虚拟交换机（vSwitch）原理及配置

厉梦如　2014-02-11 发表

云计算产品vSwitch功能的配置

一、vSwitch原理

1．概要

在物理环境之中，主机是通过pSwitch连接到网络当中。而在虚拟化环境中，则使用 vswitch。虚拟机通过vSwitch来连接网络，vSwitch是通过主机上的物理网卡作为上行链路与外界网络进行连接。

跟普通服务器设备一样，每个虚拟机有着自己的虚拟网卡(virtual NIC)，每个 virtual NIC有着自己的MAC地址和IP地址。 Virtual Switch(vSwitch)相当于一个虚拟的二层交换机，该交换机连接虚拟网卡和物理网卡，将虚拟机上的数据报文从物理网口转发出去。与物理交换机一样，vSwitch的作用就是用来转发数据。

2．vSwitch端口介绍

每个vSwitch都有两种接口，上联口和下联口。上联口用来连接物理网卡，每个上联口绑定一个物理网卡。一个vSwitch至少要有一个上联口，多个上联口可以进行捆绑。上联口可以配置IP地址，方便进行管理操作。下联口用于连接虚拟机，每个下联口连接一台虚拟机。与物理交换机不同的是，vSwitch下联口不会自动学习MAC地址，都是静态绑定的。上联口的IP和虚拟机的IP没有必然的关系，可以配置为不同的网段。

3．ACL策略

ACL策略与物理交换机的有些不同，在功能上有一定的限制。ACL策略能对流量进行基于目的IP的包过滤，因为下联口只有一台虚拟机，所以根本不需要基于源IP的规则。ACL支持基于ICMP、TCP和UDP协议进行操作。ACL策略应用于下联口的入方向。

4．网络策略模板

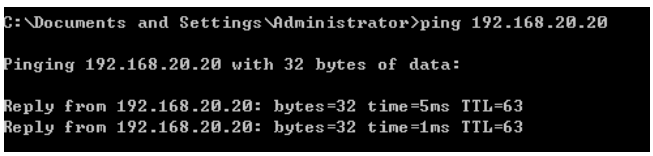
网络策略模板是对下联口和虚拟机的一些操作，可以对下联口进行出入方向的流量进行限制，规划下联口所属VLAN，设置VSI和应用ACL策略。缺省情况下，没有对下联口进行流控，端口所属VLAN为0（即数据包不带VLAN Tag）。

5．vSwitch转发模式

虚拟交换机有三种转发模式，VEB、VEPA和多通道模式。

1.1 VEB模式：

VEB是指vSwitch相对全面的网络转发功能，此模式下，同一VLAN的数据直接通过vSwitch转发，不需要通过外部网络。如果是不同的VLAN间通信，或者同一虚拟机属于VLAN但是在不同的vSwitch上，则需要通过外部网路。这样做的好处是可以减少外部网路的流量，减轻网络维护的难度，内部交换的速度也更快,只与CPU性能和内存带宽总线有关，而且对现有网络的兼容性较好。但是缺点是需要耗费更多的CPU，转发性能受制于CPU和网卡IO架构，也缺乏对流量的监控和安全控制。



1.2 VEPA模式：

VEPA是对VEB的一种修改，无论二层三层流量，统一要先发往外部网络，由物理交换机再回发到vSwitch。这种方式简化了服务器的vSwitch功能，使得内外网络相关联，服务器内部网络相当于外部网络的扩展和延伸。对于传统的交换机来说，从一个端口收到的报文是不能再从此端口发出的，这时候为了支持VEPA，交换机必须做一些修改，允许数据绕回，这种方式称为RR模式。如果是组播/广播流量，先绕回vSwitch，在服务器内部再进行数据的复制。VEPA模式可以借助于外部网络，很方便的对VM流量进行监控和安全控制管理，也减少了CPU的消耗，但是却额外增加了外部网络的流量和延迟。VEPA模式中用到了VSI（虚拟站点接口），可以看成是虚拟机网卡在交换机上的一个逻辑接口。VSI涉及到如下几个术语：

VSI管理ID：指VSI管理者的ID号，用于识别出可以访问并获取VSI类型的数据库，用IPv6地址来标识。

VSI类型ID：用来描述一个VSI的类型，对每个VSI管理者ID来说，只有唯一一个VSI类型ID。

VSI类型版本：一个整数标识符，允许VSI管理者数据库可以包含一个给定的VSI类型的多个版本。

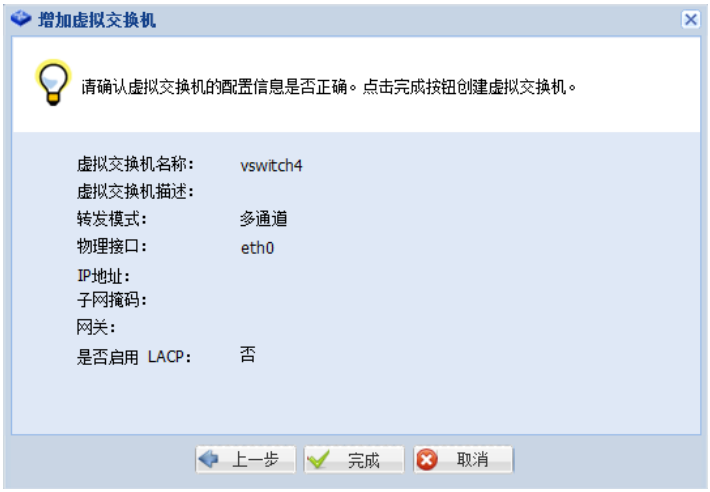
1.3 多通道模式：

多通道模式是将交换机端口或网卡划分为多个逻辑通道,称为S通道，并且各通道间逻辑隔离。每个逻辑通道可由用户根据需要定义成VEB或者VEPA。每个逻辑通道作为一个独立的到外部网络的通道进行处理。多通道借用了QINQ标准，另外增加了一个VLAN Tag，用channel ID和VLAN ID来对不同的通道进行标识，分别称为S-channel ID和S-VLAN ID，外层的Tag只会在服务器和接入交换机之间存在，而在VEPA模式下则不需要另外添加VLAN Tag。可以说，多通道模式是VEB和VEPA的结合。在这种模式下，组播/广播流量直接在物理交换机上进行数据的复制，然后通过各个通道再发给虚拟机。

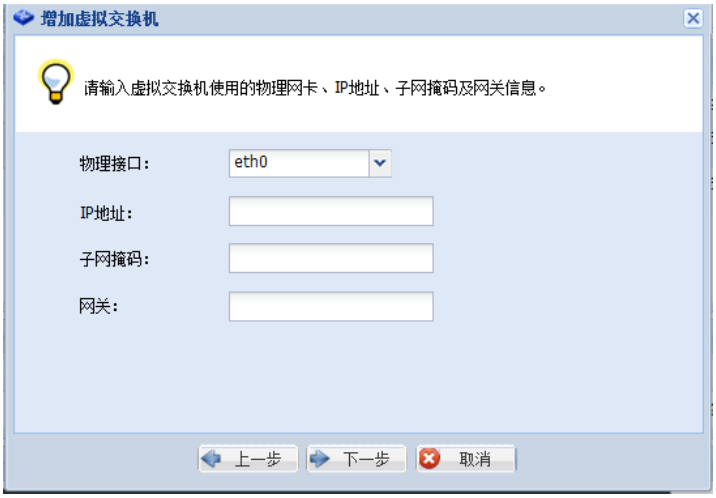
二、CAS vSwitch功能配置

1、显示流量实时监控和流量统计

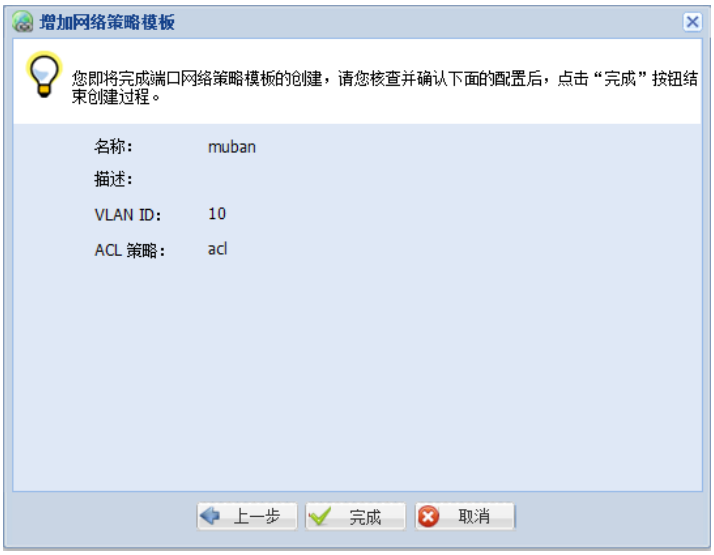
如下图是CAS管理台显示的vSwitch，闪烁绿色的端口表示下连的虚拟机处于开机状态。



在闪烁绿色的端口右击，选择查看端口流量实时监控，则会显示十分钟以内端口的流量：

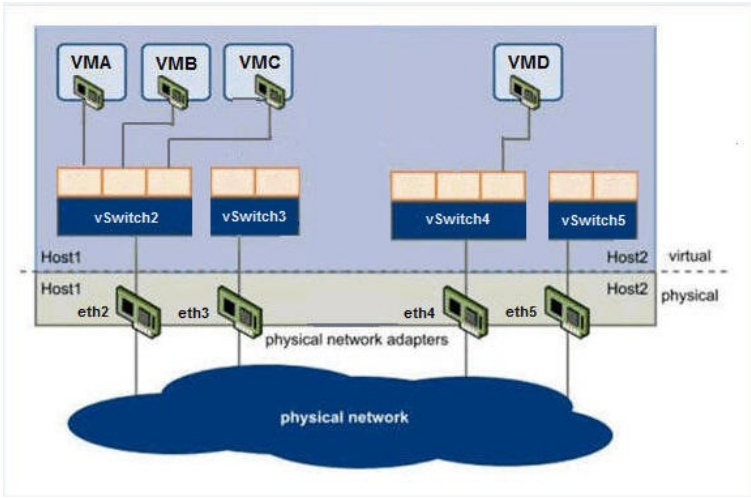


在闪烁绿色的端口右击，选择查看端口详细信息，则会显示端口从启动到目前为止流量总数：



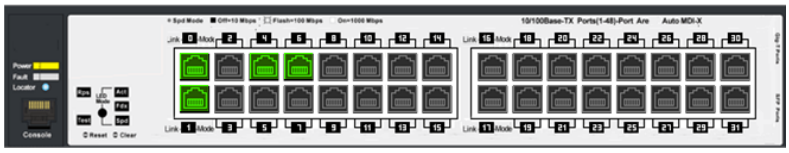
2、配置ACL策略

在导航栏的系统管理项里，选择ACL策略



然后在右侧的ACL策略栏里，选择增加ACL策略

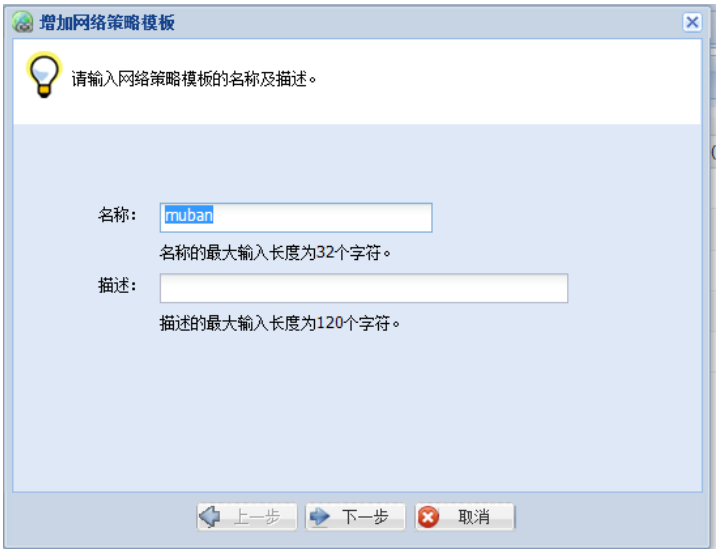
可以看到，默认的ACL动作是允许，现在添加规则：



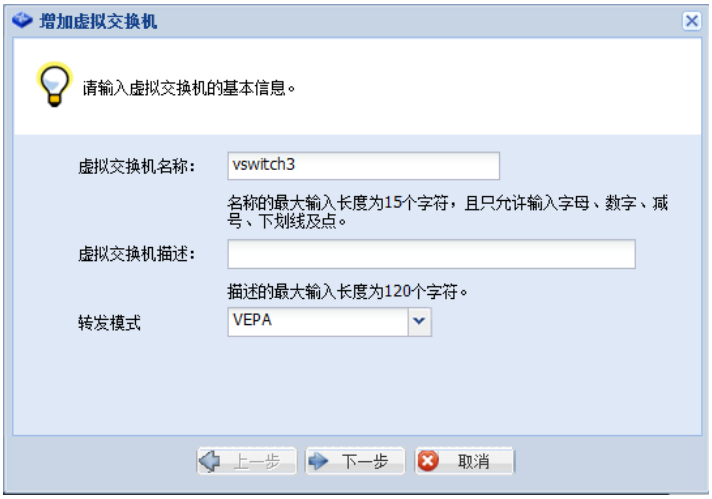
多个规则可以调整优先级：

3、配置网络策略模板

导航栏选择集群或者主机，在右面的栏里选择网络策略模板，然后按增加按钮：



输入网络策略模板名称：



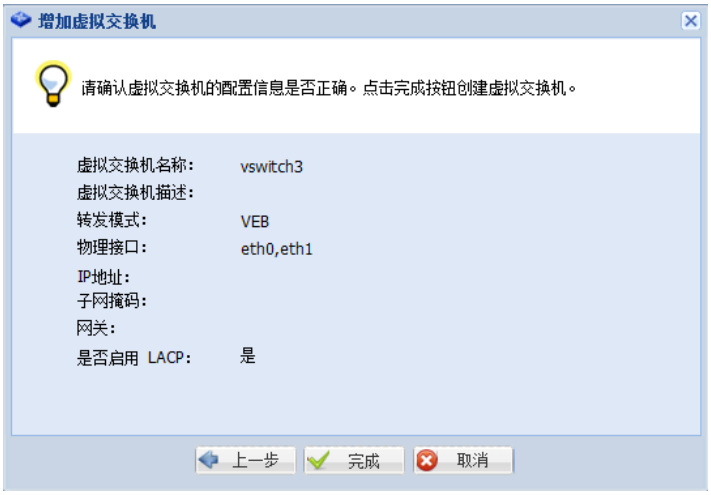
配置端口出入方向流量控制：

```
root@H3C-HZ-CVM:~# tcpdump -i eth2 -s -0 -xx host 192.168.10.10
tcpdump: WARNING: eth2: no IPv4 address assigned
tcpdump: verbose output suppressed, use -v or -vv for full protocol decode
listening on eth2, link-type EN10MB (Ethernet), capture size 65535 bytes
10:50:32.872841 IP 192.168.10.10 > 192.168.10.12: ICMP echo request, id 60451, seq 1, length 64
        0x0000:  0cda 411d 7456 0cda 411d b74a 8100 000a
        0x0010:  0800 4500 0054 0000 4000 4001 a542 c0a8
        0x0020:  0a0a c0a8 0a0c 0800 3537 ec23 0001 f7dc
        0x0030:  2651 0000 0000 eca2 0d00 0000 0000 1011
        0x0040:  1213 1415 1617 1819 1a1b 1c1d 1e1f 2021
        0x0050:  2223 2425 2627 2829 2a2b 2c2d 2e2f 3031
        0x0060:  3233 3435 3637
10:50:32.873739 IP 192.168.10.12 > 192.168.10.10: ICMP echo reply, id 60451, seq 1, length 64
        0x0000:  0cda 411d b74a 0cda 411d 7456 8100 000a
        0x0010:  0800 4500 0054 0129 4000 8001 6419 c0a8
        0x0020:  0a0c c0a8 0a0a 0000 3d37 ec23 0001 f7dc
        0x0030:  2651 0000 0000 eca2 0d00 0000 0000 1011
        0x0040:  1213 1415 1617 1819 1a1b 1c1d 1e1f 2021
        0x0050:  2223 2425 2627 2829 2a2b 2c2d 2e2f 3031
        0x0060:  3233 3435 3637
```

设置VSI信息：

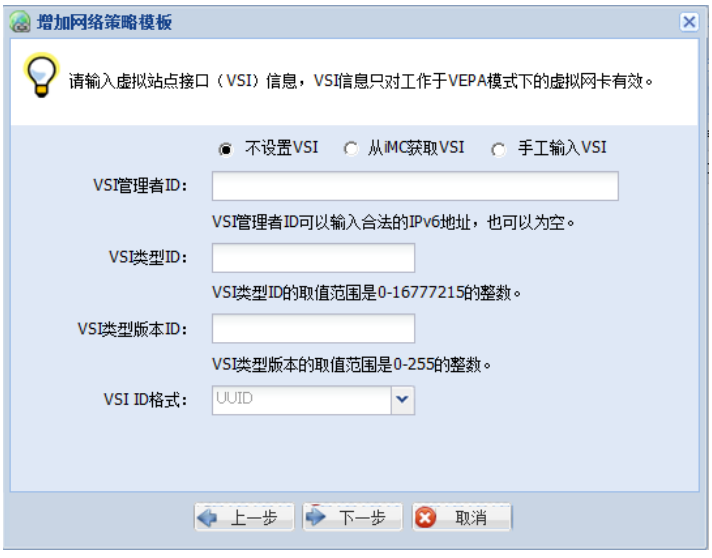
```
root@H3C-HZ-CVM:~# tcpdump -i eth2 -s -0 -xx host 192.168.0.10
tcpdump: WARNING: eth2: no IPv4 address assigned
tcpdump: verbose output suppressed, use -v or -vv for full protocol decode
listening on eth2, link-type EN10MB (Ethernet), capture size 65535 bytes
```

应用ACL策略和设置VLAN：

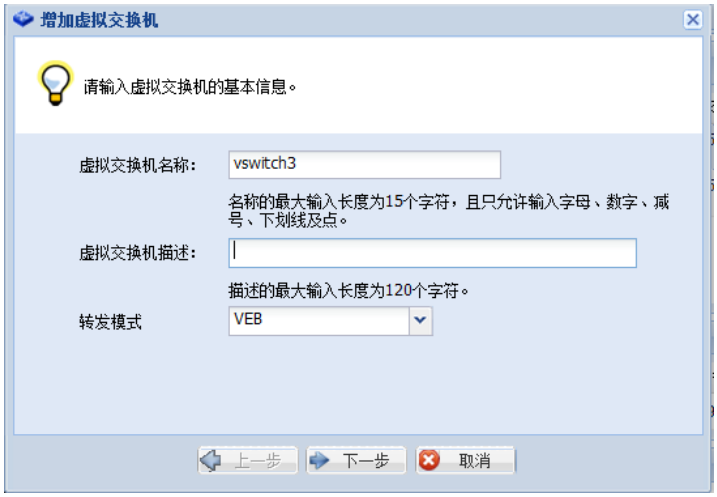


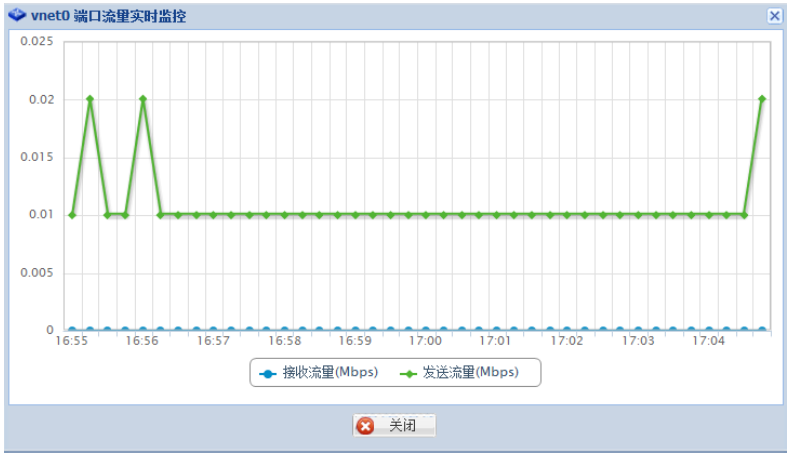
4、配置vSwitch

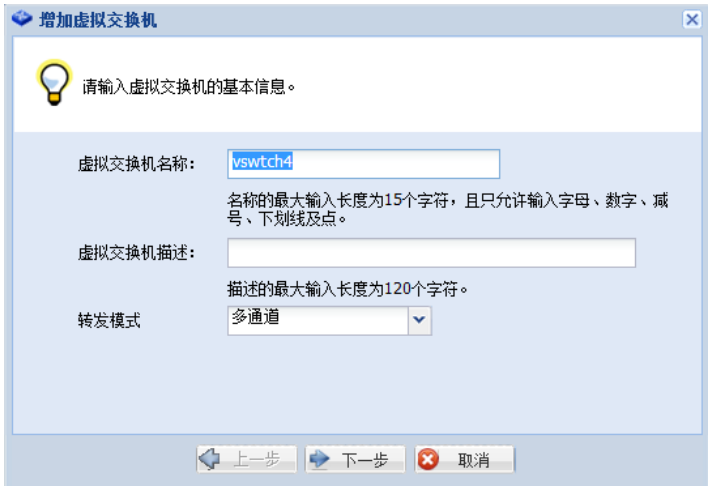
导航栏选择主机，然后主机栏选择虚拟交换机，按增加按钮：



VEB模式：

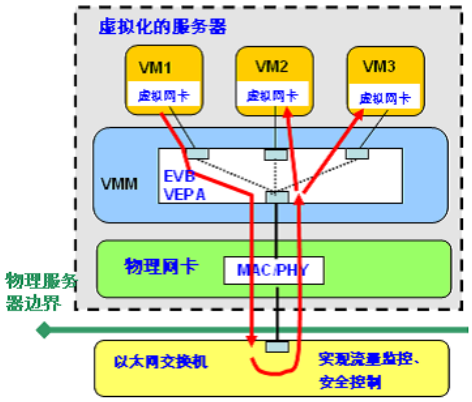




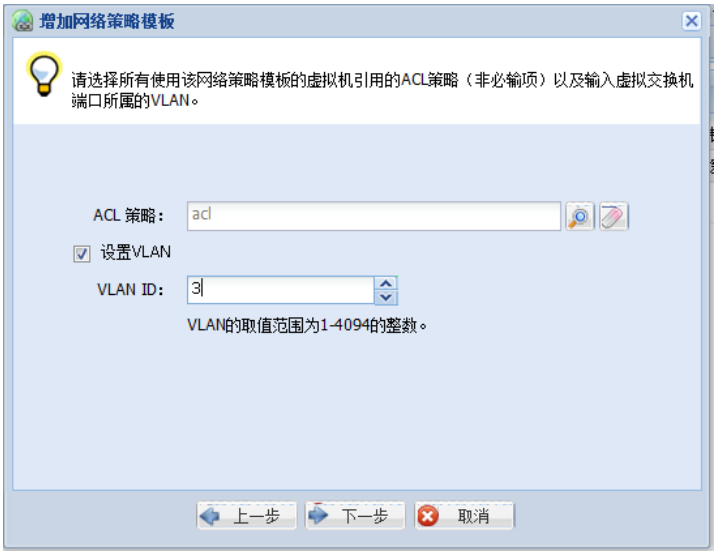


VEPA模式：

```
15:45:00.874737 IP 192.168.10.12 > 192.168.20.20: ICMP echo request, id 512, seq 2304, length 40
        0x0000:  00e0 fc6d ce8a 0cda 411d 7456 8100 000a
        0x0010:  0800 4500 003c 0171 0000 8001 99df c0a8
        0x0020:  0a0c c0a8 1414 0800 425c 0200 0900 6162
        0x0030:  6364 6566 6768 696a 6b6c 6d6e 6f70 7172
        0x0040:  7374 7576 7761 6263 6465 6667 6869
15:45:00.875664 IP 192.168.20.20 > 192.168.10.12: ICMP echo reply, id 512, seq 2304, length 40
        0x0000:  0cda 411d 7456 00e0 fc6d ce80 8100 000a
        0x0010:  0800 4500 003c 2f6f 0000 3f01 ace1 c0a8
        0x0020:  1414 c0a8 0a0c 0000 4a5c 0200 0900 6162
        0x0030:  6364 6566 6768 696a 6b6c 6d6e 6f70 7172
        0x0040:  7374 7576 7761 6263 6465 6667 6869
```



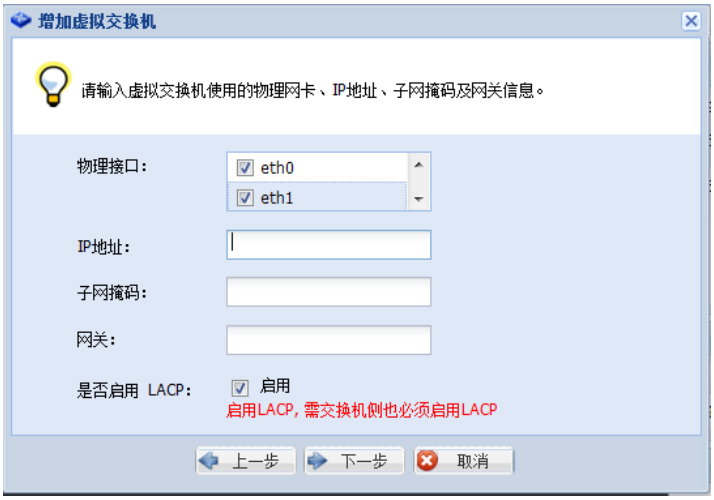
多通道模式:




```
15:50:01.824005 IP 192.168.10.10 > 192.168.20.20: ICMP echo request, id 62502, seq 1, length 64
        0x0000:  00e0 fc6d ce8a 0cda 411d b74a 8100 000a
        0x0010:  0800 4500 0054 0000 4000 4001 9b3a c0a8
        0x0020:  0a0a c0a8 1414 0800 25ad f426 0001 2823
        0x0030:  2751 0000 0000 c3e3 0c00 0000 0000 1011
        0x0040:  1213 1415 1617 1819 1a1b 1c1d 1e1f 2021
        0x0050:  2223 2425 2627 2829 2a2b 2c2d 2e2f 3031
        0x0060:  3233 3435 3637
15:50:01.824102 IP 192.168.10.10 > 192.168.20.20: ICMP echo request, id 62502, seq 1, length 64
        0x0000:  0cda 411d db82 00e0 fc6d ce80 8100 0014
        0x0010:  0800 4500 0054 0000 4000 3f01 9c3a c0a8
        0x0020:  0a0a c0a8 1414 0800 25ad f426 0001 2823
        0x0030:  2751 0000 0000 c3e3 0c00 0000 0000 1011
        0x0040:  1213 1415 1617 1819 1a1b 1c1d 1e1f 2021
        0x0050:  2223 2425 2627 2829 2a2b 2c2d 2e2f 3031
        0x0060:  3233 3435 3637
15:50:01.824896 IP 192.168.20.20 > 192.168.10.10: ICMP echo reply, id 62502, seq 1, length 64
        0x0000:  00e0 fc6d ce85 0cda 411d db82 8100 0014
        0x0010:  0800 4500 0054 08c6 0000 4001 d274 c0a8
        0x0020:  1414 c0a8 0a0a 0000 2dad f426 0001 2823
        0x0030:  2751 0000 0000 c3e3 0c00 0000 0000 1011
        0x0040:  1213 1415 1617 1819 1a1b 1c1d 1e1f 2021
        0x0050:  2223 2425 2627 2829 2a2b 2c2d 2e2f 3031
        0x0060:  3233 3435 3637
15:50:01.825003 IP 192.168.20.20 > 192.168.10.10: ICMP echo reply, id 62502, seq 1, length 64
        0x0000:  0cda 411d b74a 00e0 fc6d ce80 8100 000a
        0x0010:  0800 4500 0054 08c6 0000 3f01 d374 c0a8
        0x0020:  1414 c0a8 0a0a 0000 2dad f426 0001 2823
        0x0030:  2751 0000 0000 c3e3 0c00 0000 0000 1011
        0x0040:  1213 1415 1617 1819 1a1b 1c1d 1e1f 2021
        0x0050:  2223 2425 2627 2829 2a2b 2c2d 2e2f 3031
        0x0060:  3233 3435 3637
```

需要注意的是，VEB模式可以有多个上联口，并且可以使用LACP，但是VEPA和多通道模式则只能有一个上联口。

三、典型组网



VEB模式：

如上图所示是典型的vSwitch组网图。vSwitch存在于服务器内部，vSwitch下连虚拟机，上连物理网卡，虚拟机通过物理网卡和外界通信。其中，eth3连接管理网，eth2和eth4连接业务网，eth5连接存储网。所有vSwitch都是VEB模式的，如下是虚拟机的IP和VLAN的规划：

虚拟机	VLAN	IP	虚拟交换机	网关
VMA	10	192.168.10.10/24	vSwitch2	192.168.10.1
VMB	10	192.168.10.11/24	vSwitch2	192.168.10.1
VMC	20	192.168.20.20/24	vSwitch2	192.168.20.1
VMD	10	192.168.10.12/24	vSwitch4	192.168.10.1

网关都是在外部网络的物理交换机上。服务器和物理交换机连接的链路配置为trunk，允许VLAN10、20通过。

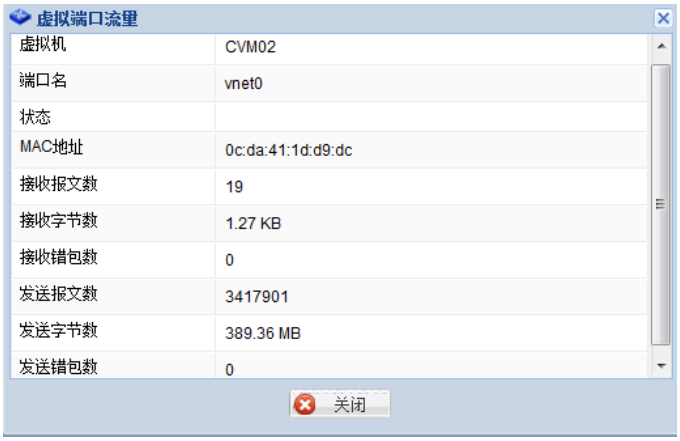
开启所有虚拟机，通过抓包来观察虚拟机通信时数据包的走向。

用VMA ping VMB，抓eth2的包：

```
[root@localhost ~]# ping 192.168.20.20
PING 192.168.20.20 (192.168.20.20) 56(84) bytes of data.
64 bytes from 192.168.20.20: icmp_seq=1 ttl=63 time=1.62 ms
64 bytes from 192.168.20.20: icmp_seq=2 ttl=63 time=0.822 ms
64 bytes from 192.168.20.20: icmp_seq=3 ttl=63 time=1.04 ms
```

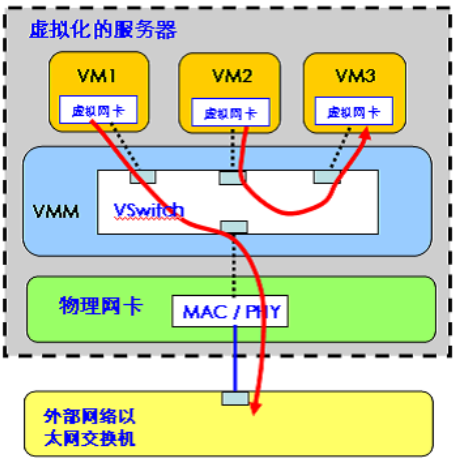
发现能够ping通，但是没有抓到任何包，说明VEB模式同一vSwitch的二层通信是只在服务器内部。

用VMA ping VMD:



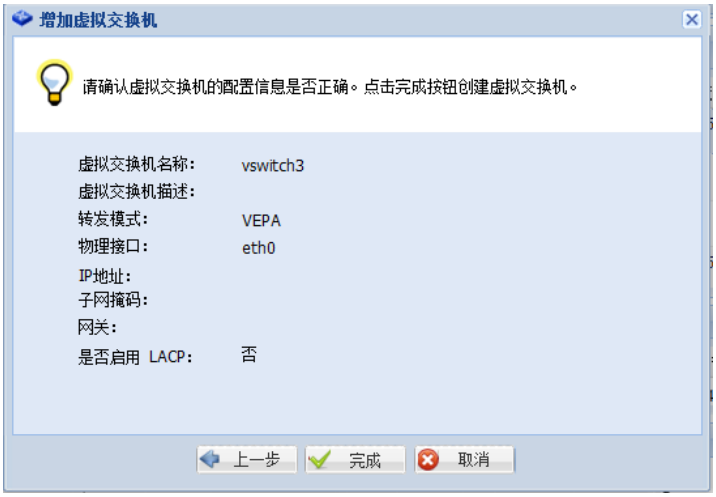
eth2抓到了ping包，并且带有VLAN 10的Tag，说明VEB模式不同vSwitch间的二层通信需要把数据先发往外部网络。

用VMA ping VMC:



eth2抓到了ping包，icmp请求报文出去时带有VLAN 10的Tag，回来时带有VLAN 20的Tag，icmp应答报文出去时带有VLAN 20的Tag，回来时带有VLAN 10的Tag说明VEB模式同一vSwitch间的三层通信需要把数据先发往外部网络。

用VMD ping VMC:



Eth4上抓包：

```
[root@localhost ~]# ping 192.168.10.11
PING 192.168.10.11 (192.168.10.11) 56(84) bytes of data.
64 bytes from 192.168.10.11: icmp_seq=1 ttl=64 time=0.789 ms
64 bytes from 192.168.10.11: icmp_seq=2 ttl=64 time=0.761 ms
64 bytes from 192.168.10.11: icmp_seq=3 ttl=64 time=5.67 ms
```

Eth2上抓包：

```
[root@localhost ~]# ping 192.168.10.12
PING 192.168.10.12 (192.168.10.12) 56(84) bytes of data.
64 bytes from 192.168.10.12: icmp_seq=1 ttl=128 time=1.64 ms
64 bytes from 192.168.10.12: icmp_seq=2 ttl=128 time=0.989 ms
64 bytes from 192.168.10.12: icmp_seq=3 ttl=128 time=0.956 ms
```

eth2和eth4抓到了ping包，eth4抓的报文带有VLAN 10的Tag，eth2抓的报文带有VLAN 20的Tag，说明VEB模式不同vSwitch间的三层通信需要把数据先发往外部网络。