阿里云 | 奥运会全球指定云服务商

阿里云如何构建高性能云原生容器网络?

王炳燊(溪恒)-阿里云容器服务

关注“阿里巴巴云原生”公众号
获取第一手技术资料

目录

阿里云 | 奥运会全球指定云服务商

CONTENT

目录

Kubernetes容器网络概述

阿里云 | 奥运会全球指定云服务商

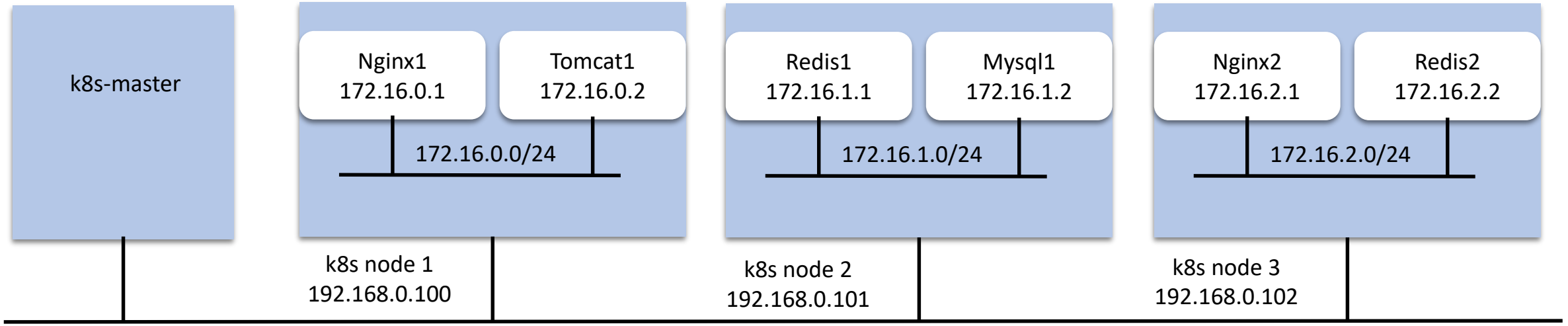
Pod网络连通性(CNI)
Kubernetes 负载均衡(Service)
Kubernetes 服务发现(Coredns)

Kubernetes容器网络概述

阿里云 | 奥运会全球指定云服务商

Pod网络连通性(CNI)

- Pod有自己独立的网络空间和IP地址
  - 不同Pod的应用可以监听同样的端口而不冲突
- Pod可以通过各自的IP地址互相访问
  - 集群中Pod可以通过它独立的IP地址访问其他网络
    - Pod和Pod的联通
    - Pod与Node的联通
    - Pod与外部网络连通
- 具体的地址分配和网络打通由网络插件(CNI)来负责



Kubernetes容器网络概述

阿里云 | 奥运会全球指定云服务商

Pod网络连通性(CNI)

- **Container Network Interface**，容器网络的**API**接口
- **CNI**插件：一系列实现了**CNI API**接口的网络插件
  - 配置**Pod**网络空间
  - 打通**Pod**之间网络连通
- **Kubelet**通过这个标准的**API**调用不同的网络插件实现配置网络

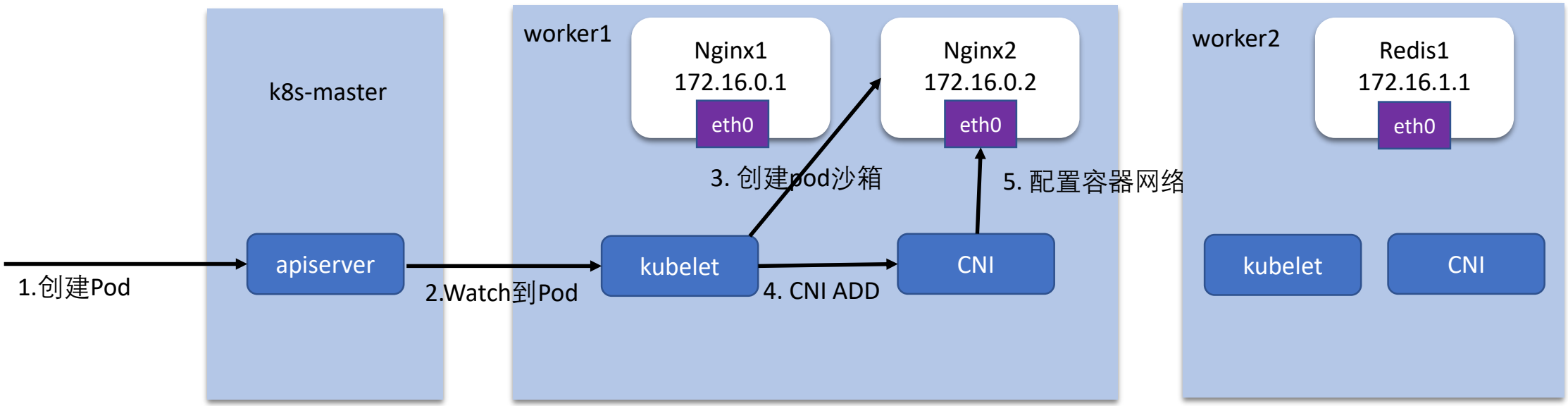
Calico

flannel

Terway

Weave Net

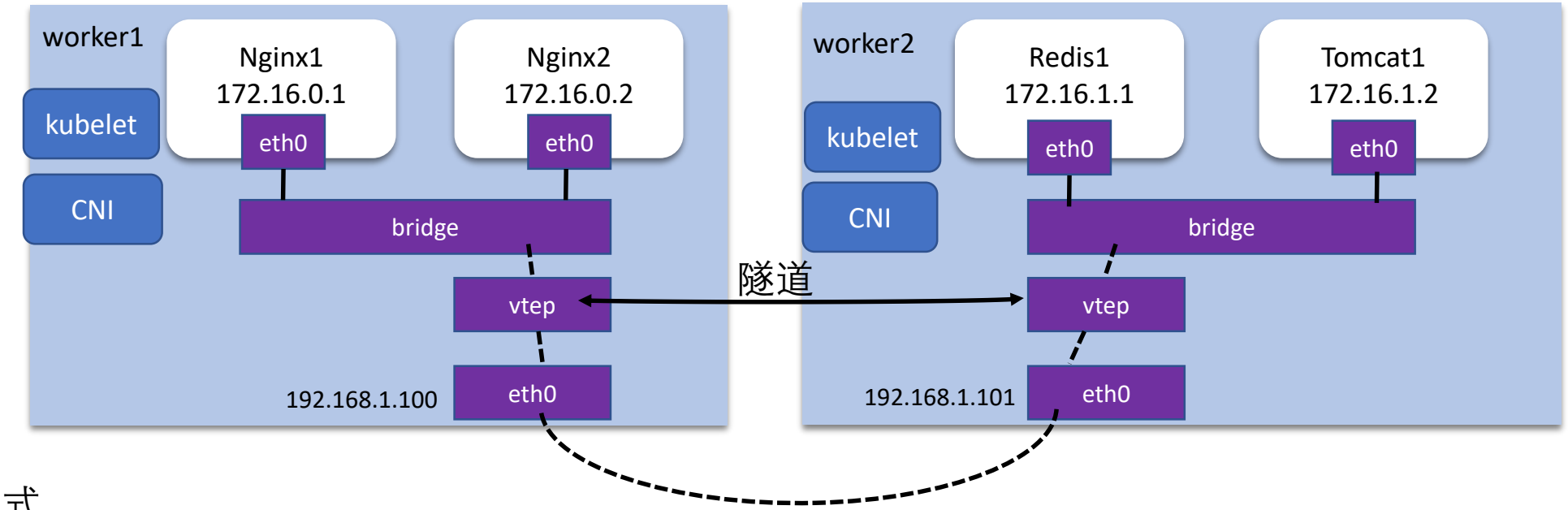
Contiv



Kubernetes容器网络概述

Pod网络连通性(CNI)

- 容器的地址跟节点不在一个平面，如何在节点之上实现容器网络通信



封包方式

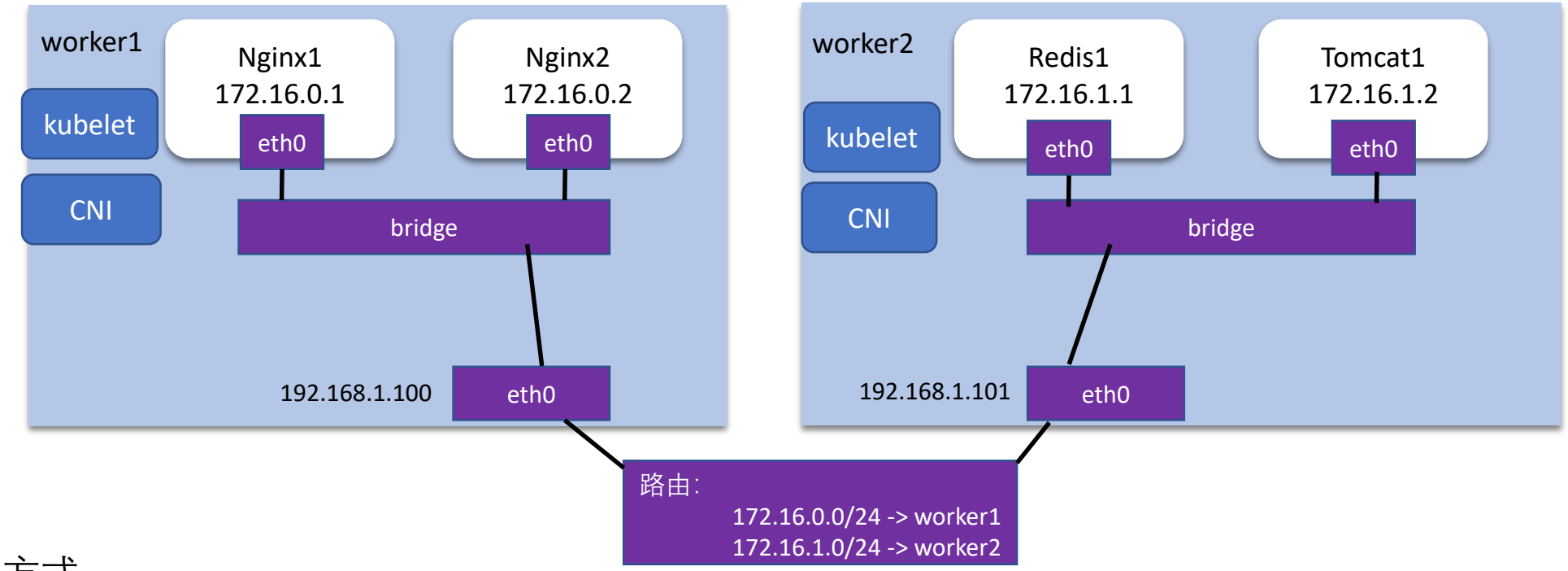
- 容器之间的通信报文封装成宿主机之间通信的报文

Kubernetes容器网络概述

阿里云 | 奥运会全球指定云服务商

Pod网络连通性(CNI)

- 容器的地址跟节点不在一个平面，如何在节点之上实现容器网络通信



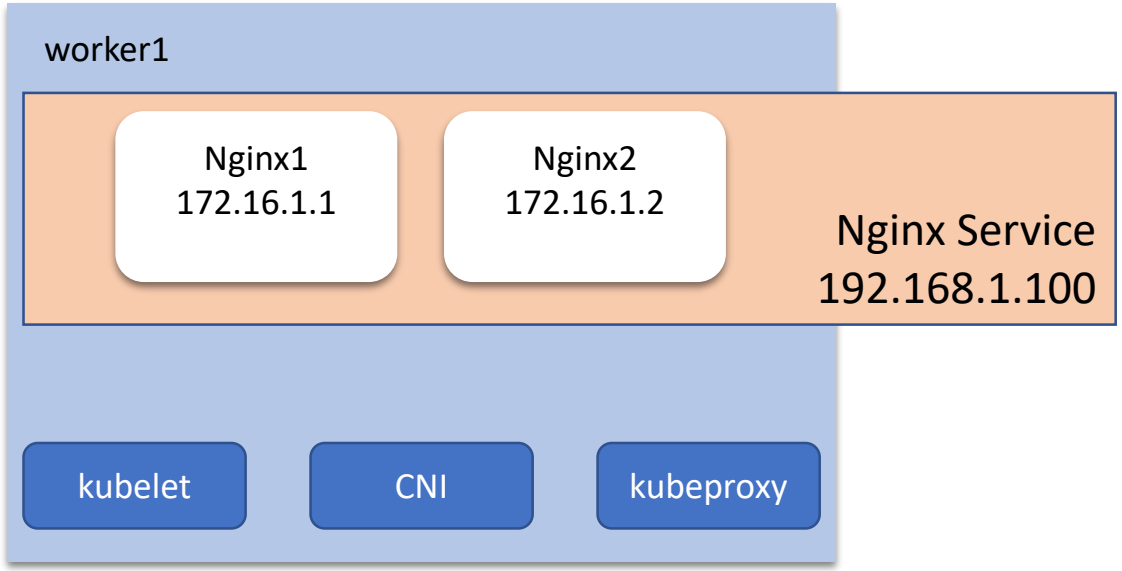
路由方式

- 容器之间的通信报文由路由表转发到对应的节点上

Kubernetes容器网络概述

Kubernetes网络负载均衡 Service

- **Pod**生命周期短暂，**IP**地址随时变化，需要固定的访问方式
- **Deployment**等的一组**Pod**组需要统一访问入口和做负载均衡



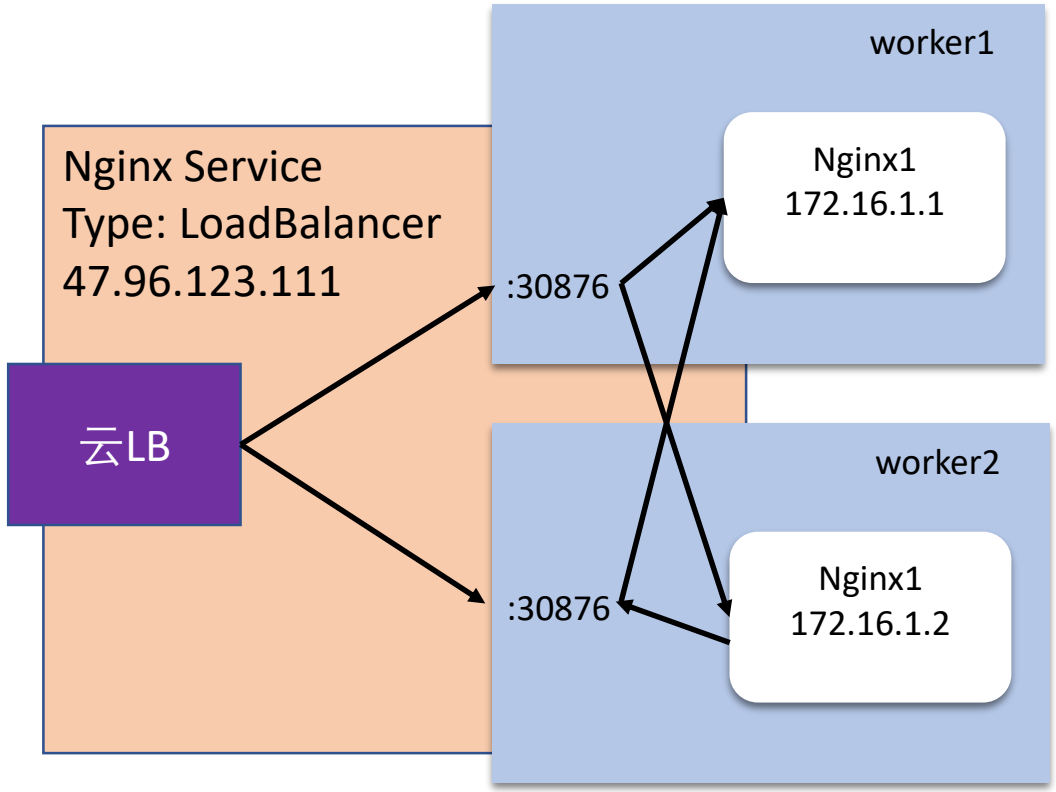

```
$ cat nginx-svc.yaml
apiVersion: v1
kind: Service
metadata:
  name: nginx
spec:
  selector:
    app: nginx
  ports:
  - protocol: TCP
    port: 80

$ kubectl get service nginx
CLUSTER-IP PORT(S)
192.168.1.100 80/TCP

$ kubectl get endpoints nginx
NAME        ENDPOINTS
nginx       172.16.1.1:80,172.16.1.2:80
```

Kubernetes容器网络概述

阿里云 | 奥运会全球指定云服务商

Kubernetes网络负载均衡 Service -- LoadBalancer

- 外部也需要固定的访问Pod的入口
- LoadBalancer类型的Service创建LB服务来暴露集群服务




```
$ cat nginx-svc.yaml
apiVersion: v1
kind: Service
metadata:
  name: nginx
spec:
  selector:
    app: nginx
  ports:
  - protocol: TCP
    port: 80
  type: LoadBalancer

$ kubectl get service nginx
EXTERNAL-IP    PORT(S)
47.96.123.111 80:30867/TCP

$ kubectl get endpoints nginx
NAME        ENDPOINTS
nginx       172.16.1.1:80,172.16.1.2:80
```

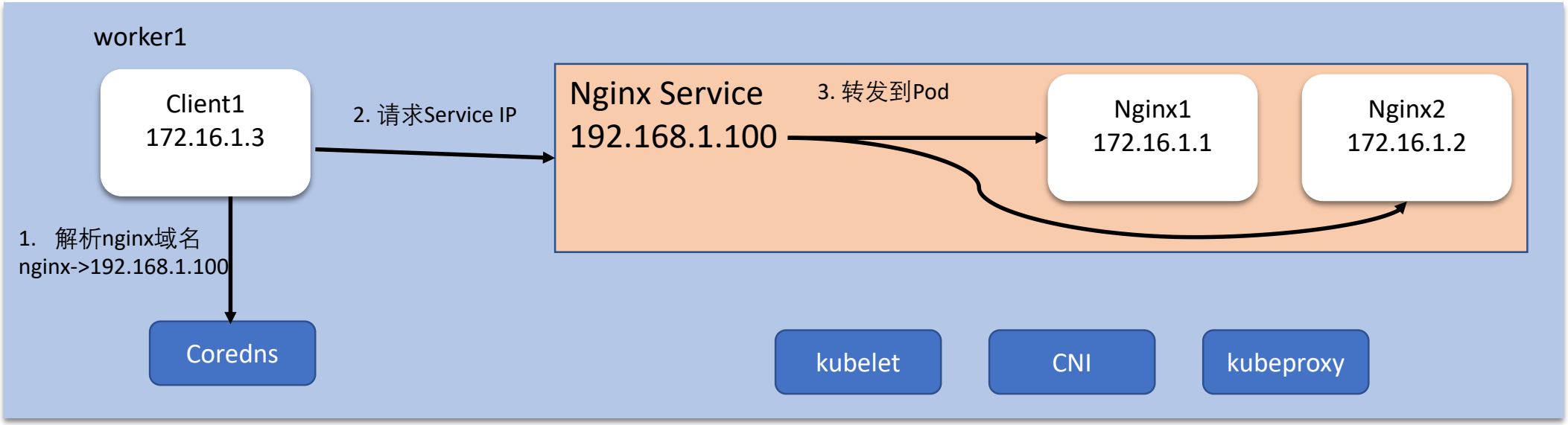
Kubernetes容器网络概述

阿里云 | 奥运会全球指定云服务商

Kubernetes网络服务发现

虽然Service固定了IP访问方式，但Service的IP在不同的namespace或者集群中是不同的

集群的Coredns会将Service名自动转换成对应的Service的IP地址，来实现不同部署环境中同样的访问入口



目录

阿里云 | 奥运会全球指定云服务商

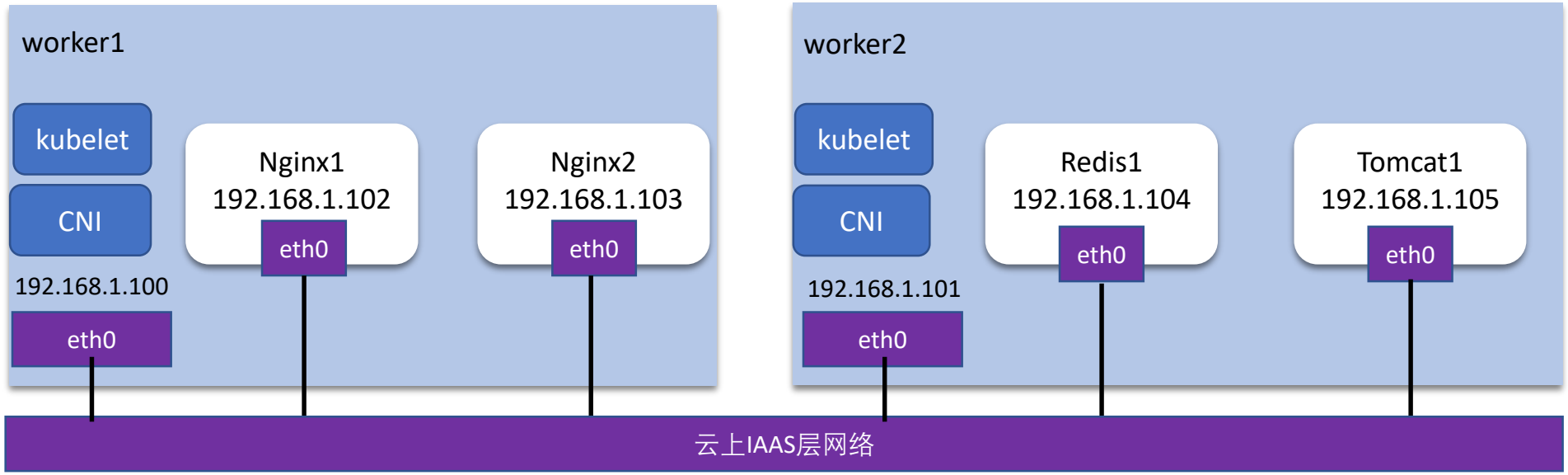
CONTENT

高性能云原生的容器网络

阿里云 | 奥运会全球指定云服务商

什么是云原生容器网络?

- 云上IAAS层网络虚拟化，在容器中再做一层网络虚拟化性能损失大
- 云原生容器网络直接使用云上原生云资源配置容器网络
  - 容器和节点同等网络平面，同等网络地位
  - POD网络可以和云产品无缝整合
  - 不需要封包和路由，网络性能和虚机几乎一致

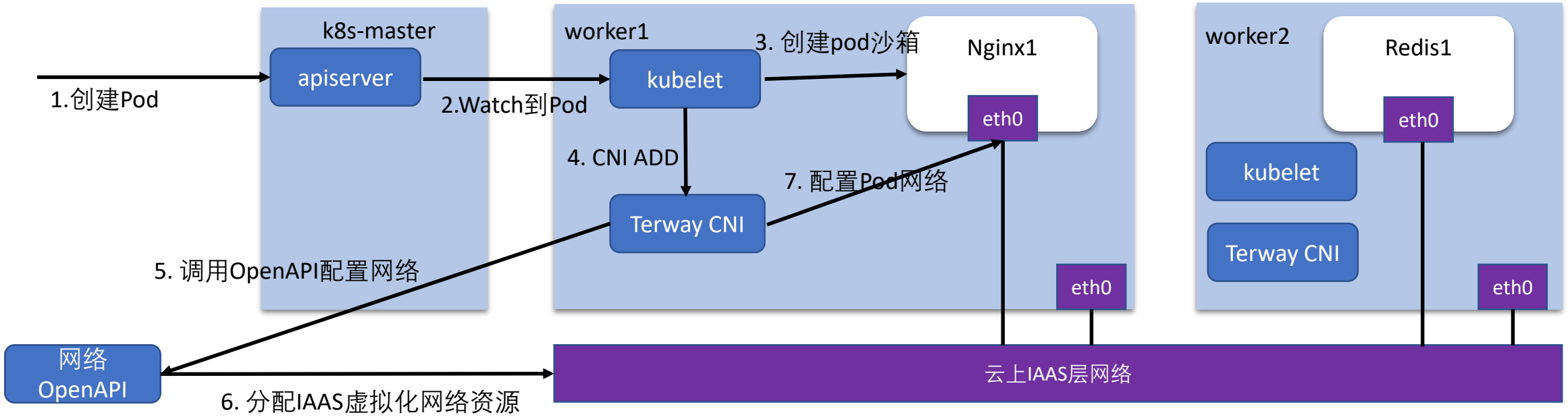


高性能云原生的容器网络

云原生容器网络原理

- 在CNI调用中调用云网络OpenAPI完成网络资源分配
  - 网络资源一般是弹性网卡，弹性网卡辅助IP等，绑定到Pod所在的节点上
- 网络资源分配出来之后CNI插件在宿主机中将资源配置到Pod的沙箱中

下图以阿里云上的Terway CNI插件为例：



高性能云原生的容器网络

使用云原生网络资源的优势

- Pod和虚拟机同一层网络，便于业务云原生化迁移
- 不依赖封包或者路由表，分配给Pod的网络设备本身可以用来通信
- 集群节点规模不受路由表或者封包的FDB转发表等Quota限制
- 不需要额外为Pod规划Overlay的网段
  - 多个集群Pod之间只要安全组放开就可以互相通信
- 可以直接把Pod挂到LoadBalancer后端，无需节点上端口再做一层转发
- NAT网关可以对Pod做SNAT，无需节点上对容器网段做SNAT
  - Pod访问VPC内资源，所带的源IP都是PodIP，便于审计
  - Pod访问外部网络不依赖conntrack SNAT，失败率降低

高性能云原生的容器网络

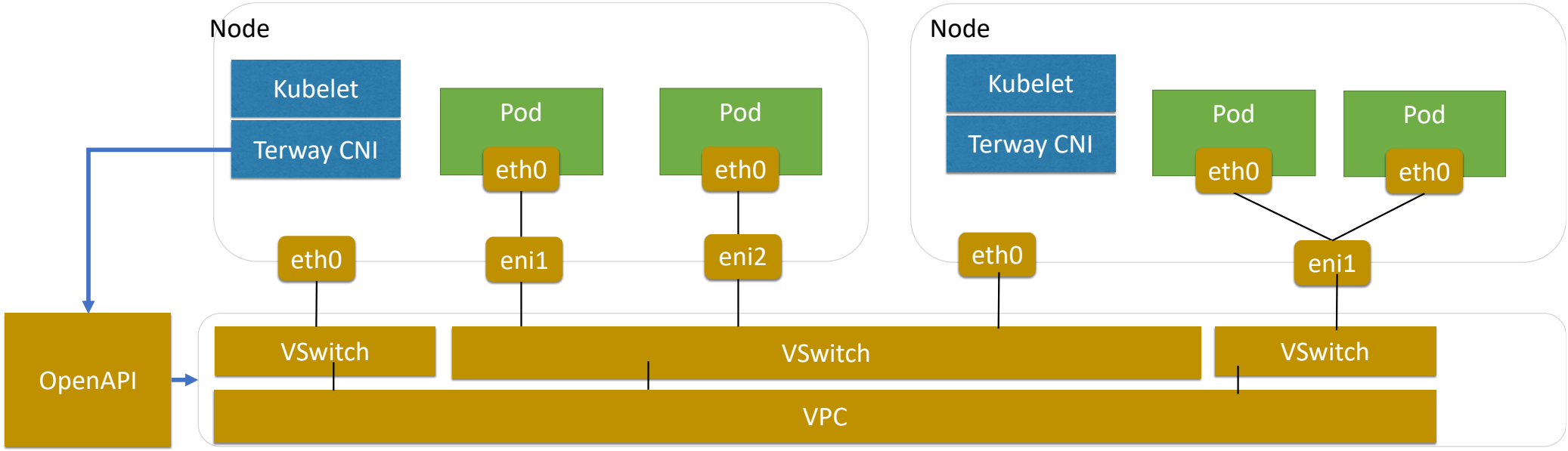
阿里云 | 奥运会全球指定云服务商

如何利用云原生资源构建容器网络

IAAS层网络资源（以阿里云为例）：

- 弹性网卡（ENI）
  IAAS层虚拟化出来的虚拟网卡，可动态分配和绑定到虚拟机上
  一般能绑定数量有限，受限于PCI-E的限制
- 弹性网卡辅助IP
  弹性网卡上通常可以绑定多个VPC的IP地址，为辅助IP
  一个弹性网卡可以绑定数十个辅助IP地址，限制较小

利用弹性网卡或者弹性网卡辅助IP分配给Pod来实现云原生容器网络



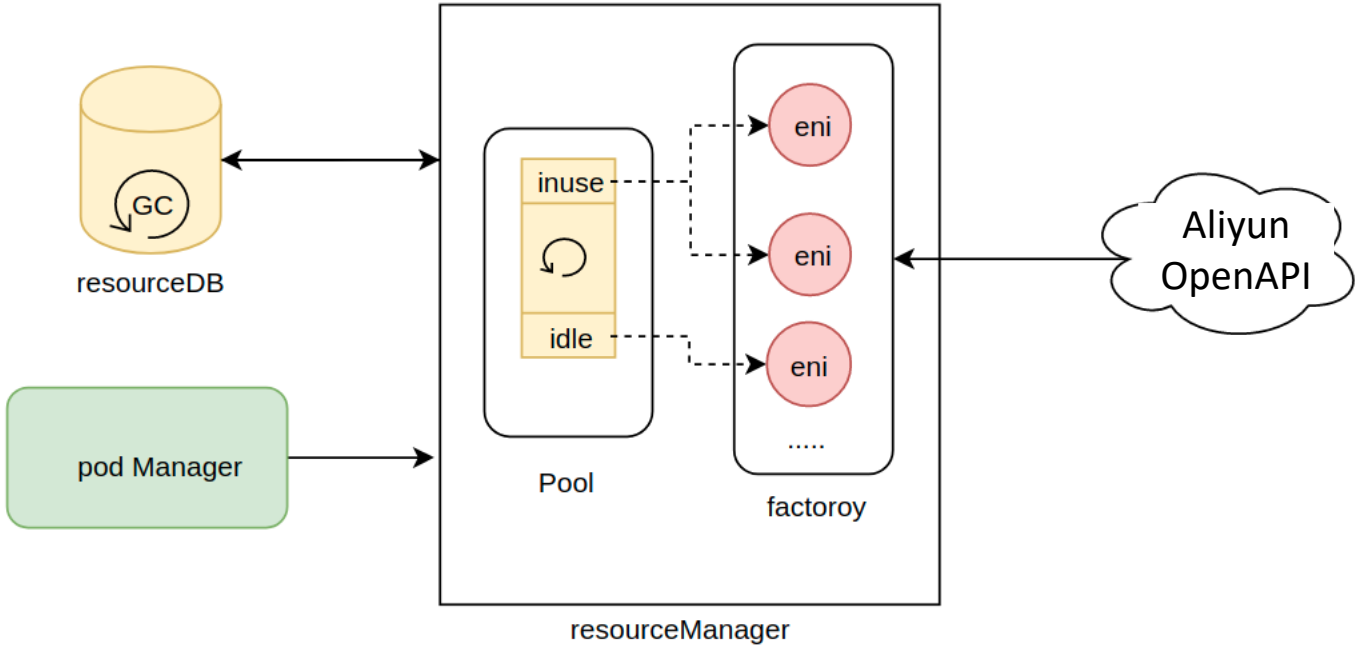
高性能云原生的容器网络

如何利用云原生资源构建容器网络 --网络资源管理

- 容器的启动速度是秒级的， 而IAAS层操作和调用一般在10s的级别

Terway通过内置资源池来缓冲资源加速启动：

- 资源池中记录分配的正在使用和空闲的资源
- Pod释放后资源会保留在资源池中供下次快速启动
- 资源池有最低水位和最高水位
  - 空闲资源低于最低水位调用API补充资源
  - 空闲资源高于最高水位调用API释放资源



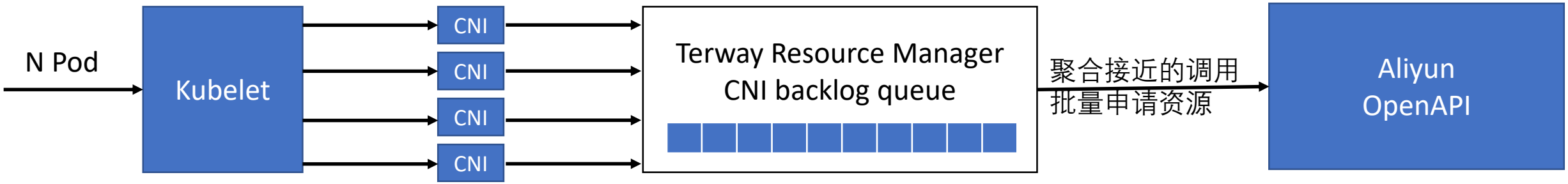
高性能云原生的容器网络

阿里云 | 奥运会全球指定云服务商

如何利用云原生资源构建容器网络 -- 网络资源管理

- 容器扩缩容操作频繁，云产品OpenAPI通常有严格的调用限流

1. Terway通过缓冲池，小于水位自动补充，来预热资源减缓峰值API调用
2. 对并行的Pod网络配置调用批量申请资源



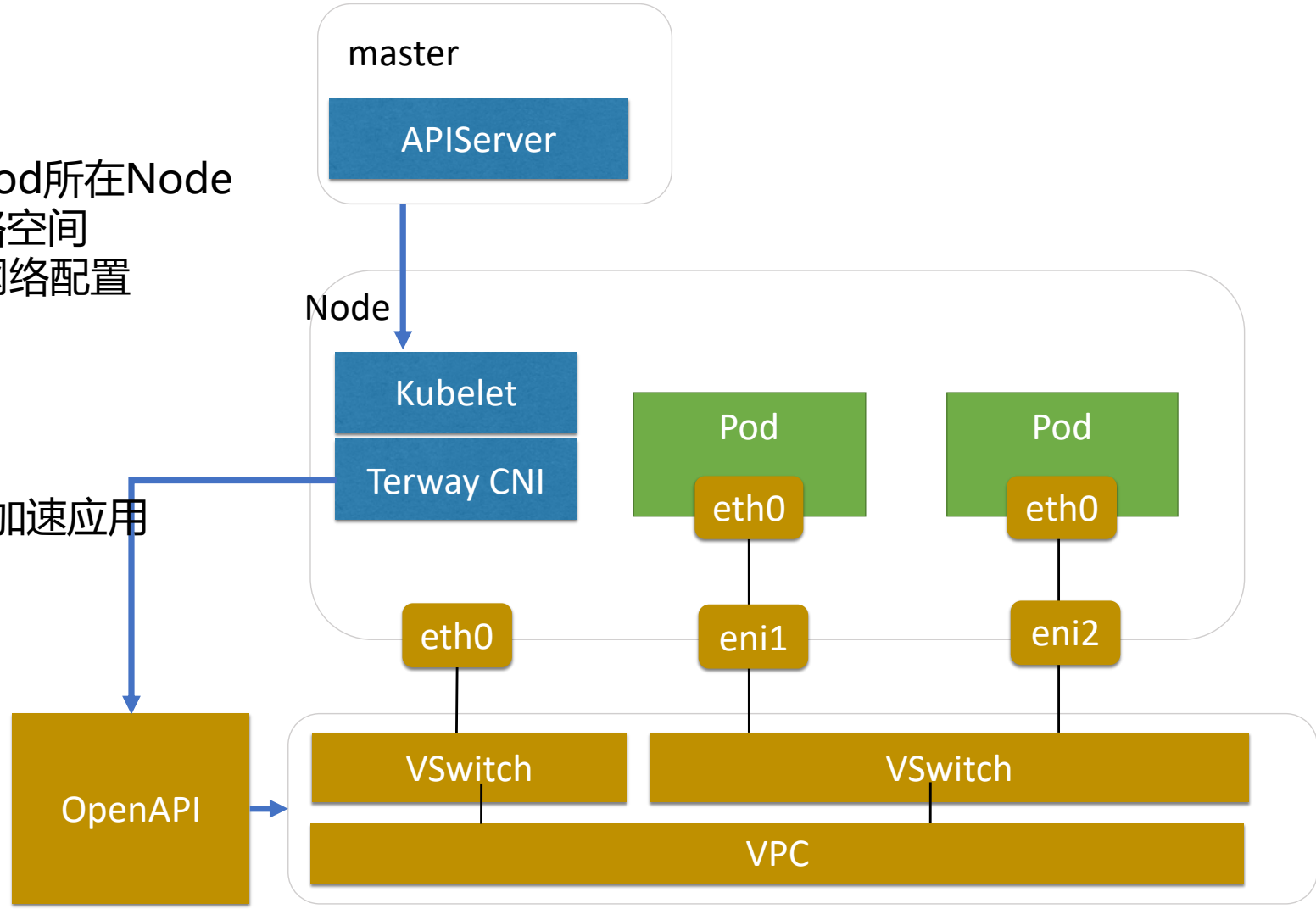
- 如何选择Pod要用的虚拟交换机，保证IP充裕
- 如何平衡每个节点上的网卡的队列和中断数，保证争抢最少
- 以及更多的资源管理器设计可参考Terway文档或代码
  https://github.com/AliyunContainerService/terway

高性能云原生的容器网络

如何利用云原生资源构建容器网络 -- 网络如何打通

Pod独占弹性网卡模式
1. 通过Terway资源管理将弹性网卡绑定Pod所在Node
2. Terway CNI将网卡直接移入到Pod网络空间
3. Terway CNI为其网卡配置IP和路由等网络配置

- Pod网络完全不经过宿主机网络栈
- 性能和ECS机器一致，无损耗
- Pod中为弹性网卡，可以使用DPDK等加速应用



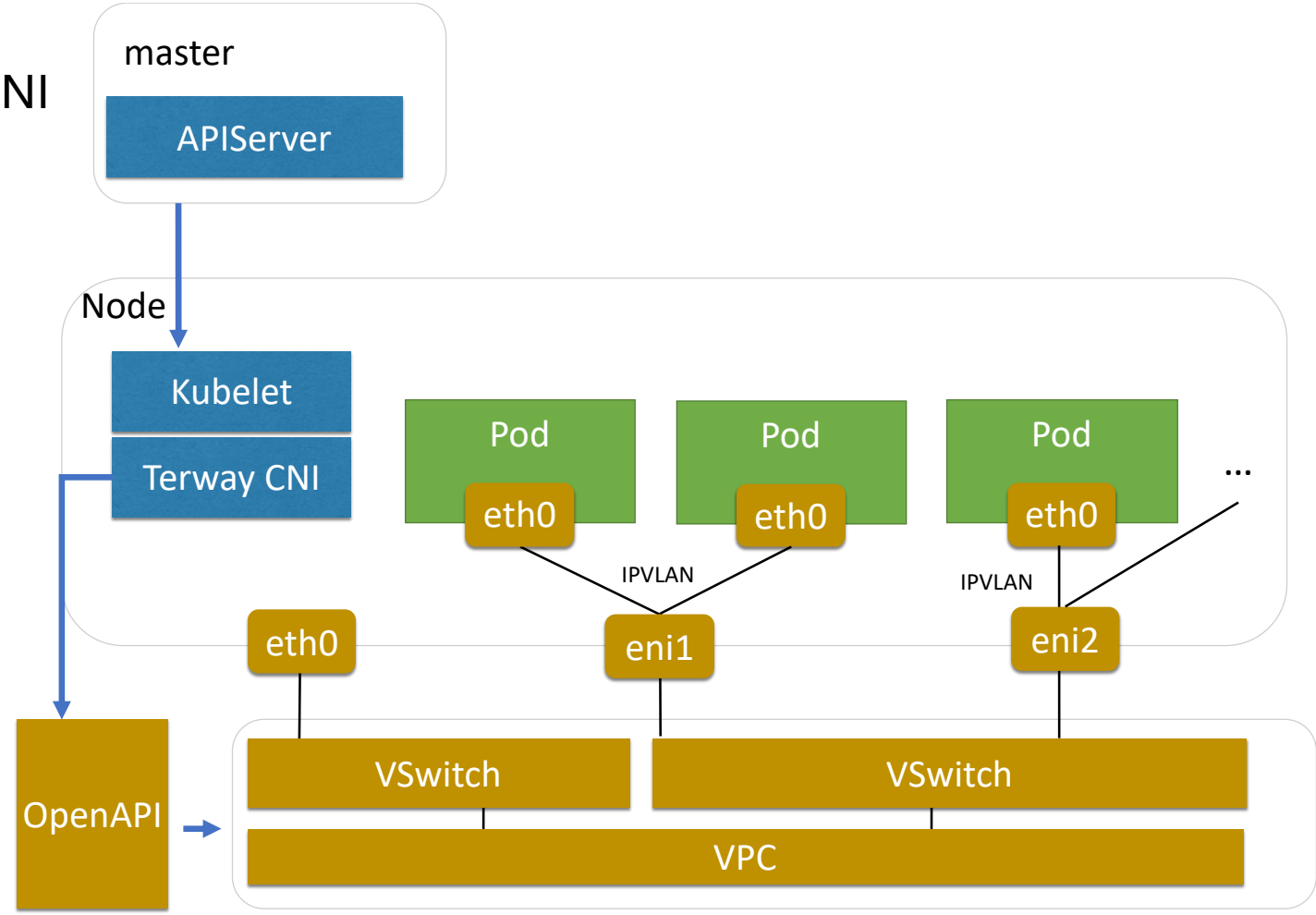
高性能云原生的容器网络

如何利用云原生资源构建容器网络 -- 网络如何打通

Pod共享弹性网卡模式

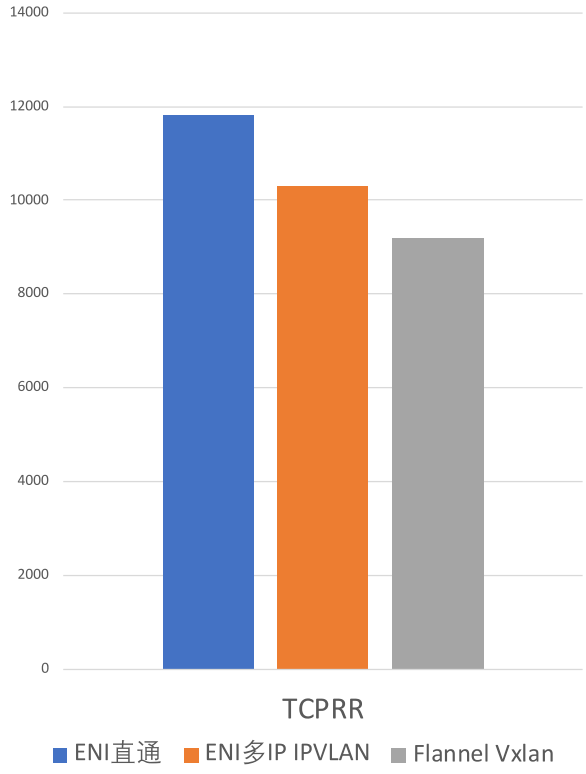
1. Terway 资源管理器根据申请的IP数量和现有ENI上IP来判断申请ENI还是辅助IP地址
2. Terway CNI在ENI上创建IPVLAN子接口
3. 将IPVLAN子接口放入到Pod网络空间
4. 配置Pod网络空间的IP和路由信息

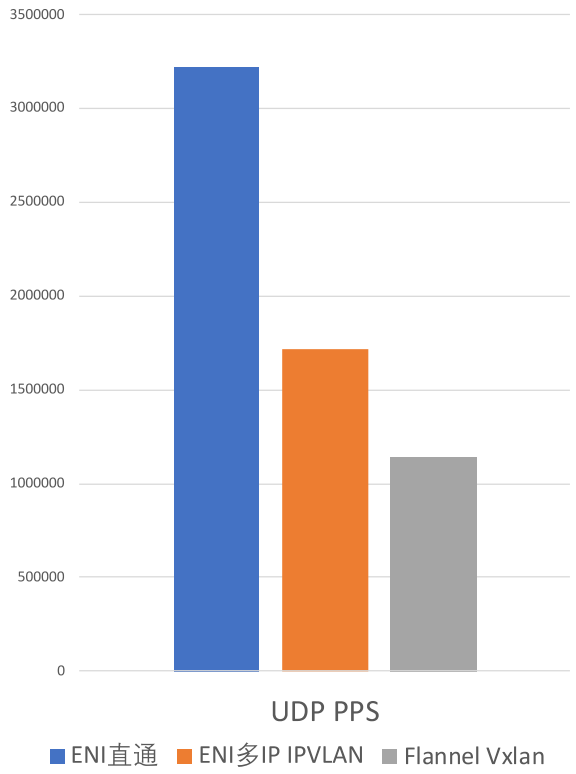
- IPVLAN网络仅部分经过网络栈，不走iptables、路由表，损耗极低
- 一个弹性网卡一般支持10~20个辅助IP，不担心部署密度

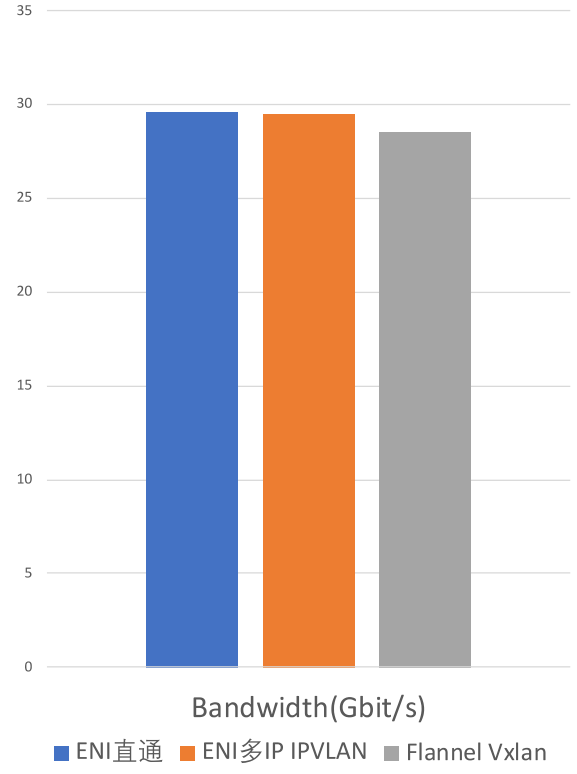


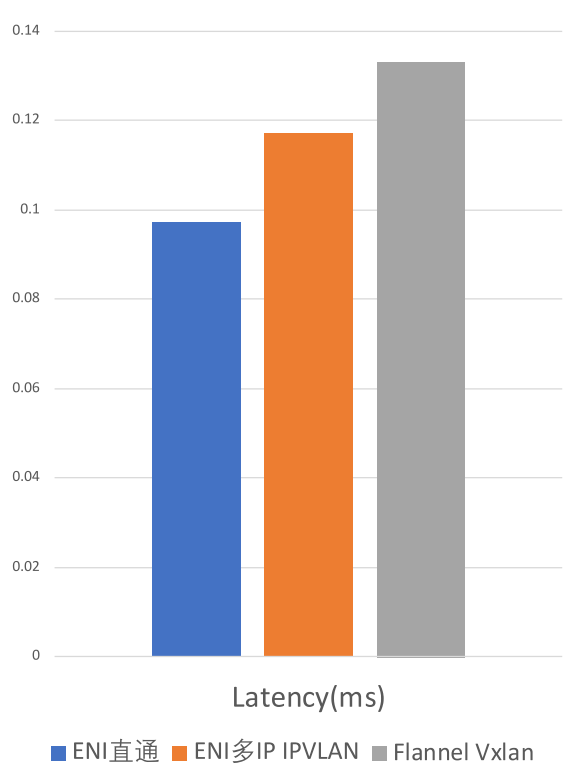
高性能云原生的容器网络

如何利用云原生资源构建容器网络 – 性能对比









- TCP RR, UDP PPS, 带宽，延时优于Vxlan Overlay方案
- 在UDP小包情况下的PPS优势明显，适合于高性能计算和游戏场景

* 测试环境 aliyun神龙 ecs.ebmg5s.24xlarge

目录

阿里云 | 奥运会全球指定云服务商

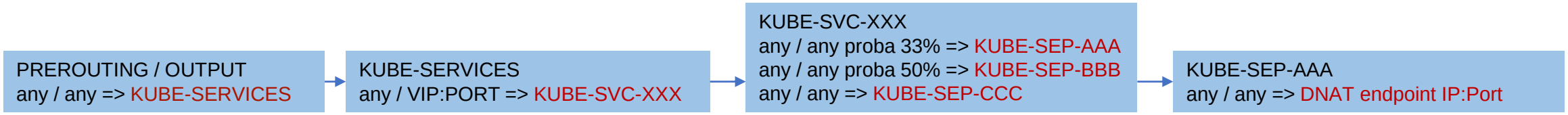
CONTENT

增强网络的扩展性和性能

阿里云 | 奥运会全球指定云服务商

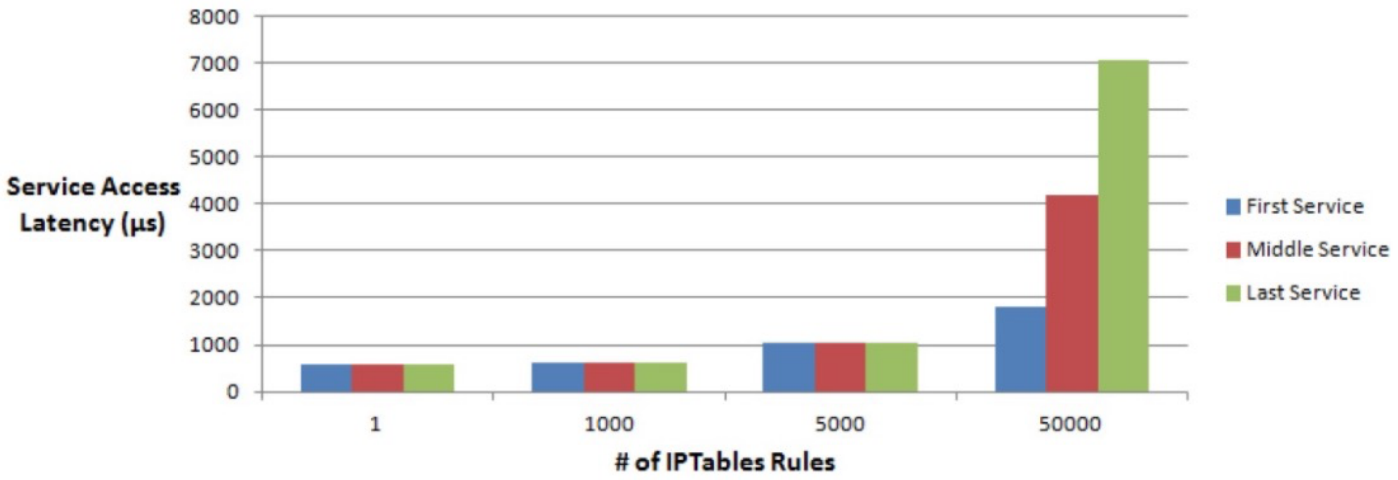
Kubernetes Service性能和扩展性问题

默认采用kube-proxy的实现在大规模中表现不佳



数量：1　　数量：service数量　　数量：endpoint(pod)数量　　数量：1

- iptables链路长，导致网络延时增加，就算是ipvs模式也依赖了iptables的规则
- 扩展性差
  - iptables规则同步是全量刷的，Service和Pod数量多了，一次规则同步都得接近1s
  - Service和Pod数量多了之后数据链路性能大幅降低



增强网络的扩展性和性能

NetworkPolicy性能和扩展性问题

NetworkPolicy是Kubernetes控制Pod和Pod间是否允许通信的规则
目前主流的NetworkPolicy实现基于iptables实现，同样有iptables的扩展性问题

- iptables线性匹配，性能不高，scale能力差
- iptables线性更新，更新速度慢

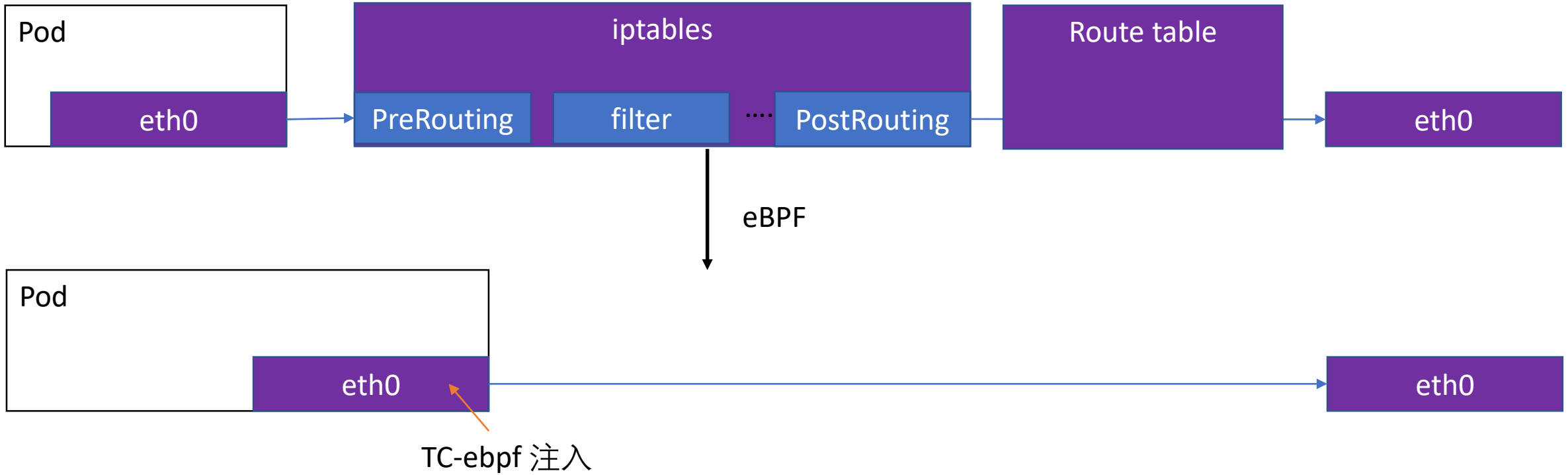
增强网络的扩展性和性能

使用eBPF加速Service和NetworkPolicy扩展性

阿里云 | 奥运会全球指定云服务商

eBPF:

- Linux在最近版本提供的可编程接口
- 通过tc-ebpf将ebpf程序注入到网卡中
- 大幅降低网络链路的长度和复杂度

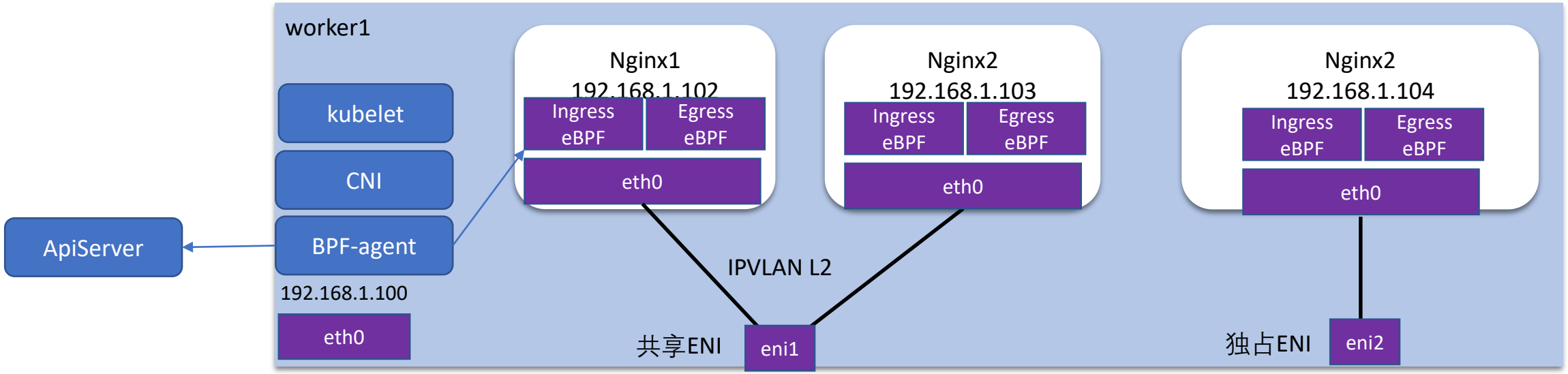


增强网络的扩展性和性能

使用eBPF加速Service和NetworkPolicy扩展性

阿里云 | 奥运会全球指定云服务商

- 每个节点上运行eBPF Agent，监听Service和NetworkPolicy，配置容器网卡的Ingress和Egress规则
- 在Egress的eBPF程序中，判断访问k8s Service IP的请求直接负载均衡到后端Endpoint
- 在Ingress的eBPF程序中，根据NetworkPolicy规则计算源IP决定是否放行



我们使用Cilium作为节点上的BPF-agent去配置容器网卡的BPF规则

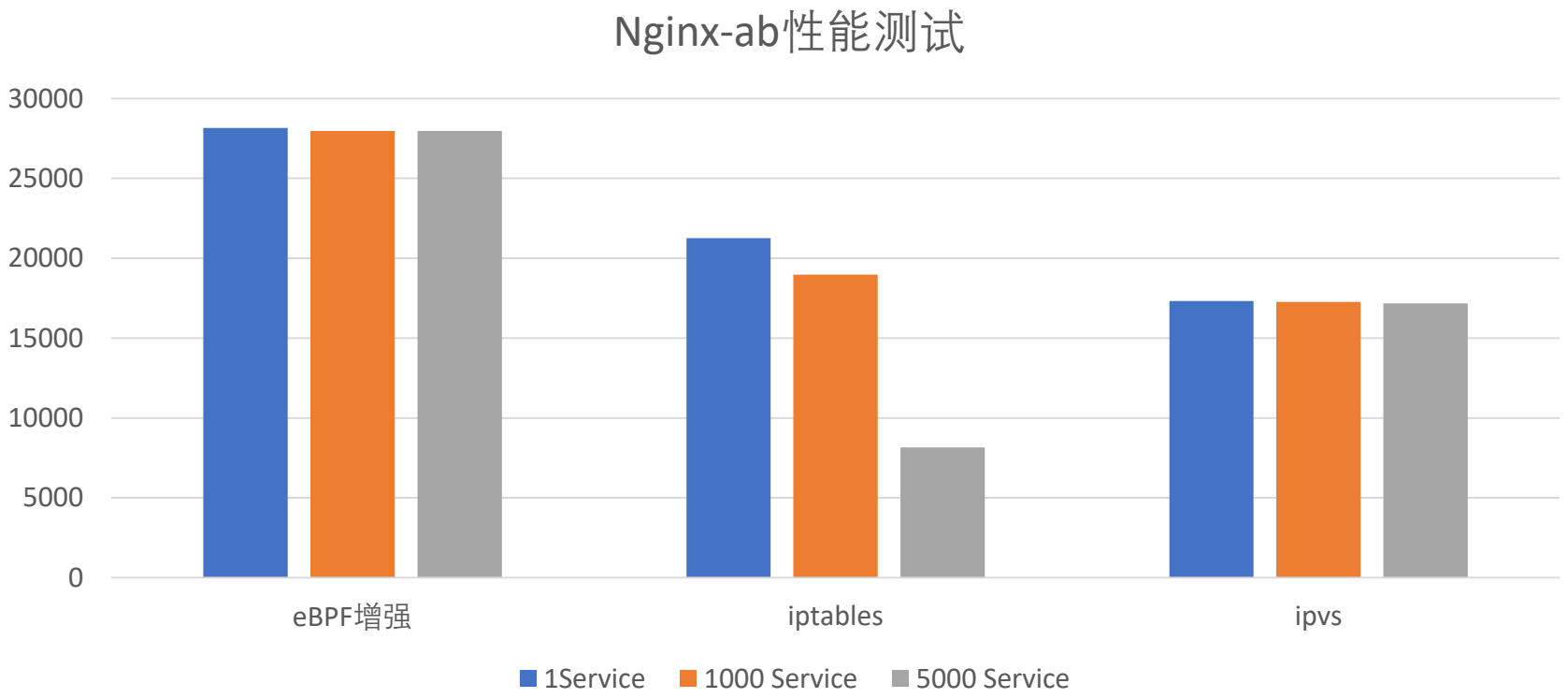
https://github.com/cilium/cilium/pull/10251

增强网络的扩展性和性能

阿里云 | 奥运会全球指定云服务商

使用eBPF加速Service和NetworkPolicy扩展性 – 性能对比

- 通过eBPF的对链路的简化，性能有明显提升
- 并且通过编程的eBPF的形式，让Service扩容时也几乎没有性能损耗



增强网络的扩展性和性能

阿里云 | 奥运会全球指定云服务商

Kubernetes Coredns性能和扩展性问题

- Kubernetes Pod解析DNS域名会search很多次，例如如下Pod中DNS配置：

```
$ cat /etc/resolv.conf
nameserver 172.21.0.10
search kube-system.svc.cluster.local svc.cluster.local cluster.local
options ndots:5
```

当它请求aliyun.com，会依次解析
aliyun.com.kube-system.svc.cluster.local -> NXDOMAIN
aliyun.com.svc.cluster.local -> NXDOMAIN
aliyun.com.cluster.local -> NXDOMAIN
aliyun.com -> 1.1.1.1

- 中心化Coredns，解析经过链路过长，又是UDP协议，导致失败率高

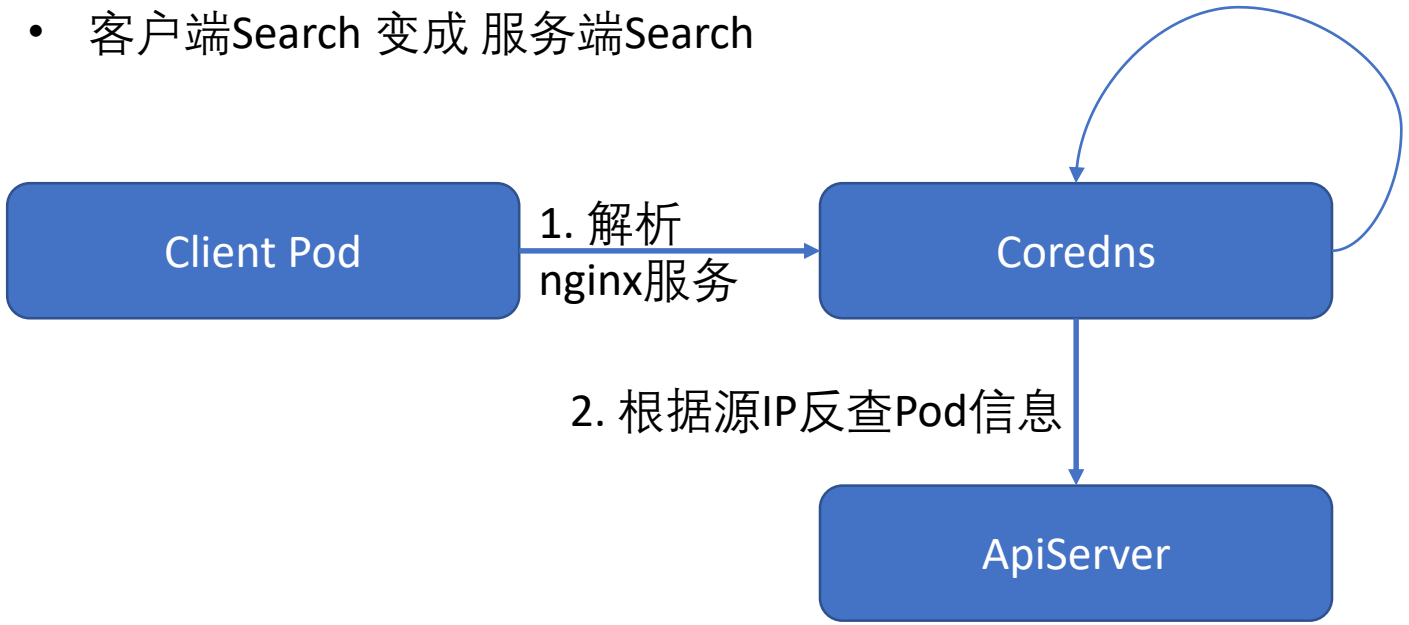
增强网络的扩展性和性能

阿里云 | 奥运会全球指定云服务商

Kubernetes Coredns性能和扩展性问题

使用Coredns Autopath大幅降低Pod DNS查询量

- 客户端Search 变成 服务端Search



3. 根据Pod的Namespace信息 search 对应服务
一次解析请求即可返回结果

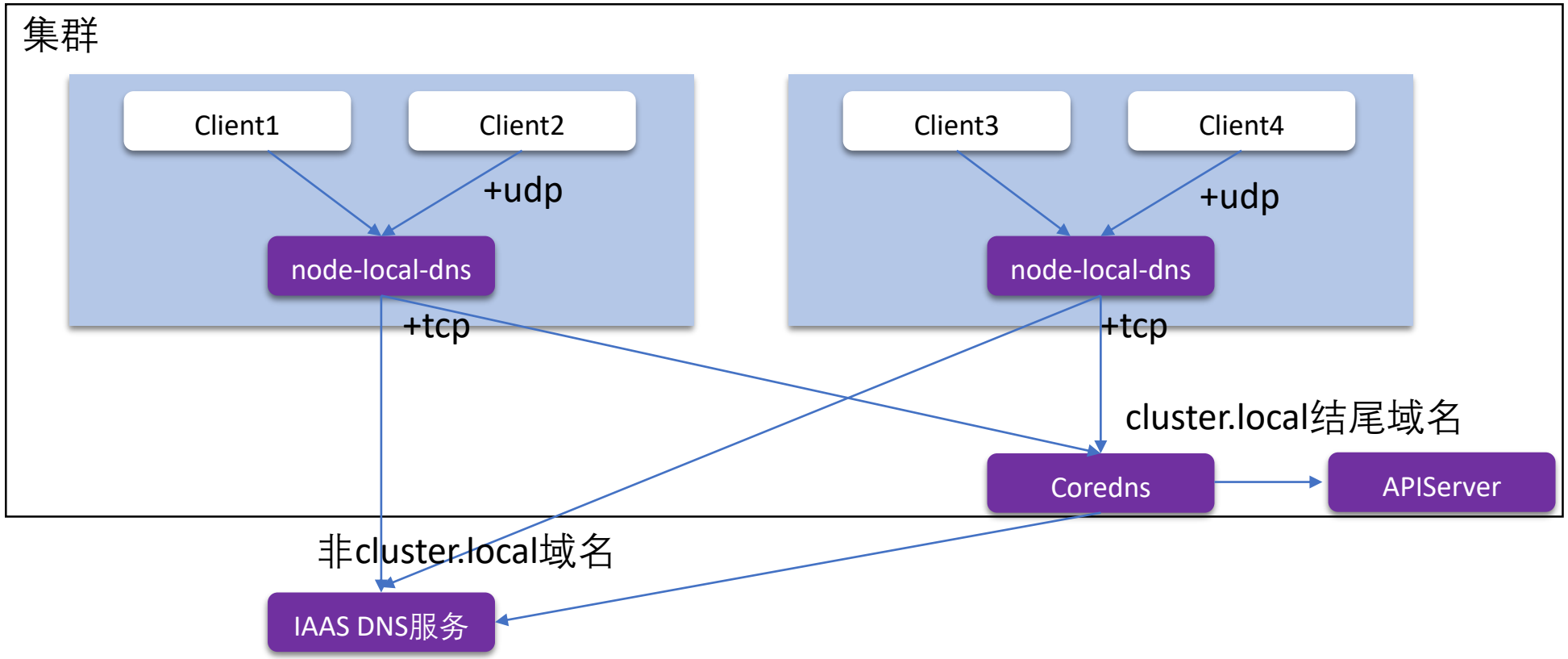
增强网络的扩展性和性能

阿里云 | 奥运会全球指定云服务商

Kubernetes Coredns性能和扩展性问题

在每个节点上使用node-local-dns：

- 将外部域名分流，外部域名请求不再请求中心化Coredns
- 中间链路使用更稳定的TCP解析
- 缓存DNS解析结果



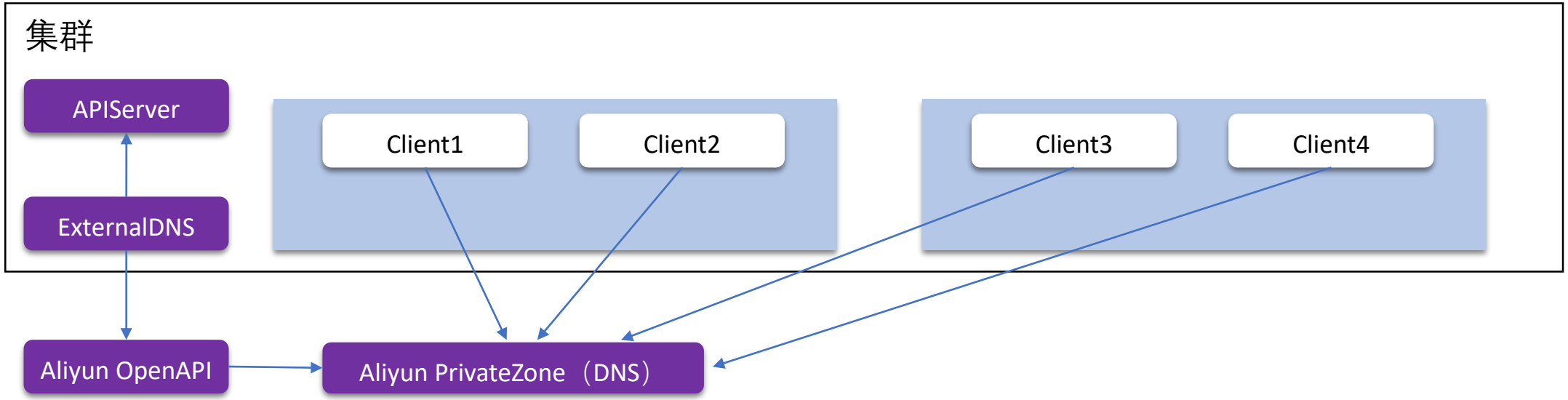
增强网络的扩展性和性能

阿里云 | 奥运会全球指定云服务商

Kubernetes dns性能和扩展性问题

ExternalDNS:

- 云原生的DNS解析，Pod直接请求云服务PrivateZone中的自定义DNS能力
- 由ExternalDNS监听到服务和Pod创建后配置PrivateZone中的域名解析
- 和原生的ECS解析同样的DNS解析性能



阿里云

奥运会全球指定云服务商

关注“阿里巴巴云原生”公众号
获取第一手技术资料