

苏宁容器云网络架构技术实现

Kubernetes + Contiv





王晓飞
苏宁云数据公司·容器网络架构师



杨川
苏宁云数据公司·高级研发工程师

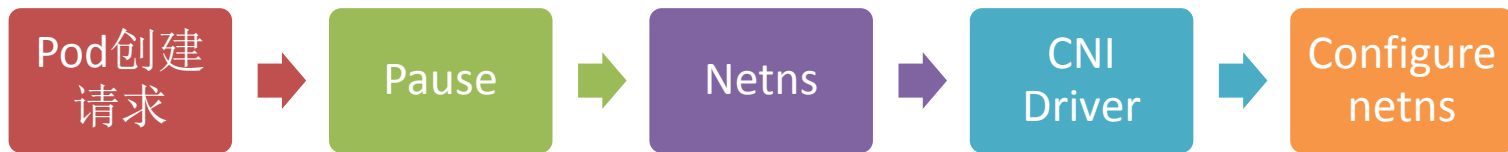
Agenda

- Kubernetes CNI 规范简介
- Kubernetes CNI 插件对比
- Contiv 网络架构
- 苏宁容器云 Contiv 与 Kubernetes 部署架构
- 苏宁容器云 Pod-IP 固定实现网络原理
- 苏宁容器云基于 Contiv 实现双向限速
- 苏宁容器云基于 Contiv 支持 etcd v3
- 苏宁容器云基于 Contiv 支持 SR-IOV
- Q&A



KUBERNETES CNI 介绍

- CNI 简介
- CNI 组成
 - CNI Plugin负责为容器配置网络，包括两个基本接口：
 - 1) 配置网络: `AddNetwork(net NetworkConfig, rt RuntimeConf) (types.Result, error)`
 - 2) 清理网络: `DelNetwork(net NetworkConfig, rt RuntimeConf) error`
 - IPAM Plugin负责给容器分配IP地址，主要实现包括host-local和dhcp
- Kubernetes Pod 中的其他容器都是Pod所属Pause下的容器网络，创建流程为：





kubernetes



CONTIV对比

方案	优点	缺点	结论
Calico	纯三层的数据中心解决方案（不依赖overlay） 对 openstack 、 kubernetes 、 aws 、 GCE都比较友好。该方案集成简单	操作管理起来较复杂。需要定制开发	calico 基于BGP协议的 overlay sdn 的解决方案，企业生产环境如果可以开启BGP协议，可以考虑calico bgp方案。
Contiv	能够和非容器环境兼容协作，不依赖物理网络具体细节，支持Policy/ACI/QoS/租户，支持物理网卡sriov和offload，该方案配置比较复杂。	集成配置较复杂，相关资料少，学习成本较高，需要定制开发。	功能强性能好，对CNI和CNM模型支持性好 集群配置较复杂，学习研究成本较高，可以考虑此方案
Macvlan	docker 原生，性能衰减少，可控性高，隔离性好。	集群规模很大时，存在arp广播风暴和交换机MAC表超限风险，另外需要自研QoS功能。 缺少IPAM、DNS、QoS、租户等概念。	初期较少的开发量即可上线（之后需要自研QoS），通过合理的设计容量规划来规避相关风险时，可以考虑此方案。

Contiv Macvlan性能对比数据

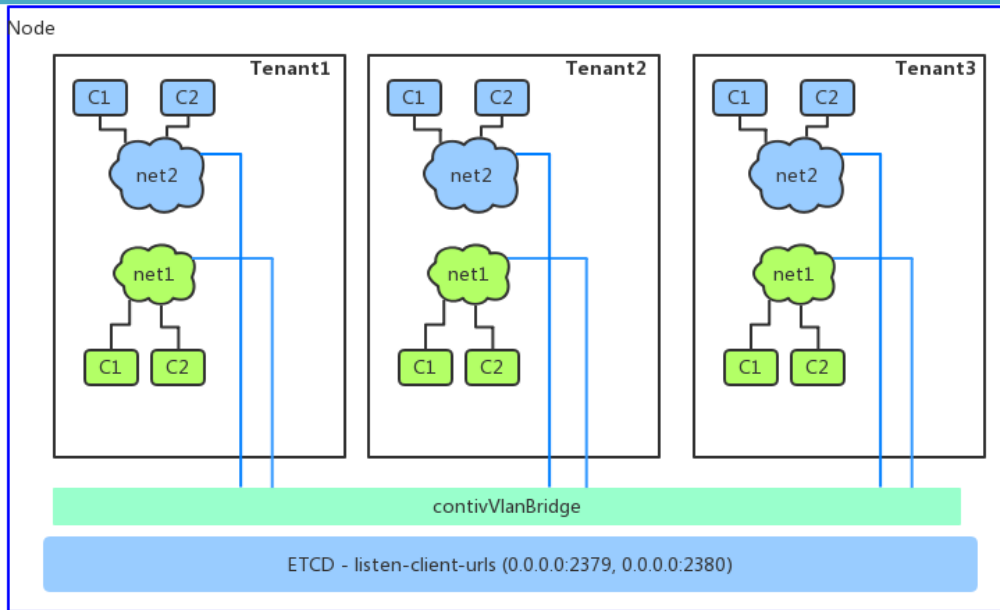
对比指标	Contiv(ovs)	Macvlan
总PPS(64Bytes)	1747459	2161710
平均PPS(64Bytes)	218432	270213
最大吞吐量(1500Bytes)	9.58Gbit/s	9.71Gbit/s

物理带宽：10Gbit/s

CONTIV 网络架构

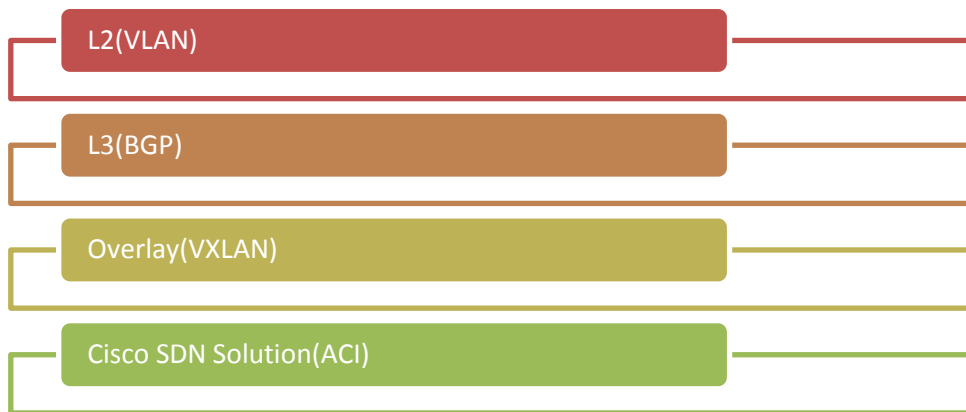
Contiv是思科开源容器网络解决方案，同时支持CNI模型以及CNM模型。支持以下功能：

支持Tenant，支持租户隔离



CONTIV 网络架构

支持多种网络模式



CONTIV 网络架构

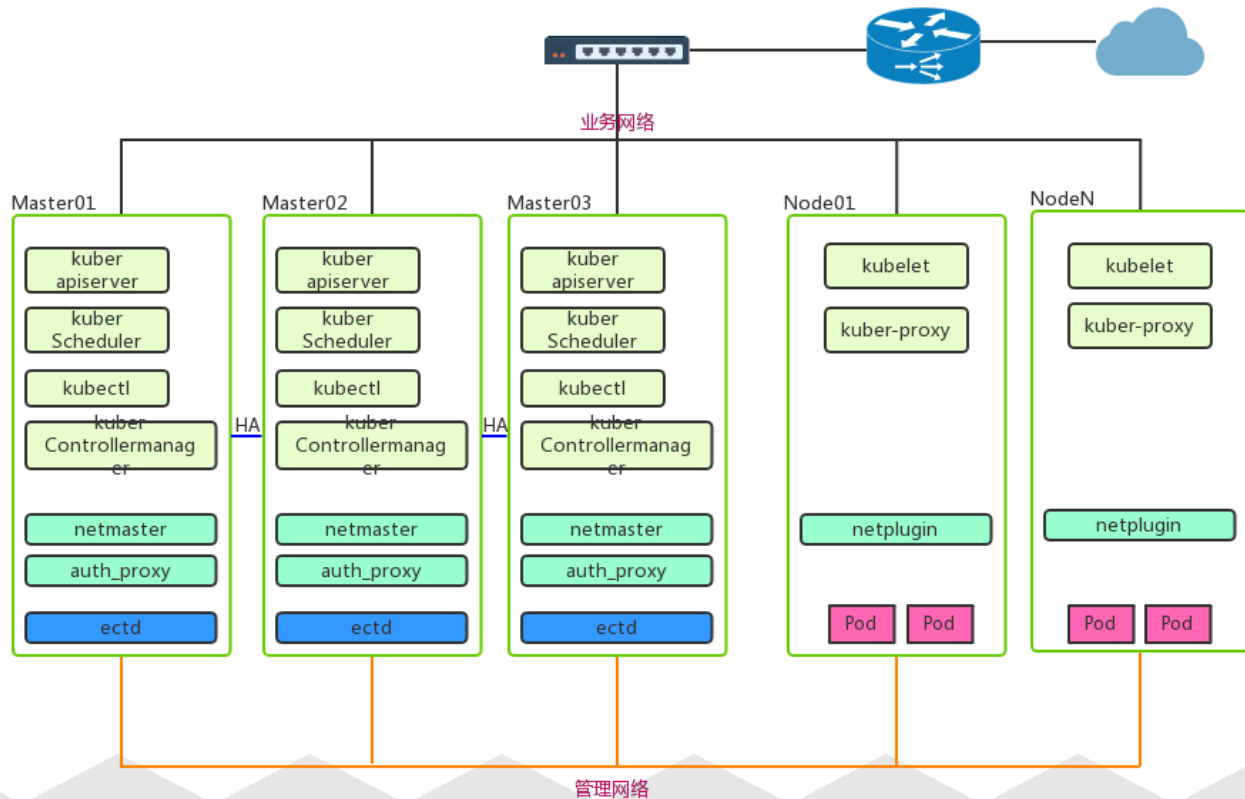
Network Policy

Network policy 描述了一组容器（Pod）上的网络资源使用、隔离策略、优先级等等。

在Contiv中有两种类型的Network policies:

- Bandwidth
- Isolation

CONTIV 与 KUBERNETES 部署架构

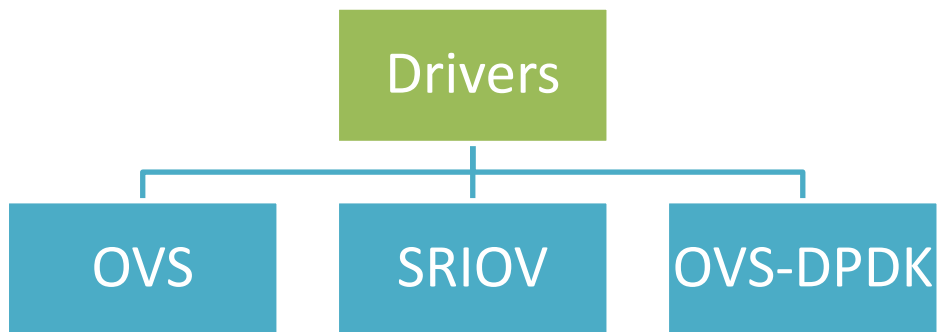


- Kubernetes集群
- Contiv集群
- Etcd集群
- 租户Pod

- 使用的Bridge VLAN模式

```
kind: ConfigMap
apiVersion: v1
metadata:
  name: contiv-config
  namespace: kube-system
data:
  contiv_mode: kubernetes
  contiv_fwdbmode: bridge
  contiv_netmode: vlan
```

- 实现多网络Driver支持



苏宁容器云POD-IP固定实现原理

实现Pod-IP固定的需求

- 内部防火墙系统需要依据IP地址进行白名单或者黑名单设置
- 内部RSF系统需要容器具有固定的IP地址
- 利于运维或者研发调试修正问题



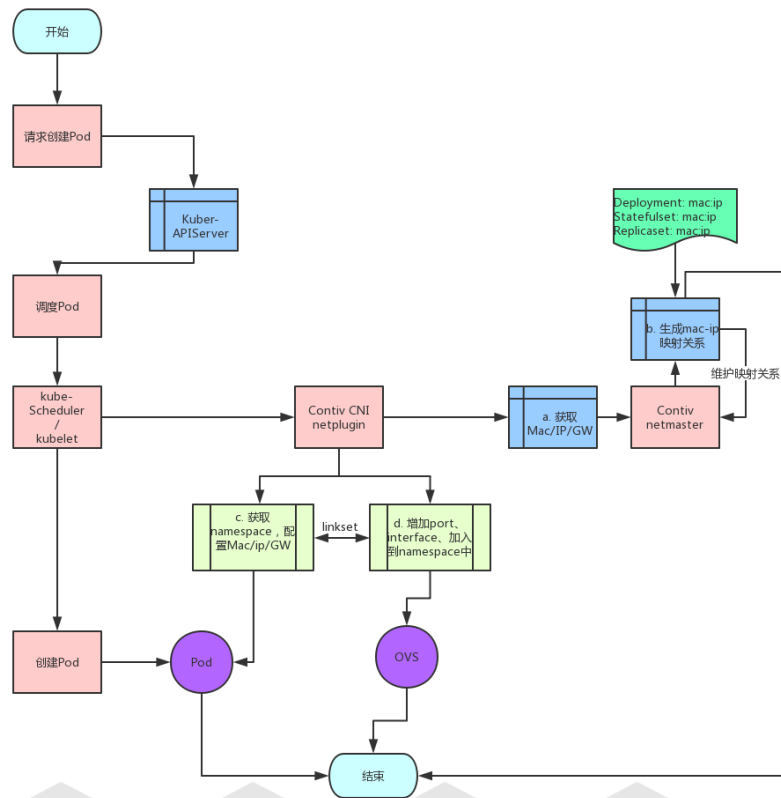
对Kubernetes支持的场景

- Deployment
- Stateful set
- Replica set

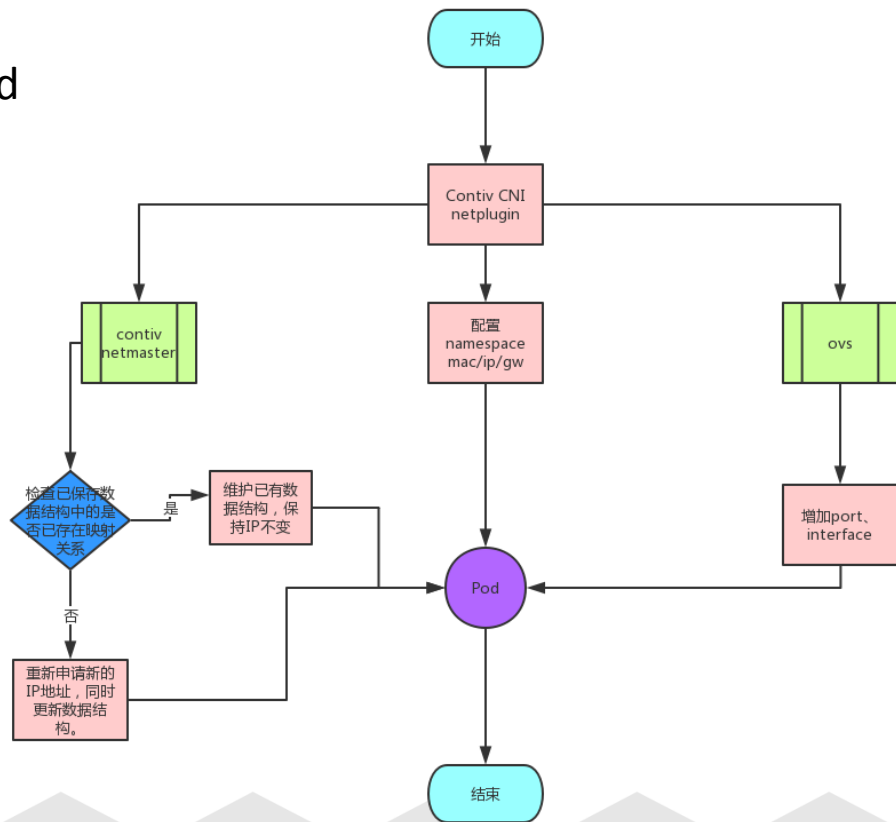
对Kubernetes不支持的场景

- DaemonSet

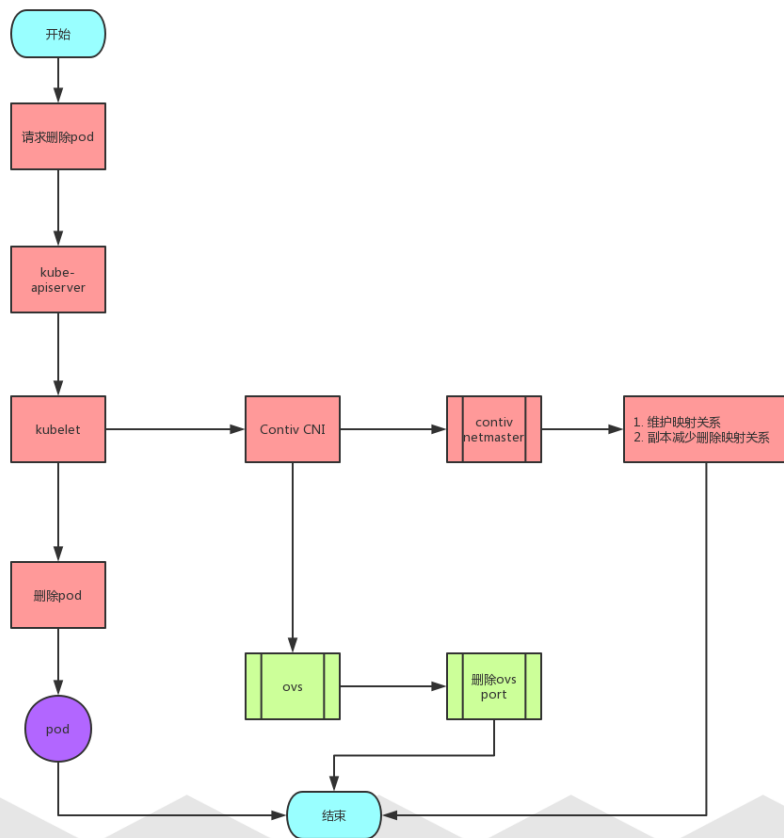
创建Pod



重新调度Pod



删除Pod



苏宁容器云基于CONTIV 双向限速实现

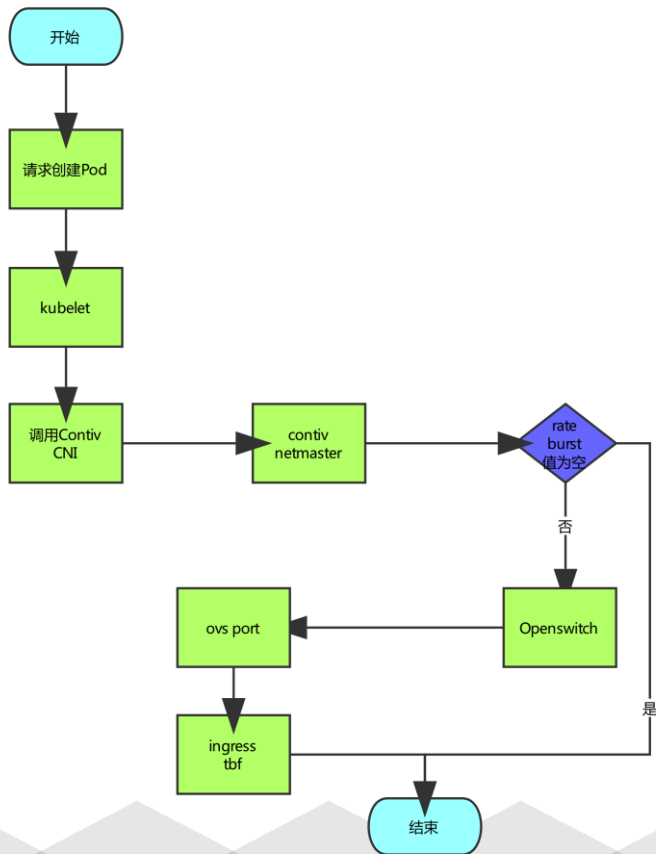
Pod 双向限速实现原理

上行

- 使用ovs中ingress方式
- 设置rate 和 burst
- 超过设置速率限制，直接简单丢弃

下行

- 使用netlink包，通过unix socket和内核通信
- 使用tbf限速方式，直接将规则作用在ovs port上



- 新建资源时，labels增加io.contiv.net-group: groupb
- 检测app group中，是否关联Bandwidth Policies
- 如果关联Bandwidth Policies，则在创建port时，增加限速规则。

苏宁容器云基于CONTIV 支持 ETCD v3

Contiv 支持多种分布式存储，consul、etcd等

Contiv默认仅支持Etcdv2 client

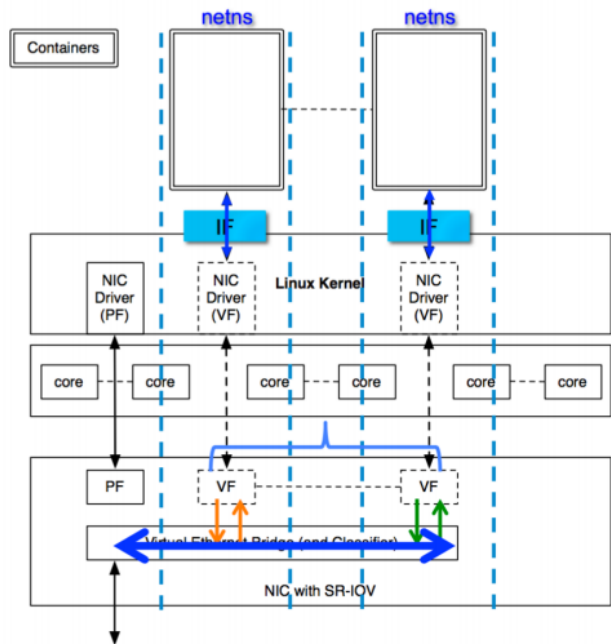
预留fakestatedriver.go,支持扩展其它client

保持方法名称不变，使用clientv3重写clientv2方法。

```
▼ state
  ► cfgtool
    /* consulstatedriver.go
    /* consulstatedriver_test.go
    /* etcdstatedriverv3.go
    /* etcdstatedriverv3_test.go
    /* fakestatedriver.go
```

苏宁容器云基于 CONTIV 支持 SR-IOV

- SR-IOV技术是基于硬件的虚拟化解决方案
- SR-IOV标准允许虚拟化高效共享PCIe设备
- 可以获取与本机媲美的性能



Contiv支持SR-IOV要从代码级别做改造:

1. 增加sriovd驱动目录
2. 仿造fakenetepdriver设计sriovd驱动
3. 需要重新修改Contivinit, 以便在初始化netplugin时, 对sr-iov进行初始化。
4. 一个netplugin仅支持一种驱动, 但是同一集群内可以有多种网络驱动。

```
— driverconstants.go
— endpointstate.go
— endpointstate_test.go
— fakenetepdriver.go
— ovsd
  |— nodeProxy.go
  |— nodeProxy_test.go
  |— ovsconstants.go
  |— ovsdbDriver.go
  |— ovsdriver.go
  |— ovsdriver_test.go
  |— ovsSwitch.go
— sriovd
  |— sriovconstants.go
  |— sriovdriver.go
  |— sriovdriver_test.go
  |— sriovns.go
  |— sriovresource.go
— vppd
  |— vppconstants.go
  |— vppdriver.go
```



苏宁数据云公司

Thank You