k8s 中 CNI(容器网络接口) 解析

作者：Leif160519
原文链接：https://ld246.com/article/1591602509491
来源网站：链滴

概念

CNI(Container Network Interface)容器网络接口：是Kubernetes提出的一个标准，解决了跨主机络通信的问题。使用CNI的约束：

1.一个pod分配一个唯一的IP， 这个IP是整个集群的唯一IP，这是保障跨主机通讯的前提，一个pod个IP

2.node可以访问任何节点pod，不限于本节点，pod之间也可以相互访问

3.pod可以访问所有的pod

CNI网络插件实现：路由方案、隧道方案

flannel

vxlan（隧道）、host-gw（路由）、udp(弃用)。默认使用隧道模式，将数据包二次封装走宿主机的层网络，然后到达目的网络

```
net-conf.json: |
  {
    "Network": "10.244.0.0/16",
    "Backend": {
      "Type": "vxlan"
    }
  }
```

caclio

也同样支持隧道和路由方案：ipip（隧道）、bgp（路由）。默认也使用隧道模式

```
            - name: IP
              value: "autodetect"
            # Enable IPIP
            - name: CALICO_IPV4POOL_IPIP
              value: "Always"
            # Enable or Disable VXLAN on the default IP pool.
            - name: CALICO_IPV4POOL_VXLAN
              value: "Never"
            # Set MTU for tunnel device used if ipip is enabled
            - name: FELIX_IPINIPMTU
              valueFrom:
```

模式选择：

取决于当前网络现状，如果在公有云上，那很多公有云会对路由模式有所限制，也就是说如果你用路模式的话，会在每台k8s node机器上写入相应的路由表，而有些云主机是不支持的，会导致云主机无通信，影响现有的网络，故选择路由模式不可取。

路由方案：对现有网络有要求，性能最好，一般要求二层可达，组件大二层网络

隧道方案：只要三层可达基本都可以通信，基于现有以太网

vxlan模式介绍：

会在每个节点上创建一个cni0网桥，还会创建一个lannel.1的隧道端点，主要是对数据包的再次封装之后传输到目标节点上，也会在每台机器上创建一些路由表

```
[root@k8s-master ~]# ip route
default via 192.168.31.1 dev ens33 proto static metric 100
10.244.0.0/24 dev cni0 proto kernel scope link src 10.244.0.1
10.244.2.0/24 via 10.244.2.0 dev flannel.1 onlink
10.244.3.0/24 via 10.244.3.0 dev flannel.1 onlink
172.17.0.0/16 dev docker0 proto kernel scope link src 172.17.0.1
192.168.31.0/24 dev ens33 proto kernel scope link src 192.168.31.61 metric 100
```

路由表解释：若数据包的目的IP匹配到其中一个IP段，则走改条路由表规则，将数据包到达flannel.1里

flannel和caclio网络是相互冲突的，在一个k8s集群中一般只有一个cni网络组建，所以下面讲解如何将lannel切换成caclio

1.删除flannel-pod

kubectl delete -f kube-flannel.yaml

上述删除只是将flannel的守护进程删除了而已，但是网桥和隧道端点并没有删除，在删除pod的时候pod之间的网络就已经不通了（若是路由方案的话，还是可以通的）

2.删除cni0网桥和flannel.1隧道端点

ip link del cni0
ip link del flannel.1

使用以下命令查看网桥，隧道端点和路由表是否删除干净，

ifconfig
ip route

确保网桥，隧道端点和路由表清除干净之后再部署caclio

3.修改caclio网段与当前集群网段一致

下载caclio.yaml

wget https://docs.projectcalico.org/manifests/calico.yaml

从前面文章可以得出，集群网段为10.244.0.0/16，所以calico.yaml中的网段也需要改成当前网段：
取消前面的注解并将IP网段从192.168改成10.244

```
            # The default IPv4 pool to create on startup if none exists. Pod IPs will be
            # chosen from this range. Changing this value after installation will have
            # no effect. This should fall within `--cluster-cidr`.
            - name: CALICO_IPV4POOL_CIDR
              value: "10.244.0.0/16"
            # Disable file logging so `kubectl logs` works.
            - name: CALICO_DISABLE_FILE_LOGGING
              value: "true"
            # Set Felix endpoint to host default action to ACCEPT.
```

在不指定caclio工作模式的情况下，默认都是隧道方案，此方案对现有网络的依赖是最小的

```
            - name: IP
              value: "autodetect"
            # Enable IPIP
            - name: CALICO_IPV4POOL_IPIP
              value: "Always"
            # Enable or Disable VXLAN on the default IP pool.
```

若将Always改成Never，则使用BGP

保存后执行：

kubectl apply -f caclio.yaml

```
[root@k8s-master k8s]# kubectl get pod -n kube-system
NAME                                       READY   STATUS    RESTARTS   AGE
calico-kube-controllers-76d4774d89-pcdrn   1/1     Running   6          7d6h
calico-node-472bs                          1/1     Running   6          7d6h
calico-node-54jjl                          1/1     Running   6          7d6h
calico-node-hpv7x                          1/1     Running   6          7d6h
coredns-7ff77c879f-cgfjw                   1/1     Running   6          7d7h
coredns-7ff77c879f-pn8qk                   1/1     Running   7          7d7h
etcd-k8s-master                            1/1     Running   6          7d7h
kube-apiserver-k8s-master                  1/1     Running   11         7d7h
kube-controller-manager-k8s-master         1/1     Running   6          7d7h
kube-proxy-grnpw                           1/1     Running   6          7d7h
kube-proxy-mshjk                           1/1     Running   6          7d6h
kube-proxy-nkkk4                           1/1     Running   6          7d6h
kube-scheduler-k8s-master                  1/1     Running   6          7d7h
```

使用场景考虑方面

- 1.集群规模
- 2.是否需要网络策略（flannel不支持）
- 3.现有网络有无限制，包括主机写路由表，bgp是否可以通信（有限制就用隧道方案）
- 4.维护成本

flannel：适合小规模集群，维护成本低。集群规模小于100台，可以使用flannel的host-gateway

calico：以上相反

注意

当网络模式切换过后，所有的pod都需要重新构建才能使用新的网络，故在生产环境切换网络模式代成本很高，所以需要运维人员在知道产生的后果之后再进行操作。