蚂蚁金服Docker网络插件开发和实践

署名：辽客
职称：高级技术专家

主办单位：杭州 Alibaba Group 阿里巴巴集团 战略合作伙伴：intel

docker网络分析

docker网络插件开发

蚂蚁金服网络插件实践

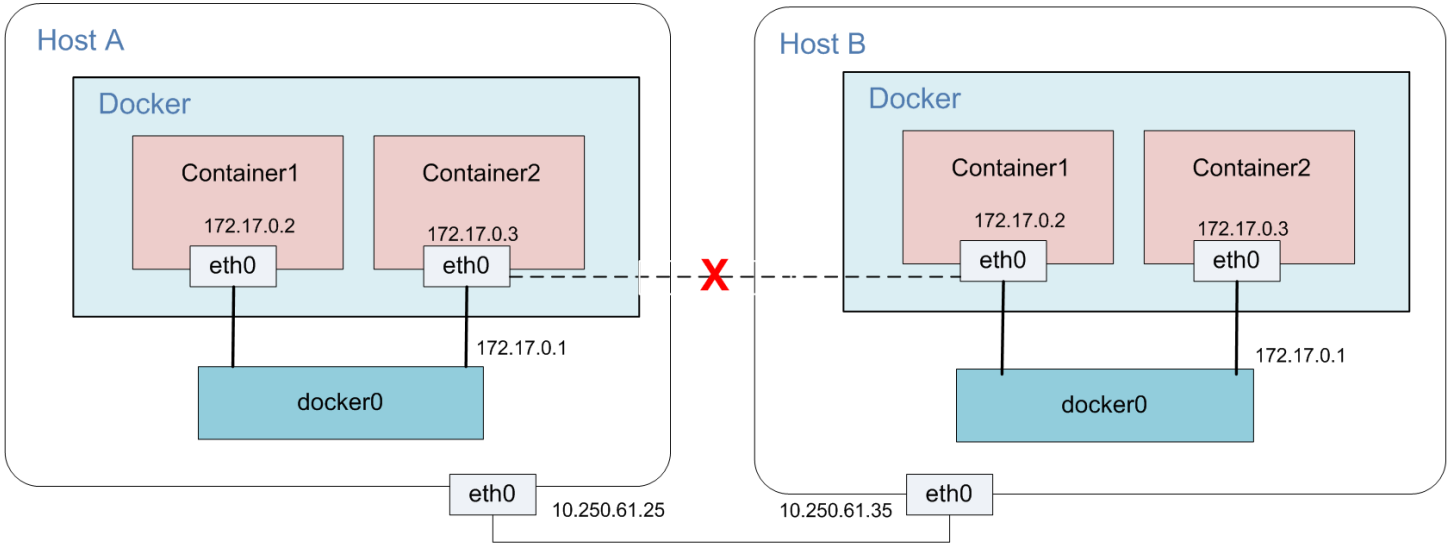
Docker 网络分析

Bridge Network

Overlay Network

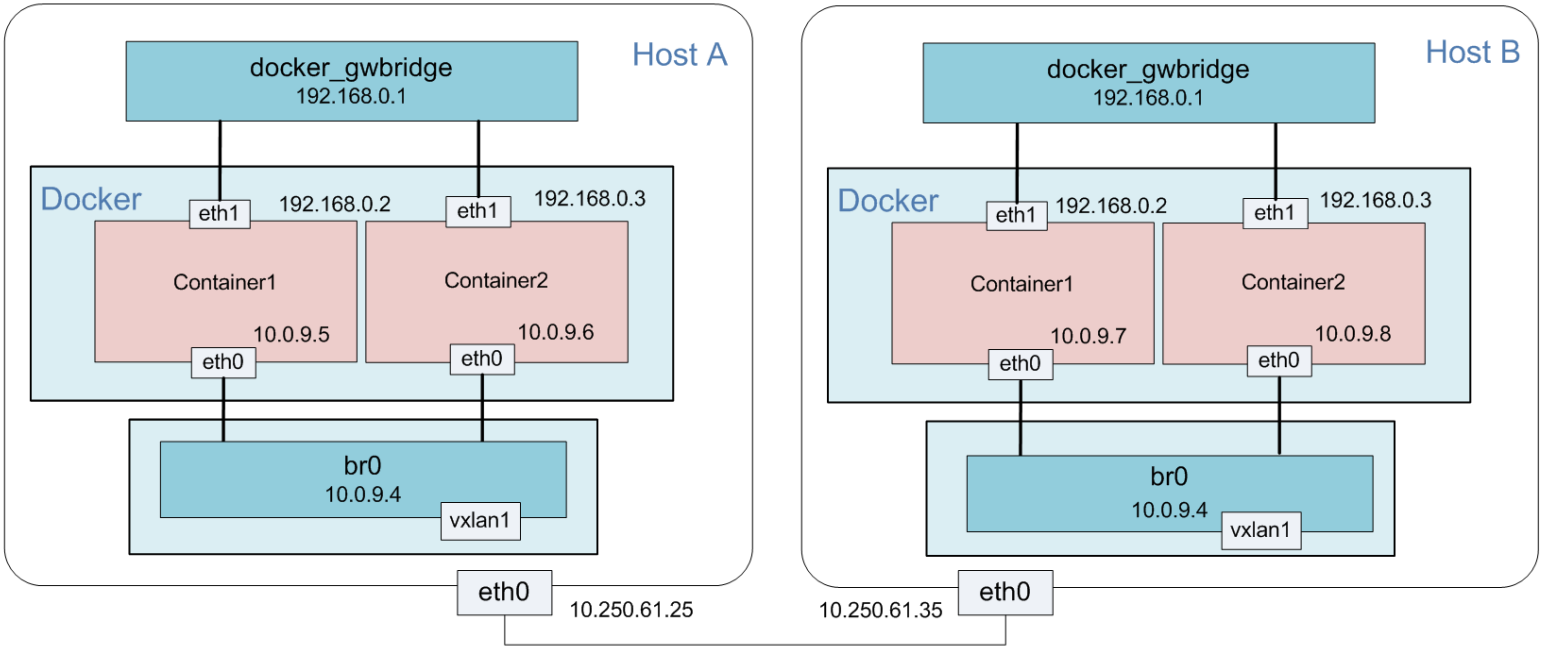
Weave Network

Docker Bridge Network

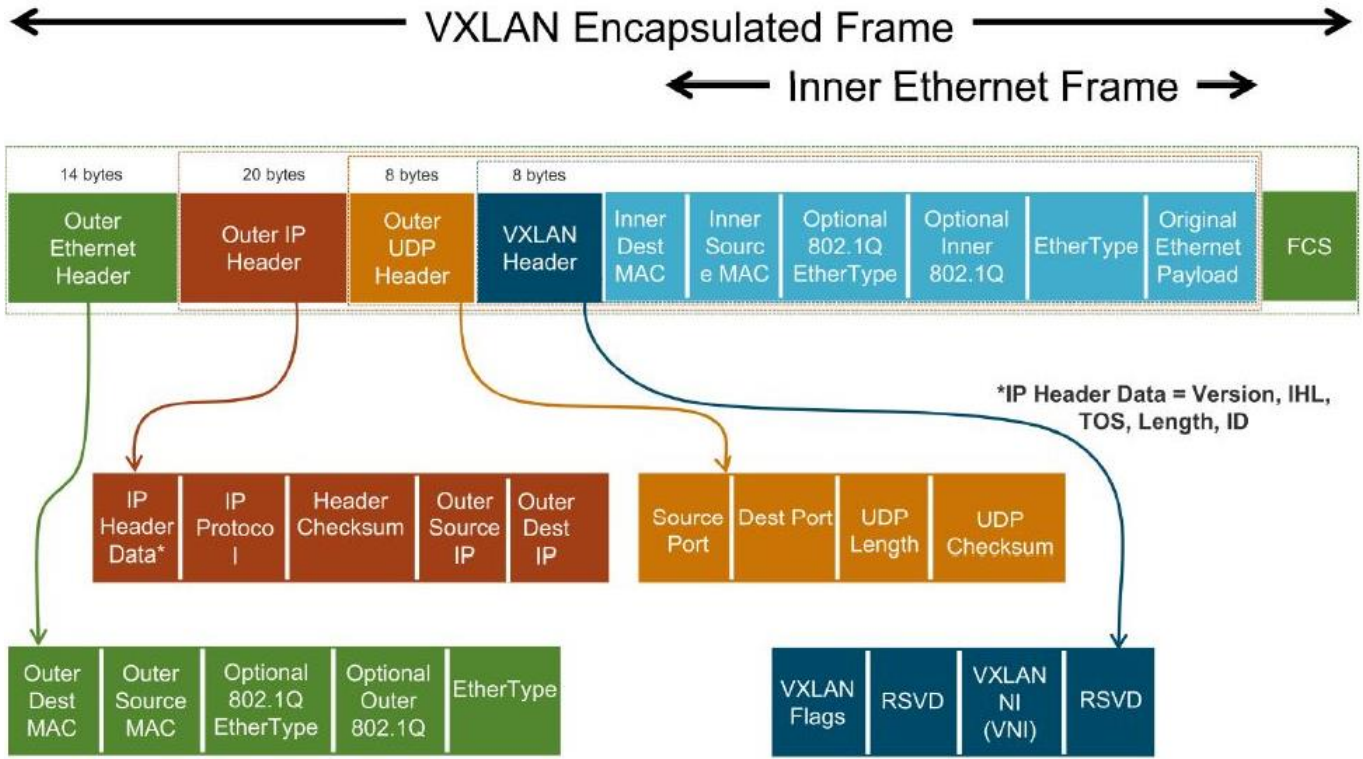


- docker启动会在宿主机上创建docker0的网桥作为容器的网关
- 不同宿主机的容器无法直接通信
- 通过iptables SNAT/DNAT和容器外网络通信

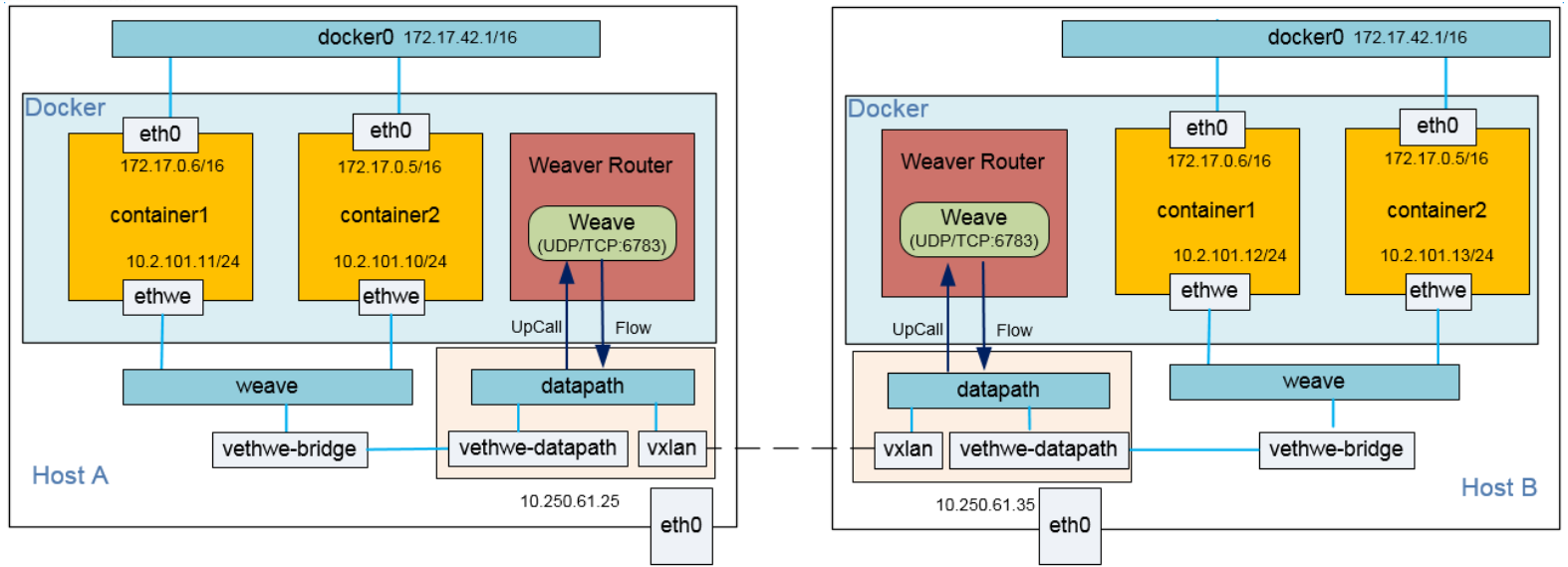
Docker Overlay Network



- 每个overlay网络有独立namespace和br0网桥
- ARP：使用gossip协议管理节点成员关系（Serf）、广播L3Miss
- 不同宿主机的容器通过vxlan隧道直接通信


VXLAN Encapsulated Frame
Inner Ethernet Frame
14 bytes
20 bytes
8 bytes
8 bytes
Outer Ethernet Header
Outer IP Header
Outer UDP Header
VXLAN Header
Inner Dest MAC
Inner Source MAC
Optional 802.1Q EtherType
Optional Inner 802.1Q
EtherType
Original Ethernet Payload
FCS
*IP Header Data = Version, IHL, TOS, Length, ID
IP Header Data*
IP Protocol
Header Checksum
Outer Source IP
Outer Dest IP
Source Port
Dest Port
UDP Length
UDP Checksum
Outer Dest MAC
Outer Source MAC
Optional 802.1Q EtherType
Optional Outer 802.1Q
EtherType
VXLAN Flags
RSVD
VXLAN NI (VNI)
RSVD

Weave Network



- 主机间通过weave 容器建立连接（gossip协议）
- 同一主机内的容器间直接通过weave 网桥通信，无需经过weave router
- 跨主机的容器通过datapath, vxlan隧道通信
- Weaver 容器处理datapath上的FlowMiss, 把转发规则注入到datapath

扫码观看大会视频

小结：

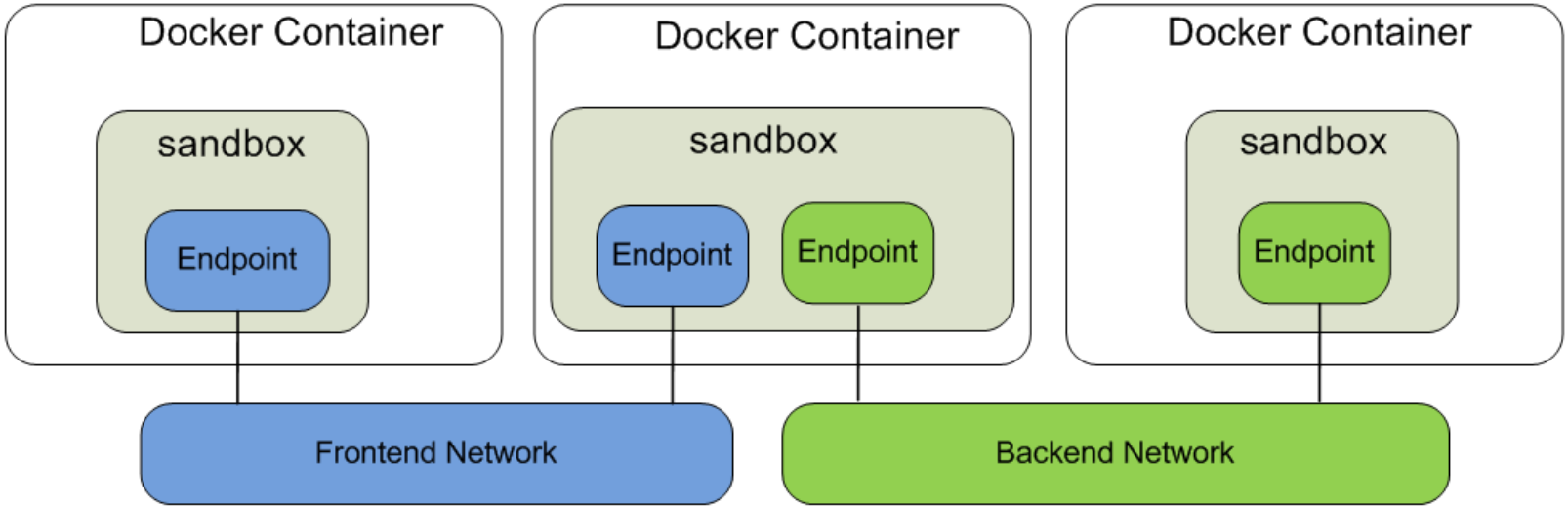
- 容器的跨主机通信的需求
- ARP的广播增加网络负担
- Gossip协议的运维复杂度
- 多种docker网络环境（aliyun经典VM/VPC 混合云）
- 高性能网络业务（带宽延迟）

Docker 网络插件开发

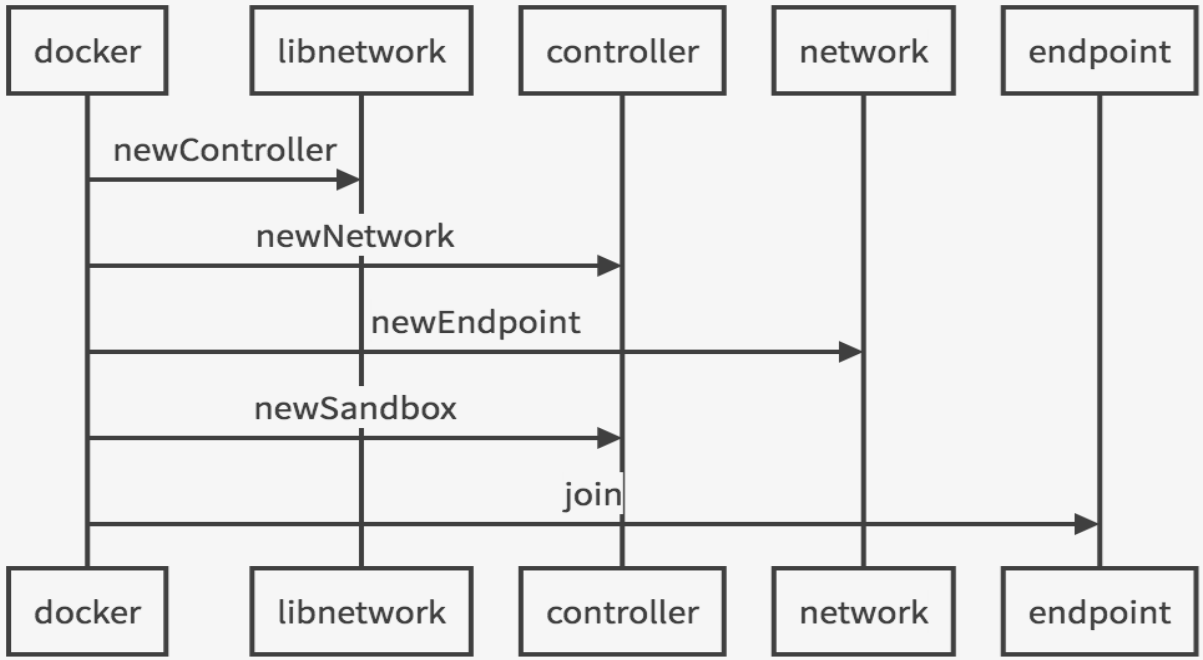
CNM模型

插件架构和开发

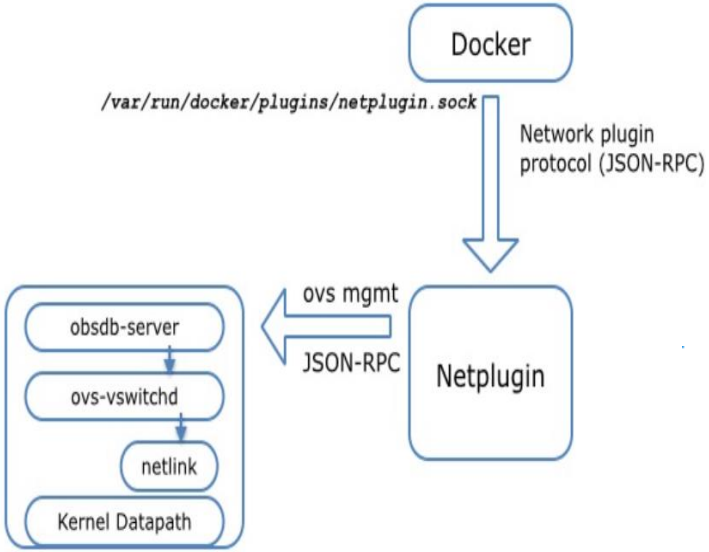
CNM 模型



- Sandbox: 一个隔离的网络配置环境
- Endpoint：隶属于某一个network的通讯端口，多个Endpoint也可以在一个sandbox中共存。
- Network： 一个唯一的、可识别的endpoint组，组内endpoint可以相互通讯。你可以创建一个Frontend和Backend network，然后这两个network是完全隔离的。


docker
libnetwork
controller
network
endpoint
newController
newNetwork
newEndpoint
newSandbox
join
docker
libnetwork
controller
network
endpoint

插件架构和开发



- /var/run/docker/plugins
  .sock/.spec/.json
- Docker plugin API
  https://docs.docker.com/engine/extend/plugin_api/
- 开发docker plugin的利器
  https://github.com/docker/go-plugins-helpers

Example using Unix sockets:

```
import "github.com/docker/go-plugins-helpers/network"

d := MyNetworkDriver{}
h := network.NewHandler(d)
h.ServeUnix("root", "test_network")
```

扫码观看大会视频

```
// Driver represent the interface a driver must fulfill.
type Driver interface {
    GetCapabilities() (*CapabilitiesResponse, error)
    CreateNetwork(*CreateNetworkRequest) error
    CreateEndpoint(*CreateEndpointRequest) (*CreateEndpointResponse, error)
    EndpointInfo(*InfoRequest) (*InfoResponse, error)
    Join(*JoinRequest) (*JoinResponse, error)
    Leave(*LeaveRequest) error
    DiscoverNew(*DiscoveryNotification) error
    DiscoverDelete(*DiscoveryNotification) error
}
```

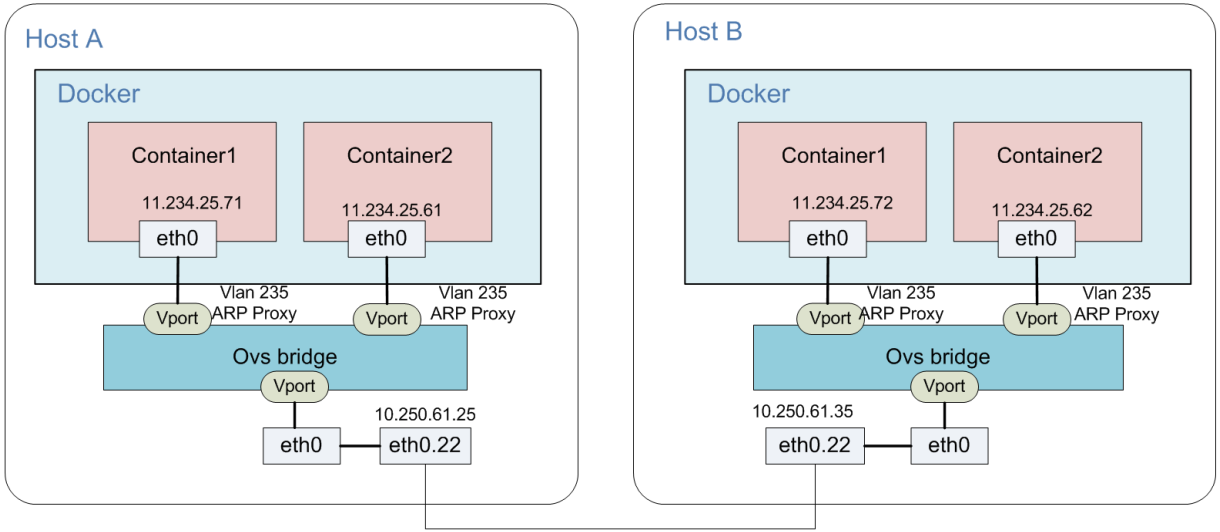
蚂蚁金服网络插件

Vlan/SRIOV Driver

VPC Driver

Vxlan Driver(Smartnic)

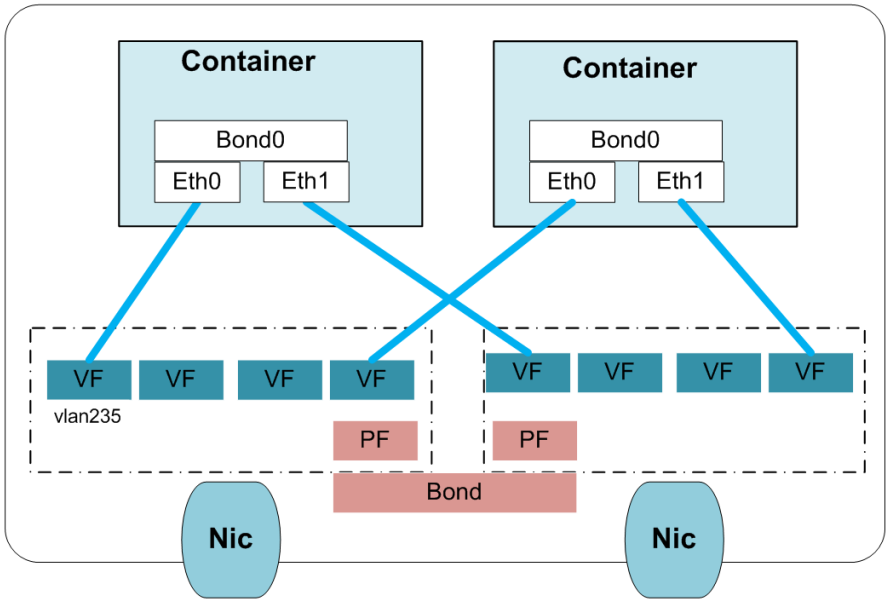
Vlan Driver



- 宿主机的管控流量和容器的流量隔离在两个不同的vlan
- ARP广播风暴，交换机PPS压力（ARP 带回）
- 大二层，交换机MAC表象有限（MACNAT）
- 最适配现有物理网络，对业务影响最小（企业内部，物理机）

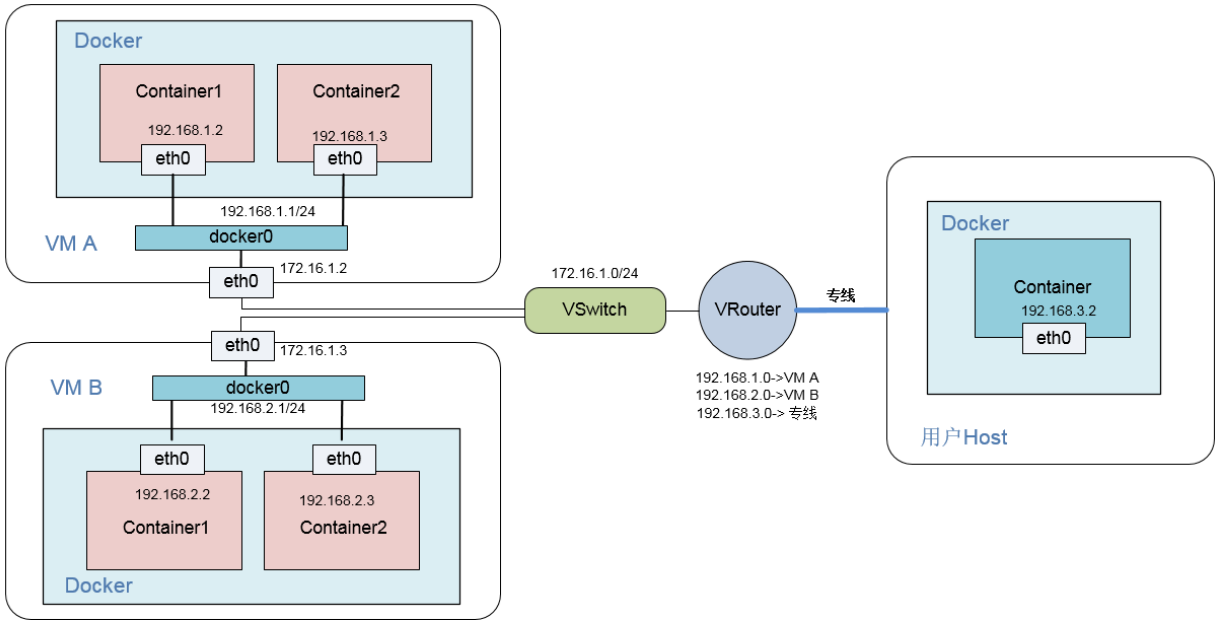
扫码观看大会视频

SRIOV Driver



- VF/PF 不同vlan隔离容器和宿主机流量
- VF 直接assign给Container， 低延迟高带宽
- PF Lacp bonding， VF XOR 双活bonding
- 适用于对网络带宽，延迟有较高要求的业务场景（DB等）

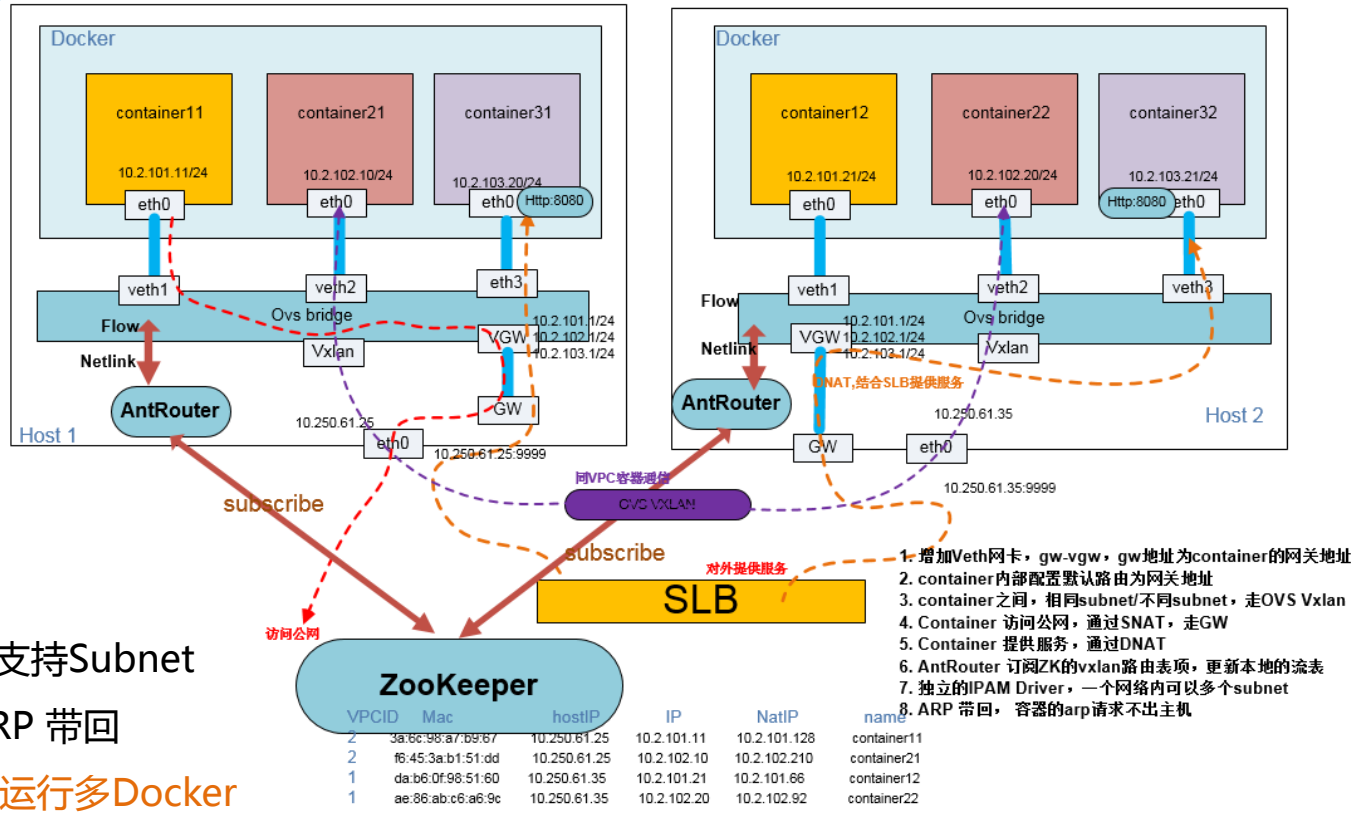
VPC Driver



- 自定义路由的方式实现网络可达，阿里云openapi支持
- 利用专线可以实现和用户中心的Docker互联
- 适用于在阿里云的VPC上使用docker（公有云业务，公司业务上云）

扫码观看大会视频

Vxlan Driver



- 网络拓扑静态
- 独立的IPAM，支持Subnet
- MAC NAT，ARP 带回
- 阿里云经典VM运行多Docker
- 内部测试网络，CICD环境

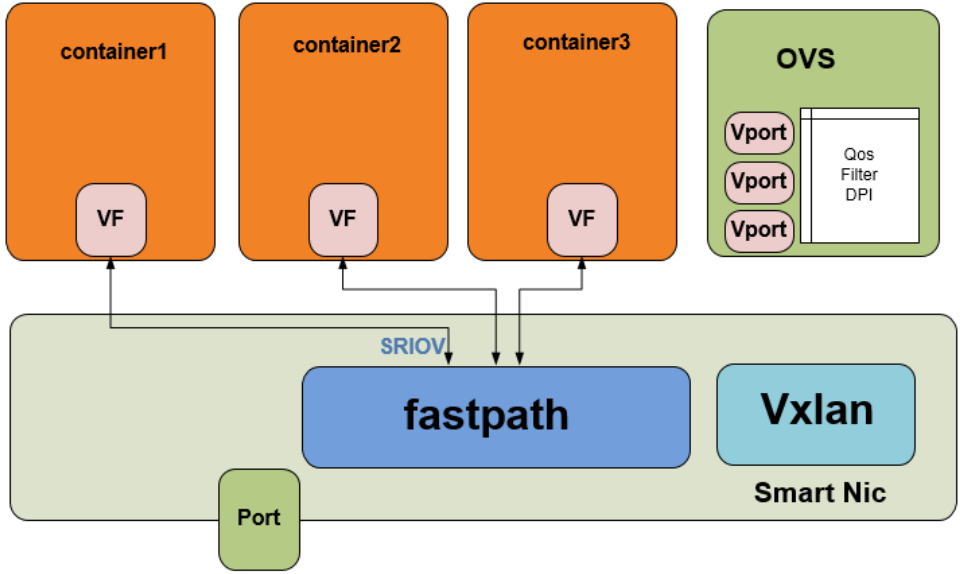
扫码观看大会视频

- 多Subnet示例：

```
docker network create -d vxlan --ipam-driver=vxlan --
gateway=192.168.0.1 --subnet=192.168.0.0/24 --opt VxlanId=110 --
ipam-opt VxlanId=110 vxlan11.1

docker network create -d vxlan --ipam-driver=vxlan --
gateway=192.168.1.1 --subnet=192.168.1.0/24 --opt VxlanId=110 --
ipam-opt VxlanId=110 vxlan11.0
```

Smart Nic Driver



- 性能损耗（Vxlan卸载，流表查询）
- 万兆网络普及
- 统一调度，资源整合

- SRIOV
- Qos 功能，离在线业务混部
- Fastpath offload
- Bonding/Session/NAT

扫码观看大会视频

2016 The Computing Conference

THANKS