



QCon 全球软件开发大会
INTERNATIONAL SOFTWARE
DEVELOPMENT CONFERENCE

BEIJING 2017

容器技术在千万用户级企业的实践及 网络方案优化

SPEAKER / 王翺宇



促进软件开发领域知识与创新的传播



关注InfoQ官方信息
及时获取QCon软件开发者
大会演讲视频信息



扫码，获取限时优惠

ArchSummit
全球架构师峰会 2017 [深圳站]

2017年7月7-8日 深圳·华侨城洲际酒店
咨询热线：010-89880682

QCon

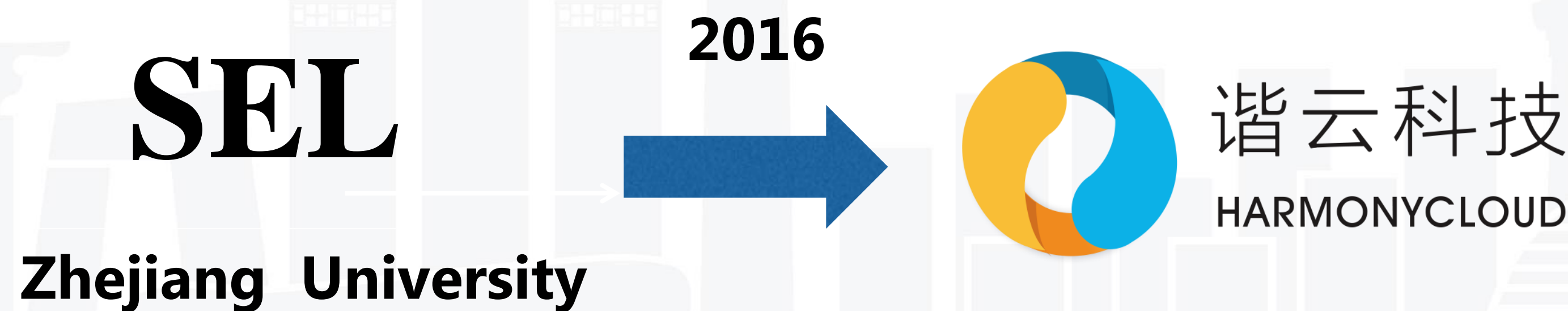
全球软件开发大会 [上海站]

2017年10月19-21日
咨询热线：010-64738142

自我介绍 - 王翱宇



- 谐云科技CEO & 联合创始人
- 来自浙江大学SEL实验室，曾就职于美国道富银行，浙大网新。
- 2005年毕业于浙江大学计算机学院



演讲内容简介

- 研究方向和产品定位
- 落地案例介绍
- 容器落地方案
 - 多集群方案
 - CI/CD
 - 应用迁移
 - 网络改造与优化

SEL实验室与容器技术书籍

SEL

Zhejiang
University



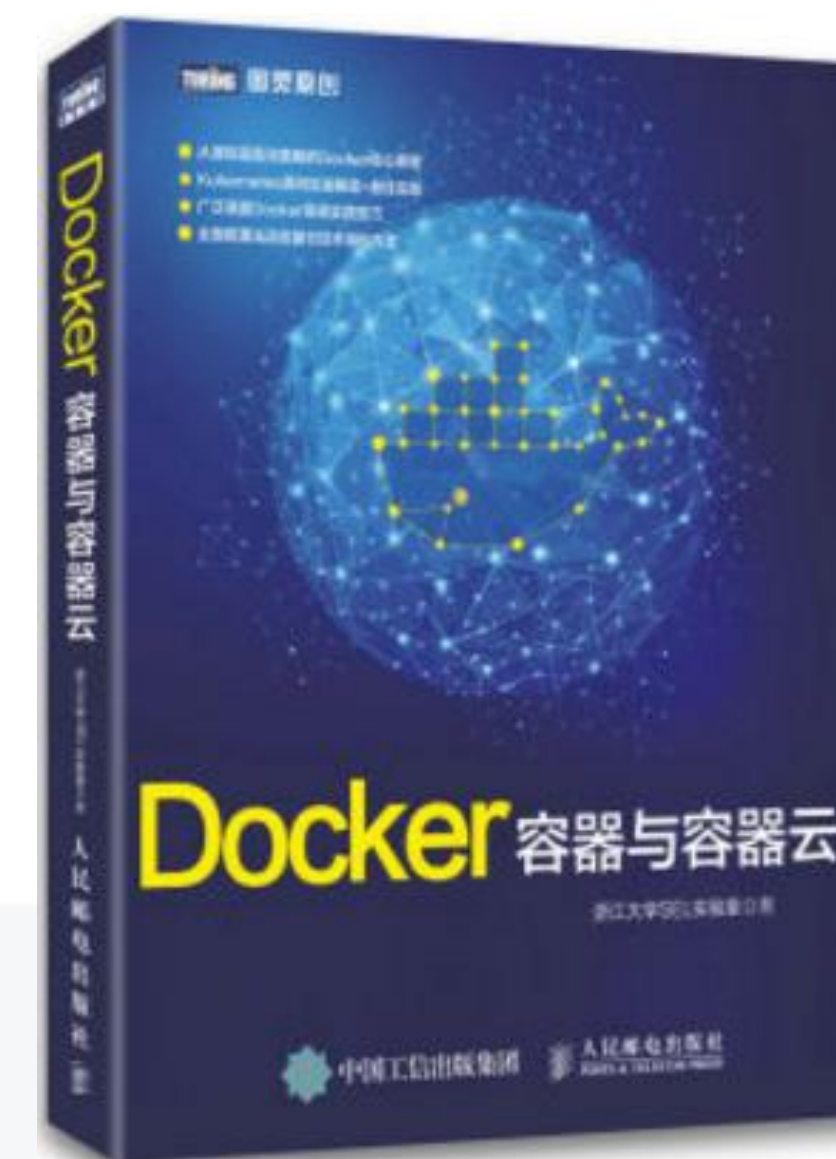
40+浙大博士
/硕士

2011

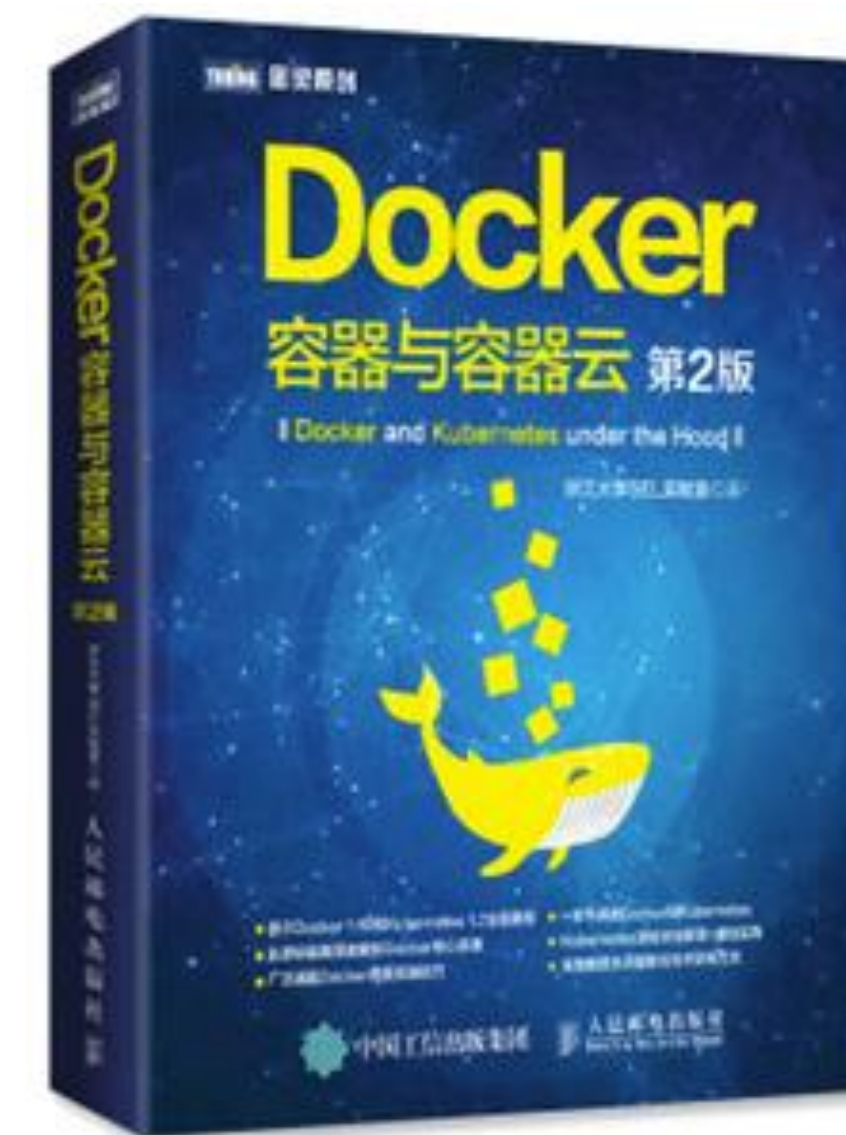
成立于2011年



专注开源
PaaS/CaaS技术



2015

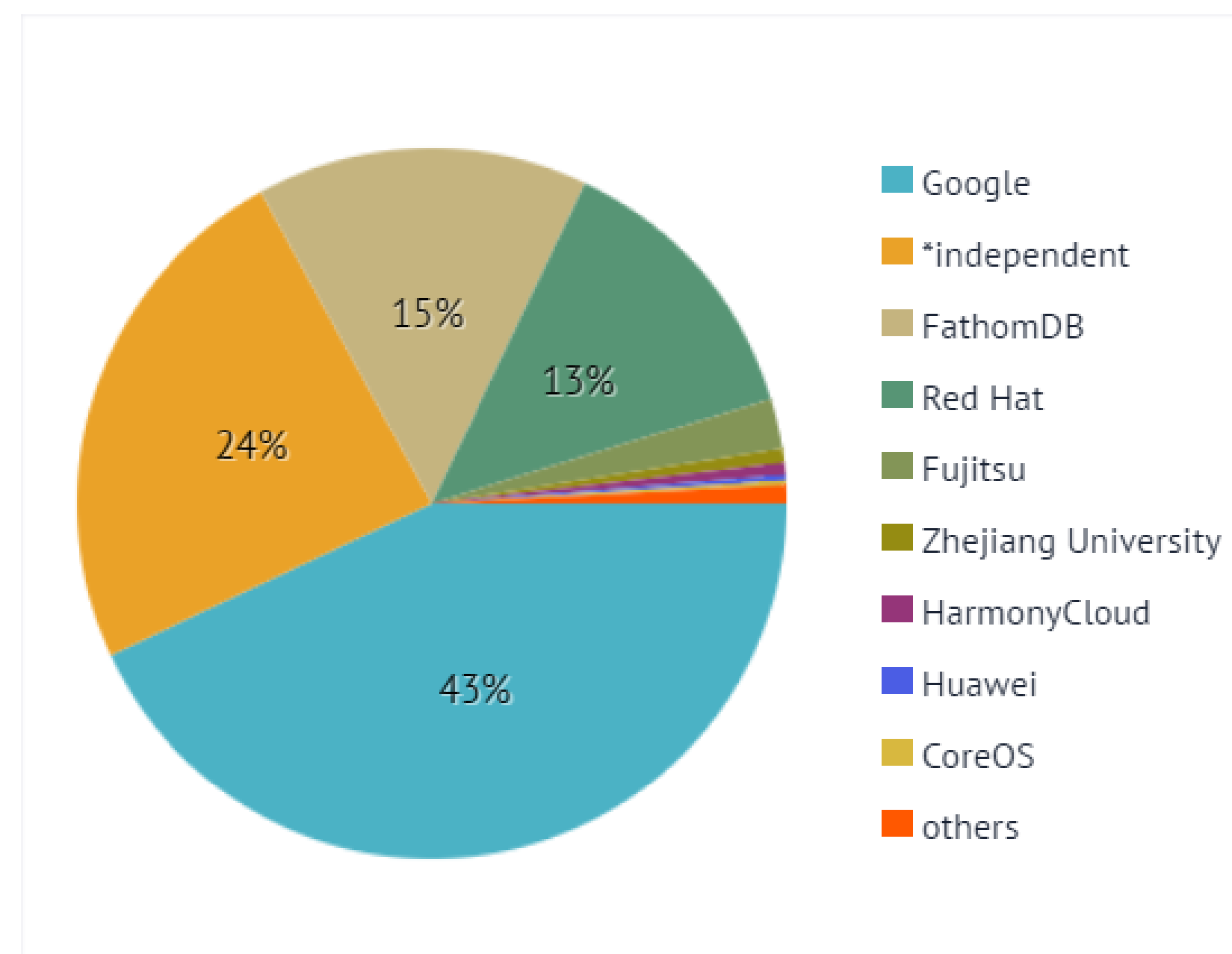


2016

积极参与贡献社区

- 贡献了**493**个Patch到Kubernetes，总共**142**万行代码
- 2**个Kubernetes feature maintainer

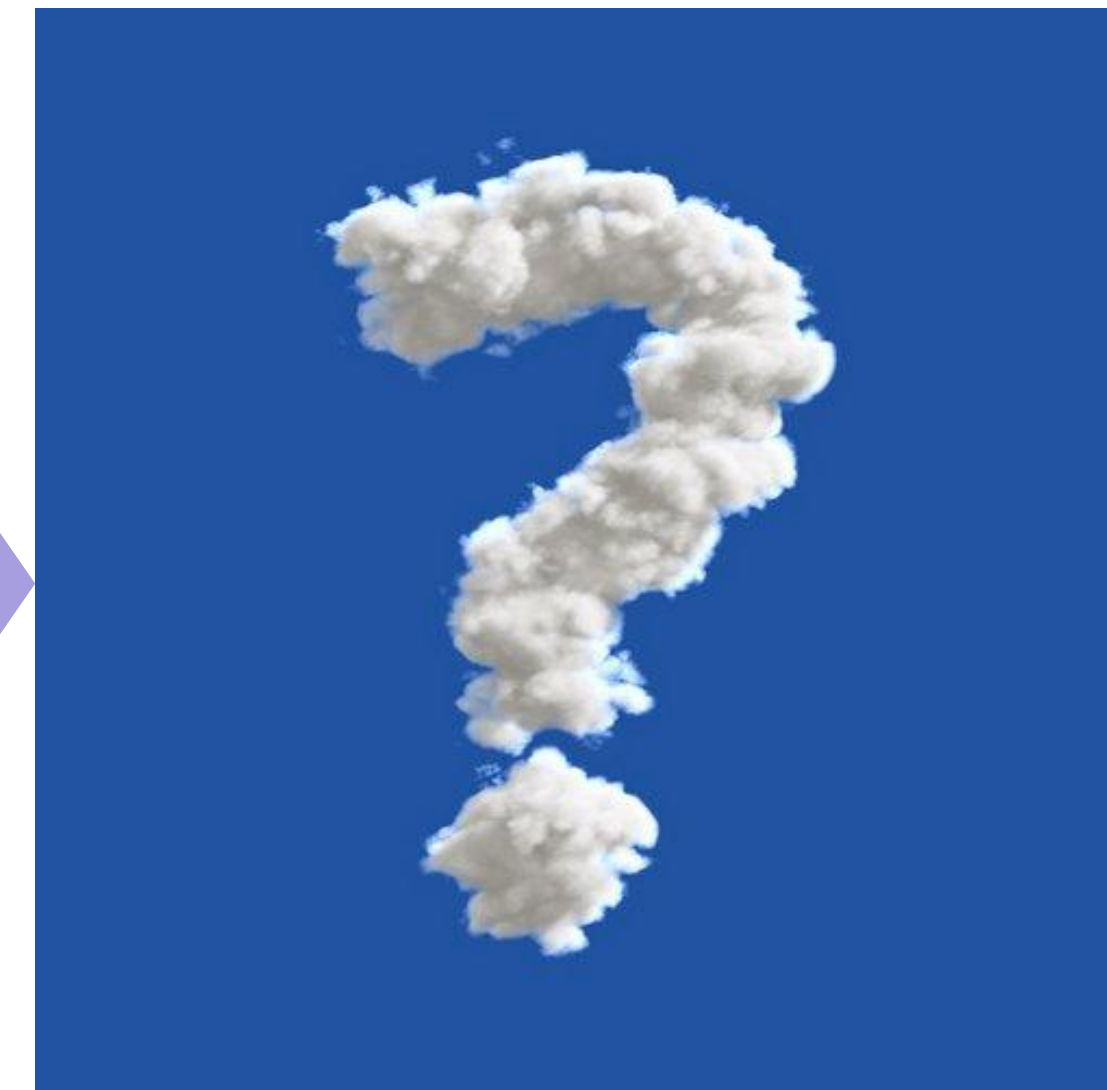
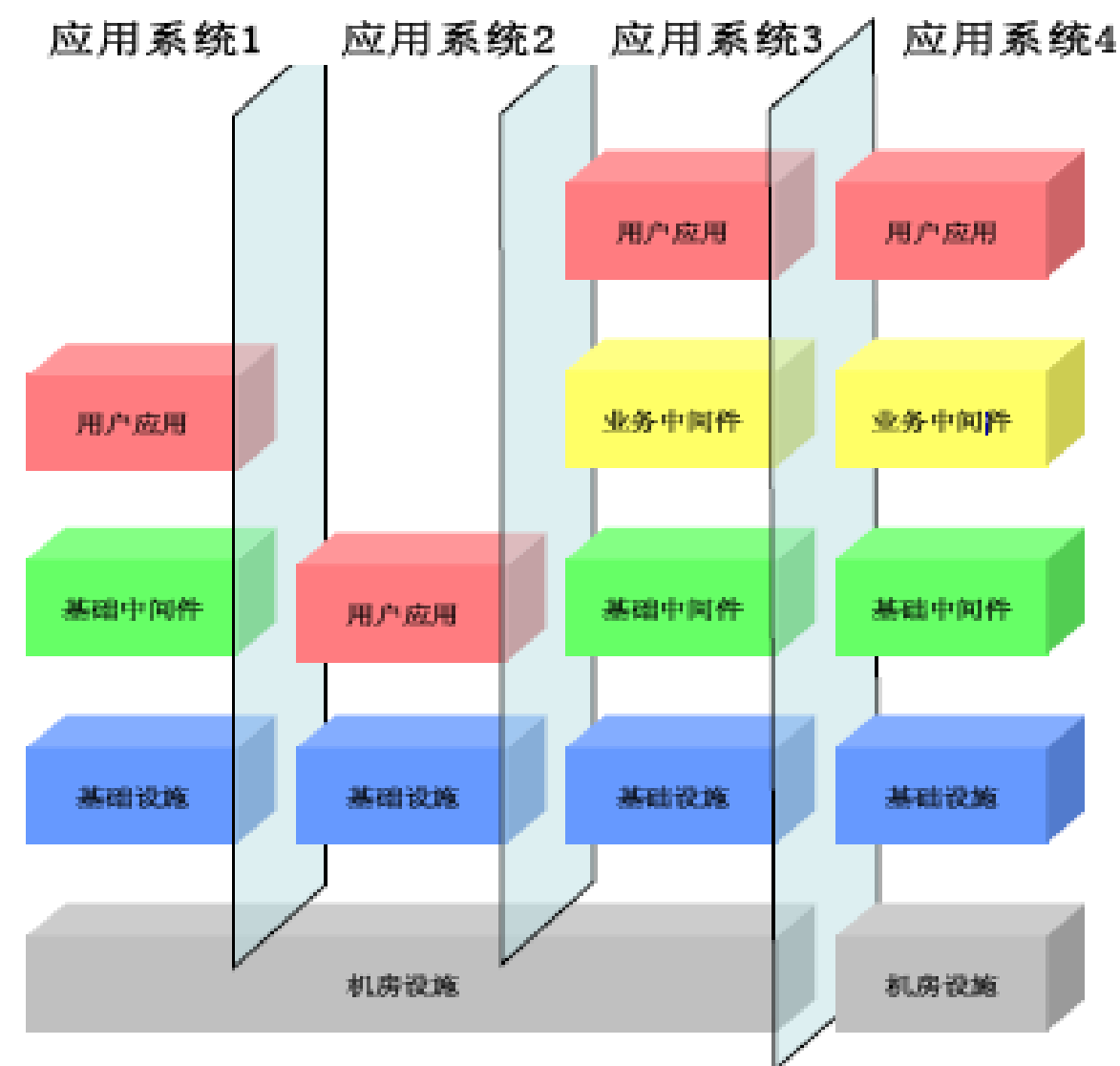
Contribution by companies



#	Company	Lines of code
1	Google	51550613
	*independent	28806338
2	FathomDB	18119611
3	Red Hat	15711748
4	Fujitsu	2755657
5	Zhejiang University	764824
6	HarmonyCloud	657697
7	Huawei	295721
8	CoreOS	253536
9	Intel	223840

IT向云端迁移的趋势与挑战

企业IT架构演进



竖井化、巨石型应用式IT系统架构

虚拟化？云平台化？

一个好的云平台

01

保障业务长期稳定、
高效的运行

02

快速响应业务上线

03

灵活的资源池，根据
客户业务量进行应用
调度

04

容错，及时甚至预测
性地发现、定位、诊
断故障，自我修复的
能力

落地案例-业务介绍

某互联网电视平台

在线视频

互联网电视在线视频点播平台，目前已与爱奇艺、腾讯视频、华数、ICNTV、优酷、优朋等视频服务商一起构建国内影视资源片库，为用户提供好片、大片、新片

在线教育

互联网电视在线教育平台，有超30万小时专业丰富的教学资源，97.2%的小初高同步教学课程来自国内重点名校一线教师，并覆盖幼教启蒙、品质生活兴趣课程，各类资源一应俱全，真正实现专业的全龄教育。

在线购物

互联网电视打造的大屏购物平台，覆盖电视、手机、白电等多终端，为用户提供一站式电视购物服务，支付便捷，并有专业的客服人员提供良好的售后服务。

在线娱乐

中国第一个精品电视游戏平台，囊括电视、AR、健身娱乐等各种类型游戏及周边外设，适合全家互动娱乐。游戏用户数稳步增长。

落地案例-业务介绍

国内最大的开放智能终端云平台

服务超过**3300万**智能电视和
智能家具用户

平台对终端和服务开放

服务稳定性达到**99.99%**

最丰富的运营服务内容

视频内容超过**100万**小时

国内最大的电视端教育资源库

国内应用数量最多的应用商店

互联网

终端

用户

服务

全球最大的互联网电视运营商

截至2017年3月底

该客户互联网电视用户数突破**2541万**

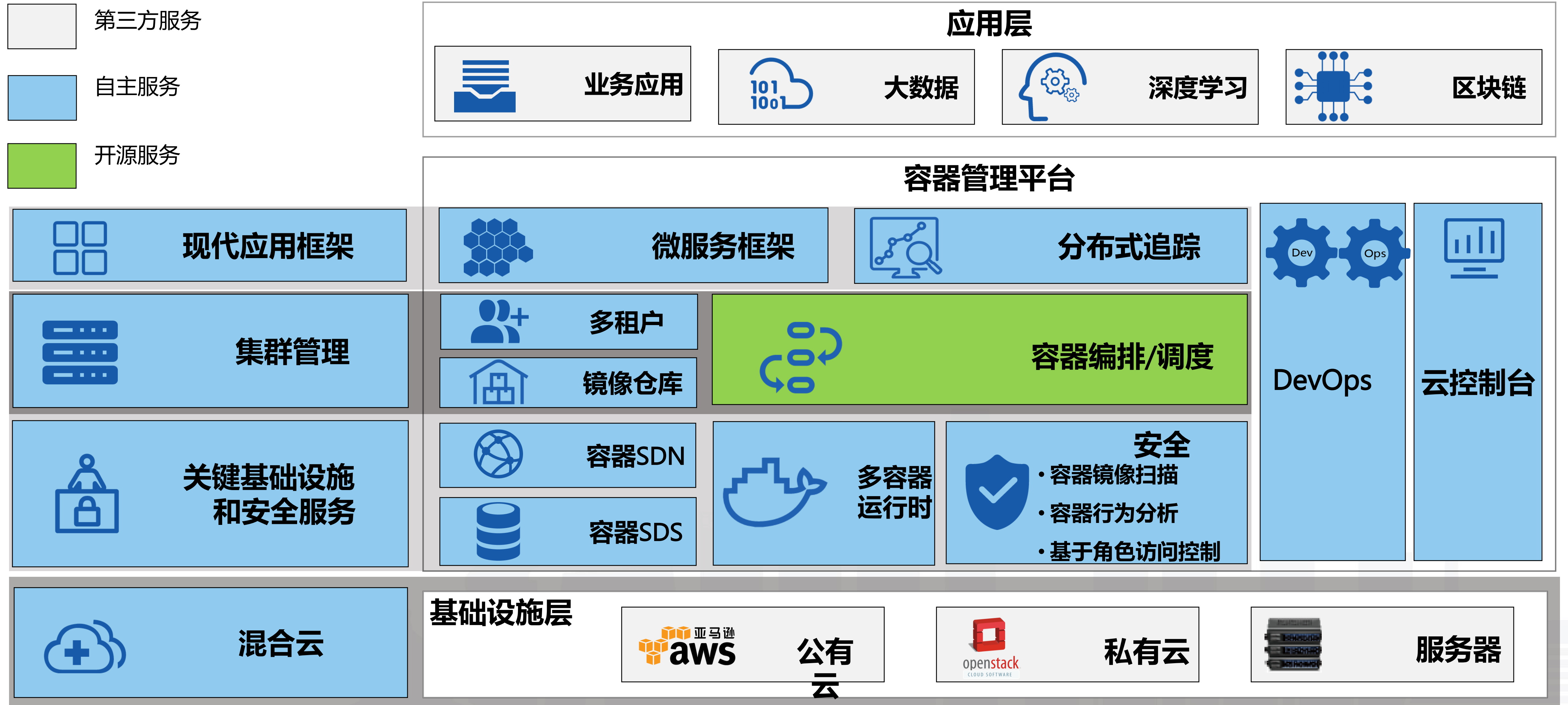
国内**2138万**海外**403万**

国内最活跃的线上智能终端用户社群

国内互联网电视用户日活跃DAU超**44.80%**

国内最活跃的用户社群

基于K8S的容器管理平台 - 观云台



多集群方案

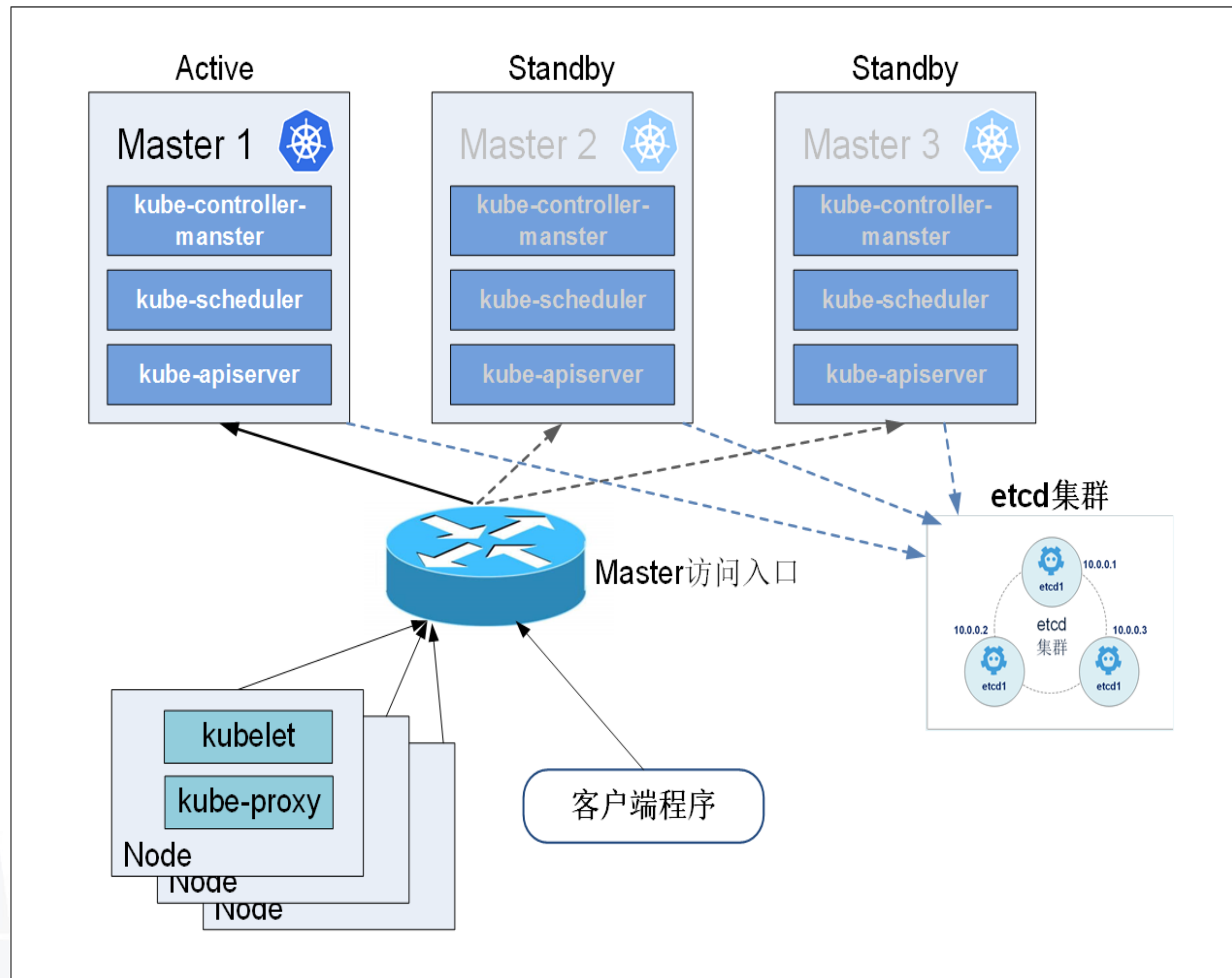


K8S集群高可用

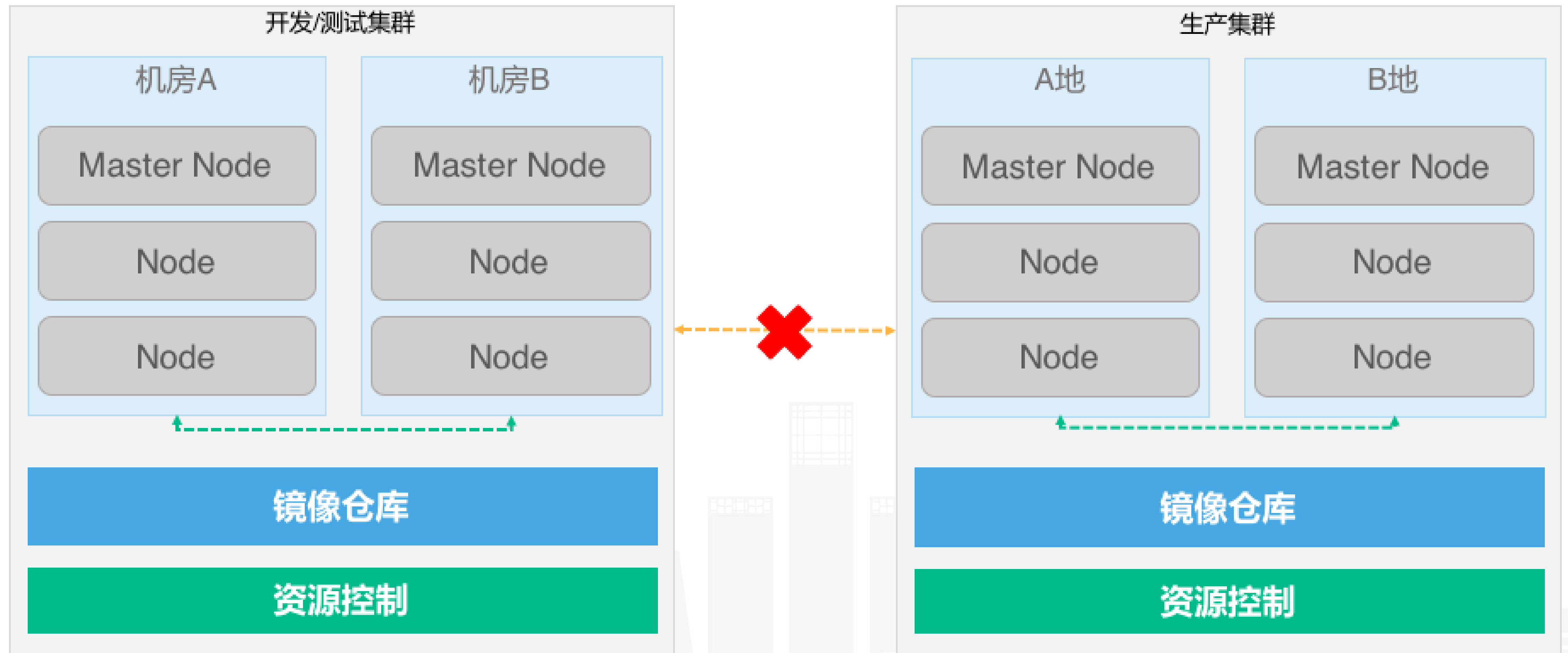
keepalived子进程崩溃？

第一种策略：主节点挂掉以后立即将虚拟IP漂移到备用节点，由备用节点提供服务。当主节点恢复时，虚拟IP重新漂移到主节点，由主节点提供服务；

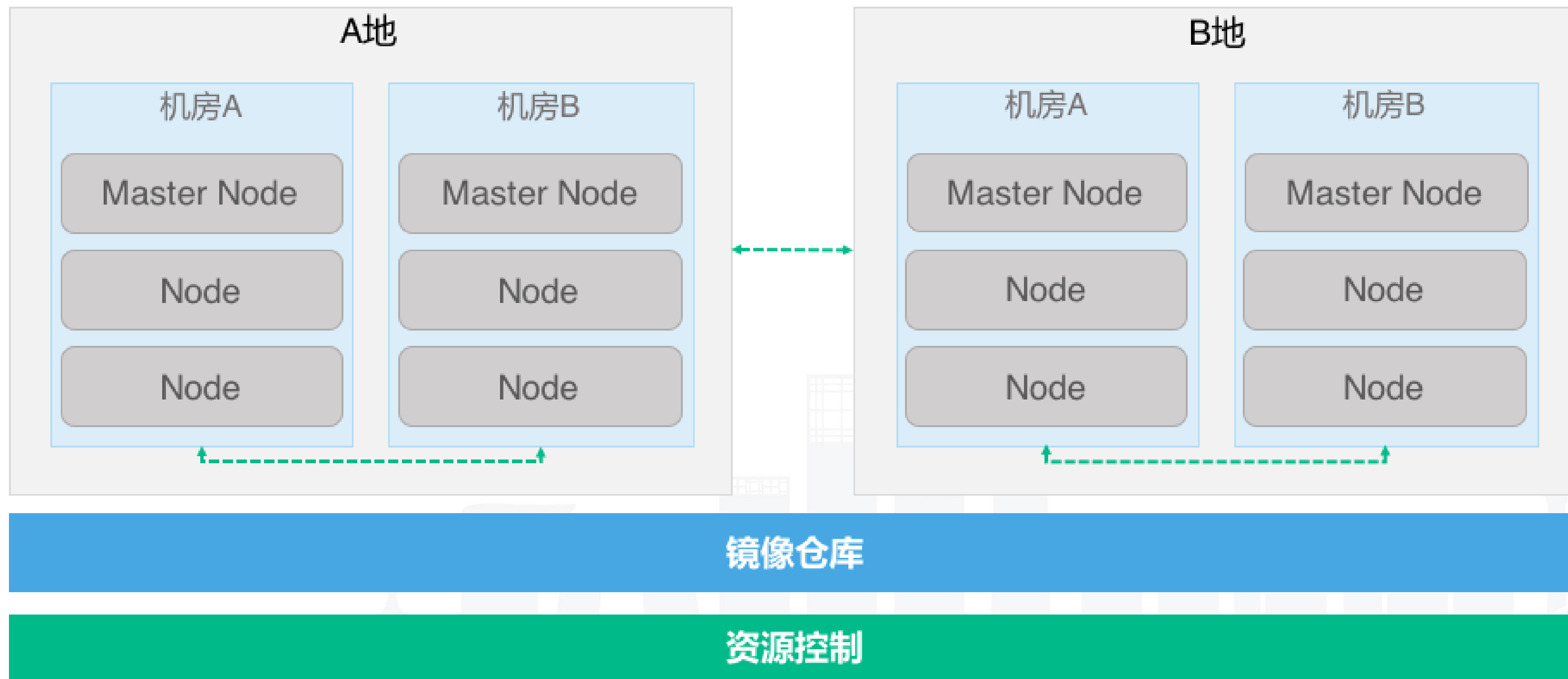
第二种策略：主节点挂了以后尝试拉起挂掉的服务，尝试N次，如果服务拉起失败，则漂移虚拟IP到备用节点，由备用节点提供服务，主节点恢复以后，并不马上漂移虚拟IP，继续由备用节点提供服务。



多集群架构



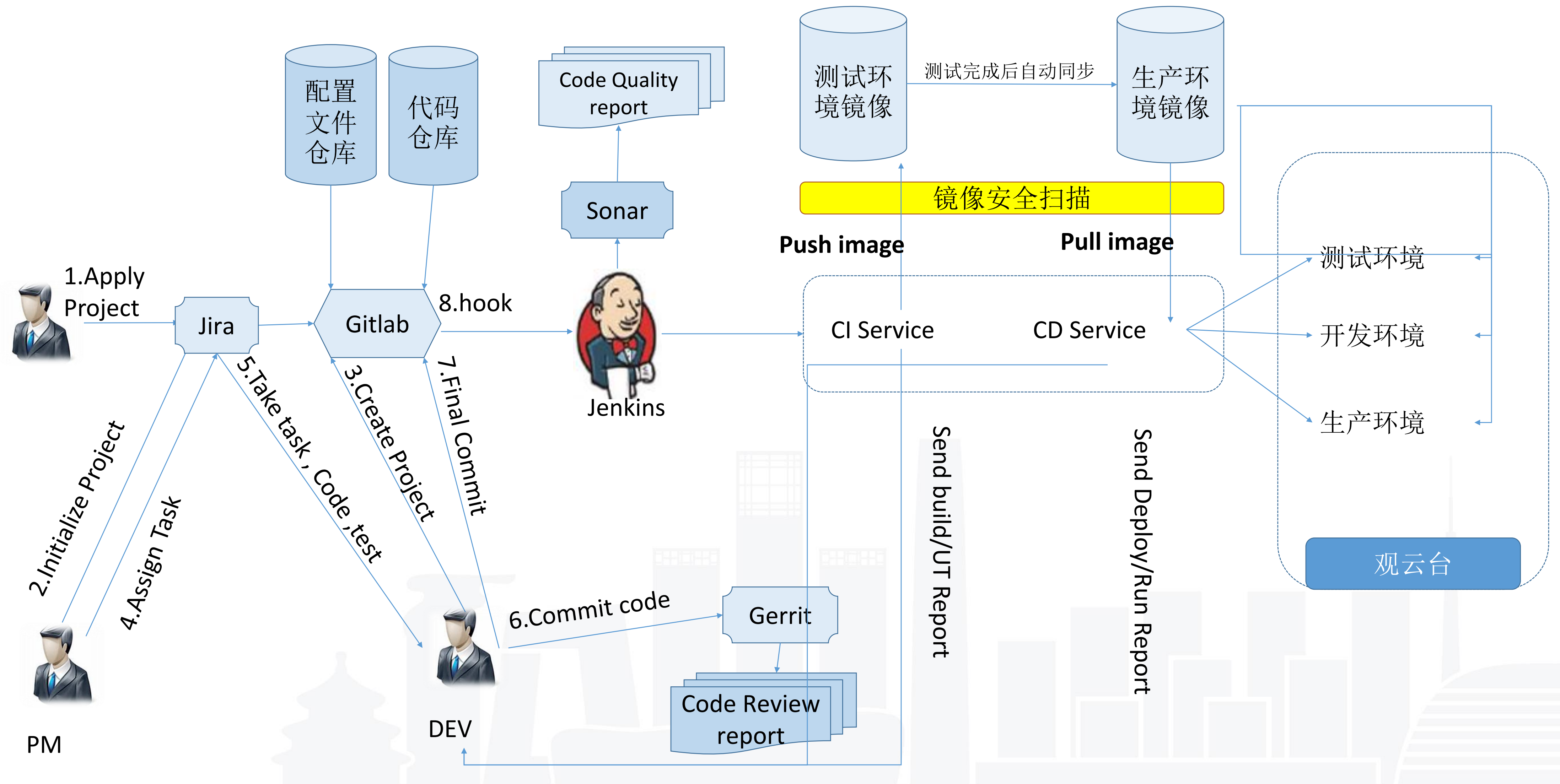
跨数据中心单集群



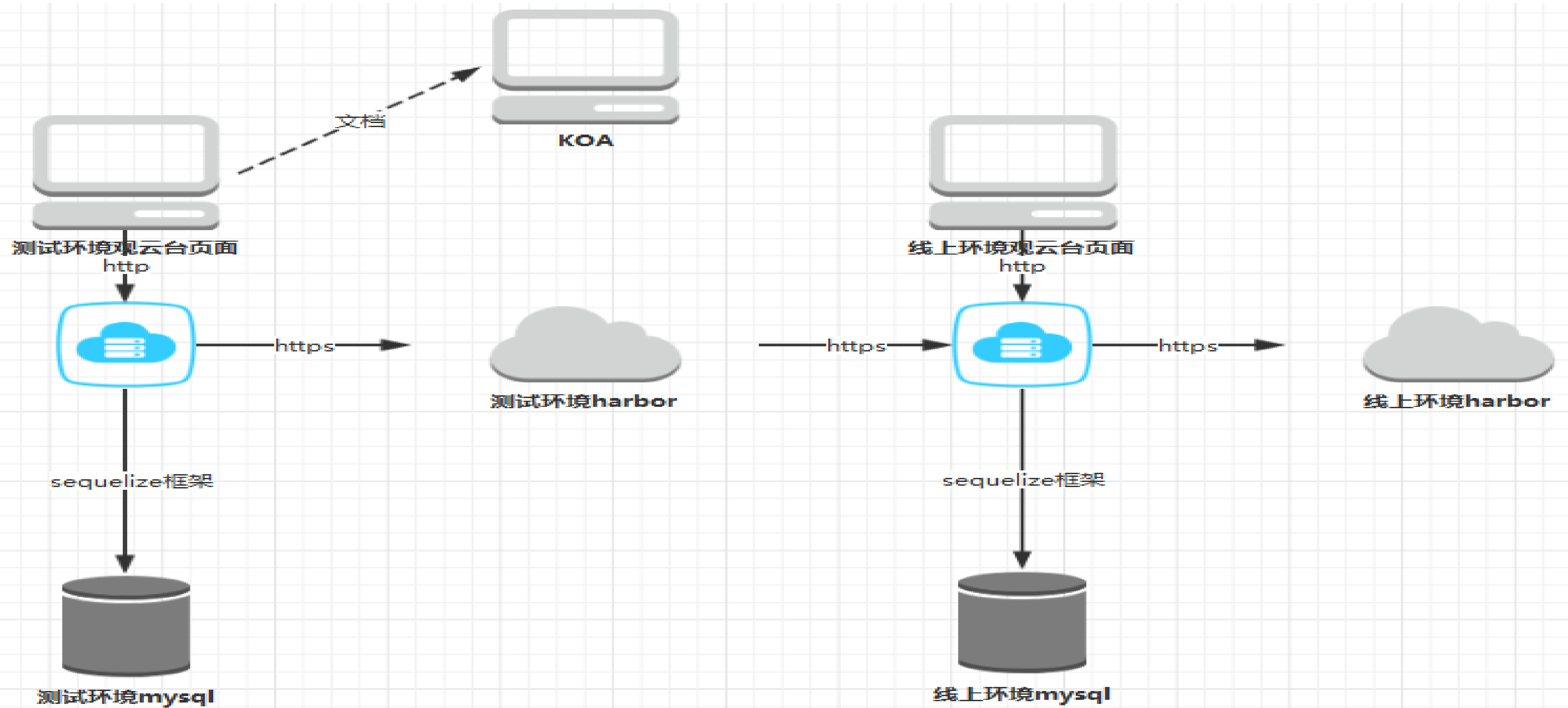
CI/CD



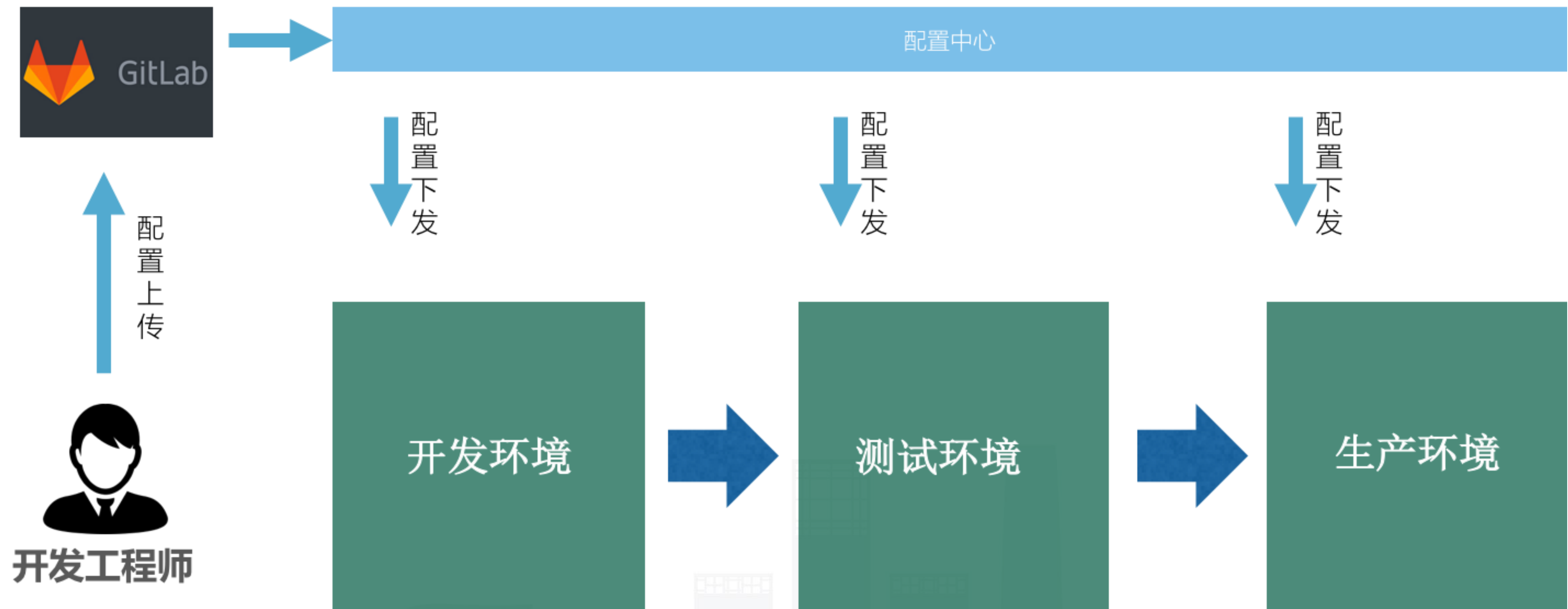
持续集成持续部署



镜像仓库同步



配置中心



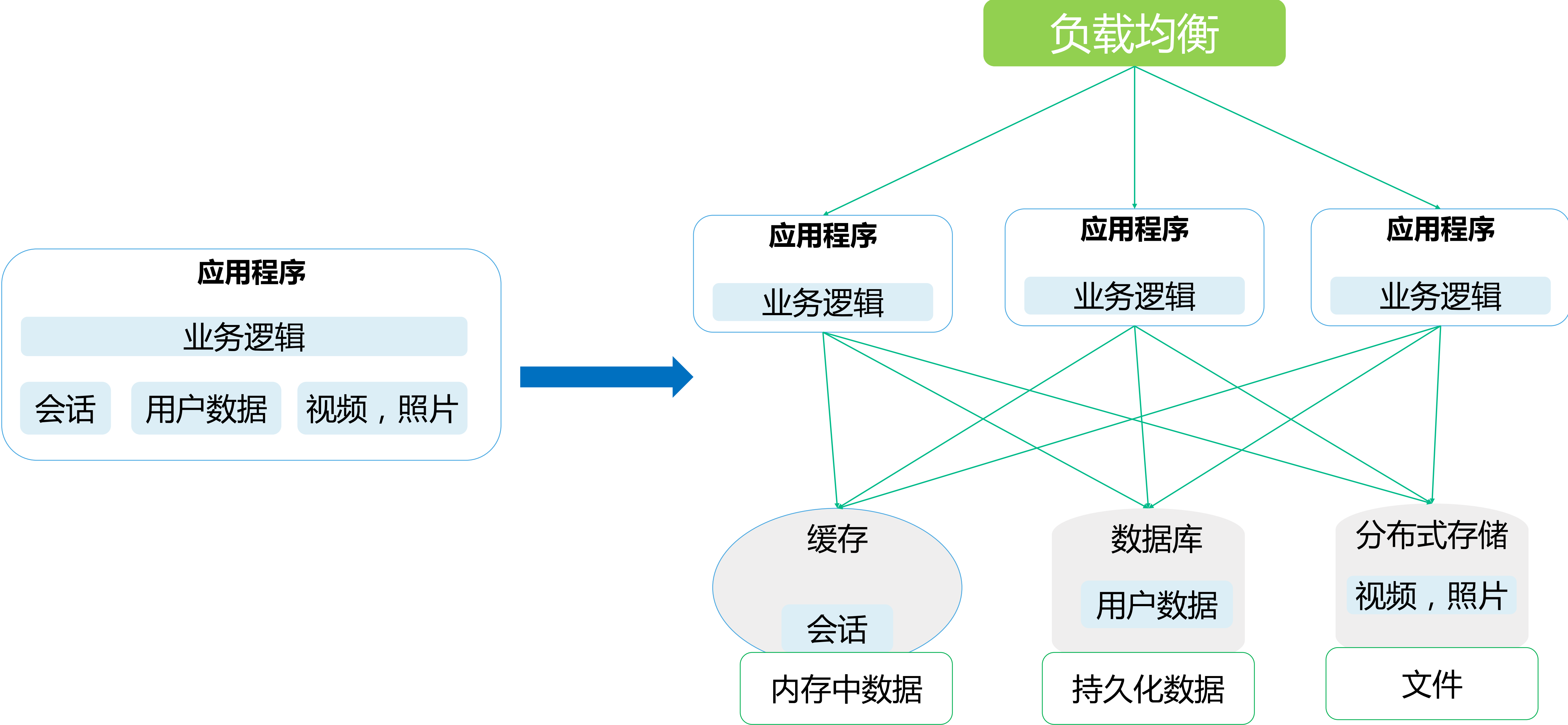
应用迁移



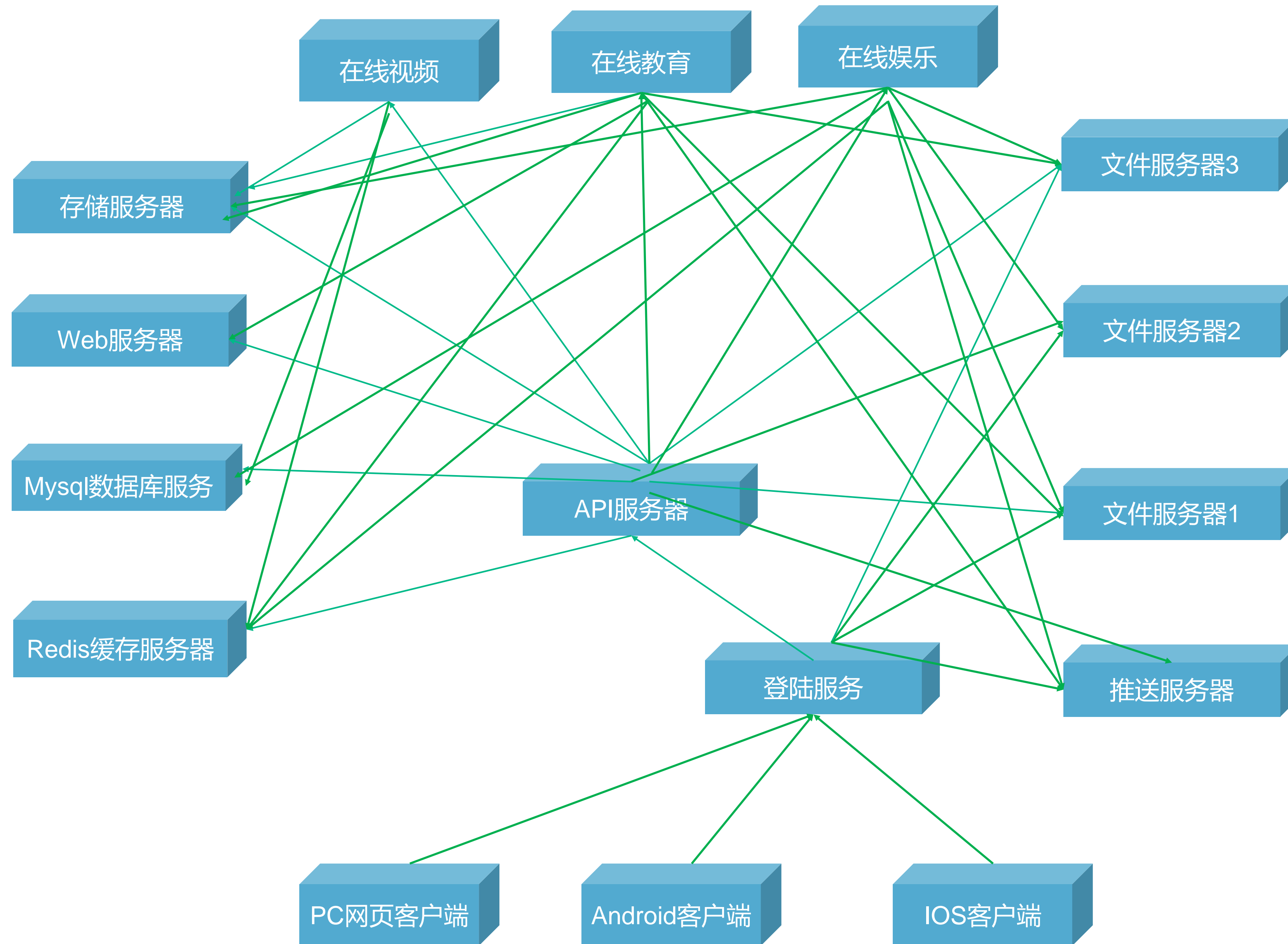
应用迁移改造

应用类型	特征	收益	改造
无状态应用	普通的web应用程序	快速部署 弹性伸缩 自动容错 高利用率	容器化,配置分离
有状态应用	采用Redis或Memcached共享用户状态	快速部署 快速扩容 可能提升利用率	做无状态改造
本地持久化应用	生成本地文件	快速部署 (可能需同步数据)	需要分布式存储支持
服务注册调用	解耦配置依赖	自动化部署，配置	服务发现

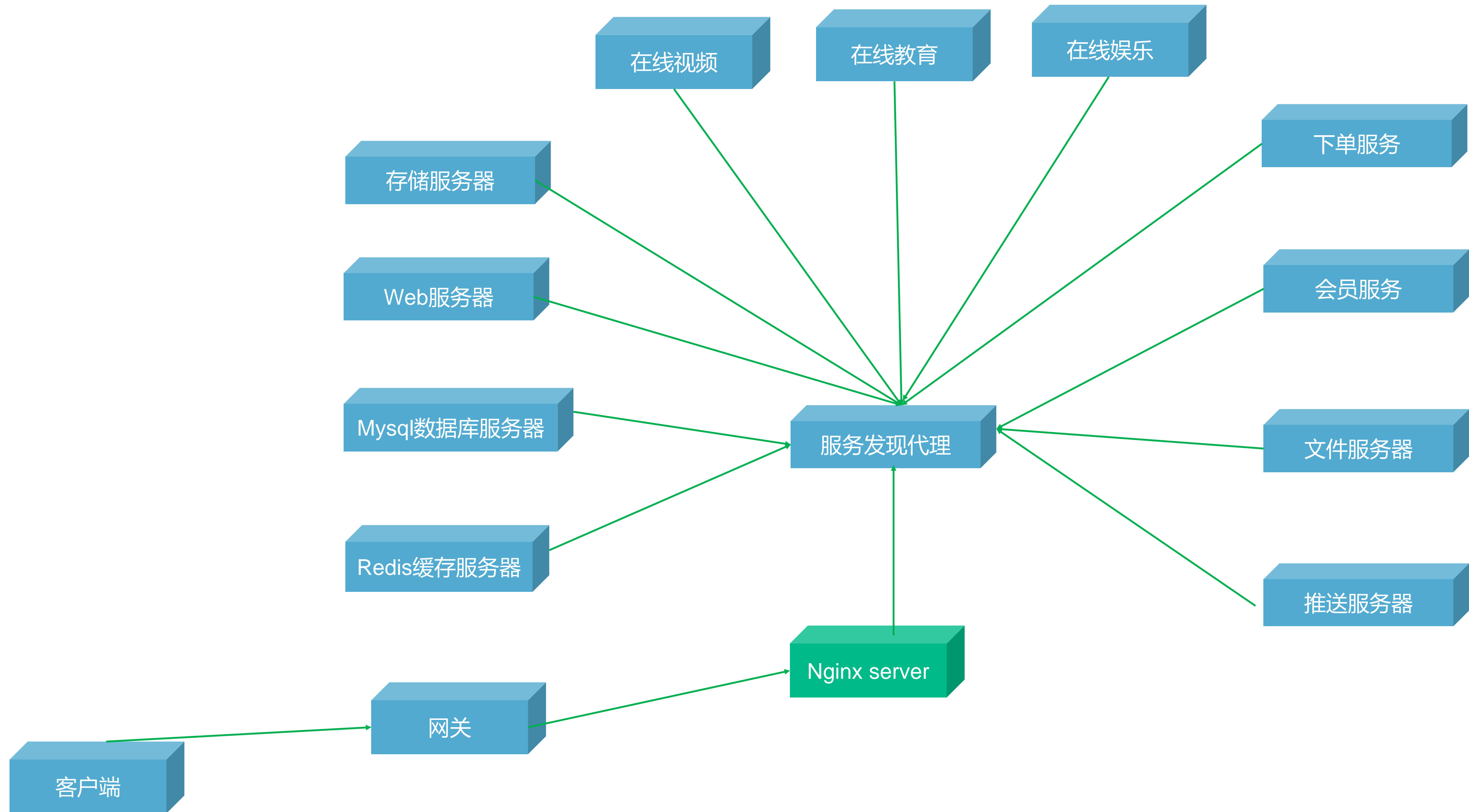
应用迁移改造



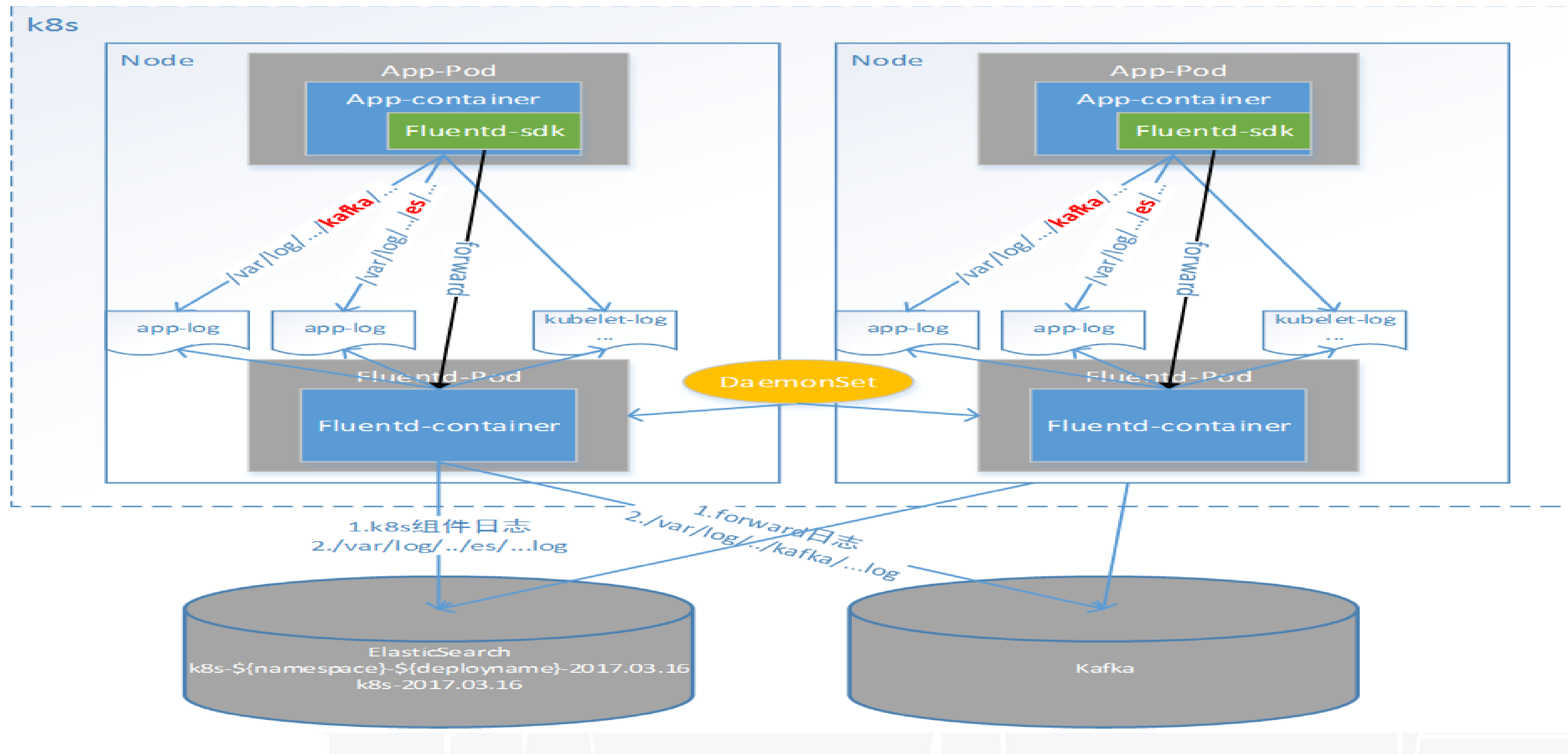
微服务化及服务管理改造前



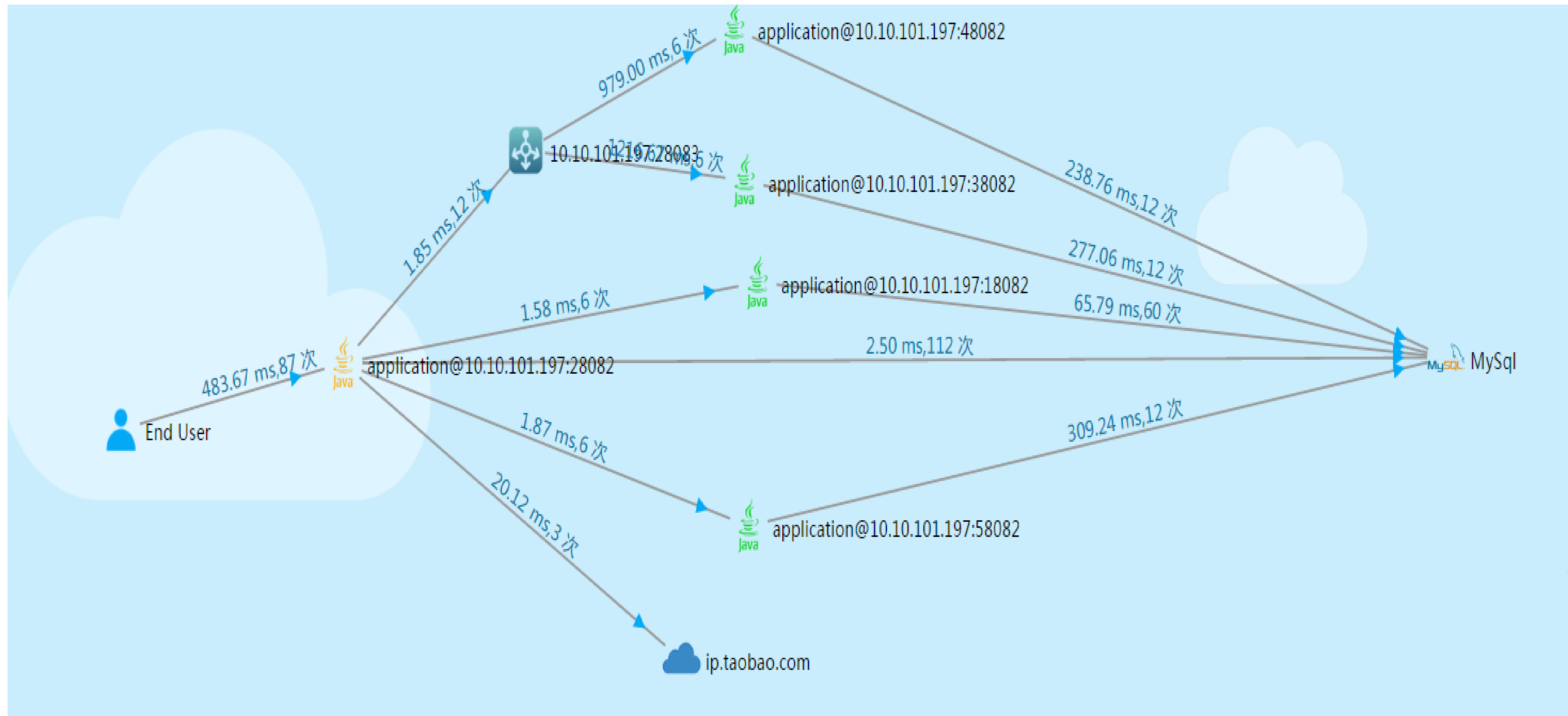
微服务化及服务管理改造后



日志过滤收集



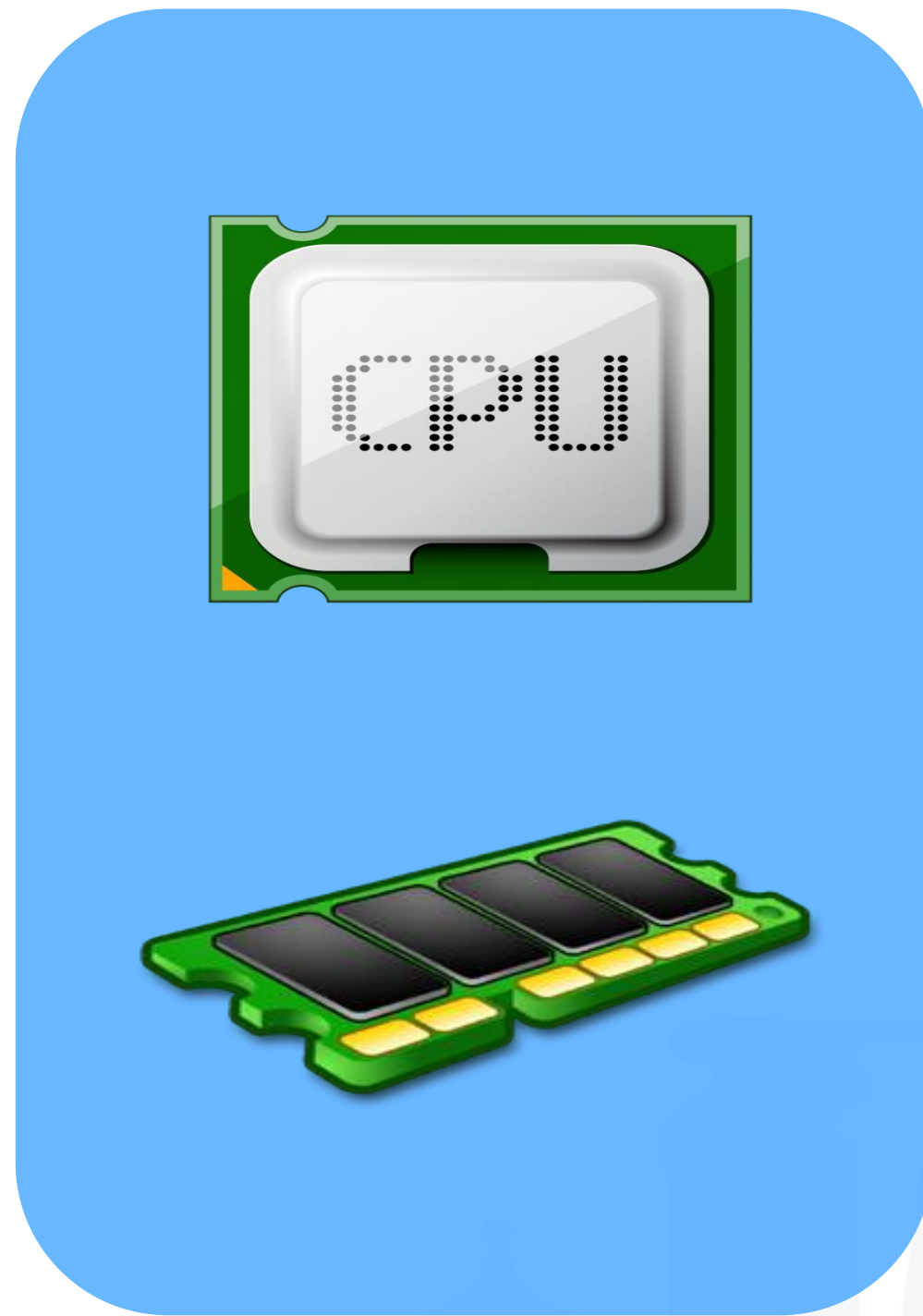
应用拓扑自发现



自动构建应用拓扑，发现全链路性能瓶颈，重现业务场景，快速定位问题

基于业务指标的弹性伸缩

标准弹性指标



基于业务弹性指标



在线视频移动端监控

崩溃

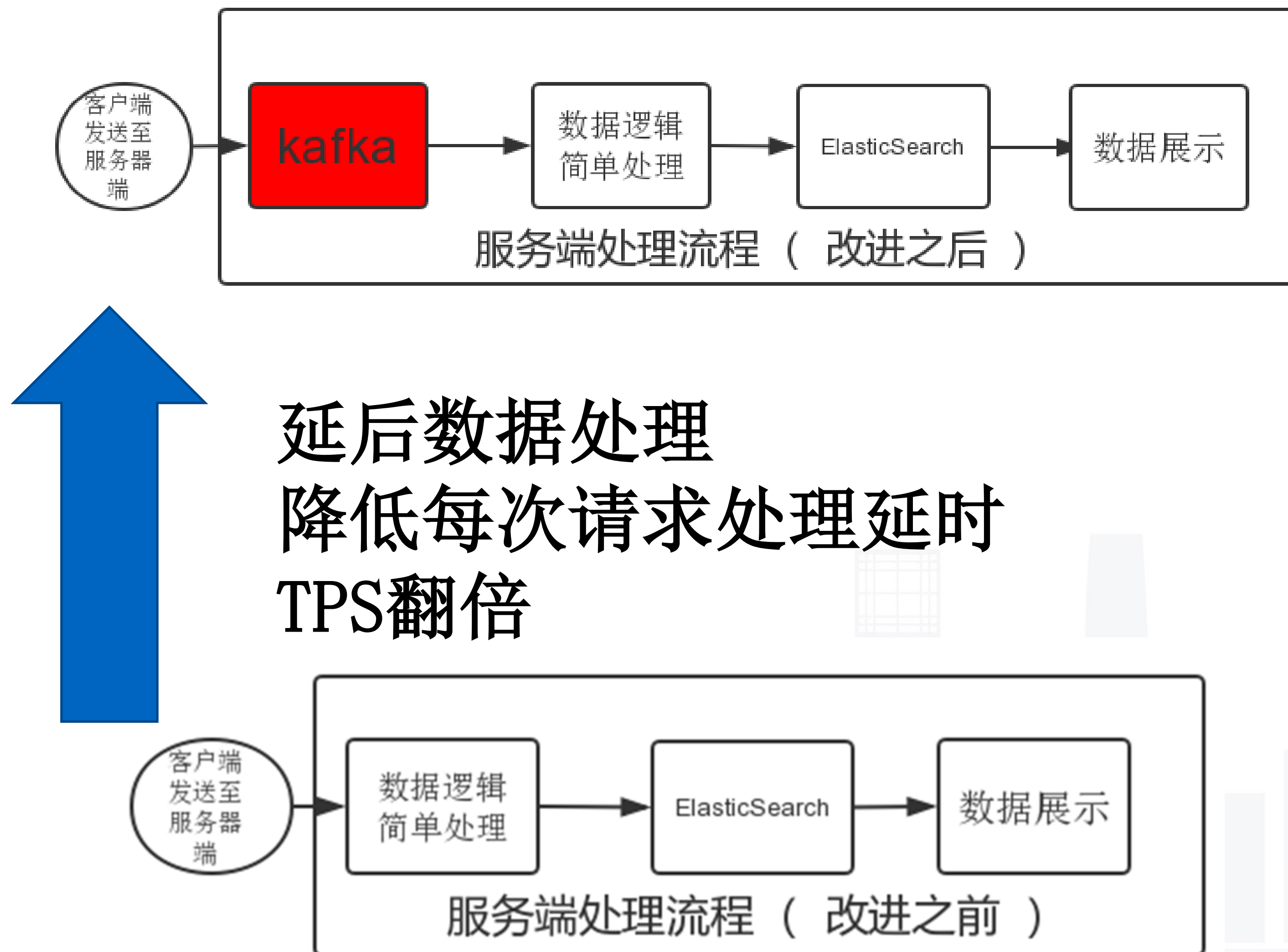
网络错误

卡顿...



- 崩溃
- ANR
- 网络监控
 - 网络错误
 - 网络性能
- 劫持监控
- 慢动作&慢交互监控
- 各机型综合监控

监控后台优化



网络方案选择及优化

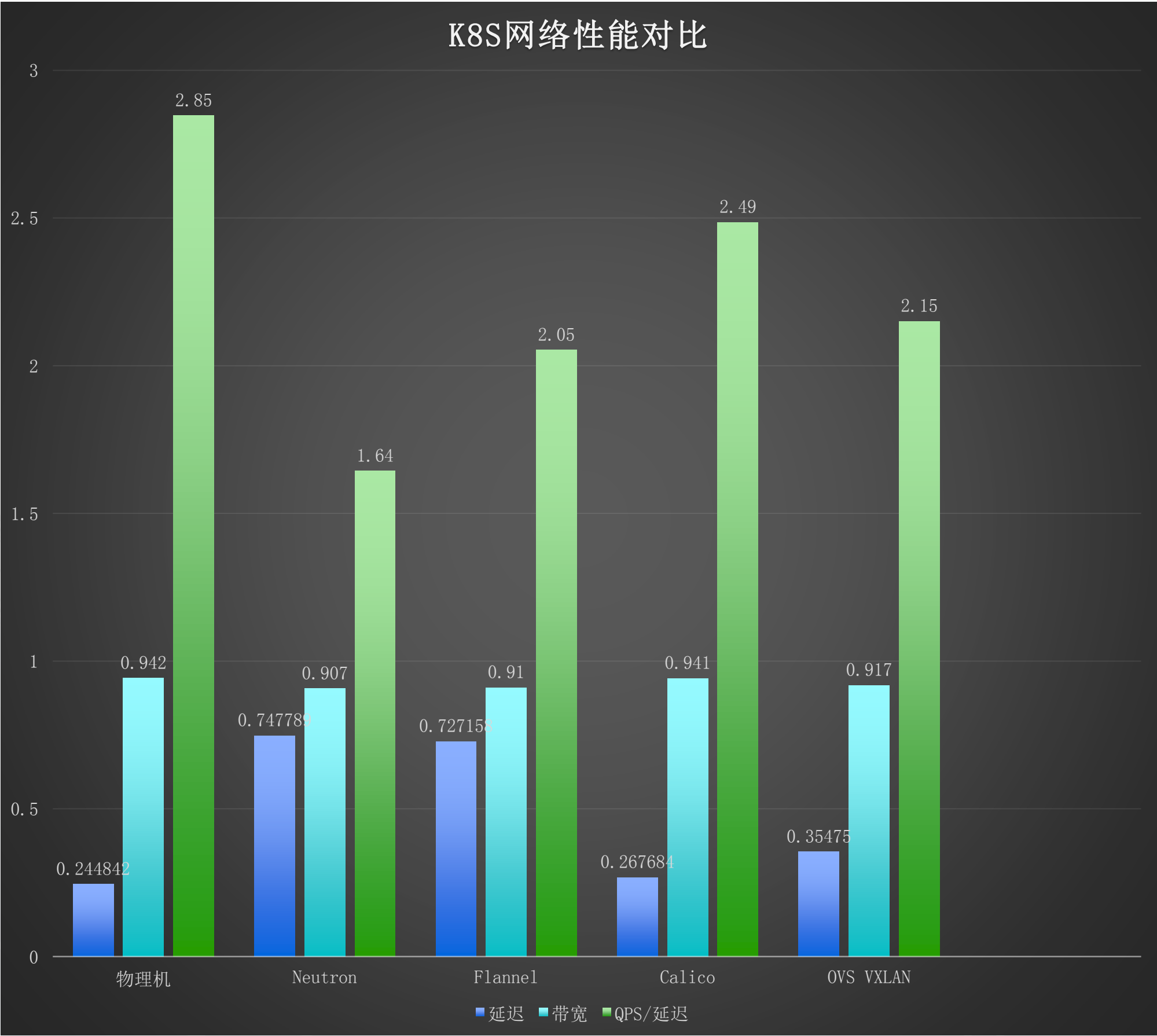
K8S容器网络方案对比

网络方案功能对比

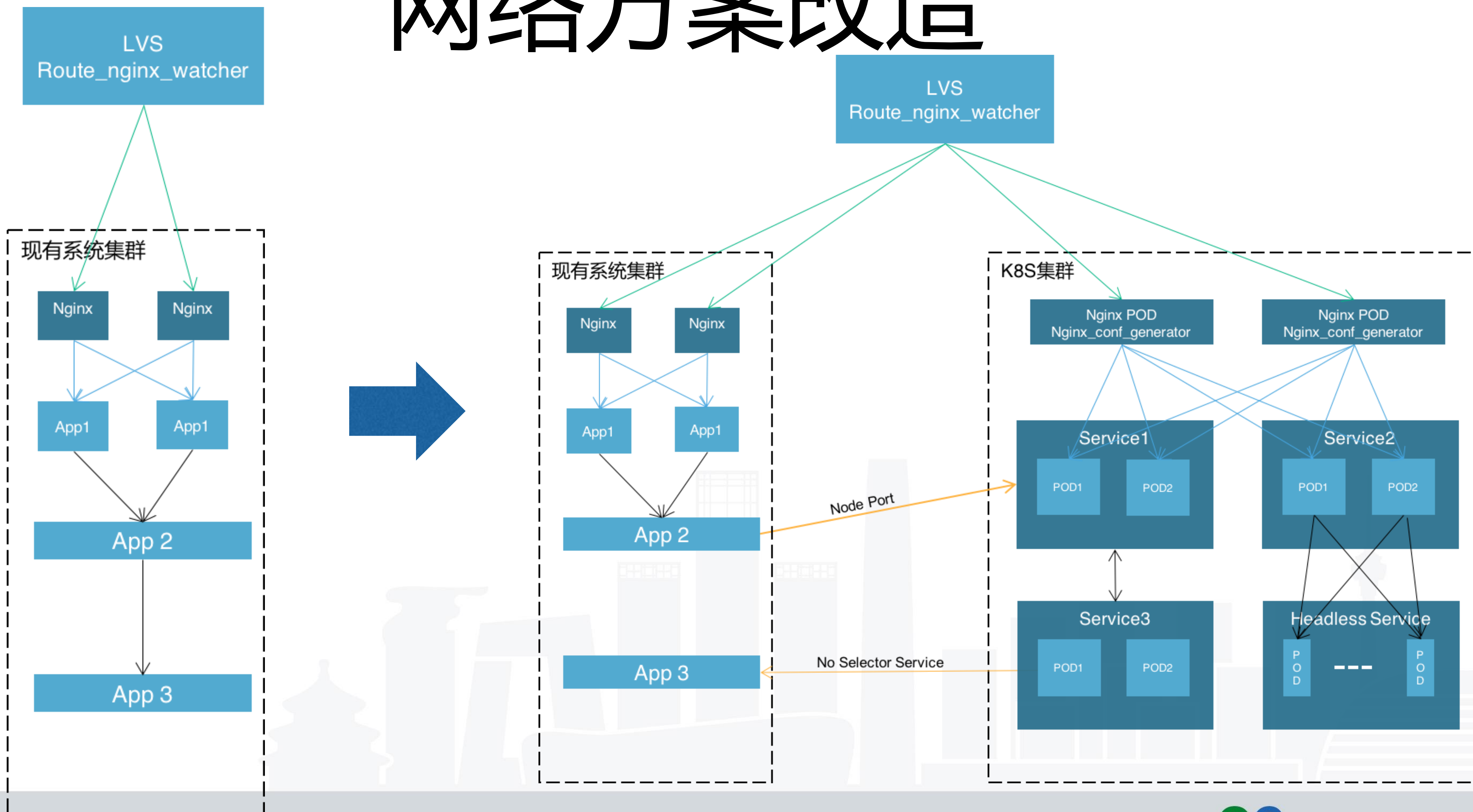
方案	技术	多节点	多租户	高可用	高级网络功能	社区支持	适用场景
隧道方案	Flannel	支持	不支持	不支持	不支持	一般	无依赖
	OVS	支持	不支持	支持	支持	一般	无依赖
	Neutron	支持	支持	支持	支持	较好	无依赖
路由方案	Calico	支持	支持	支持	部分支持	较好	大二层

网络方案性能对比

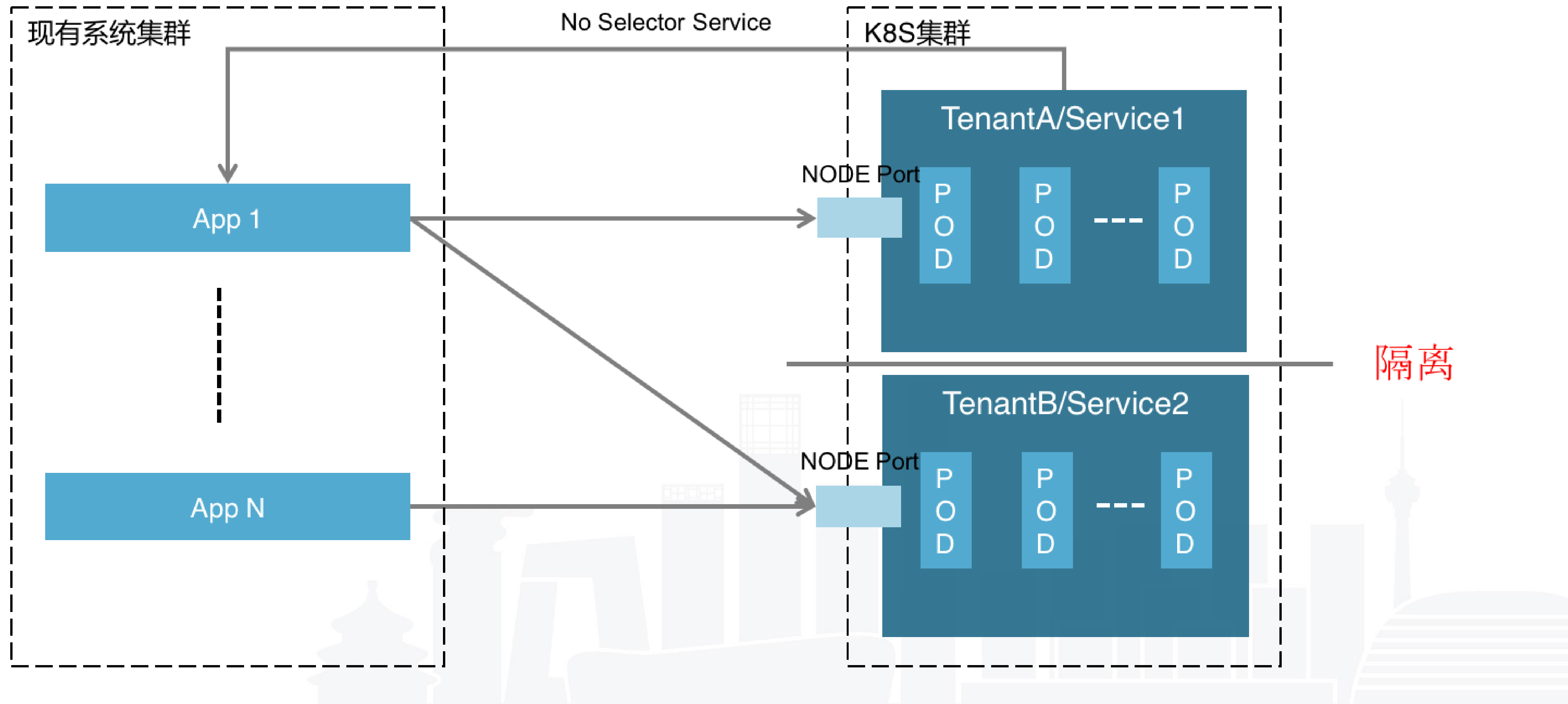
网络类型	延迟 (ms)	带宽 (Mb/s)	Nginx (QPS/延迟)
物理	0.244842	942	11439.61/4.019
Neutron VXLAN	0.747789	907	9065.38/5.515
Flannel VXLAN	0.727158	910	10031.23/4.886
Calico	0.267684	941	11004.44/4.428
OVS VXLAN	0.35475	917	10457.32/4.864



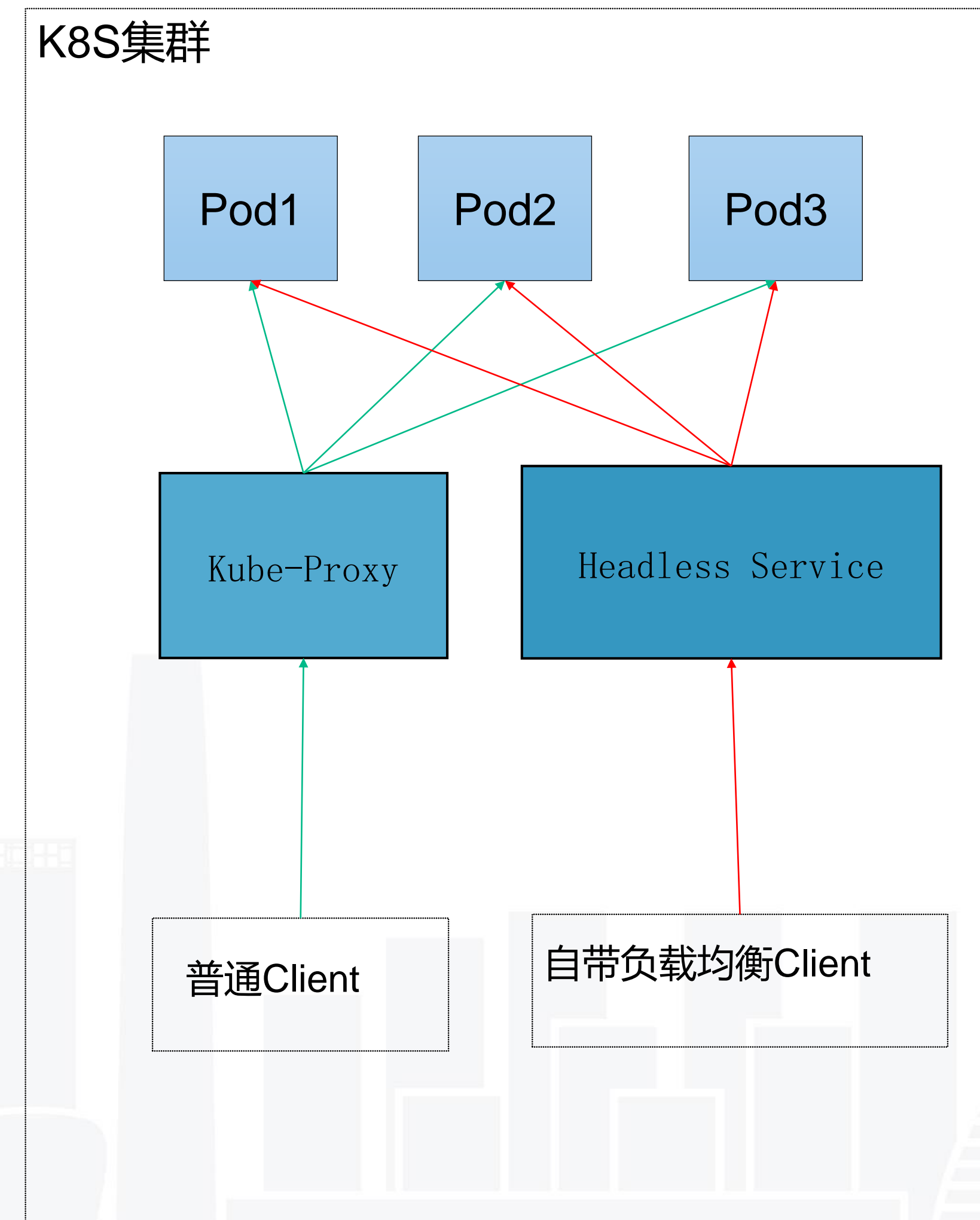
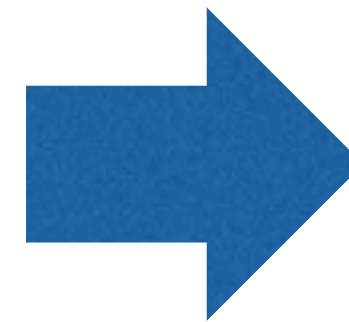
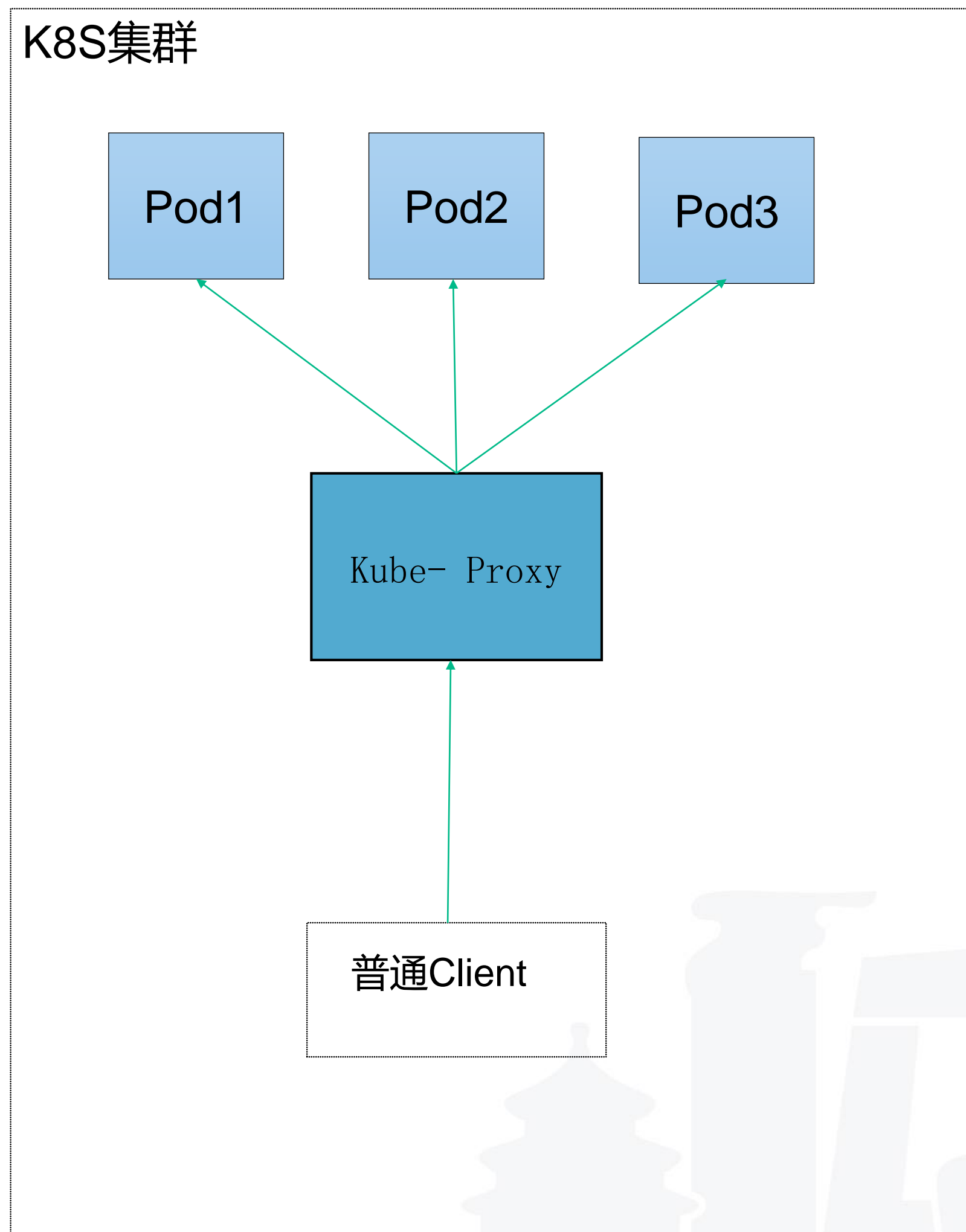
网络方案改造



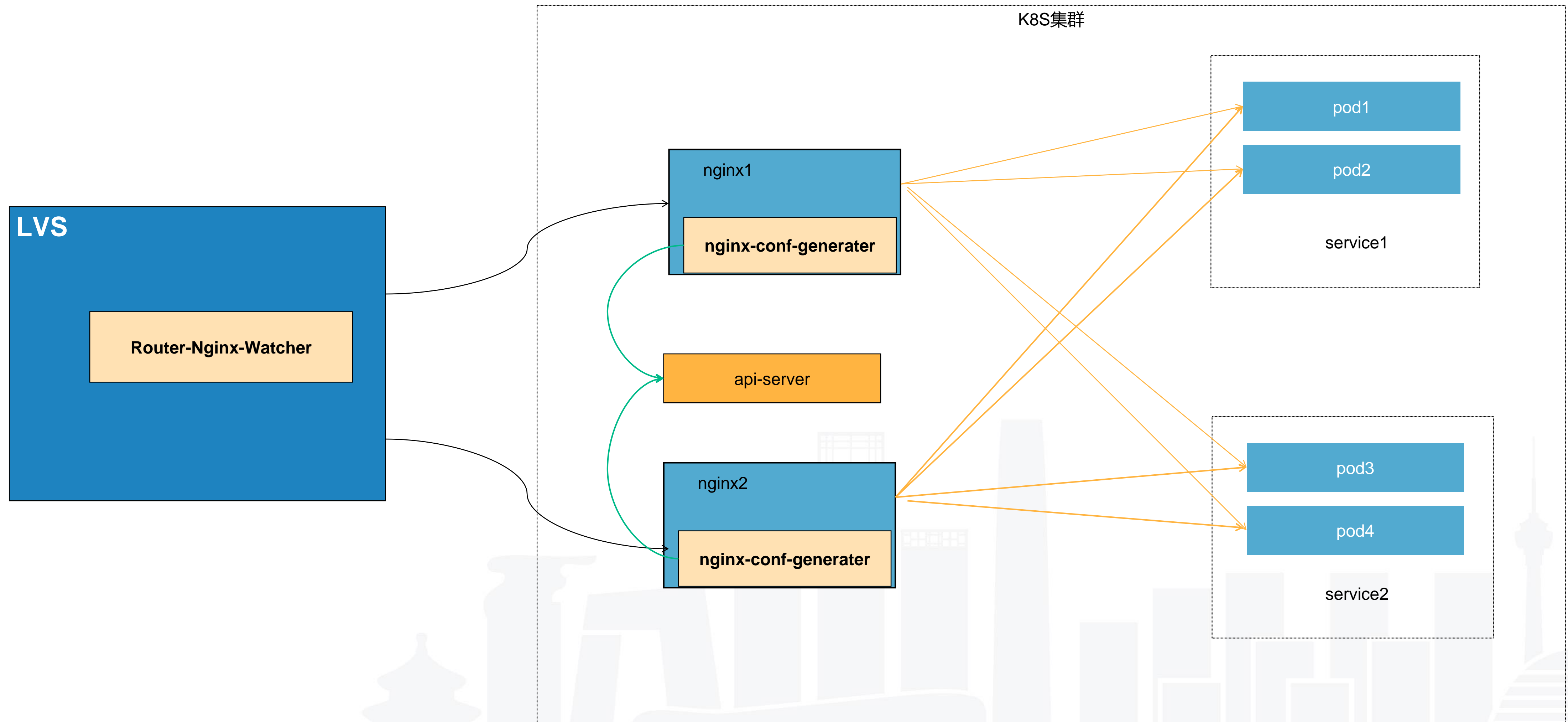
集群内外通信



客户端负载均衡



对外服务负载均衡优化



Calico网络方案实践

方案难点分析

1. nginx流量出去的时候源ip会被realserver改为LVS的VIP，所以会被felix-FORWARD链给drop掉。需要添加source 为VIP的规则。

```
Chain felix-FORWARD (1 references)
pkts bytes target prot opt in out source destination ctstate
0 0 RETURN all -- * * 10.10.102.67 0.0.0.0/0
0 0 RETURN all -- * * 10.10.102.66 0.0.0.0/0
0 0 DROP all -- cali+ * 0.0.0.0/0 0.0.0.0/0 ctstate INVALID
0 0 DROP all -- * cali+ 0.0.0.0/0 0.0.0.0/0 ctstate INVALID
0 0 ACCEPT all -- cali+ * 0.0.0.0/0 0.0.0.0/0 ctstate RELATED,ESTABLISHED
0 0 ACCEPT all -- * cali+ 0.0.0.0/0 0.0.0.0/0 ctstate RELATED,ESTABLISHED
0 0 felix-FROM-ENDPOINT all -- cali+ * 0.0.0.0/0 0.0.0.0/0
0 0 felix-TO-ENDPOINT all -- * cali+ 0.0.0.0/0 0.0.0.0/0
0 0 ACCEPT all -- cali+ * 0.0.0.0/0 0.0.0.0/0
0 0 ACCEPT all -- * cali+ 0.0.0.0/0 0.0.0.0/0
```

2. 关闭nginx的interface的反向路由 `/proc/sys/net/ipv4/conf/X/rp_filter=0`

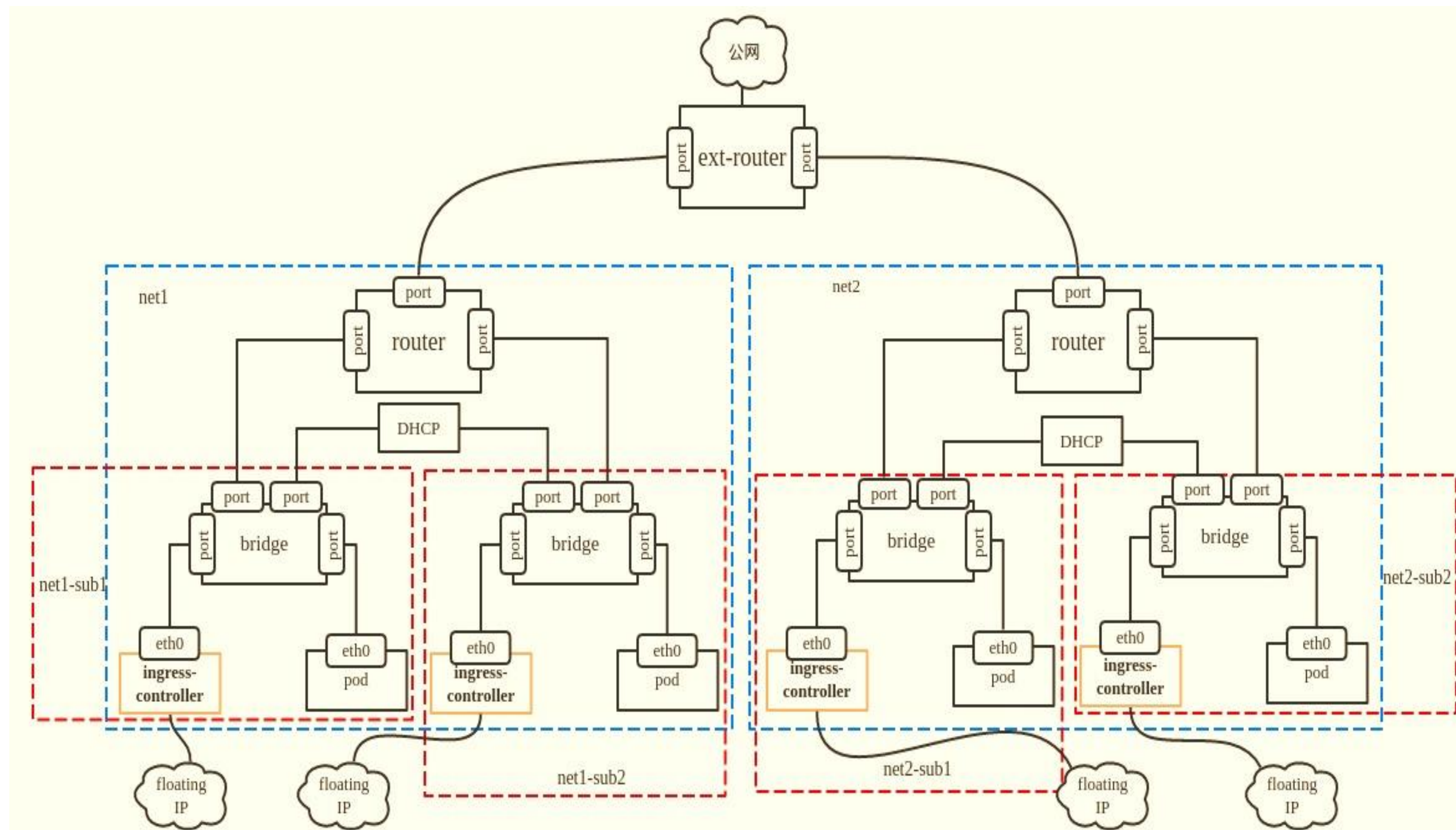
3. nginx-conf-generator容器从api-server定时获取nginx代理的upstream的pod的变化，并更新对应的nginx 配置文件。

隔离更好的Neutron方案



Neutron网络方案实践

Neutron与K8S集成方案



Neutron网络方案实践

Neutron与K8S集成组件

- Neutron CNI

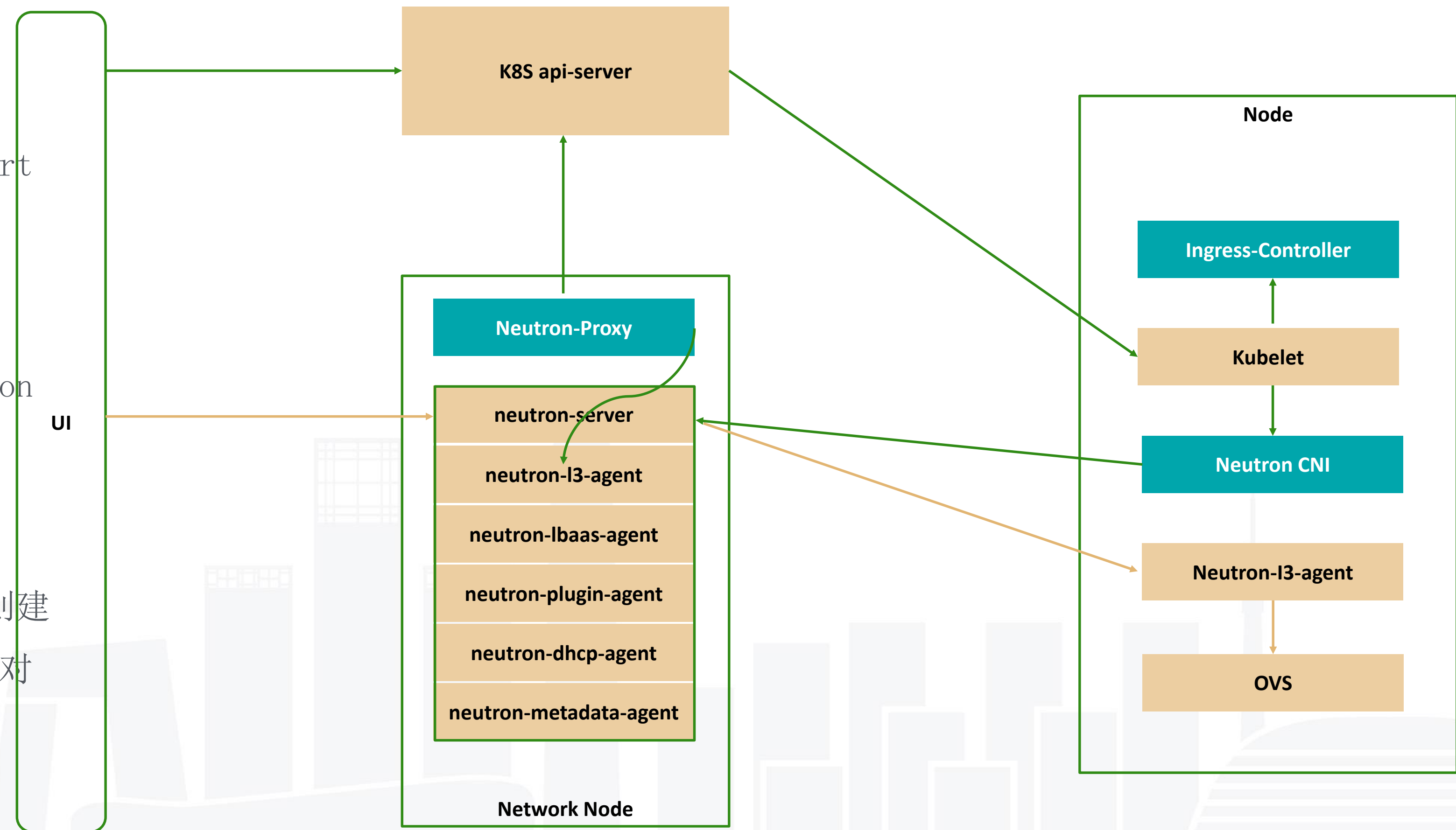
负责在创建/删除POD过程中给POD分配Neutron的Port和IP

- Neutron-Proxy

负责监听K8S Service创建信息，并在对应的Neutron网络路由器中写入对应规则

- Ingress-Controller

是一个具有TCP/HTTP代理功能的POD，每个子网在创建过程中自动回创建这个POD，同时通过Floating IP对外提供服务



某互联网电视私有容器云平台



总结

- 分享了容器技术基于观云台落地的过程包括多集群搭建，CI/CD对接，应用迁移和网络的改造对接，落地对接现有方案，过程平滑。
- 效果：
 - 基于物理机，节约了成本和资源
 - 加快了应用发布的速度
 - 提升了弹性伸缩的能力
 - 提升了应用监控的水平
- 展望：持续优化



关注QCon微信公众号，
获得更多干货！

Thanks!



主办方 **Geekbang** > **InfoQ**
极客邦科技