Alauda EE，以DevOps为理念
面向微服务应用的新一代PaaS

Kubernetes 网络更进一步

刘梦馨-基础架构

- Kubernetes 网络模型
- Kubernetes 网络模型问题
- 网络的改进

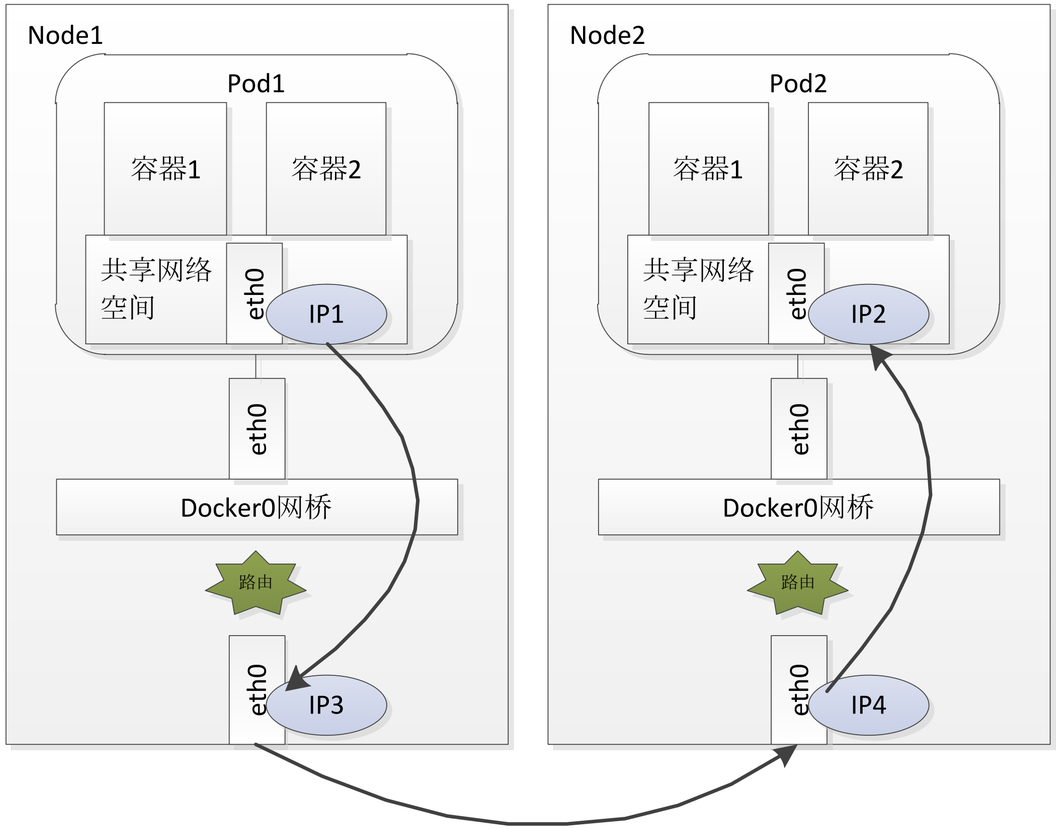
© ALL RIGHTS RESERVED

Kubernetes 网络模型

灵雀云
alauda.cn

Pod 网络

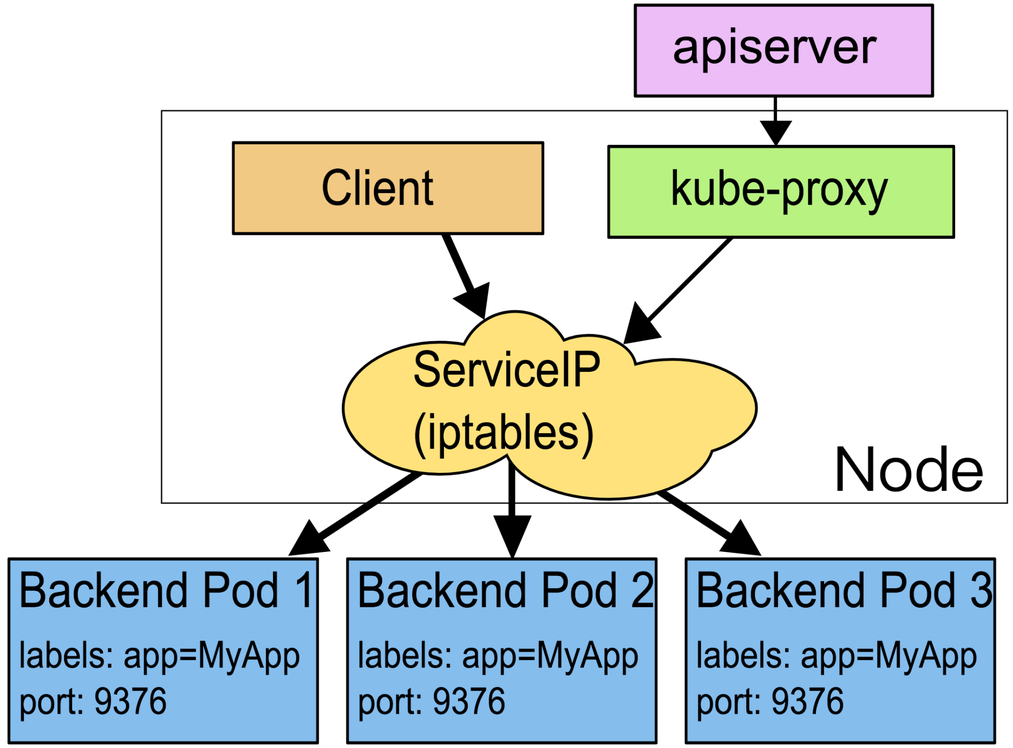
- 一个 Pod 多个容器间
  - 共享 network ns
- 同主机 Pod 间
  - 网桥（docker0,cni0,calico0）
- 跨主机 Pod 间
  - IP 可直接通信（overlay，routing，underlay）



© ALL RIGHTS RESERVED www.alauda.cn

Pod 网络

- Pod IP 要不要固定？
  - 没有规范，大部分网络实现随机分配
- IP 不固定，如何访问服务
  - 集群内：Cluster IP，DNS
  - 集群外：NodePort，Ingress
  - Kube-proxy + iptables/ipvs



© ALL RIGHTS RESERVED www.alauda.cn

Kubernetes 网络模型问题

- 功能
- 性能
- 稳定性

灵雀云
alauda.cn

Kubernetes 网络模型问题

- 功能
  - IP 不固定，带来的运维难度
  - 服务发现机制规则少，不灵活

© ALL RIGHTS RESERVED www.alauda.cn

功能

- IP 不固定
  - 无法对 IP 资源进行精细管控
  - 基于 IP 的监控无法使用
  - 基于 IP 的安全策略无法使用
  - 一些 IP 发现的服务部署困难

© ALL RIGHTS RESERVED www.alauda.cn

功能

- 服务发现功能少
  - Cluster IP 基于四层转发，没有复杂规则
  - 负载均衡策略少 round robin, ip hash
  - Ingress 转发策略少，host，url
  - Ingress 无法使用多个端口
  - 流量没有日志和监控

© ALL RIGHTS RESERVED www.alauda.cn

性能

- 1000 Cluster IP，1000 NodePort，3W+ iptables rule
- Iptables 没有增量更新功能，需要整体 flush，更新时间长
- Iptables 规则串行，最坏情况，流量需要经过所有规则匹配
- NodePort 需要经过多一次转发，流量翻倍

© ALL RIGHTS RESERVED

稳定性

- 网络分区，网络问题导致转发异常
- 缺少健康检查功能
- NodePort 屏蔽了 Pod 的直接访问，上层健康检查失效

© ALL RIGHTS RESERVED www.alauda.cn

网络的改进

- 固定IP
- Headless service
- 自研 ingress

灵雀云
alauda.cn

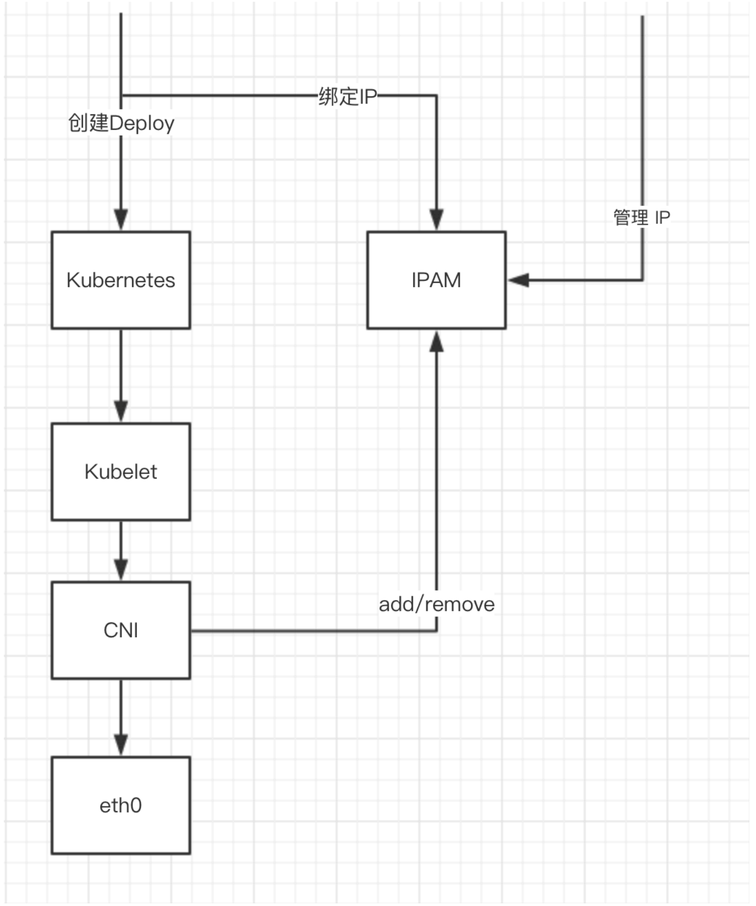
固定IP

- 理念之争（无状态vs有状态）
- 存不存在完全无状态的服务？
- 固定 DNS 存在缓存的问题

© ALL RIGHTS RESERVED www.alauda.cn

固定IP

- IP 作为头等公民来对待
- IP 导入，权限管理，路由，网关，dns...
- 自研 CNI IPAM 插件
- IP 回收保证一致性



© ALL RIGHTS RESERVED

固定IP

- 多网卡
- 固定 mac
- IP，路由动态更新

```
{
  "cniVersion": "0.3.1",
  "interfaces": [                                            (this key omitted by IPAM plugins)
      {
          "name": "<name>",
          "mac": "<MAC address>",                            (required if L2 addresses are meaningful)
          "sandbox": "<netns path or hypervisor identifier>" (required for container/hypervisor interfaces, empty/omi
      }
  ],
  "ips": [
      {
          "version": "<4-or-6>",
          "address": "<ip-and-prefix-in-CIDR>",
          "gateway": "<ip-address-of-the-gateway>",          (optional)
          "interface": <numeric index into 'interfaces' list>
      },
      ...
  ],
  "routes": [                                                (optional)
      {
          "dst": "<ip-and-prefix-in-cidr>",
          "gw": "<ip-of-next-hop>"                           (optional)
      },
      ...
  ]
  "dns": {
    "nameservers": <list-of-nameservers>                     (optional)
    "domain": <name-of-local-domain>                         (optional)
    "search": <list-of-additional-search-domains>            (optional)
    "options": <list-of-options>                             (optional)
  }
}
```

© ALL RIGHTS RESERVED www.alauda.cn

Headless Service

- 没有 Cluster IP，没有 iptables
- DNS 直接返回后端 pod IP
- 同样存在 dns 缓存问题
- Kubernetes 1.8 之后引入 IPVS，可以考虑尝试

© ALL RIGHTS RESERVED www.alauda.cn

自研 Ingress

- Ingress 功能太少，Openresty
  - DSL 定义流量分发规则
  - 直接 watch endpoint 对接后端 Pod IP
  - 复杂的负载均衡策略
  - 健康检查
  - 日志监控

© ALL RIGHTS RESERVED www.alauda.cn

自研 Ingress

- 规则：域名是 www.baidu.com 或者 baidu.com，路径是 /search, 源 IP 为 114.114.114.114，header 中 uid 在 100 到 999 或者 10000 到 11000 之间的请求

- （AND (IN HOST www.baidu.com baidu.com) (EQ URL /search) (EQ SRC_IP 114.114.114.114) (OR (RANGE HEADER uid 100 999) (RANGE HEADER uid 10000 11000)))

© ALL RIGHTS RESERVED

总结

- 动态 IP 运维管理困难 -> 固定 IP 的 IPAM 插件
- Iptables带来性能下降 -> Headless Service, IPVS
- Ingress 功能太少 -> 自研 Openresty Ingress

© ALL RIGHTS RESERVED www.alauda.cn

灵雀云
alauda.cn

Q&A