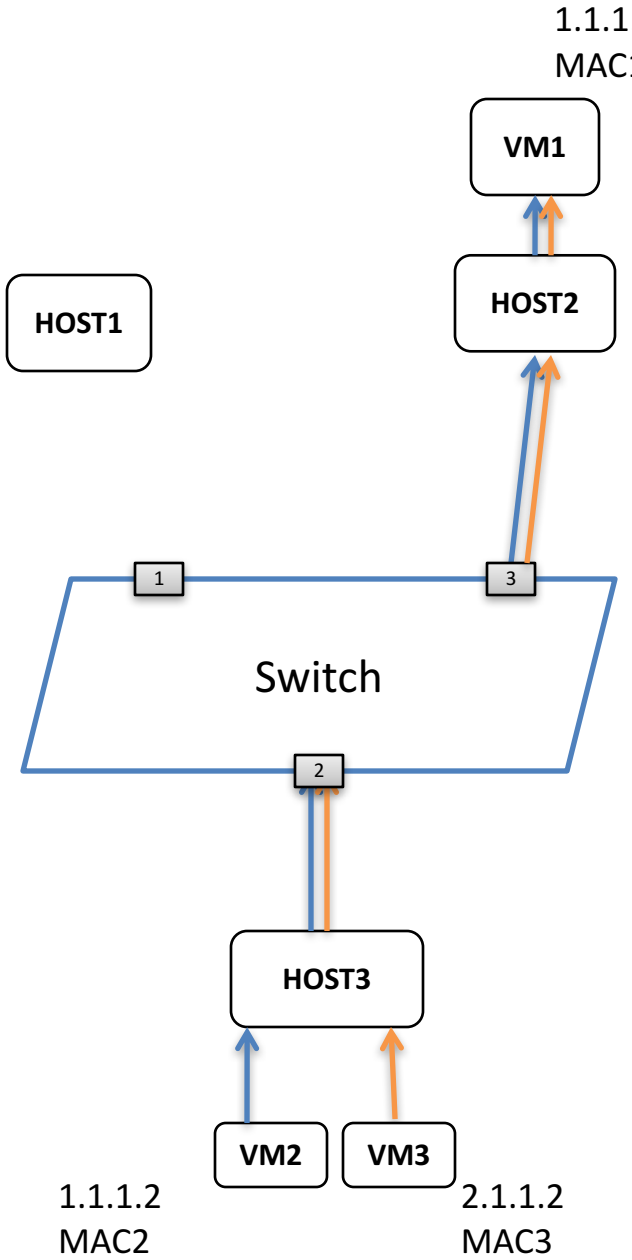
FDB:

VLAN	MAC Address	Type	age	Ports
* 100	MAC1	dynamic	100	3
* 100	MAC2	dynamic	100	2
* 200	MAC3	dynamic	100	2

ARP Cache

Address Age MAC Address Interface
1.1.1.2 00:16:02 MAC2 VLAN100
2.1.1.2 00:16:01 MAC3 VLAN200

: Traffic from VM2 to VM1
: Traffic from VM3 to VM1



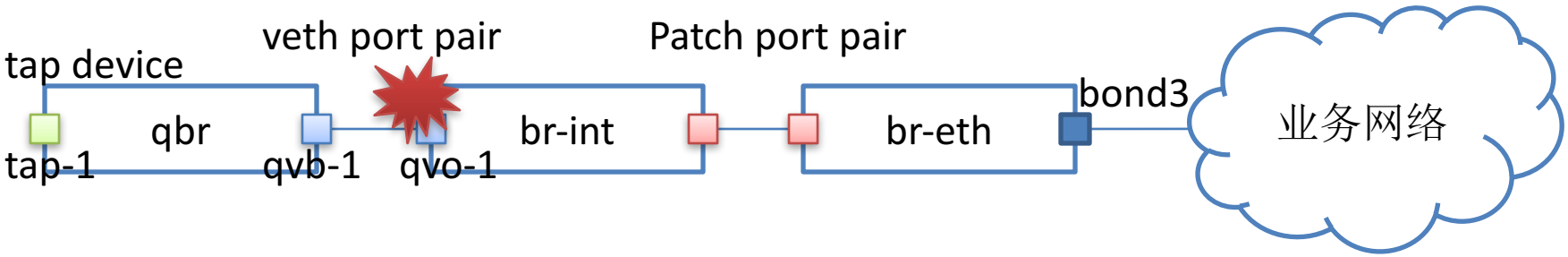
问题出现后，等待一段时间。

同子网/跨子网访问VM1恢复正常。

问题：迁移后的虚拟机，为什么访问不通？

: Traffic from VM2 to VM1

: Traffic from VM3 to VM1



正常情况下， rarp报文从tap-1发出，经过qvb-1 -> qvo-1 -> bond3 达到接入交换机。
在出问题情况下，rarp报文从tap-1发出，经过qvb-1 -> qvo-1 终止，原因在于此时qvo-1的internal tag配置尚未生效，而未匹配上br-int流表中特定的VLAN规则而被丢弃。
实际测试发现，该问题并非必现，10次当中最多出现1,2次。

解决该问题可以从两个方面着手。
第一个方面是从nova，neutron交互的角度出发，通过及时的事件通知驱动ovs-agent在虚拟机在目的计算节点上运行前配置好网络。

第二个方面是从Hypervisor角度出发，增加通告报文发送次数。

Q&A

Thank You!