2017 China Kubernetes End User Conference
kubernetes 中国用户大会
KEUC 2017
2017.10.15 / 中国·杭州

网易云如何基于大规模Kubernetes集群支撑高并发应用

刘 超 网易云解决方案总架构师

主办：CLOUD NATIVE COMPUTING FOUNDATION
caicloud 才云

容器的使用场景

大规模云原生应用的支撑痛点

Kubernetes的性能问题

公有云的支撑痛点

Kubernetes的规模问题

容器化的本质?

容器	虚拟机
我秒级启动	你没算应用吧，我也秒级
我一台机器上启动1000个nginx	然并卵
我有镜像，原子升级回滚	说的谁好像没有似得
我秒级自修复	你确认这一秒没丢啥
我服务发现	Dubbo和springcloud笑而不语，熔断，降级
我弹性伸缩	你听说过autoscaling group么?

容器化的本质?

如果是个传统应用，启动慢，进程少，不更新，IaaS就够了

容器化的本质?

如果你遇到了以下问题

变化快　　扛不住

拆

微服务

容器化的本质?

微服务

扛不住，进程多

变化快，常更新

容器：100个进程，每天一个镜像

虚拟机：我有点大

虚拟机：可不可以不用镜像，Ansible也能部署

容器化的本质?

微服务

扛不住，进程多　　变化快，常更新

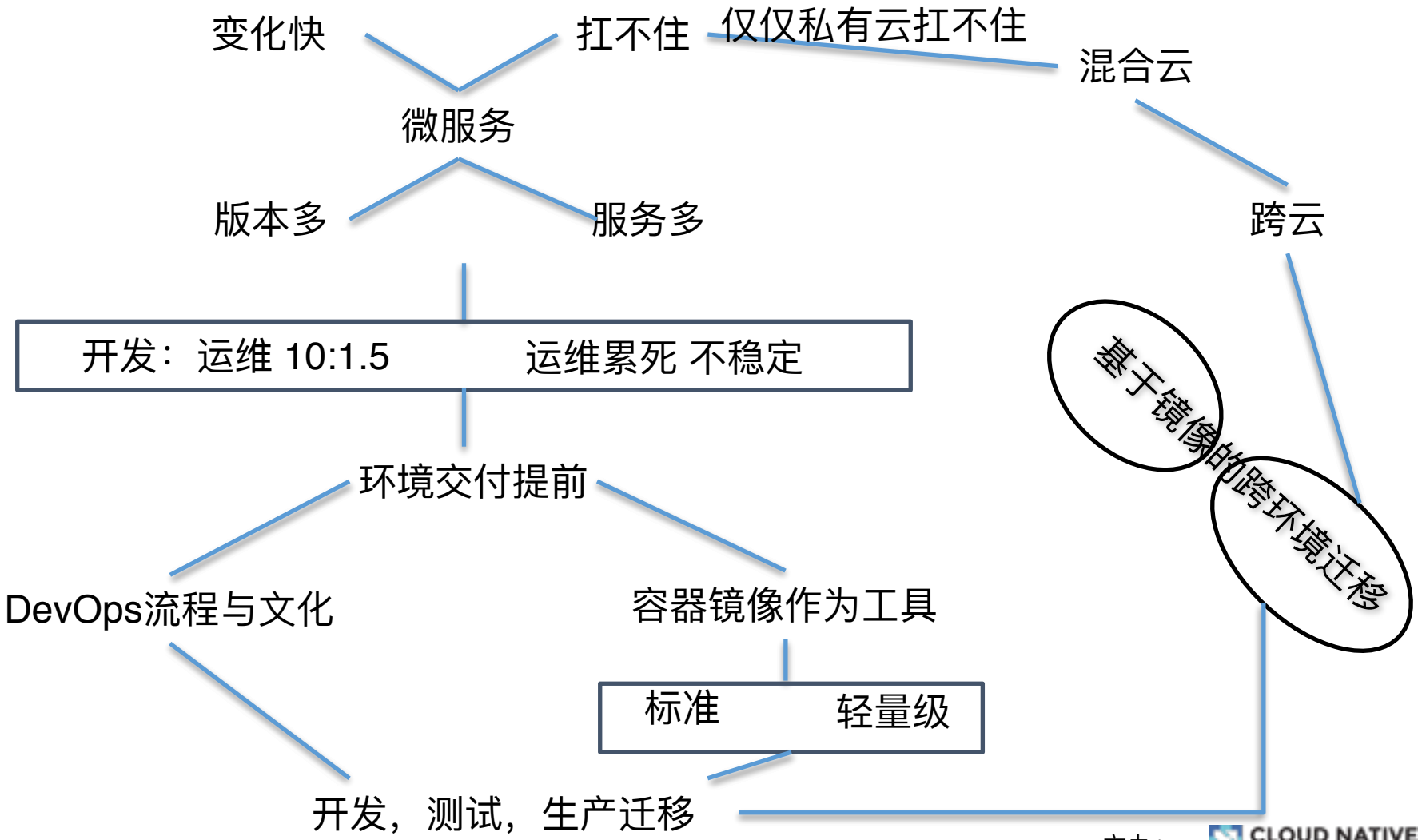
开发：运维 10:1.5

运维：开发写完代码就不管了，这么多环境都是我的

容器镜像的本质：
环境交付提前，每个研发5%的工作量，换取运维200%的工作量，增加稳定性

DevOps文化与流程

容器的本质?


变化快
扛不住
仅仅私有云扛不住
混合云
微服务
版本多
服务多
跨云
开发：运维 10:1.5
运维累死 不稳定
基于镜像的跨环境迁移
环境交付提前
DevOps流程与文化
容器镜像作为工具
标准
轻量级
开发，测试，生产迁移

容器的使用场景

大规模云原生应用的支撑痛点

Kubernetes的性能问题

公有云的支撑痛点

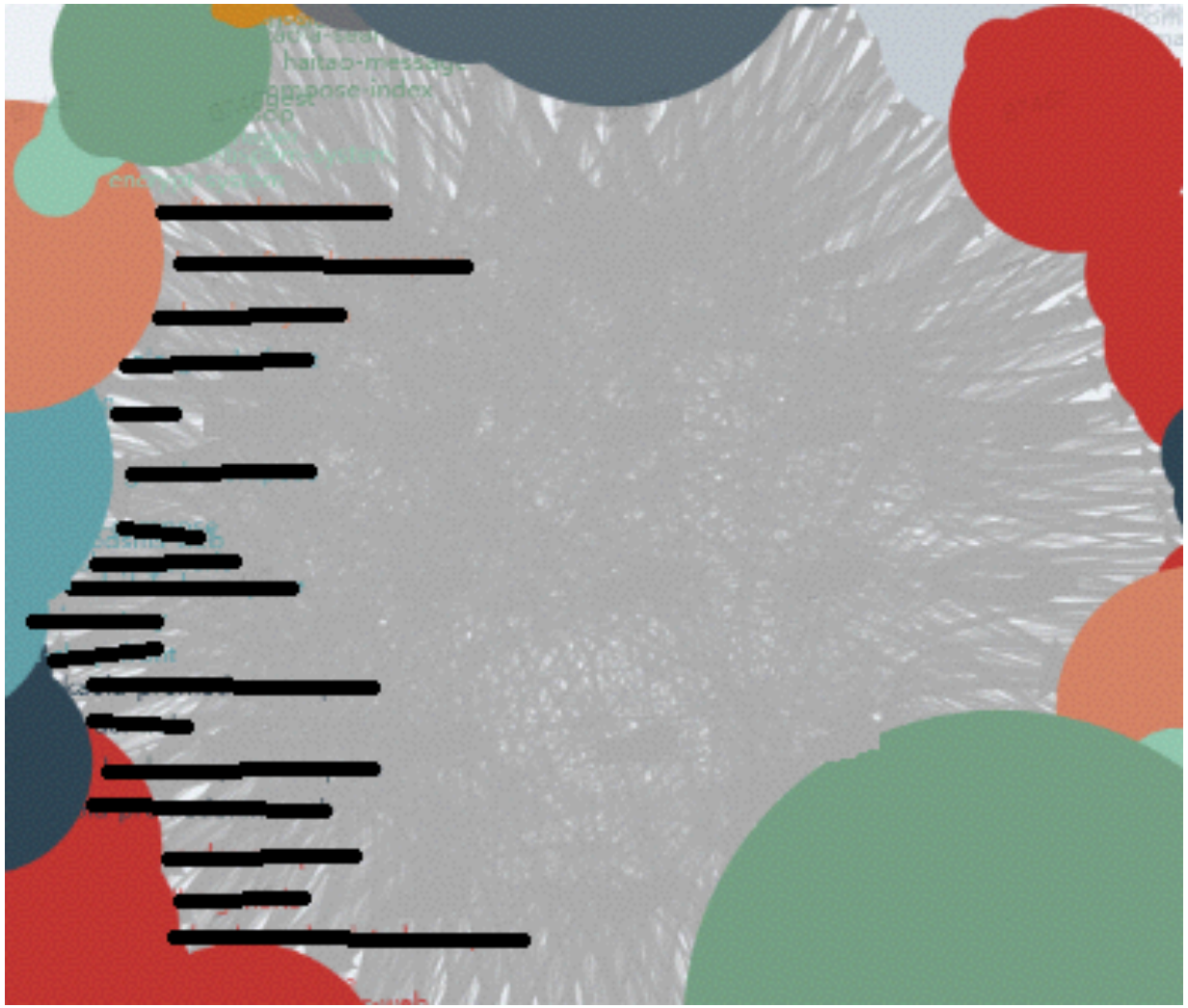
Kubernetes的规模问题

主办：

大规模云原生应用的支撑痛点

主办：

大规模云原生应用的支撑痛点



- 入口：高性能负载均衡器
- 互访：高性能私有网络
- 访问PaaS平台：容器网络和虚拟机网络互通
- 高可用：高性能机房网络
- 服务发现：灵活策略，熔断，降级
- 配置复杂：统一配置中心

容器的使用场景

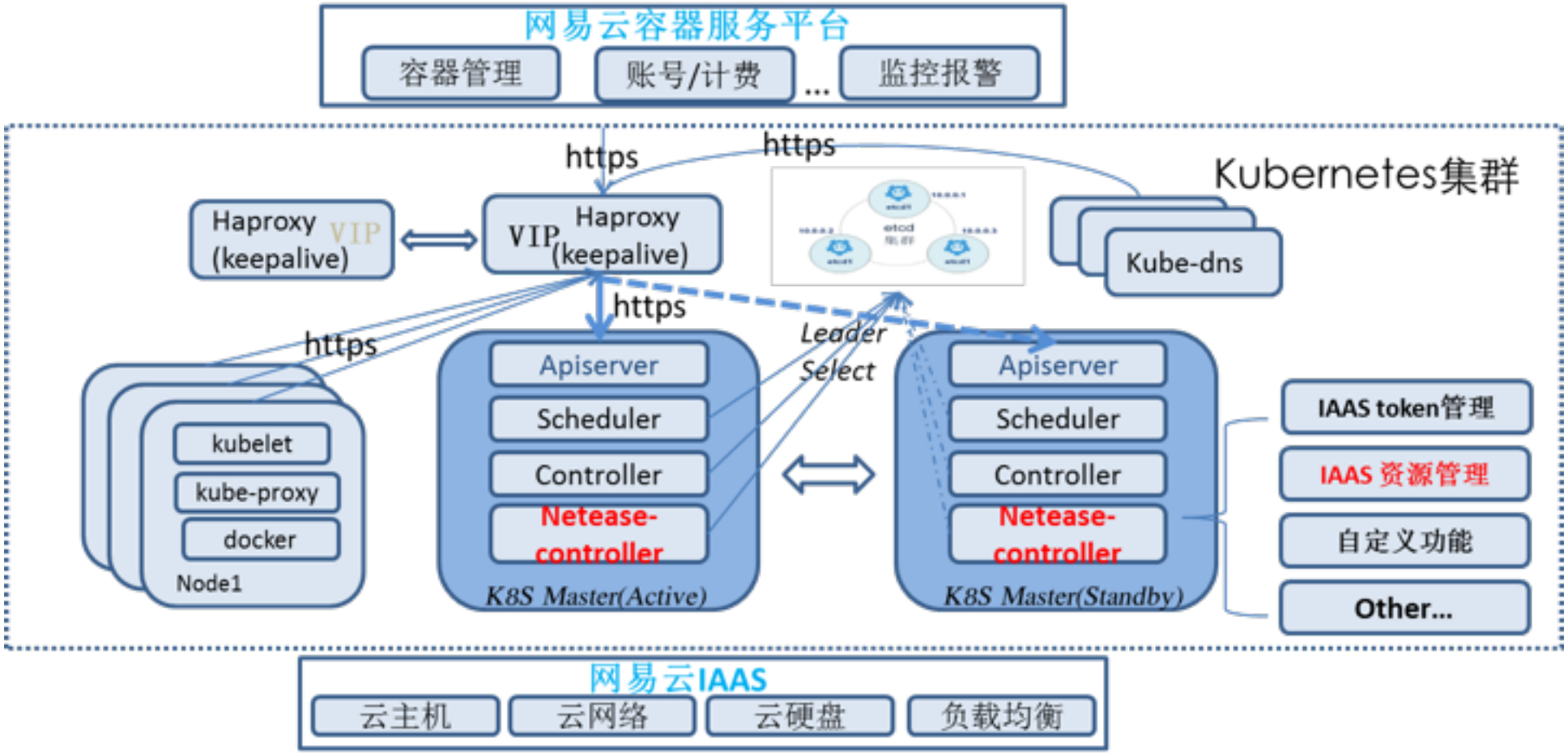
大规模云原生应用的支撑痛点

Kubernetes的性能问题

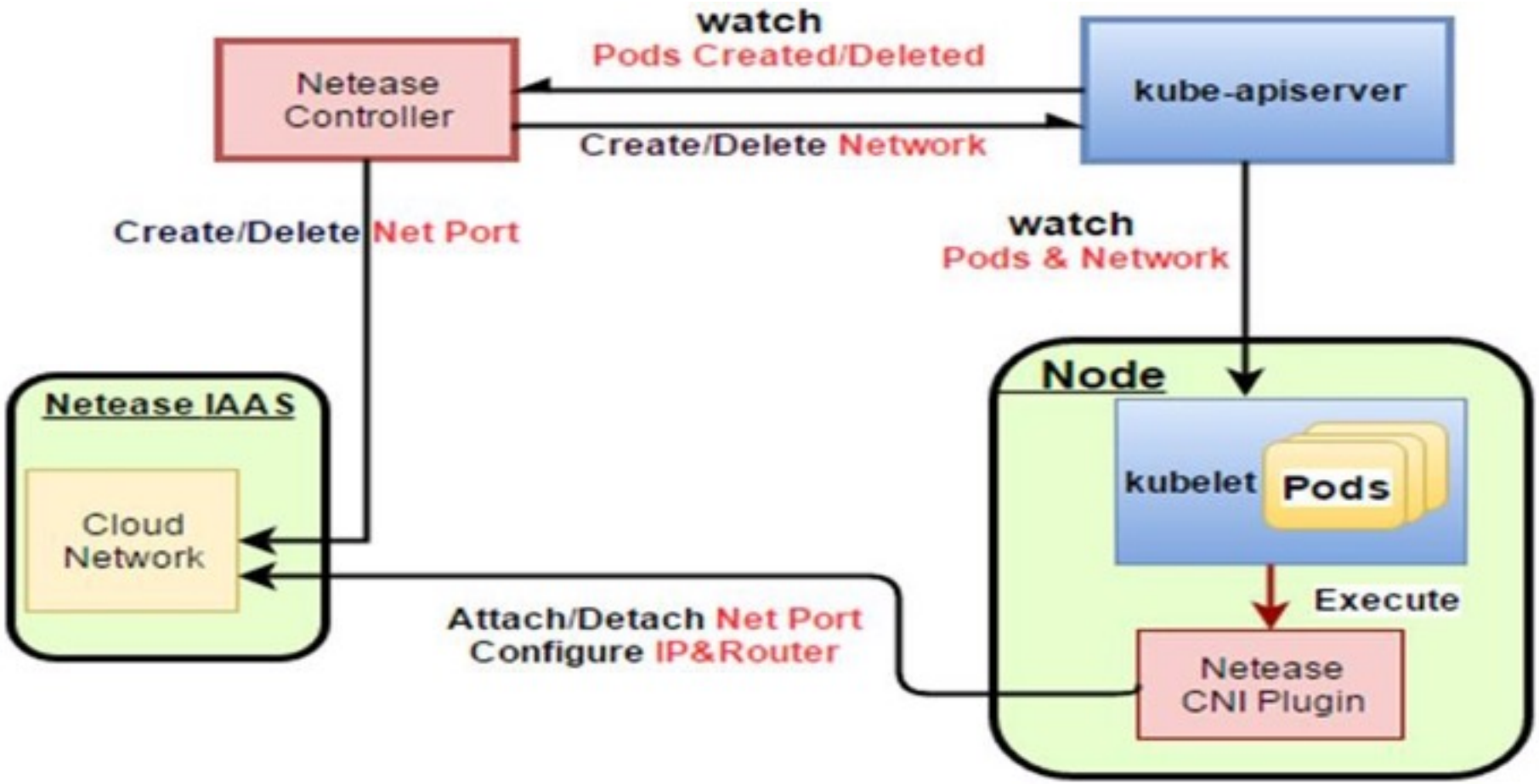
公有云的支撑痛点

Kubernetes的规模问题

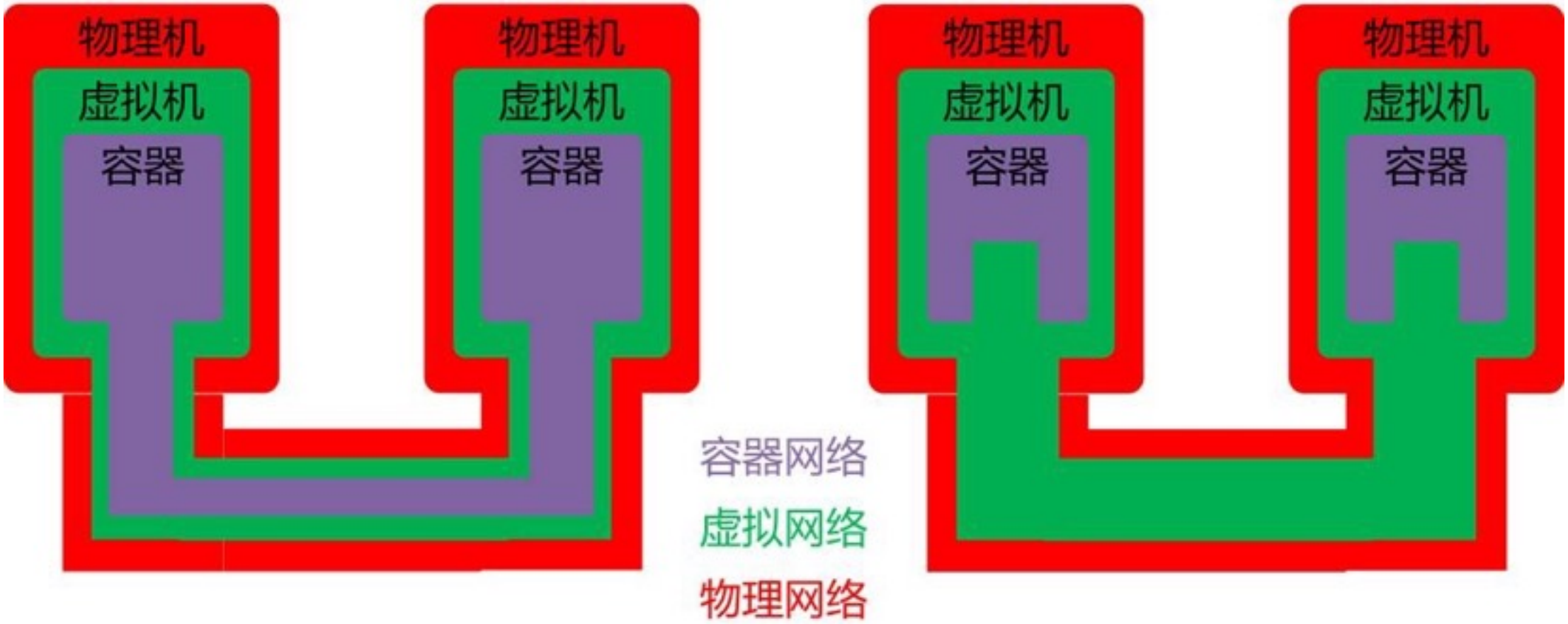
网易云容器服务平台架构



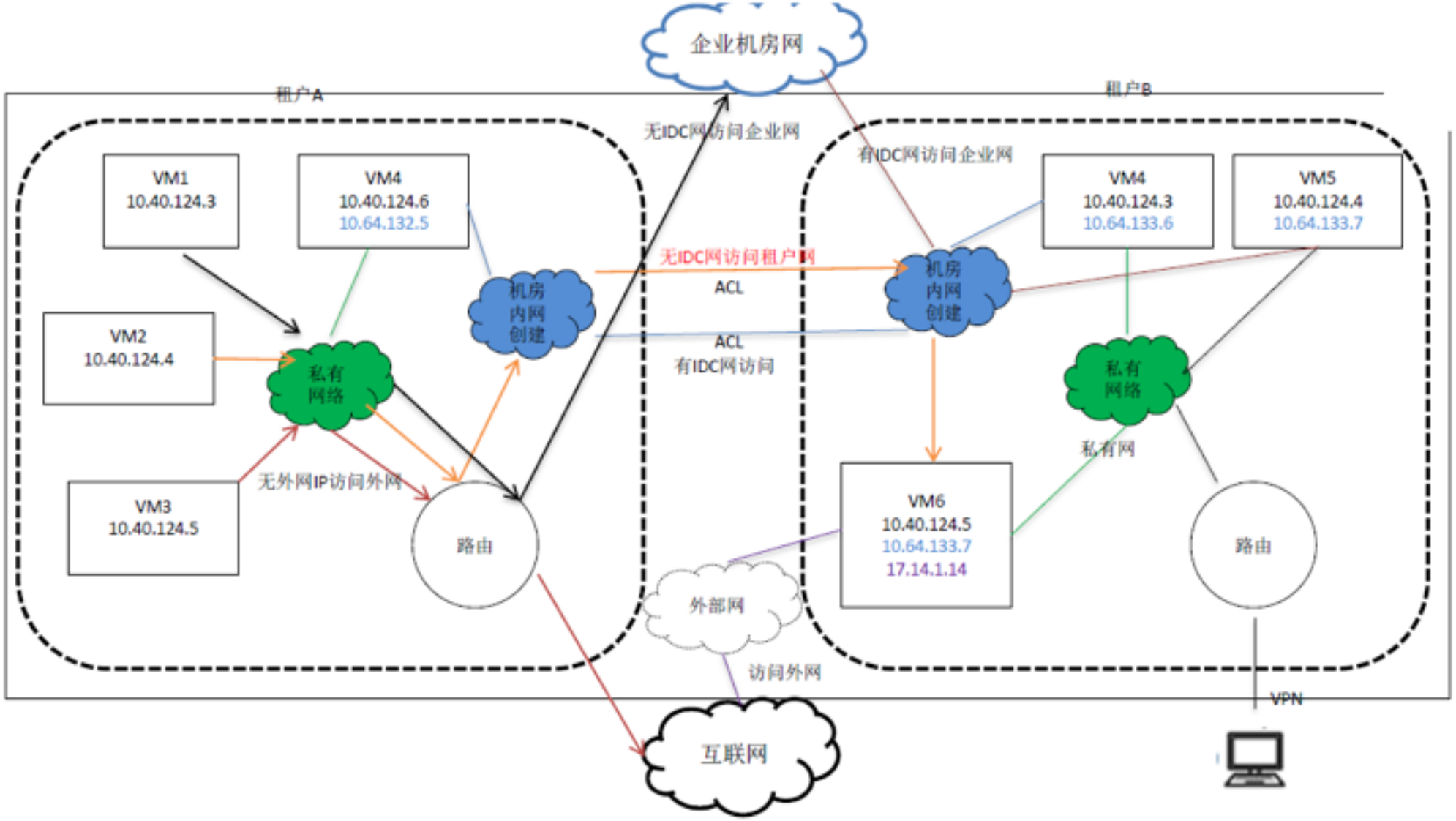
云网络


watch
Pods Created/Deleted
Netease Controller
kube-apiserver
Create/Delete Network
Create/Delete Net Port
watch
Pods & Network
Node
Netease IAAS
kubelet
Pods
Cloud Network
Execute
Attach/Detach Net Port
Configure IP&Router
Netease CNI Plugin

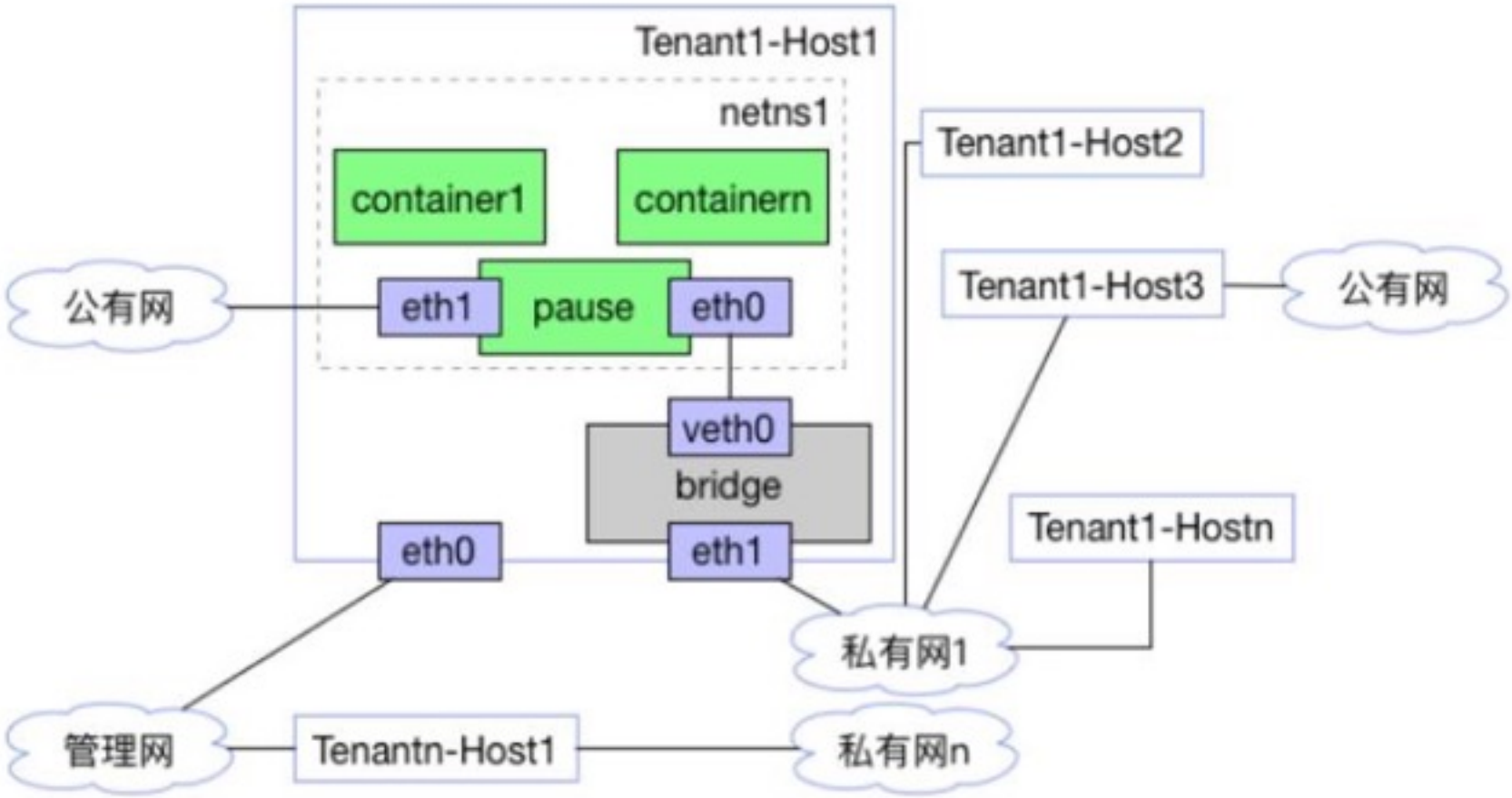
云网络



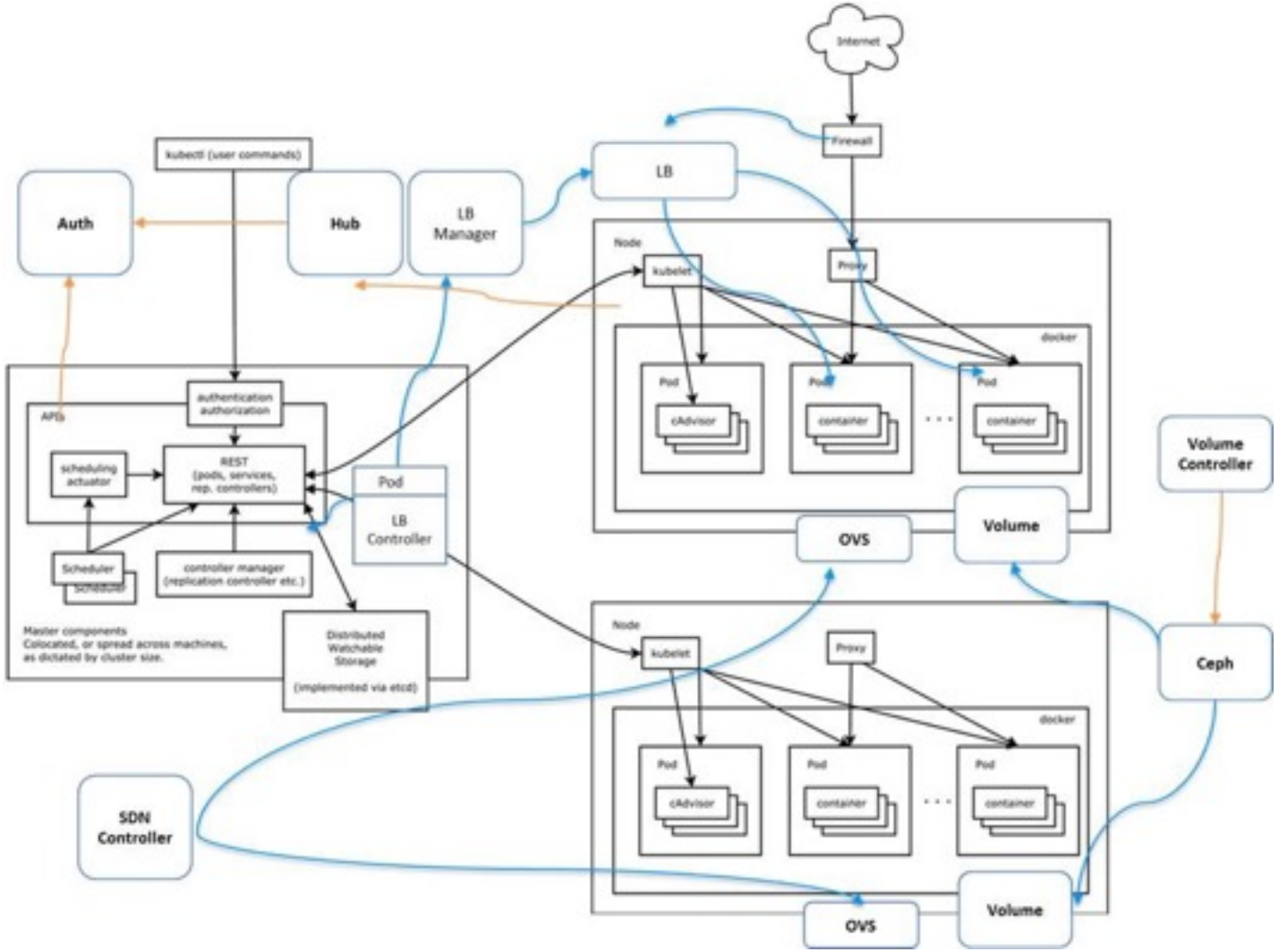
云网络


企业机房网
租户A
租户B
无IDC网访问企业网
有IDC网访问企业网
VM1
10.40.124.3
VM4
10.40.124.6
10.64.132.5
VM2
10.40.124.4
VM3
10.40.124.5
私有网络
机房内网创建
无外网IP访问外网
路由
无IDC网访问租户网
ACL
ACL
有IDC网访问
机房内网创建
VM4
10.40.124.3
10.64.133.6
VM5
10.40.124.4
10.64.133.7
私有网络
私有网
VM6
10.40.124.5
10.64.133.7
17.14.1.14
路由
外部网
访问外网
互联网
VPN

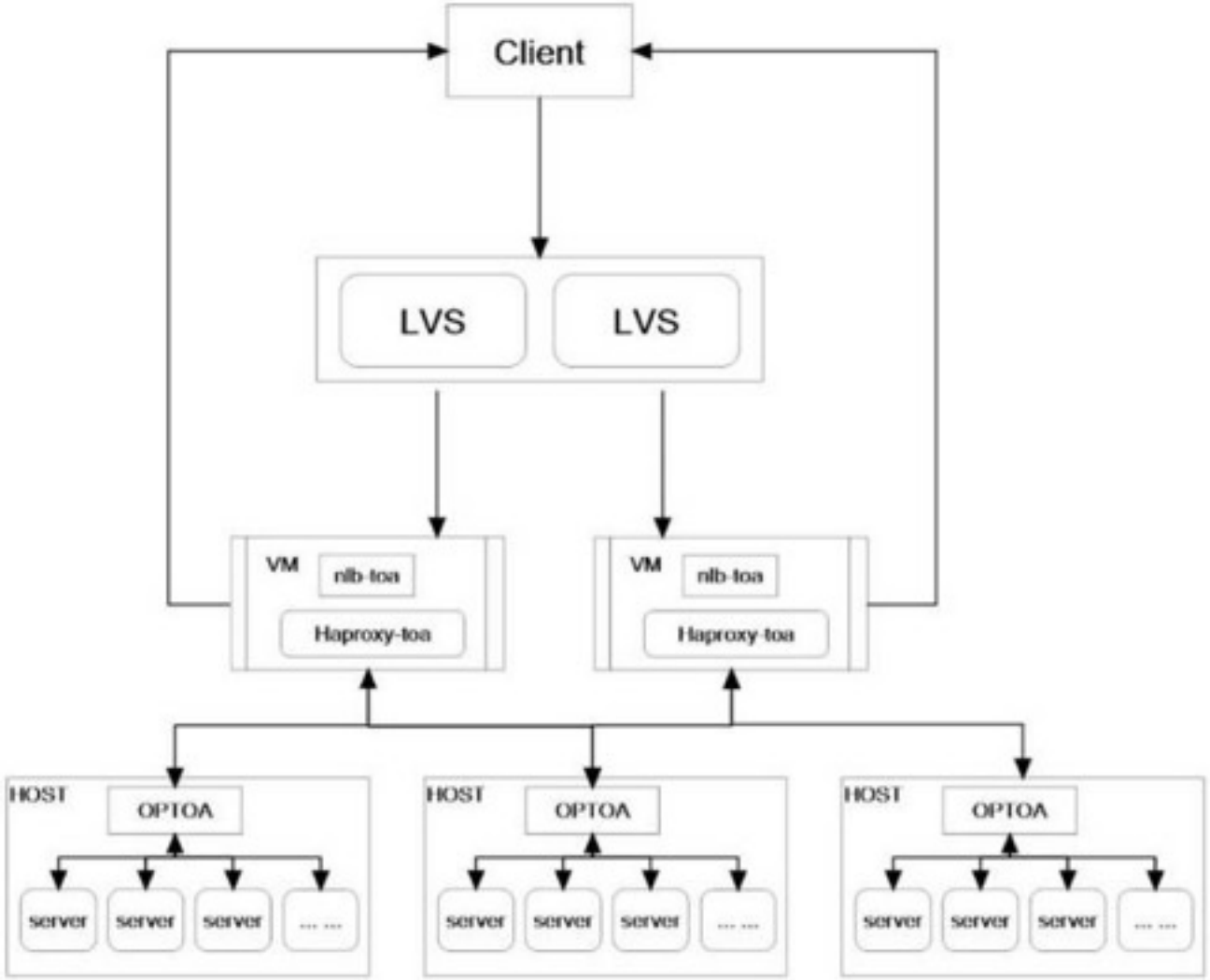
容器网络


Tenant1-Host1
netns1
container1
containern
pause
eth1
eth0
veth0
bridge
eth0
eth1
公有网
Tenant1-Host2
Tenant1-Host3
公有网
Tenant1-Hostn
私有网1
管理网
Tenantn-Host1
私有网n

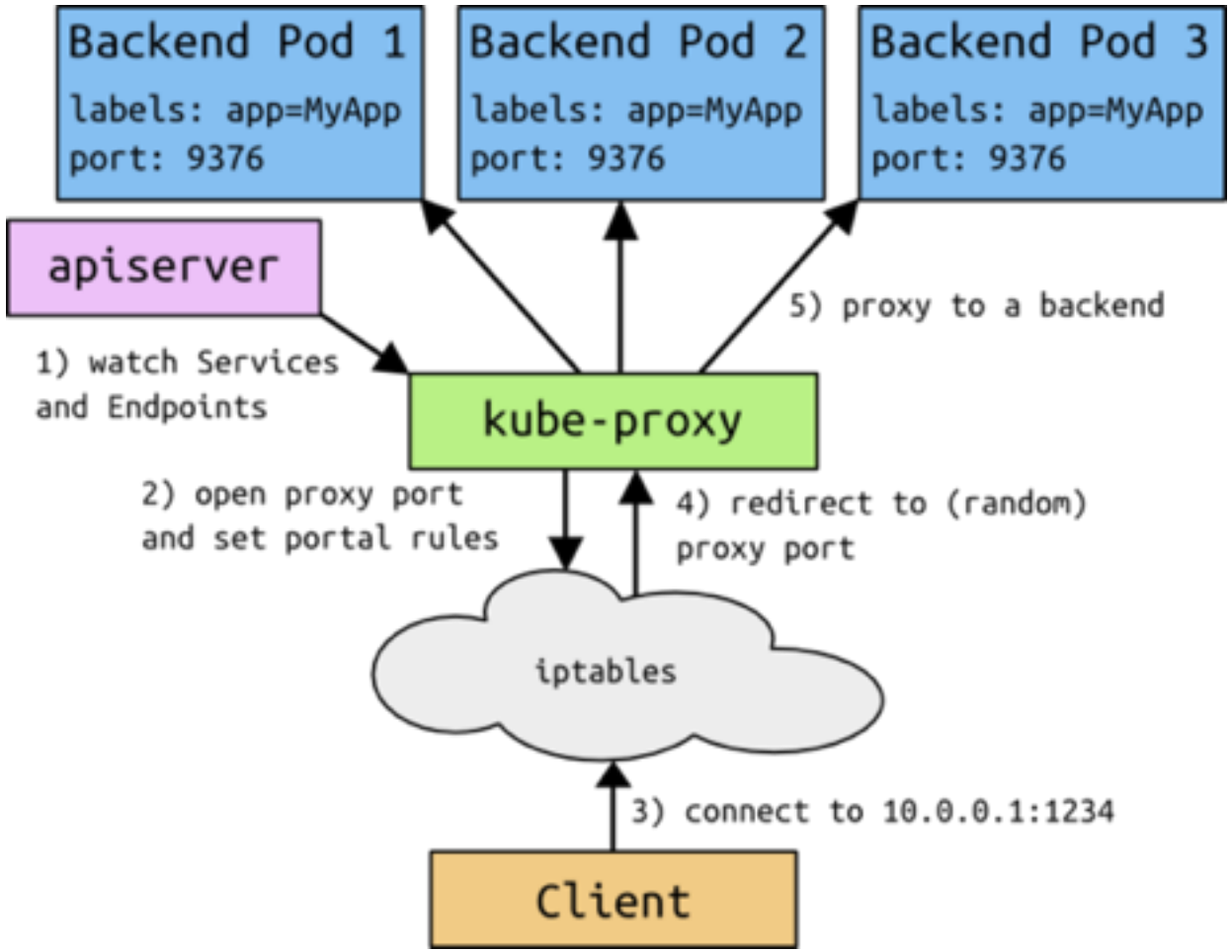
负载均衡


Internet
Firewall
kubectl (user commands)
Auth
Hub
LB Manager
LB
Node
kubelet
Proxy
docker
Pod
cAdvisor
container
Volume Controller
authentication authorization
APIs
scheduling actuator
REST (pods, services, rep. controllers)
Pod
LB Controller
OVS
Volume
Scheduler
controller manager (replication controller etc.)
Master components Colocated, or spread across machines, as dictated by cluster size.
Distributed Watchable Storage
(implemented via etcd)
Ceph
SDN Controller

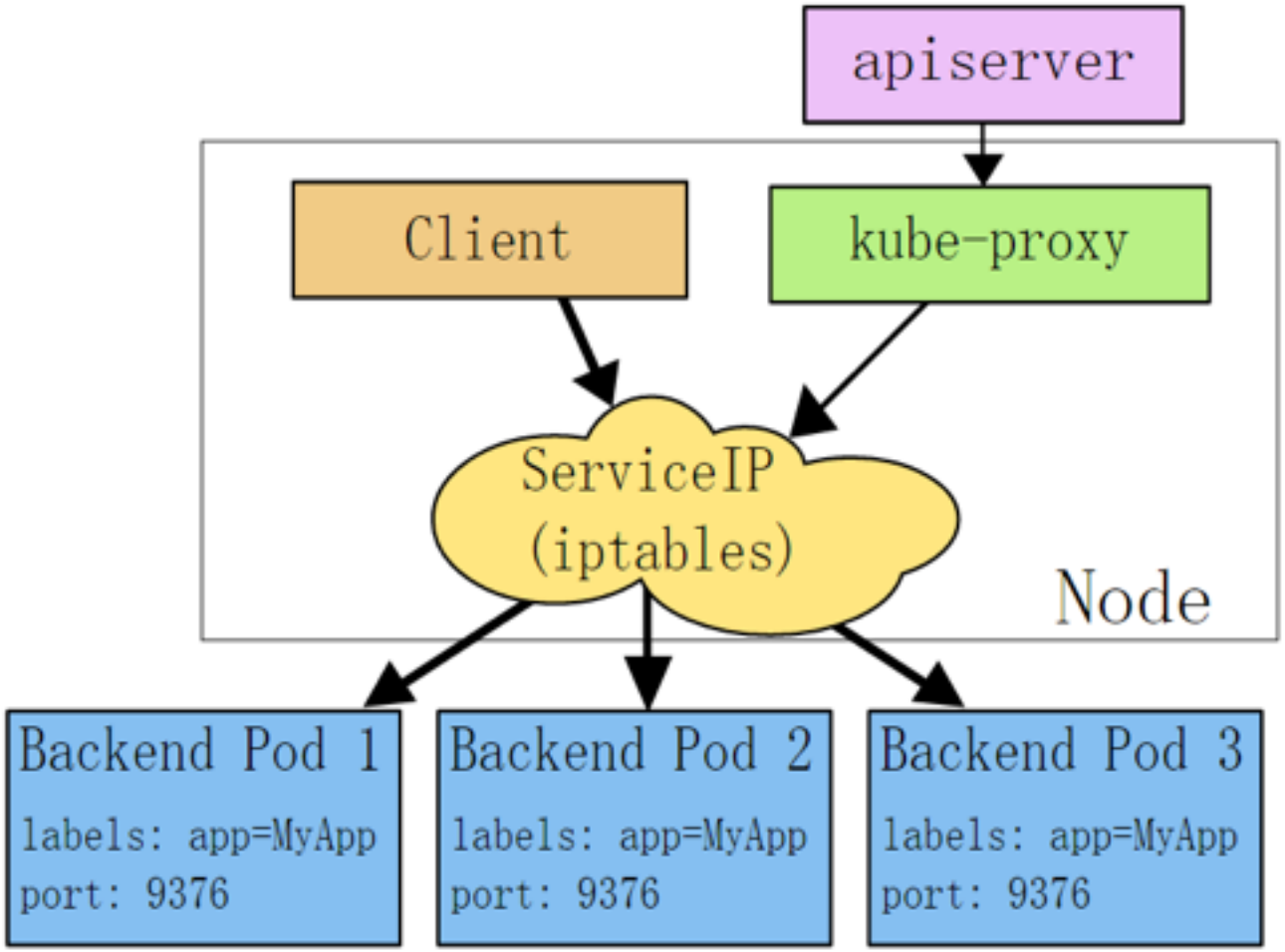
负载均衡


Client
LVS
LVS
VM
nlb-toa
Haproxy-toa
VM
nlb-toa
Haproxy-toa
HOST
OPTOA
server
server
server
... ...
HOST
OPTOA
server
server
server
... ...
HOST
OPTOA
server
server
server
... ...

服务发现


Backend Pod 1
labels: app=MyApp
port: 9376
Backend Pod 2
labels: app=MyApp
port: 9376
Backend Pod 3
labels: app=MyApp
port: 9376
apiserver
1) watch Services
and Endpoints
kube-proxy
5) proxy to a backend
2) open proxy port
and set portal rules
4) redirect to (random)
proxy port
iptables
3) connect to 10.0.0.1:1234
Client

服务发现



租户之间容器网络完全隔离，无需配置多余的转发规则

只watch本租户的Service，生成iptable 规则

服务发现

摘要：微服务化是当前电商产品演化的必然趋势，网易考拉海购通过微服务化打破了业务爆发增长的架构瓶颈。本文结合网易考拉海购引用的开源Dubbo框架，分享支持考拉微服务工作的基本原理。文章分析了使用Dubbo过程中遇到的问题，讲解了团队所做的一些问题修复和功能集成工作，在此基础上最终形成了考拉内部持续维护升级的Dubbok框架。

本文背景还要从网易考拉海购（下文简称“考拉”）微服务化说起，现在任何大型的互联网应用，尤其是电商应用从Monolithic单体应用走向微服务化已经是必然趋势。微服务化是一个比较宽泛的概念，涉及到一个产品生命周期的多个方面，首先它作为一个指导原则指引业务划分、架构解耦等；技术层面实施微服务需要开发测试阶段、运行阶段、发布阶段、部署阶段等一系列基础框架的支撑。我们在享受服务化易扩展易部署等便利性的同时，也面临新的问题，如数据一致性、分布式调用链路追踪、异常定位、日志采集等。

Headless services

With selectors

For headless services that define selectors, the endpoints controller creates `Endpoints` records in the API, and modifies the DNS configuration to return A records (addresses) that point directly to the `Pods` backing the `Service`.

Without selectors

For headless services that do not define selectors, the endpoints controller does not create `Endpoints` records. However, the DNS system looks for and configures either:

- CNAME records for `ExternalName`-type services.
- A records for any `Endpoints` that share a name with the service, for all other types.

客户端服务发现

简化外部配置

容器的使用场景

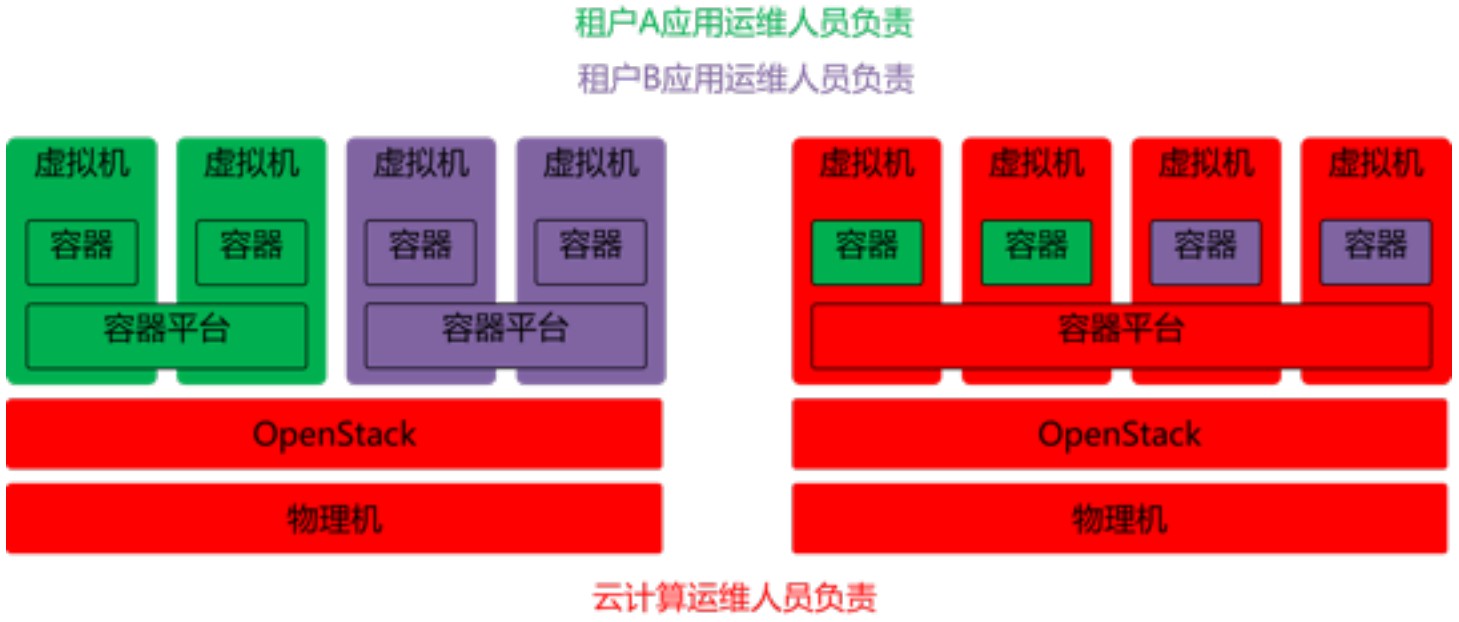
大规模云原生应用的支撑痛点

Kubernetes的性能问题

公有云的支撑痛点

Kubernetes的规模问题

公有云的支撑痛点



- 一个集群 or 多个集群
- Docker内核隔离不好
- Node无租户隔离
- 虚拟机启动速度问题
- Kube-proxy转发规则太多
- 集群规模问题

容器的使用场景

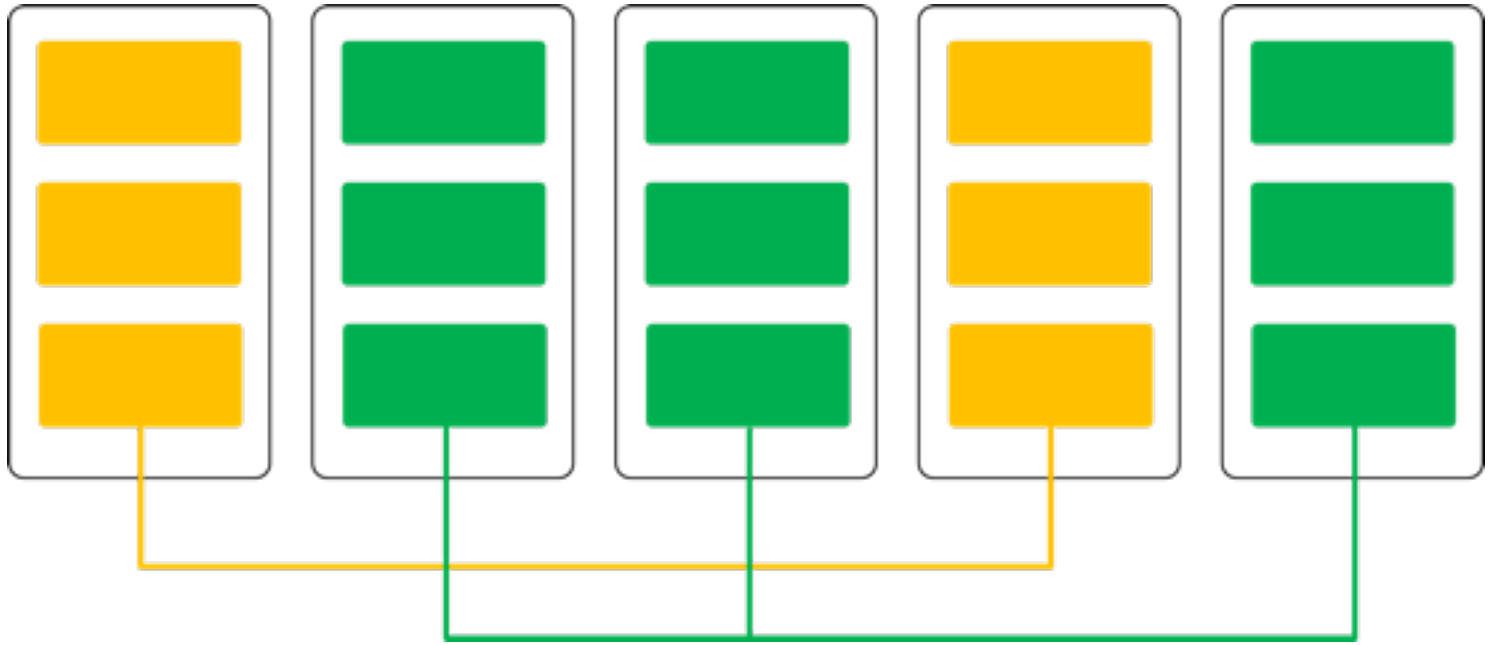
大规模云原生应用的支撑痛点

Kubernetes的性能问题

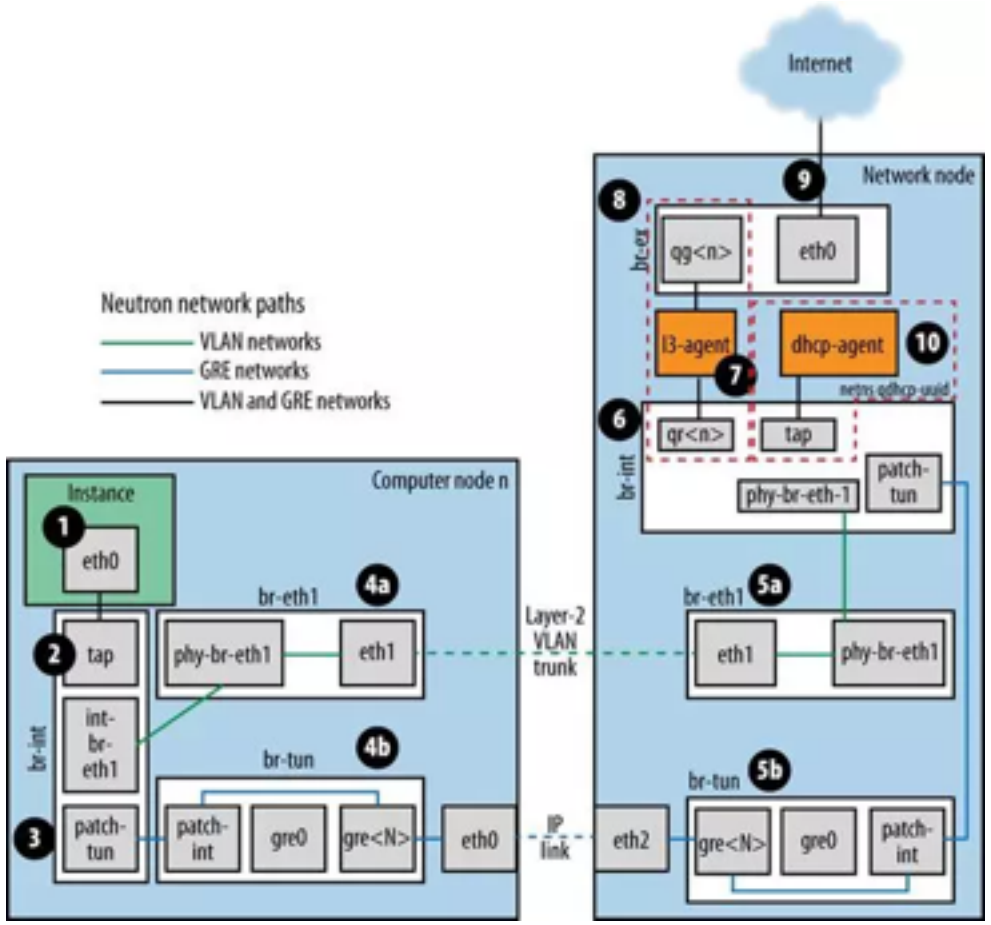
公有云的支撑痛点

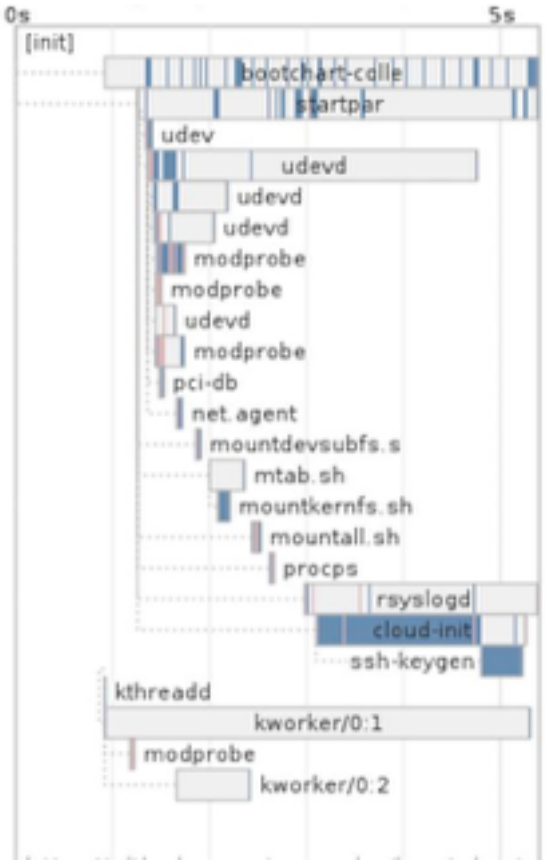
Kubernetes的规模问题

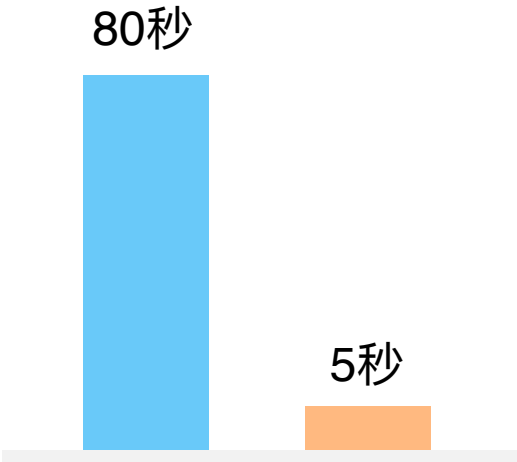
Kubernetes的kubelet运行容器



启动速度优化

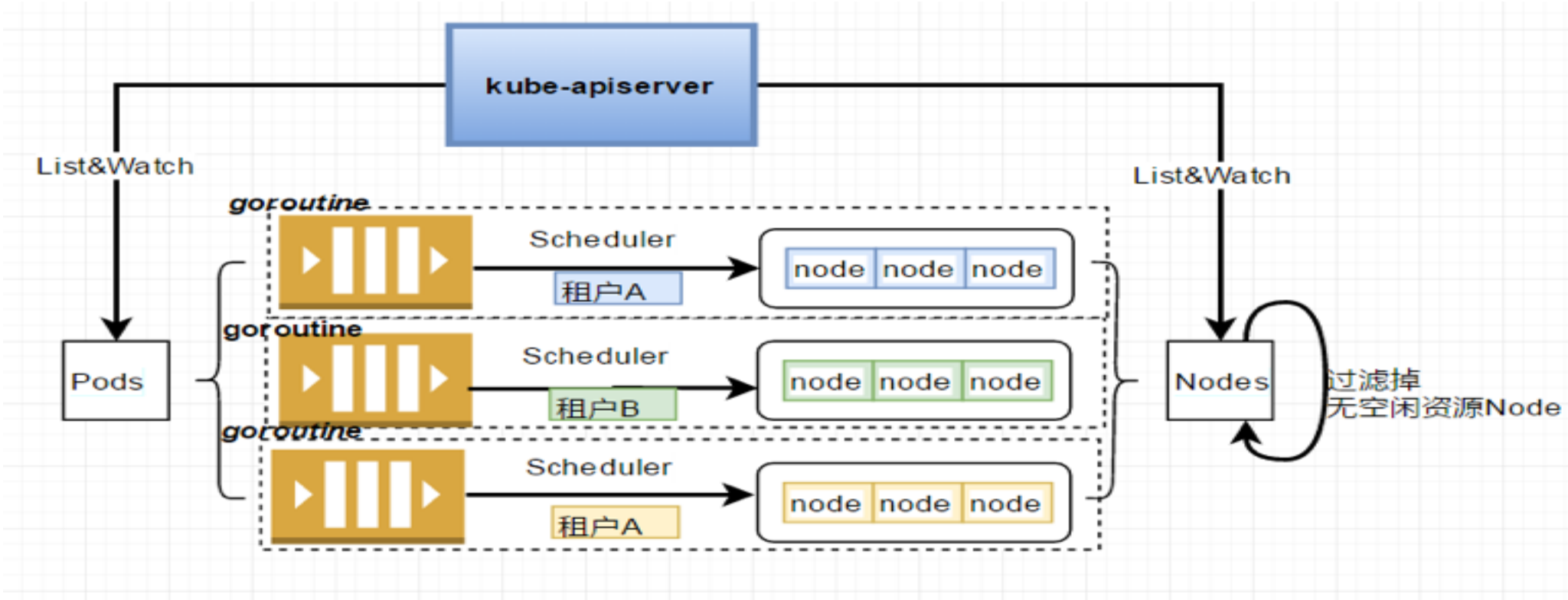






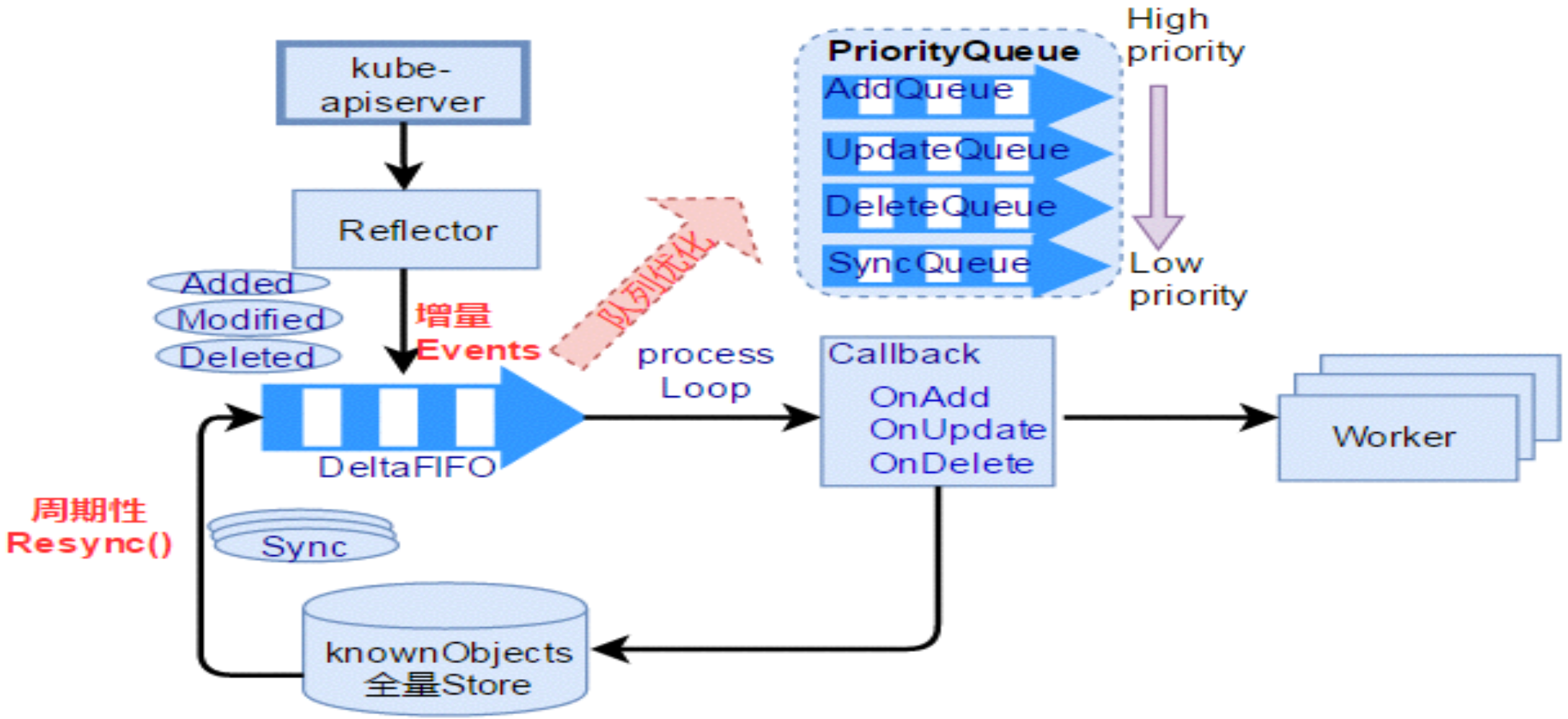
1.网卡IP初始化
2.网络路由注入
3.DNS服务IP配置
4.网卡udev规则

Scheduler优化



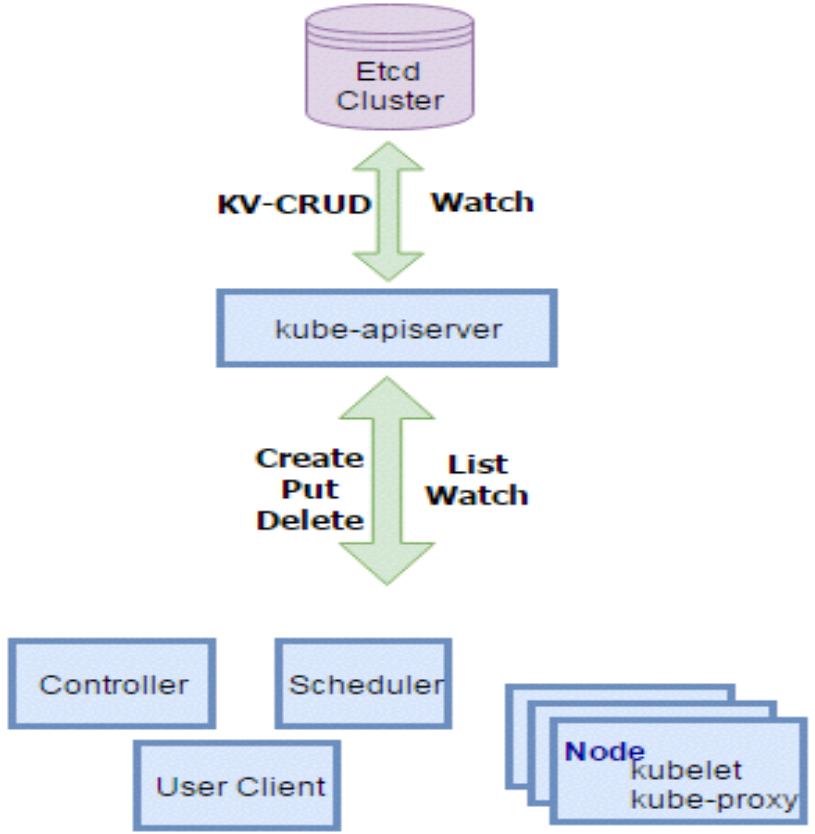
公有云租户间资源完全隔离，天然适合并行调度
预过滤无空闲资源的node
调整predicate调度算法快速过滤

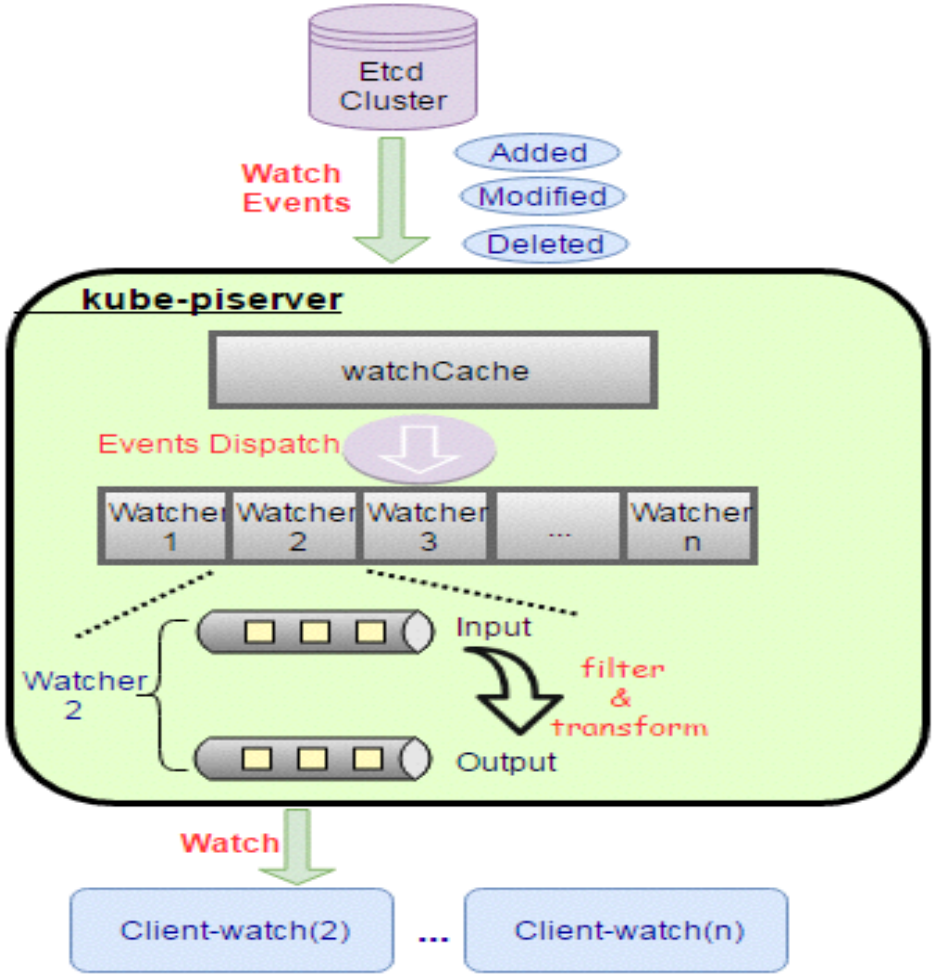
Controller优化



按事件类型进入多优先级workqueue（Add > Update > Delete > Sync）

集群规模优化





Apiserver 就是一个proxy代理，而且goroutine对web服务完美支持，最终性能瓶颈在对 etcd 的访问上

2017 China Kubernetes End User Conference
kubernetes 中国用户大会
KEUC 2017
2017.10.15 / 中国·杭州

Q&A

Thank you for your time

主办：

CLOUD NATIVE COMPUTING FOUNDATION

caicloud 才云

中国社区