小米容器平台网络实战

刘威波，小米资深运维研发工程师

2017.08

1 背景

网络性能

- 高吞吐
- 低延迟
- 无单点

容器联通性

- 容器 ←→ 容器
- 容器 ←→ 非容器服务器（物理机，云主机等）

2 总体进度

2014年6月 自建机房实现 “bridge + vlan + dhcp” 网络解决方案

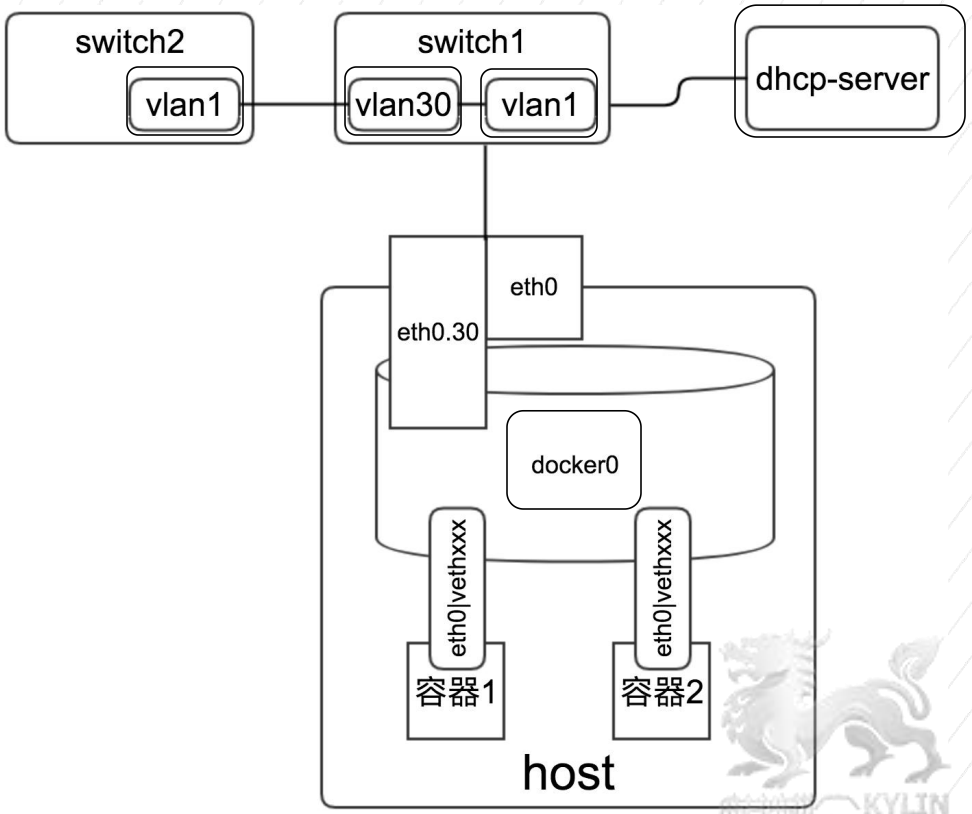
2017年4月 阿里云实现“flannel + iptables ” 网络解决方案

2017年1月 调研云厂商容器网络解决方案： calico / flannel

3 自建机房容器网络

bridge + vlan + dhcp 网络方案

- 真实内网IP，便于标识和定位
- 与原有物理网络天然互通
- 吞吐、延迟与纯物理网络差别很小



4 CALICO VS FLANNEL

路由转发 > 隧道

路由转发

- **Calico : bgp > Flannel : host-gw**
- **Flannel : ali、aws、gec**

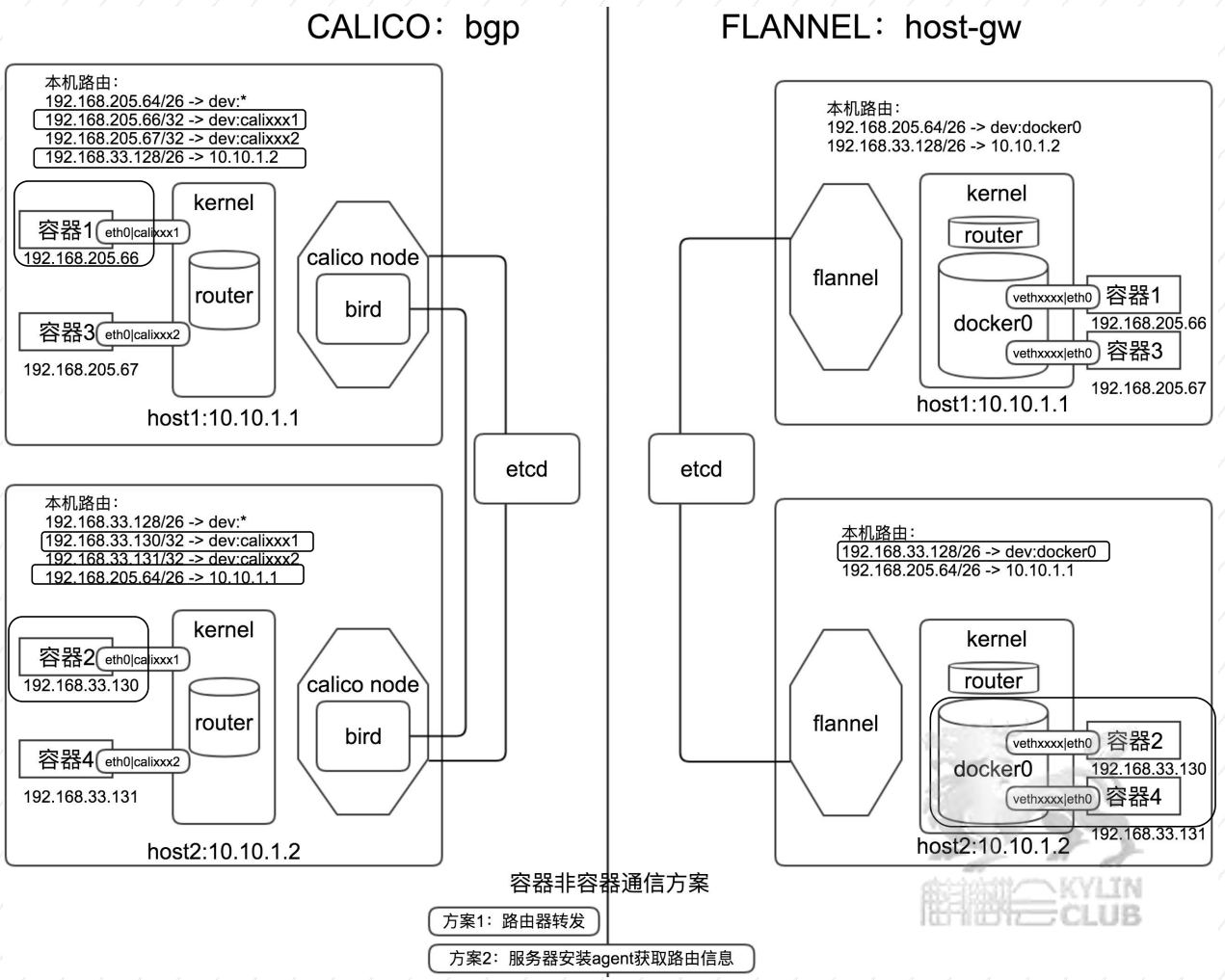
隧道

- **Calico : ipip VS Flannel : vxlan**

4 CALICO VS FLANNEL

路由转发

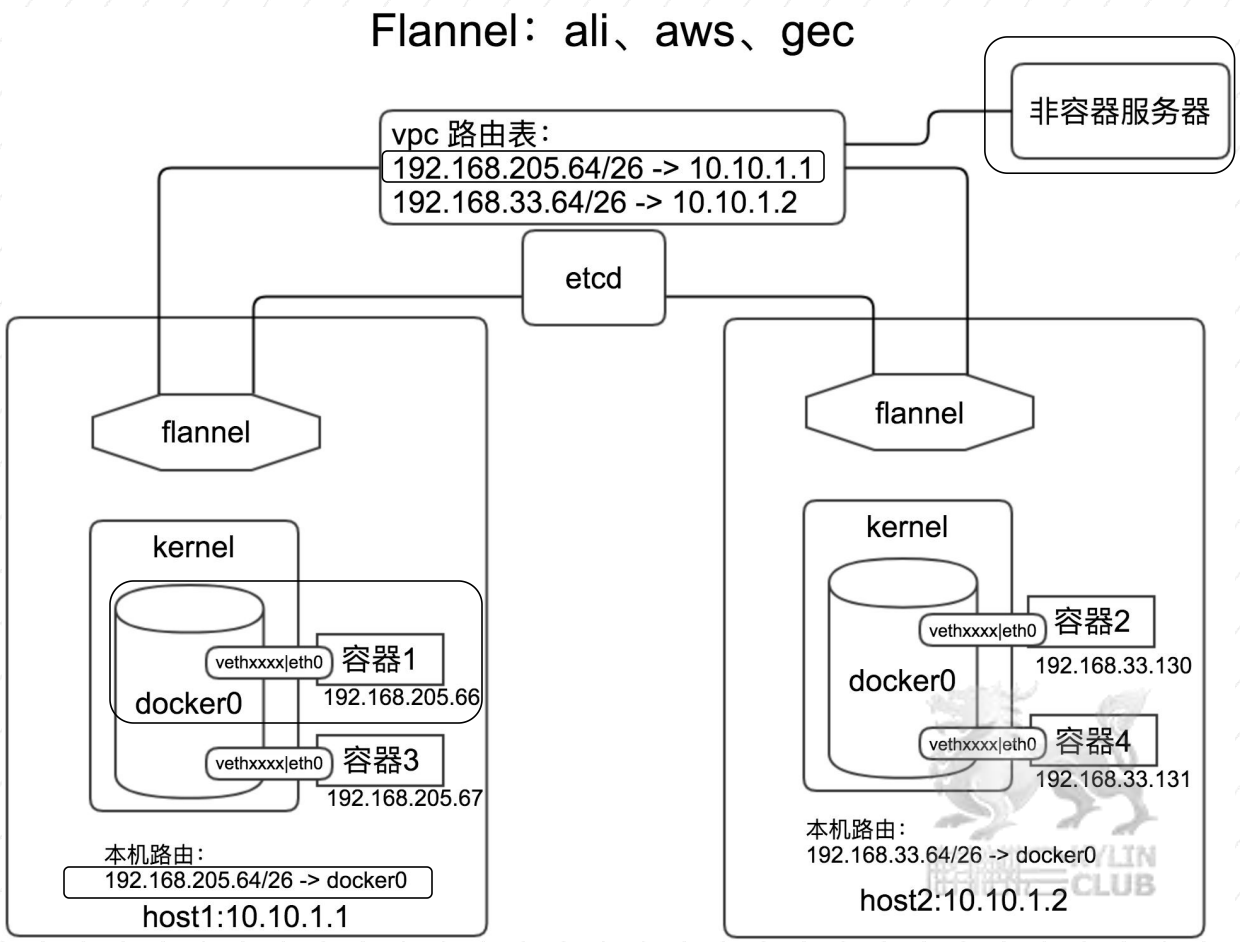
- 是否使用网桥
  - Calico 否
  - Flannel 是
- 是否支持宿主机跨广播域
  - Calico 是
  - Flannel 否
- 是否支持路由器转发流量
  - Calico 是
  - Flannel 否
- 非容器服务器是否可以通过安装agent获取容器路由信息
  - Calico 是
  - Flannel 是



4 CALICO VS FLANNEL

路由转发

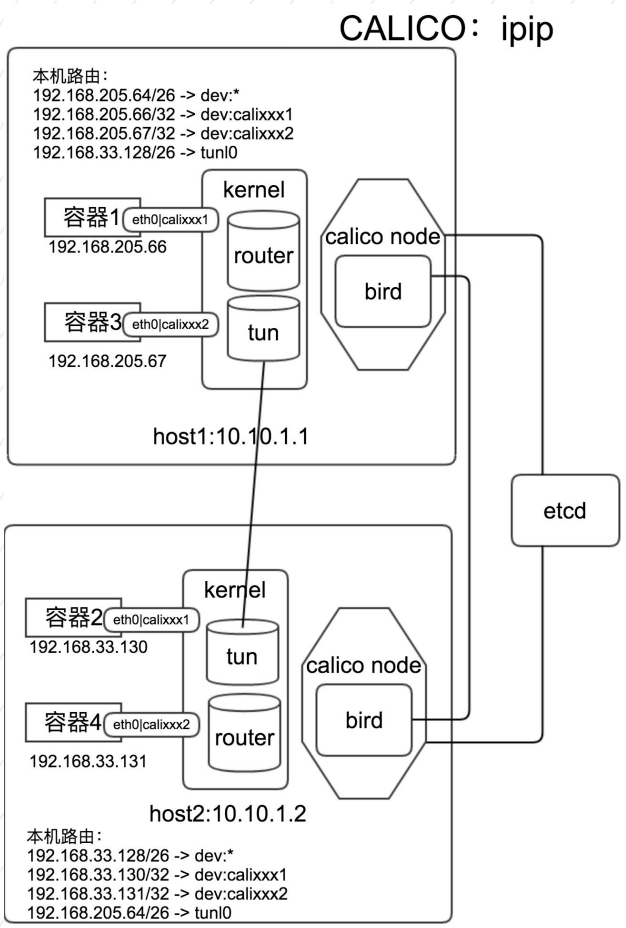
- 实现：
  - Flannel使用云厂商vpc路由表实现高吞吐、低延迟、无单点的网络
- 限制
  - vpc路由表条目数量

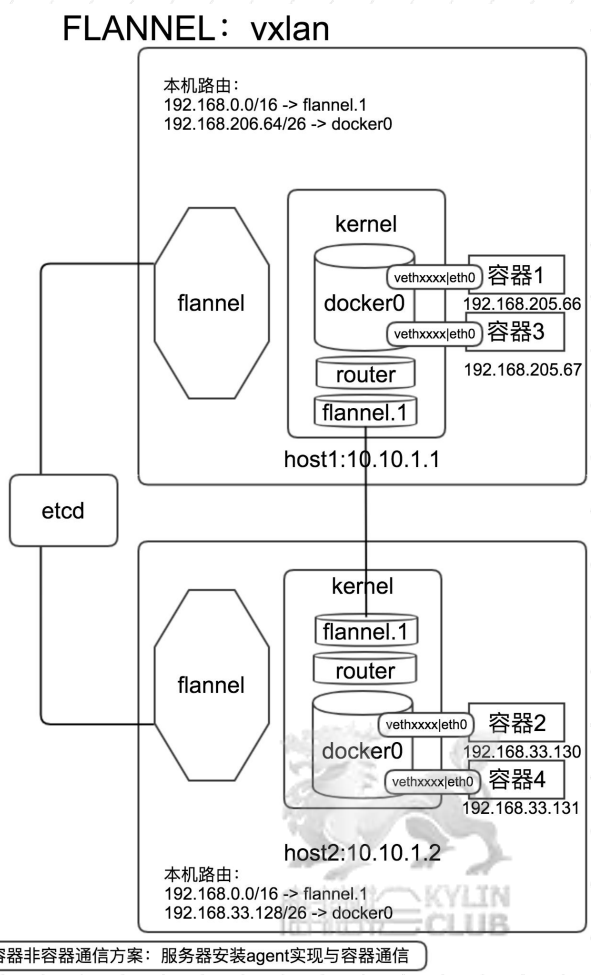


4 CALICO VS FLANNEL

隧道

- 是否使用网桥
  - Calico 否
  - Flannel 是
- 是否支持宿主机跨广播域
  - Calico 是
  - Flannel 是
- 隧道实现网络层次：
  - Calico 网络层
  - Flannel 传输层
- 是否支持容器与非容器服务器通信
  - Calico 否
  - Flannel 是，非容器服务器安装agent实现与容器通信

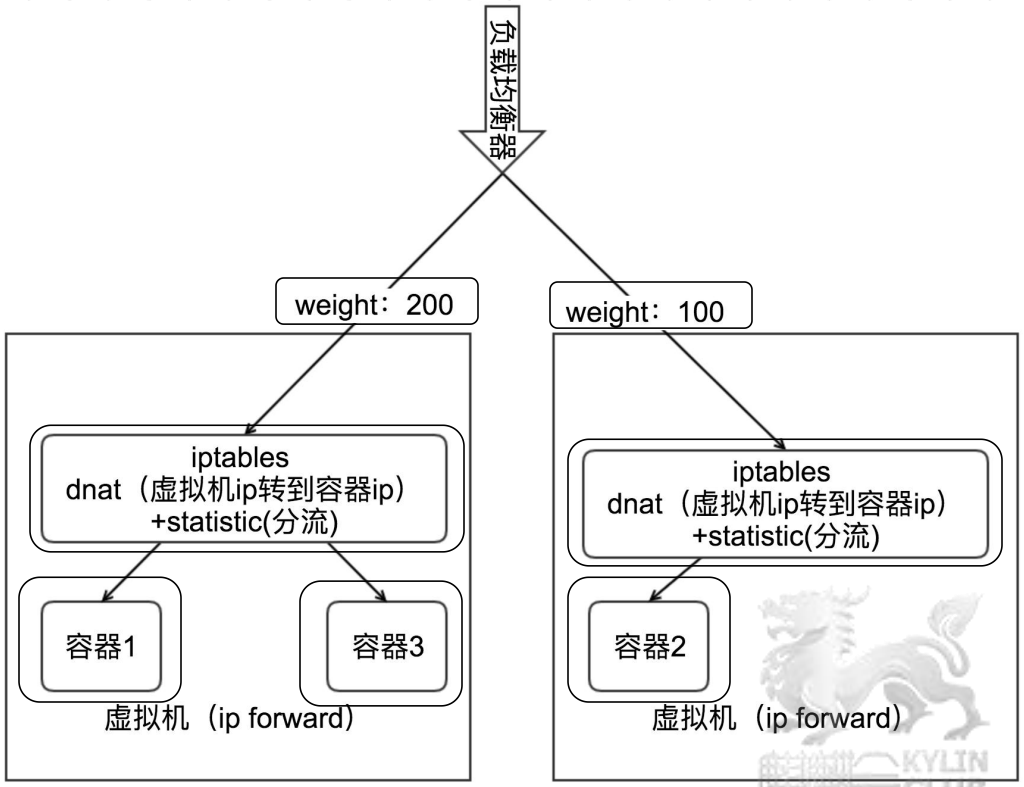




5 阿里云Flannel落地

外网流量入容器

- 前提：
  - 负载均衡器后端只能是虚拟机
  - 负载均衡器支持虚拟机获取客户端ip
- 实现
  - 使用iptables（dnat、statistic）转发负载均衡器流量到容器
- 优点：
  - 支持容器获取客户端ip
  - Iptaes转发效率高于haproxy等软件



5 阿里云Flannel落地

容器流量出外网

- 前提：
  - Flannel支持容器访问外网，但容器和非容器服务器通信时，容器流量也使用ip伪装
- 实现
  - 使用iptables snat单独实现容器访问外网

```
1 manager.go:166] Defaulting external address to interface address (192.168.0.234)
1 ipmasq.go:47] Adding iptables rule: -s 172.16.0.0/16 -d 172.16.0.0/16 -j RETURN
1 ipmasq.go:47] Adding iptables rule: -s 172.16.0.0/16 ! -d 224.0.0.0/4 -j MASQUERADE
1 ipmasq.go:47] Adding iptables rule: ! -s 172.16.0.0/16 -d 172.16.0.0/16 -j MASQUERADE
1 manager.go:250] Lease acquired: 172.16.0.0/24
```

```
Chain POSTROUTING (policy ACCEPT)
target     prot opt source               destination
RETURN     all  --  [redacted]/16        10.0.0.0/8
RETURN     all  --  [redacted]/16        172.16.0.0/12
RETURN     all  --  [redacted]/16        192.168.0.0/16
RETURN     all  --  [redacted]/16        224.0.0.0/4
MASQUERADE  all  --  [redacted]/16        0.0.0.0/0
```

6 未来计划

Ipvlan

- 使用ipvlan替换bridge，提上容器网络效率

CNM

- 开发cnm插件，替换dhcp分配ip，更有效管理ip资源

FAQ