openshift私有云网络实战

# flanel host-gw

<https://bugzilla.redhat.com/show_bug.cgi?id=1564955>

参见

<https://mp.weixin.qq.com/s/2on8yDlklbsTtNwD1prCDg>

做容器或者说弹性调度平台，网络是一个避不开的话题，小米在做弹性调度的时候网络有以下几方面的考虑：

1. 要有独立、真实的内网IP，便于识别和定位，无缝对接现有的基础设施；
2. 要与现有的物理机网络打通；
3. 要能保证最小化的网络性能损耗（这一点基本上使我们放弃了overlay的网络方式）。

下面简单介绍下Ocean在Flannel+hostgw上的实践。

1. Ocean和网络组协商，规划了一个Ocean专用的大网段；
2. 网络组同学为Ocean平台提供了动态路由添加、删除的接口，即提供了路由、三层交换简单OpenAPI能力；
3. Ocean平台规范每台宿主机的网段（主要是根据宿主机配置，看一台宿主机上启动多少实例，根据这个规划子网掩码位数）；
4. 每台容器宿主机上启动Flanneld，Flanneld从etcd拿宿主机的子网网段信息，并调用网络组提供的动态路由接口添加路由信息（下线宿主机删除路由信息）；
5. Dockerd用Flanneld拿到的网段信息启动Docker daemon。
6. 容器启动是根据bip自动分配IP。

这样容器的每个IP就分配好了。容器的入网和出网流量，都依赖于宿主机的主机路由，所以没有overlay额外封包解包的相关网络消耗，只有docker0网桥层的转发损耗，再可接受范围内。