深入 Kubernetes 的“无人区”
—— 蚂蚁金服双十一的调度系统

曹寅

目 录
contents

一、蚂蚁金服的Kubernetes现状

发展历程与落地规模


平台研发
2018年下半年开始投入 Kubernetes 及其配套系统研发
灰度验证
2019年初于生产环境开始灰度验证，对部分应用做平台迁移
云化落地
2019年4月完成云化环境适配，蚂蚁金服云上基础设施全部采用 Kubernetes 支撑618
规模化落地
2019年7月到双十一前完成全站 Kubernetes 落地，超过90% 的资源通过 Kubernetes 分配，核心链路100%落地支撑大促。

大促规模

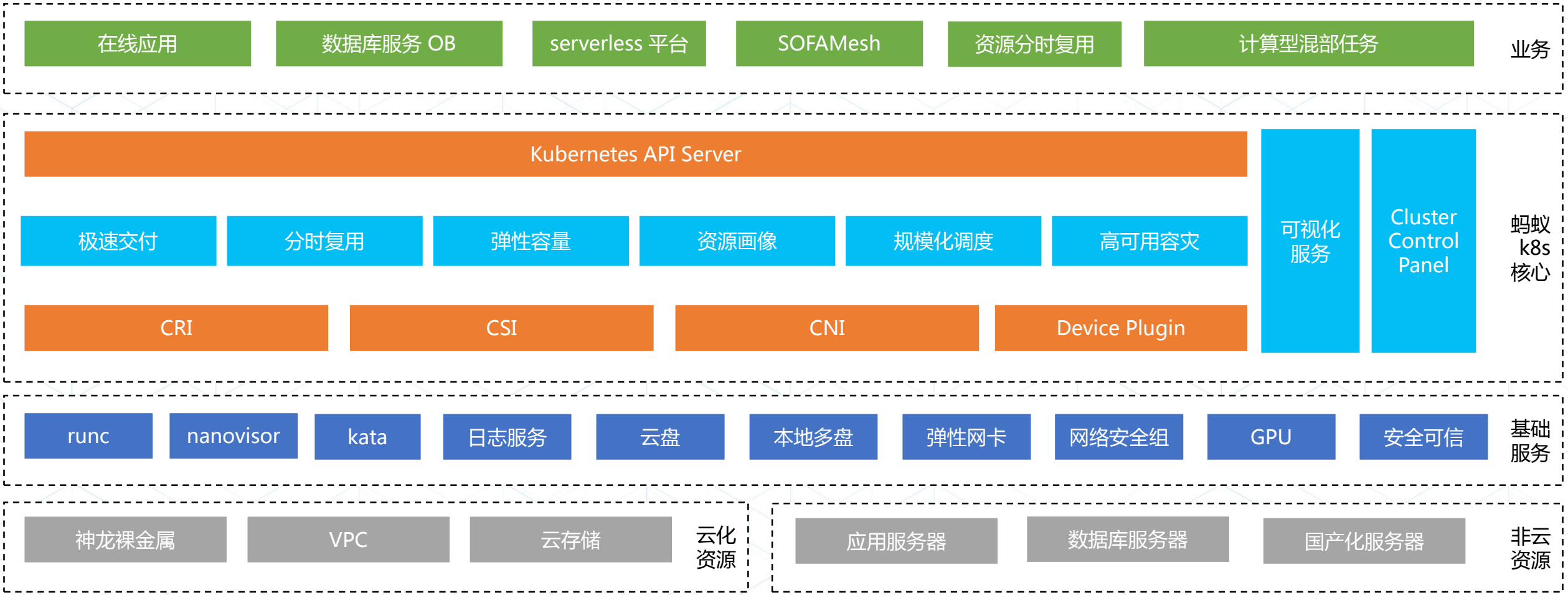
数万台
服务器和ECS

超一万
单集群规模

90%+
应用服务

数十万
应用 Pods

统一资源调度架构


在线应用
数据库服务 OB
serverless 平台
SOFAMesh
资源分时复用
计算型混部任务
业务
Kubernetes API Server
极速交付
分时复用
弹性容量
资源画像
规模化调度
高可用容灾
CRI
CSI
CNI
Device Plugin
可视化服务
Cluster Control Panel
蚂蚁k8s核心
runc
nanovisor
kata
日志服务
云盘
本地多盘
弹性网卡
网络安全组
GPU
安全可信
基础服务
神龙裸金属
VPC
云存储
云化资源
应用服务器
数据库服务器
国产化服务器
非云资源

二、双十一 Kubernetes 实践

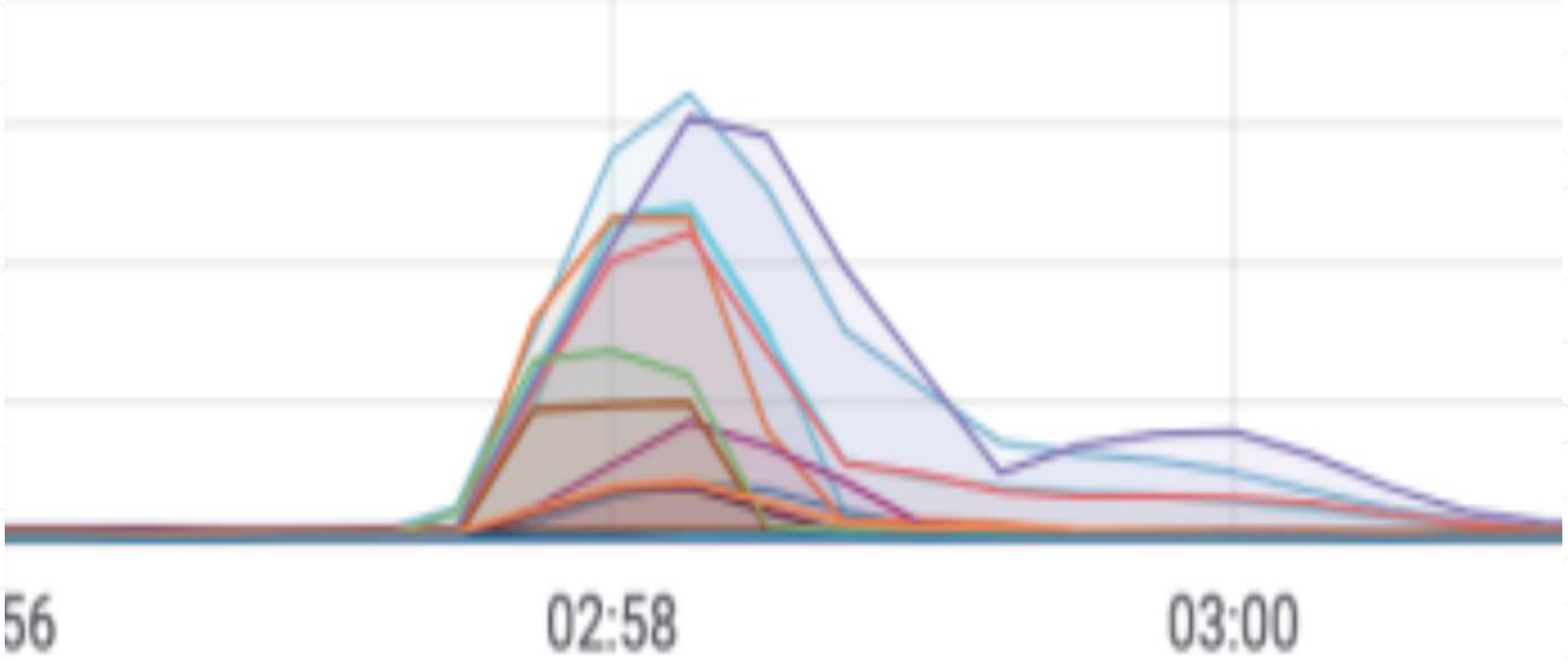
资源分时调度

快速腾挪的问题

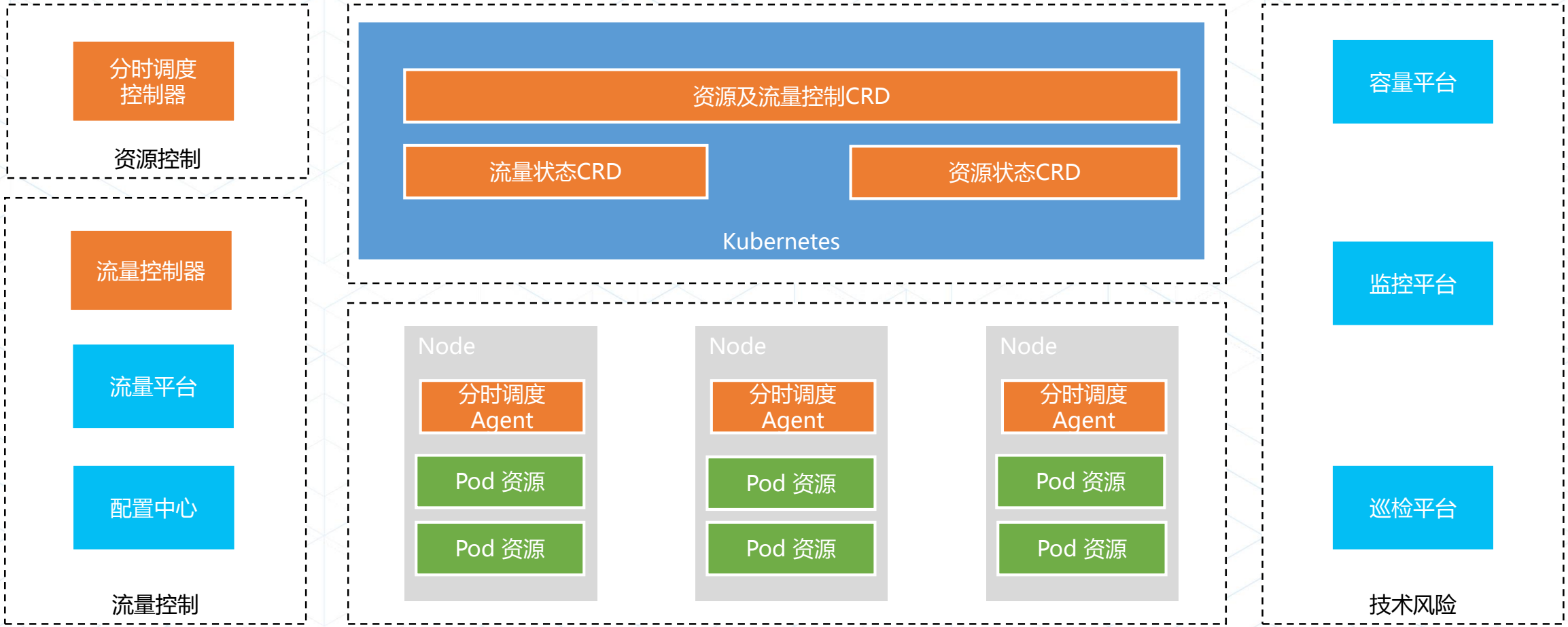
1.实例上下线需要预热

2. 腾挪耗时不可控

3. 大规模腾挪的稳定性



资源分时链路切换


分时调度
控制器
资源控制
流量控制器
流量平台
配置中心
流量控制
资源及流量控制CRD
流量状态CRD
资源状态CRD
Kubernetes
Node
分时调度
Agent
Pod 资源
Pod 资源
Node
分时调度
Agent
Pod 资源
Pod 资源
Node
分时调度
Agent
Pod 资源
Pod 资源
容量平台
监控平台
巡检平台
技术风险

分时切换效果

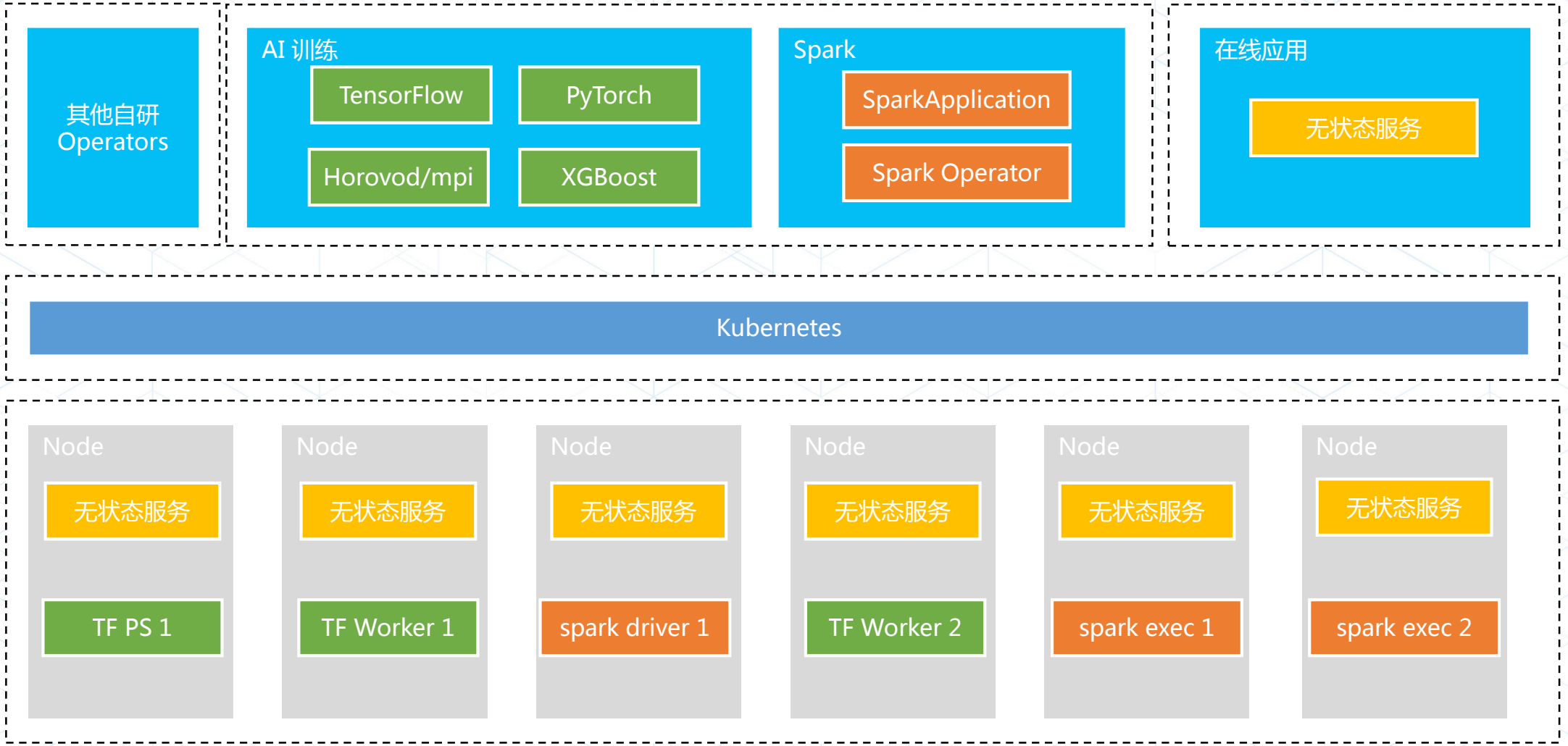
数万台
应用 Pods

分钟级
链路切换

数万核
CPU资源节省

100%
分时切换成功率

计算型任务混部


其他自研
Operators
AI 训练
TensorFlow
PyTorch
Horovod/mpi
XGBoost
Spark
SparkApplication
Spark Operator
在线应用
无状态服务
Kubernetes
Node
无状态服务
TF PS 1
Node
无状态服务
TF Worker 1
Node
无状态服务
spark driver 1
Node
无状态服务
TF Worker 2
Node
无状态服务
spark exec 1
Node
无状态服务
spark exec 2

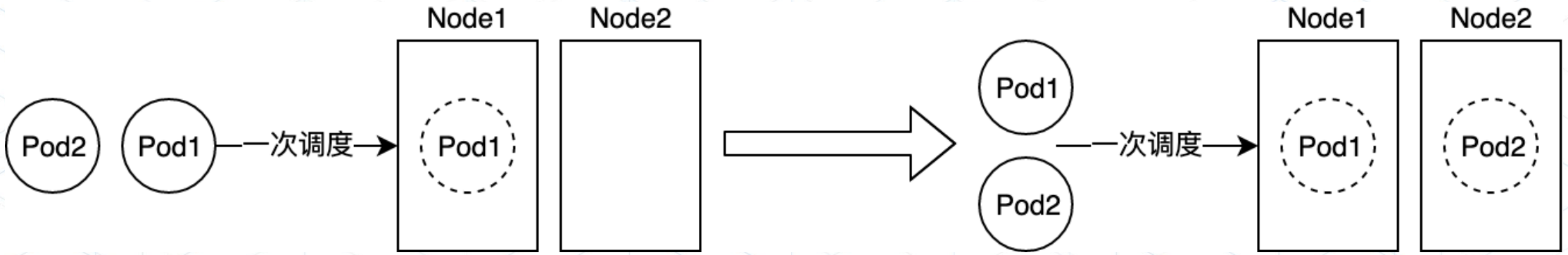
随着业务越来越多…

Czarkuberneteski @pczarkowski · 9小时

Our secret to running Kubernetes in production.

> **Czarkuberneteski** @pczarkowski · 2016年10月7日
>
> Our secret to running openstack in production.

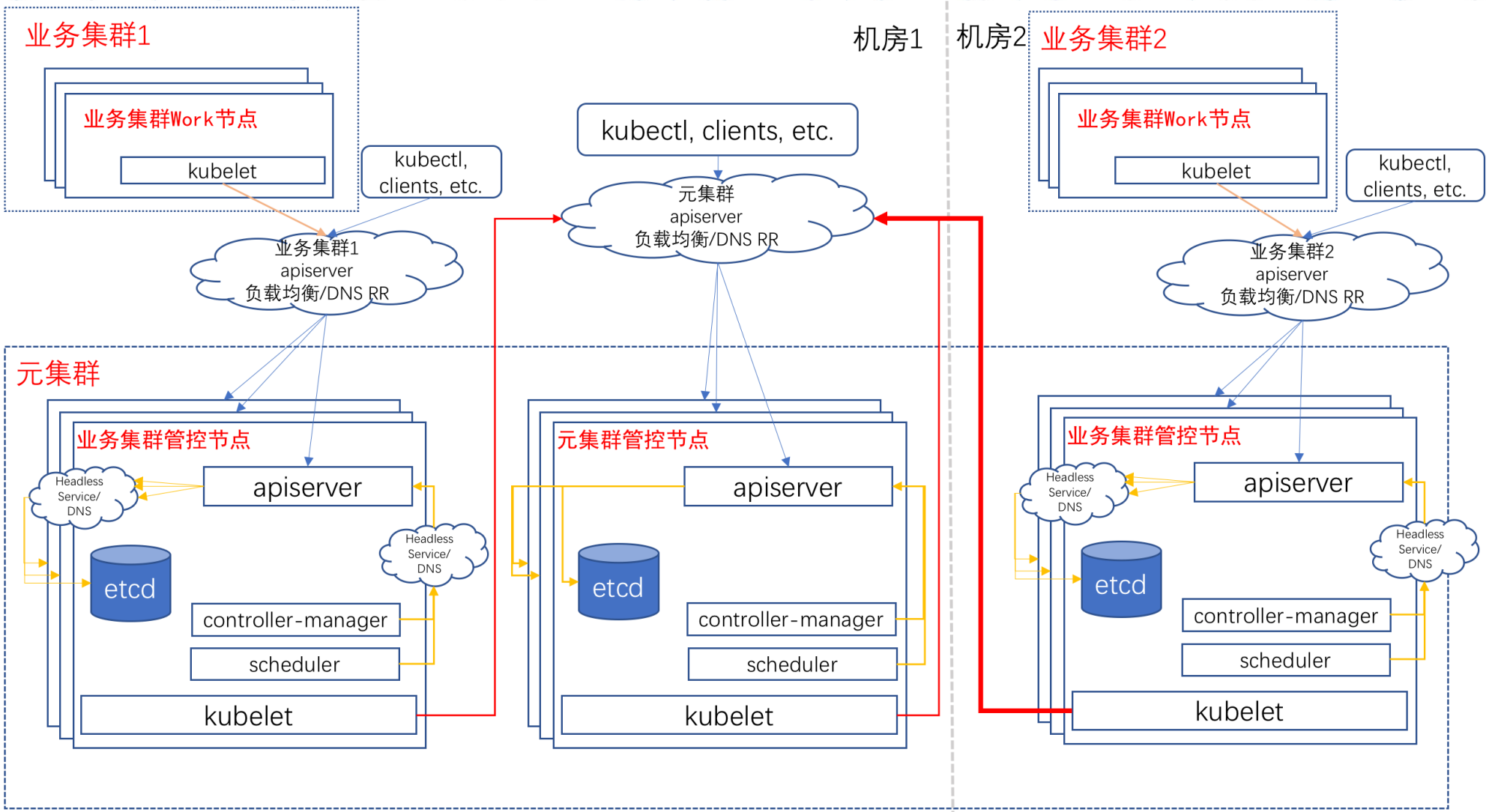
调度性能优化


Node1
Node2
Pod2
Pod1
一次调度
Pod1
Pod1
Pod2
Node1
Node2
一次调度
Pod1
Pod2

Operator开发者最佳实践

- CRD 在定义时需要明确未来的最大数量，大量CR 业务最好采用 aggregate-apiserver 进行扩展
- CRD 必须 Namespaced scope，以控制影响范围
- MutatingWebhook + 资源 Update 操作会给运行时环境带来不可控破坏，尽量避免使用这种组合
- 任何 controllers 都应该使用 informers，并且对写操作配置合理限流
- DaemonSet 非常高阶，尽量不要采用这类设计，如果必需请在 Kubernetes 专家的辅导下使用；

弹性资源建站


业务集群1
业务集群Work节点
kubelet
kubectl, clients, etc.
业务集群1 apiserver 负载均衡/DNS RR
机房1
机房2
kubectl, clients, etc.
元集群 apiserver 负载均衡/DNS RR
业务集群2
业务集群Work节点
kubelet
kubectl, clients, etc.
业务集群2 apiserver 负载均衡/DNS RR
元集群
业务集群管控节点
Headless Service/ DNS
apiserver
etcd
Headless Service/ DNS
controller-manager
scheduler
kubelet
元集群管控节点
apiserver
etcd
controller-manager
scheduler
kubelet
业务集群管控节点
Headless Service/ DNS
apiserver
etcd
Headless Service/ DNS
controller-manager
scheduler
kubelet

三、展望未来，迎接挑战

平台与多租户

Kubernetes设计的多租户

实际Kubernetes集群里的多租户

自动化运维 – 技术风险

欢迎关注 SOFAStack 公众号
获取分布式架构干货

使用钉钉扫码入群
第一时间获取活动信息

蚂蚁金服 | 金融科技