基于kubernetes的网易云容器服务的持续升级实践

娄超

网易云容器编排服务

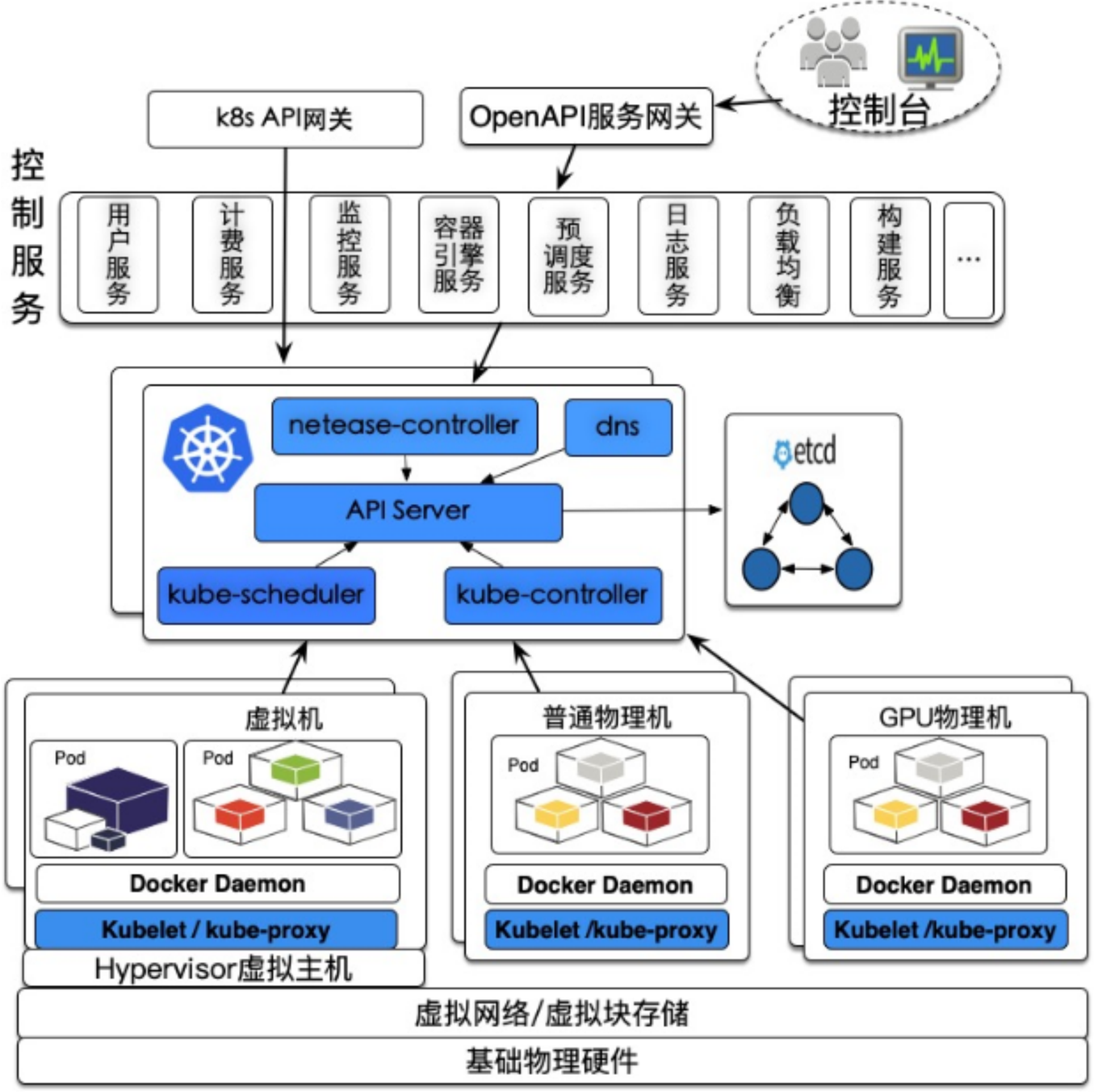
TABLE OF

CONTENTS 大纲

- 网易云容器服务介绍
- 网易云容器服务的发展演进
- 版本持续升级面临的挑战
- 网易云容器如何在线升级
- 经验教训

网易云容服务介绍

- 前身蜂巢，网易云第一批上线的云服务
- 第一个基于kubernetes（k8s）提供多租户的微服务容器平台
- 与网易云IAAS无缝融合，单集群虚拟机、物理机和GPU异构节点混部
- 最早提供无服务器架构（serverless）的容器平台
- 在线上持续运行超过1000天


控制台
k8s API网关
OpenAPI服务网关
控制服务
用户服务
计费服务
监控服务
容器引擎服务
预调度服务
日志服务
负载均衡
构建服务
...
netease-controller
dns
API Server
kube-scheduler
kube-controller
etcd
虚拟机
Pod
Pod
Docker Daemon
Kubelet / kube-proxy
Hypervisor虚拟主机
普通物理机
Pod
Docker Daemon
Kubelet /kube-proxy
GPU物理机
Pod
Docker Daemon
Kubelet /kube-proxy
虚拟网络/虚拟块存储
基础物理硬件

网易云容器发展演进

- **初期：网易容器服务1.0**
  - 网易云首个为开发者打造的容器云CAAS平台
  - 基于k8s v1.0二次开发，提供无状态/有状态两种workload
  - 支持多租户，集成IAAS虚拟机、经典虚拟网络、外网负载均衡
  - 容器调度所需Node由集群自动实时创建

网易云容器发展演进

- **发展期：网易容器服务2.0**
  - 基于k8s v1.3/v1.6二次开发
  - 服务发现，L7/L4(定制) Ingress，经典虚拟网和VPC两种网络
  - 单用户多Namespace，定制的有状态容器
  - 新增高性能物理机和GPU容器
  - 全面性能优化和架构重构

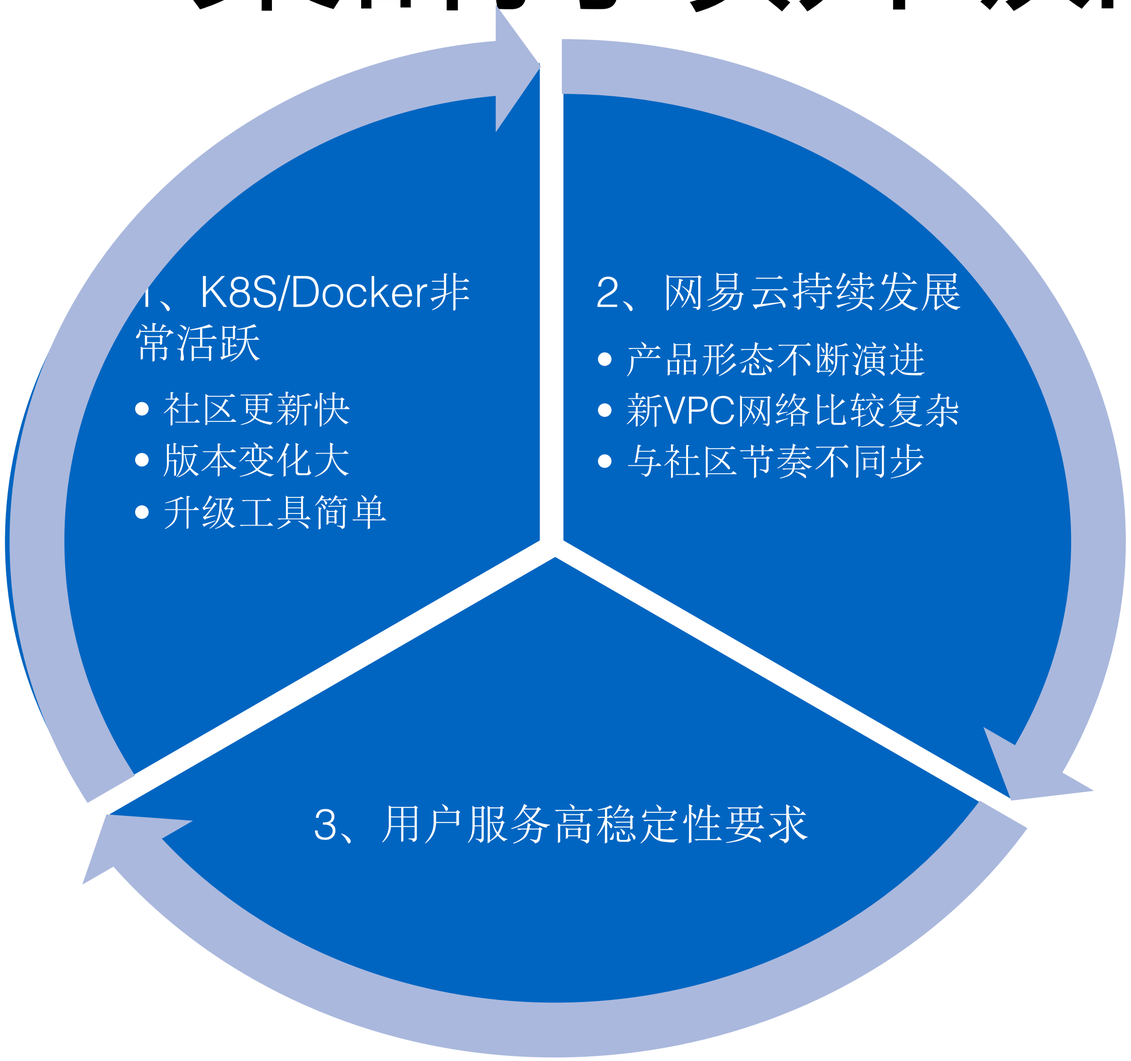
网易云容器发展演进

- **融合期：网易容器服务3.0**
  - 符合 Kubernetes conformance 统一异构容器平台
  - 同时支持k8s API和Serverless模式OpenAPI
  - 控制台开放node/pv外的k8s资源
  - 单用户支持多k8s集群
  - 容器网络支持VPC自定义路由模式

开源系统线上升级关键问题

- 何时升级、升到哪个版本
- 新老版本详细差异
- 内部定制部分新老版本如何兼容
- 是否有成熟升级工具
- 线上老系统用户有哪些，有哪些用法
- 升级对业务有什么影响

定制k8s集群持续升级的挑战



基于k8s自研的网易云容器版本持续演变升级

演进阶段	社区版本	主要新增特性	关键特色	架构缺点
早期版本1.0 （初期）	K8s 1.0 docker 1.7.2	ReplicateController / Pod IAAS计算、网络、负载均衡 keystone认证	多租户支持 容器需要计算、网络资源 自动扩缩	对k8s代码修改太随便， 扩展性差 需要改用新版插件，独立出 netese-controller
微服务版本 2.0 （发展期）	K8s 1.3 docker 1.12	Deployment / Service / Ingress / PersistenceVolume / Namespace 保存镜像、容器重启，容器垂直扩容 有状态容器rootfs故障迁移	自研基于规则访问控制 容器rootfs/ip持久保持 创建流程性能优化 （如Node/PV动态资源池）	iaas资源动态扩缩使得调度 器逻辑复杂、维护困难 需要将IAAS资源动态分配回 收拆分为IAAS资源控制器
内部优化完 善版本2.5 （发展期）	K8s 1.6 Docker 1.12	自定义Statefulpod vpc网络/裸机/GPU PV改用Flevolume 容器proc性能显示纠正 自定义Task管理API（类似1.7的CRD）	rootfs优化、有状态 StatefulPod 全链路优化和规模提升 （apiserver索引、并行调度、 租户分组）	多租户、VPC网络、有状态 容器扩展与原生API不兼容
标准化版本 3.0 （融合期）	K8s 1.9 Docker 1.13	支持OpenAPI和k8s API两种模式 VPC自定义容器网络 Node证书控制器生成 多租户授权改用RBAC	符合Conformance 基本实现代码无侵入	/

1.0升级到2.0

2.0升级到2.5

2.5升级到3.0

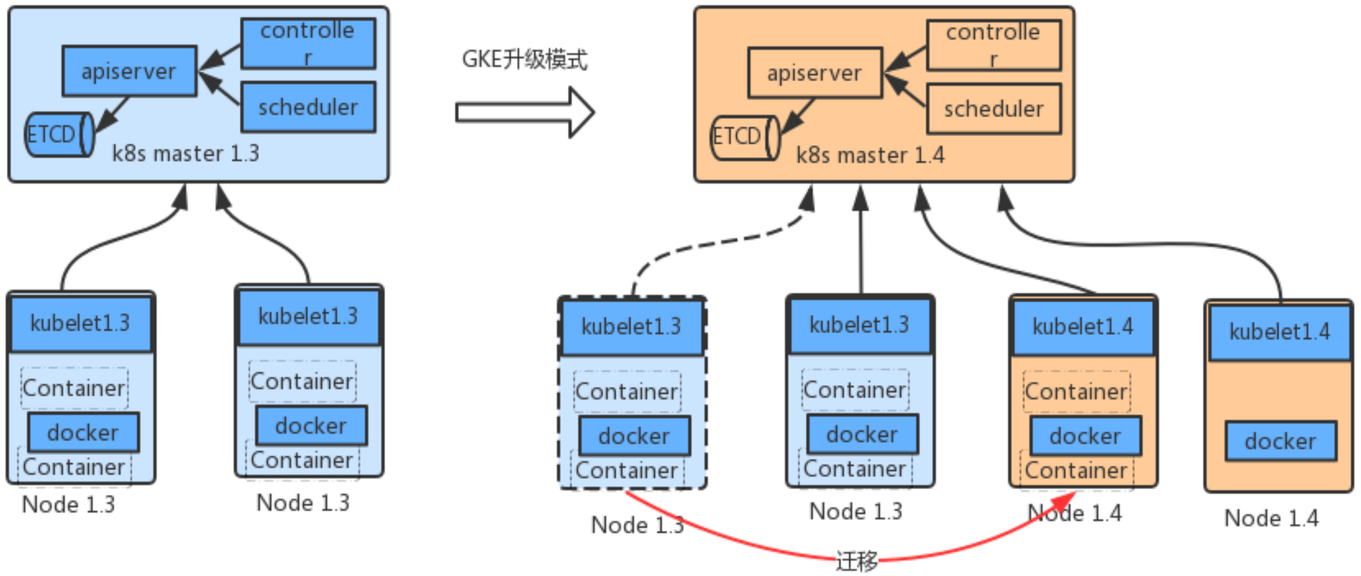
kubernetes社区升级建议

- K8s社区design-proposals
  - no upgrading more than two minor releases at a time
- 相关运维工具
  - kubeadm、juju（Ubuntu）、kops（AWS）、kubespray
  - 不建议生产环境用
- API兼容与格式转化
  - kubectl convert

GKE 的升级模式

- 升级顺序
- Master 一键升级或自动升级
  - 类似更新static pod manifest的配置
- Nodes 用户手动滚动更新或配置自动升级
  - *kubectl cordon <node_name>*
  - *kubectl drain <node_name> --force*

GKE 的升级模式特点

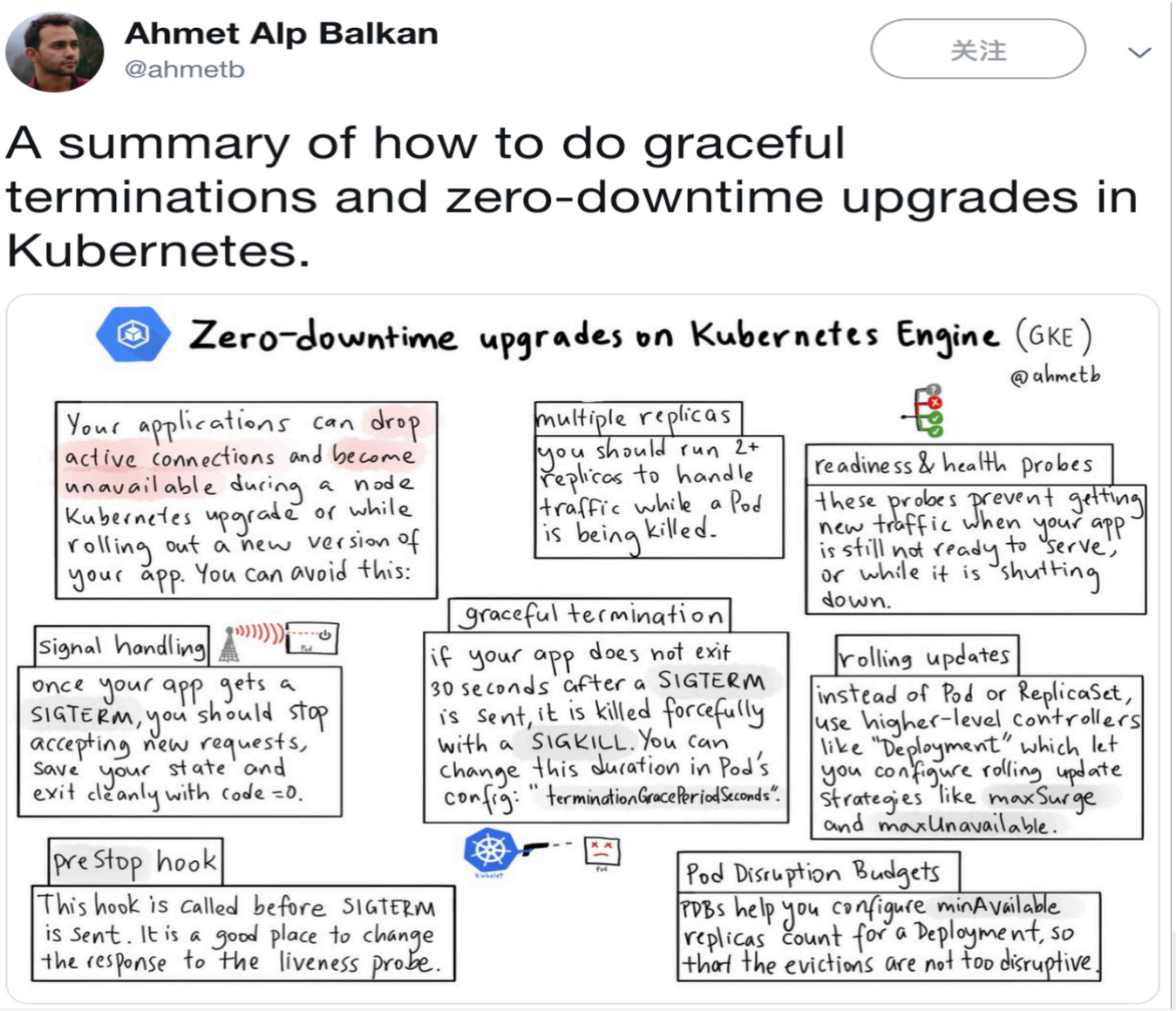


- Master与Nodes 需跨版本并存 (component-skew)
- 驱赶所有老Node上用户容器

对普通容器有影响

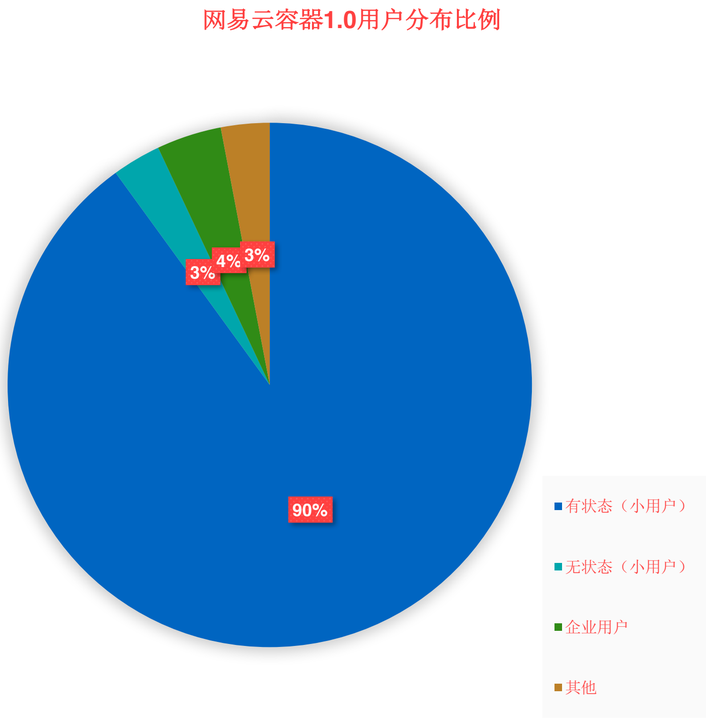
How ?

- Zero-downtime 升级
  - 多Pods
  - 健康探活
  - SIGTERM
  - 优雅终止时间
  - 容忍度
  - Pods跨node分布
  - ...



1.0升级到2.0

- 网易云容器初期v1.0线上用户情况：
  - 有状态单体容器占90%，用法随意
  - 个人小用户最多
  - 企业用户很少，新业务为主（三拾众筹）



容器“非主流”现象

- 容器要能ssh，登录修改、手动启停程序
- 超大镜像磁盘直接占满
- 直接访问IP地址
- 直接开容器作开发机、测试机
- 不挂盘数据直接写rootfs
- 镜像Tag覆盖重用
- …

1.0升级到2.0

- K8s 1.0与1.3兼容问题：
  - K8s 1.3要求docker>=1.9
  - Dockerd v1.9+ 无法接管v1.7容器实例
  - k8s的 RC（ReplicationController）废弃
  - API不兼容

升级Node容器必须重建

- 网易云容器v1.0升级v2.0主要变化：
  - 无状态的workload从RC改为Deployment
  - 与NLB负载均衡对接的Ingress
  - 与NBS块存储对接的PV云盘
  - 外网IP及容器IP绑定
  - 容器网络支持Service服务发现

1.0升级到2.0

- Case1：无状态容器
  - 新版本新建容器关联NLB，负载均衡跨k8s集群流量切换
- Case2：自带Consul/Dubbo服务注册服务发现
  - K8s新老集群底层虚拟网络互通
- Case3：有数据的有状态容器
  - 一键升级工具：commit替换本地镜像，再升级docker/kubelet

2.0升级到2.5

- K8s 1.3到1.6升级主要变化：
  - 后端数据库：etcd v2升级为v3
  - 元数据存储编码：json升级为protubuf
  - API通信：http升级为http2.0（GRPC）
  - …
- 网易云容器v2.0升级v2.5主要变化：
  - 定制有状态StatefulPod控制器
  - 容器网络支持新版VPC
  - 单集群混合异构计算Node
  - 集群最大规模优化
  - …

Master上Pod等元数据手动运维

是否存在热升级方案？

- GKE模式的局限性
  - 1、要求API兼容邻近版本
    - 1.X to 1.X+1
  - 2、容器先停止再异地重建
    - 依赖用户服务架构
    - 无本地依赖
  - 3、完全社区版本

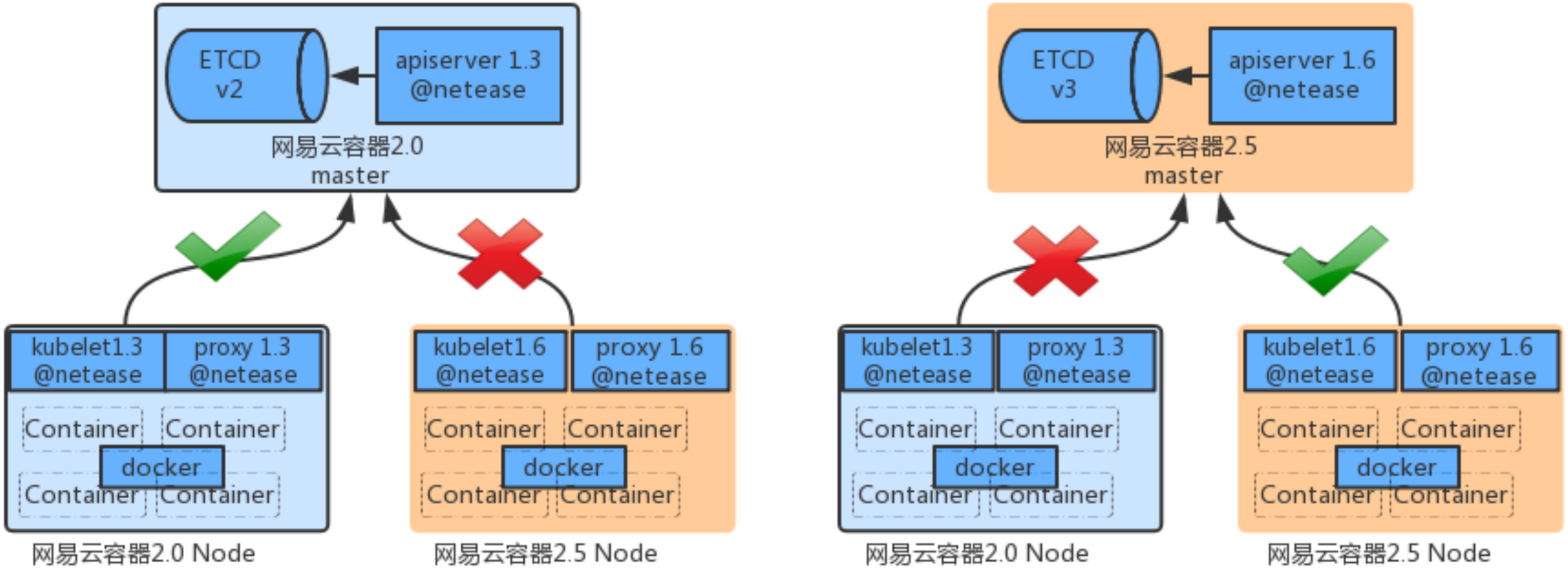
- 如何cover所有容器，真正无感知？
  - Node不迁移
  - 容器不能停

Node In-Place 升级

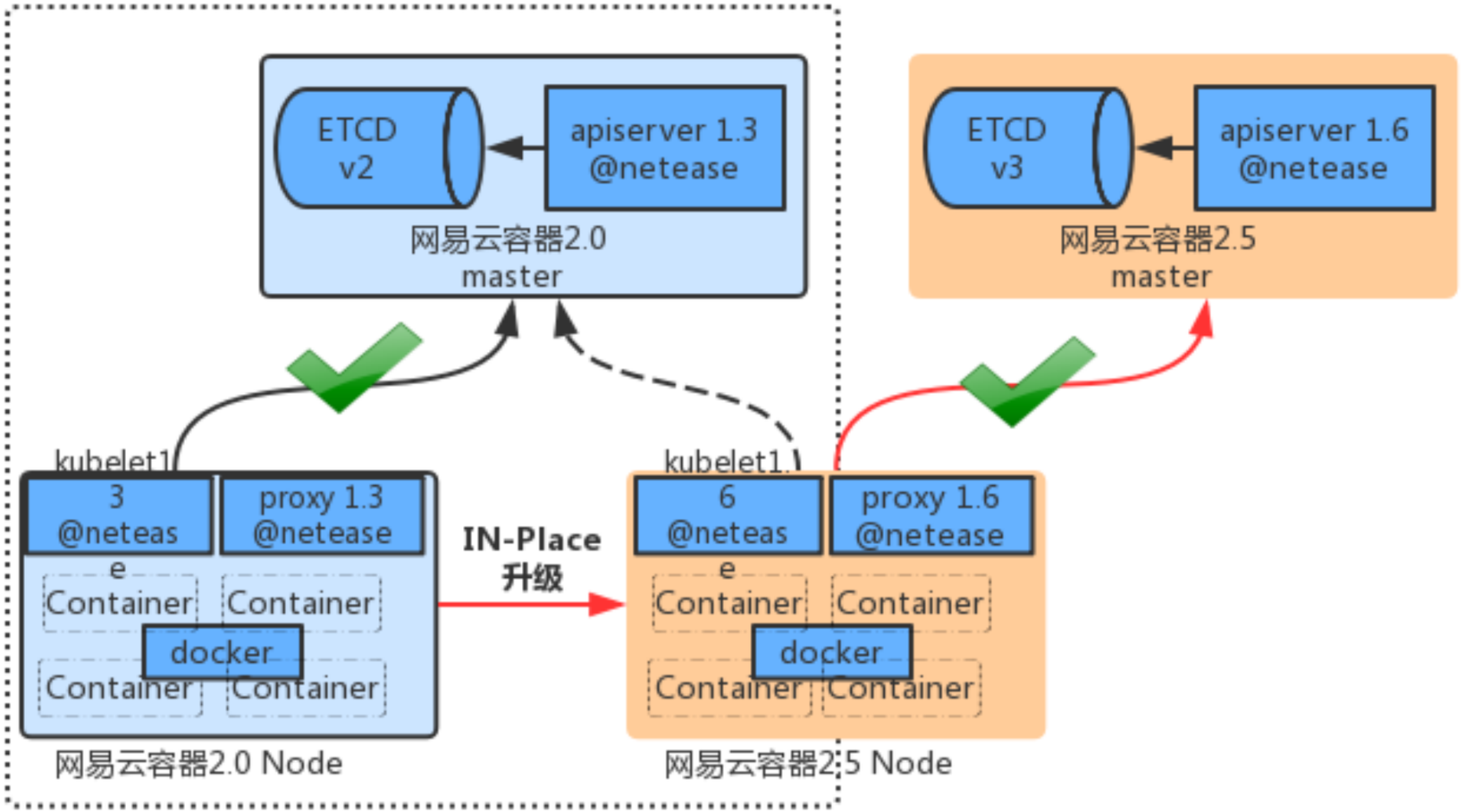
Node In-Place 热升级

- 如何容器不重启？
  - Docker Daemon 无缝升级
  - Kubelet v1.3无缝接管v1.0老Pod（包括pause）
- 如何网络不中断？
  - 容器网络方案兼容
  - kube-proxy转发模式兼容
- 如何容器本地数据不变？
  - 容器rootfs数据保持

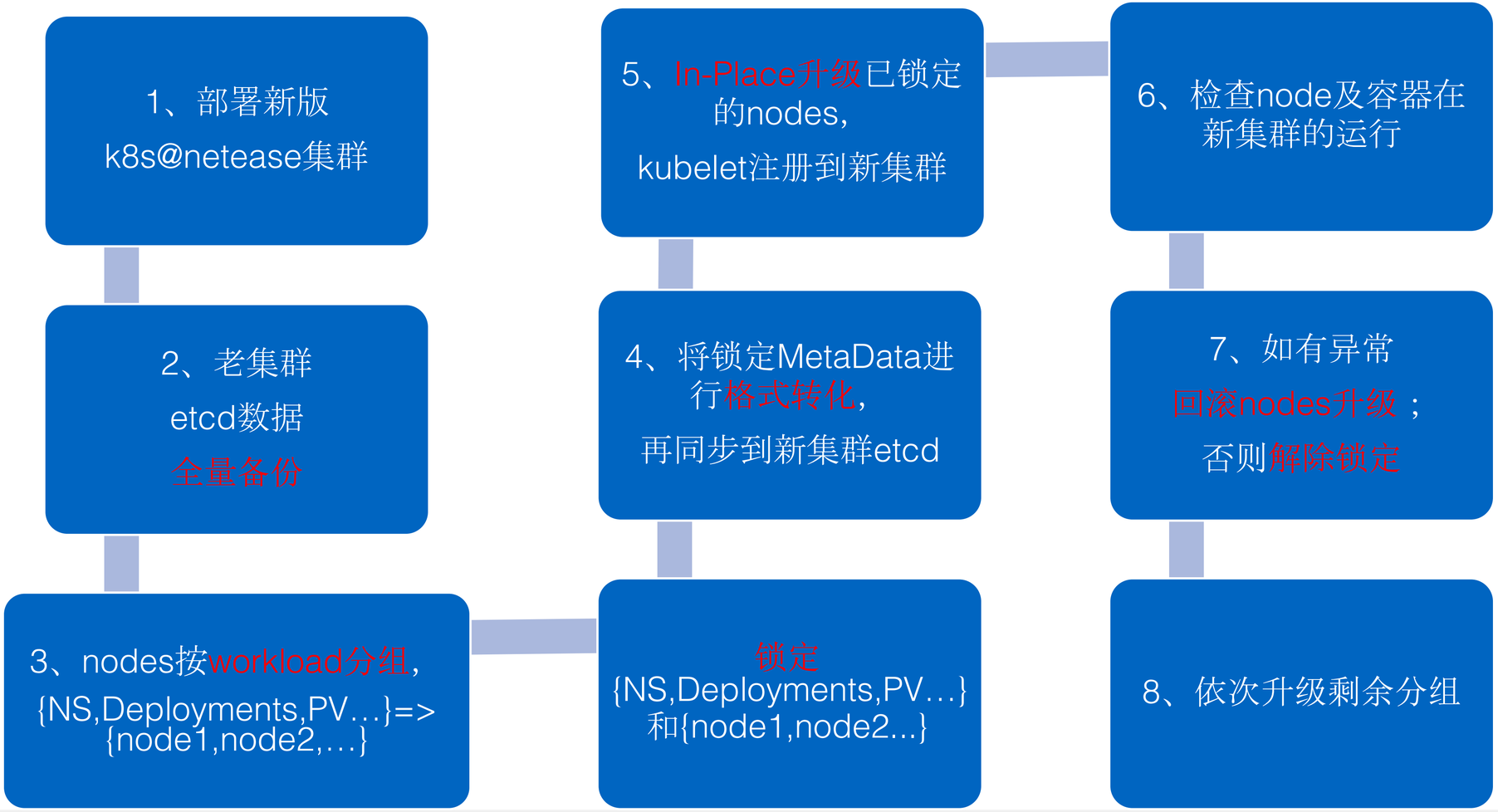
Master与Node不兼容



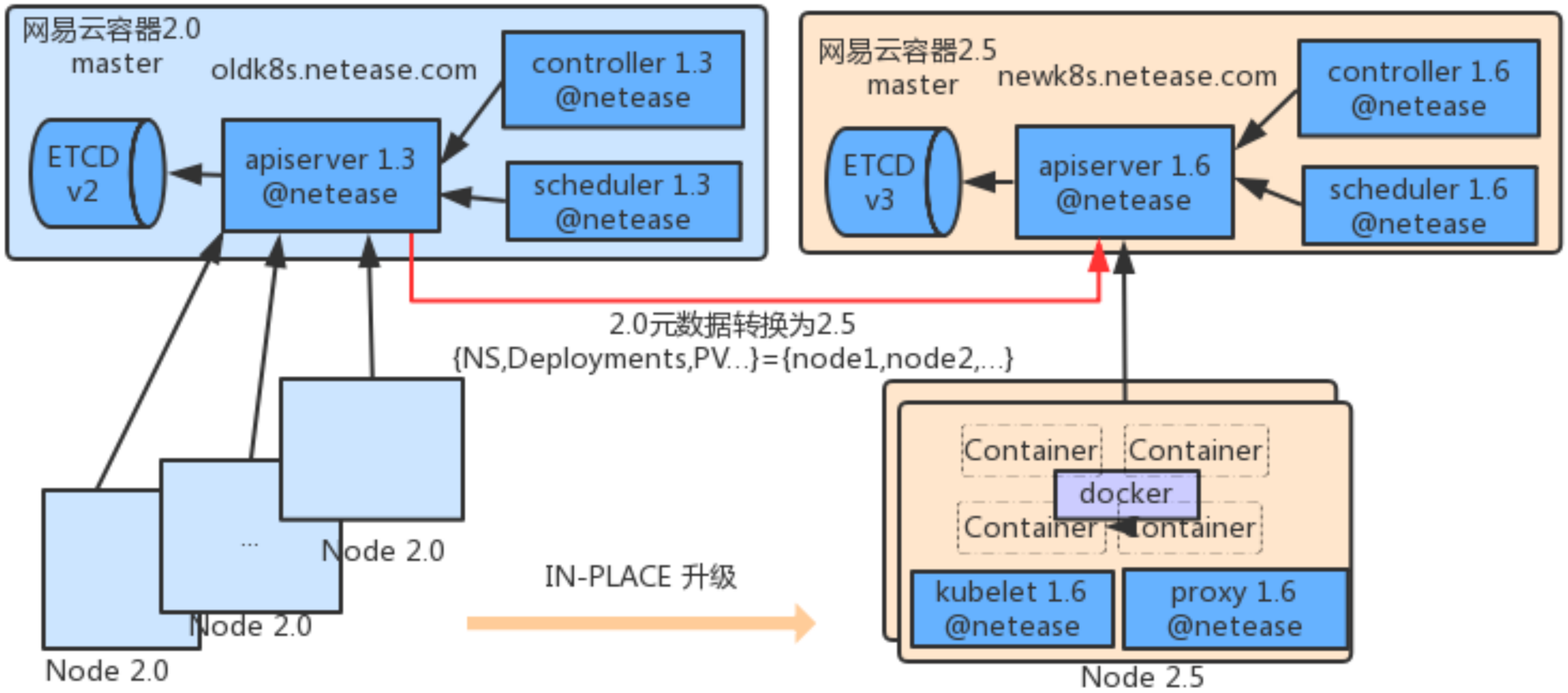
如何兼容


ETCD
v2
apiserver 1.3
@netease
网易云容器2.0
master
ETCD
v3
apiserver 1.6
@netease
网易云容器2.5
master
kubelet1
3
@neteas
proxy 1.3
@netease
e
Container
Container
docker
Container
Container
网易云容器2.0 Node
IN-Place
升级
kubelet1.
6
@neteas
proxy 1.6
@netease
e
Container
Container
docker
Container
Container
网易云容器2.5 Node

集群灰度升级流程


1、部署新版
k8s@netease集群
2、老集群
etcd数据
全量备份
3、nodes按workload分组，
{NS,Deployments,PV…}=>
{node1,node2,…}
锁定
{NS,Deployments,PV…}
和{node1,node2...}
4、将锁定MetaData进行格式转化，
再同步到新集群etcd
5、In-Place升级已锁定的nodes，
kubelet注册到新集群
6、检查node及容器在新集群的运行
7、如有异常
回滚nodes升级；
否则解除锁定
8、依次升级剩余分组

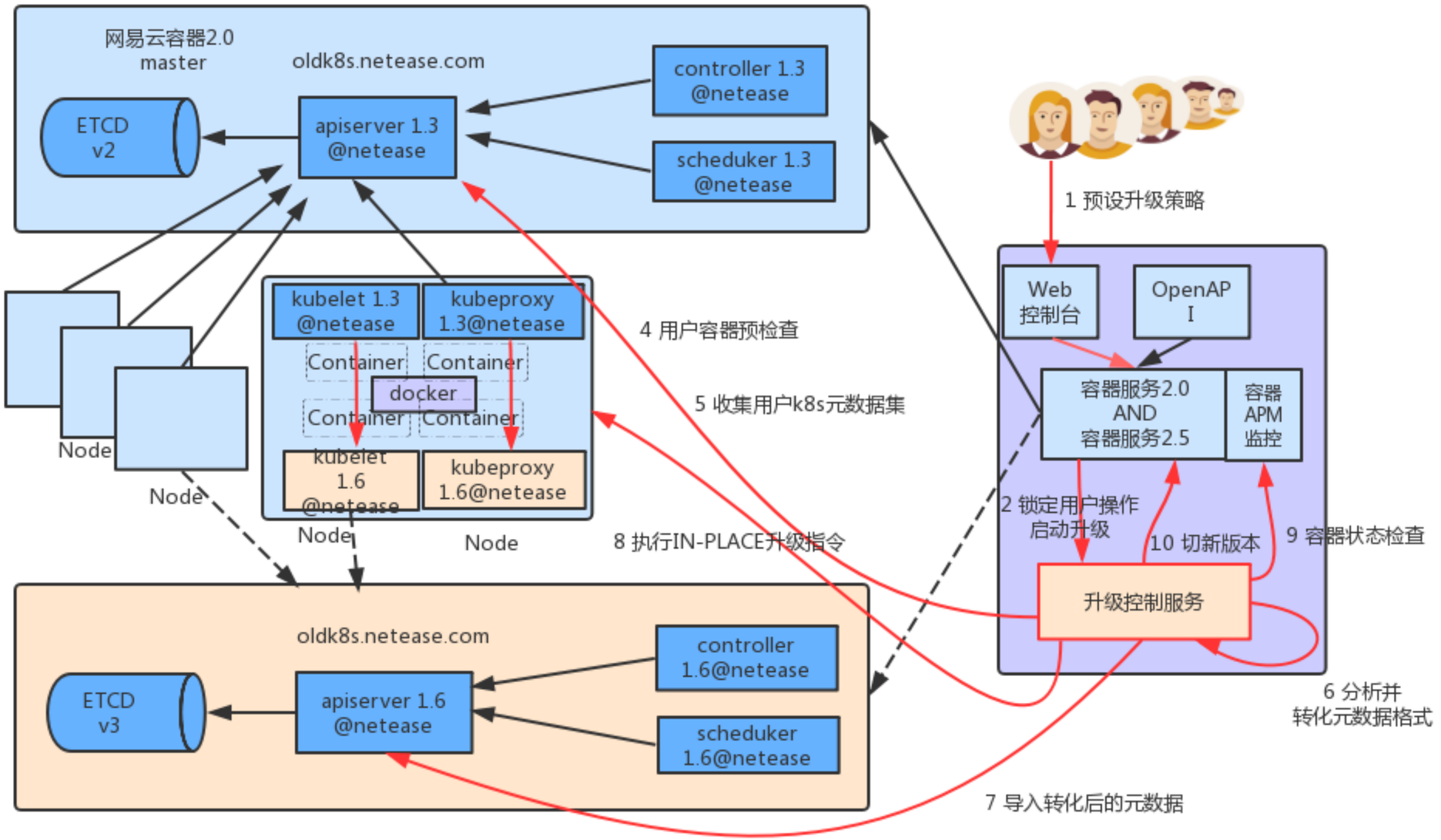
集群灰度升级示意图


网易云容器2.0
master
oldk8s.netease.com
controller 1.3
@netease
ETCD
v2
apiserver 1.3
@netease
scheduler 1.3
@netease
网易云容器2.5
master
newk8s.netease.com
controller 1.6
@netease
ETCD
v3
apiserver 1.6
@netease
scheduler 1.6
@netease
2.0元数据转换为2.5
{NS,Deployments,PV...}={node1,node2,...}
...
Node 2.0
Node 2.0
Node 2.0
IN-PLACE 升级
Container
Container
docker
Container
Container
kubelet 1.6
@netease
proxy 1.6
@netease
Node 2.5

k8s升级采坑经历

- 升级kubelet 时上面老容器各种被Kill
- 老集群Node升级期间离线导致没法回滚
- kubelet的root-dir被mount覆盖问题
- 新版本kubelet新增参数问题
- 容器网络模型很变动要特别谨慎
- 集群升级时机与用户冲突

网易云容器服务到3.0升级



经验教训

- 1、开源选型
- 2、开源还是自研
- 3、开源上无侵入扩展

THANKS!