云原生产业大会 2021 CLOUD NATIVE INDUSTRY CONFERENCE
2021/05/26 × 中国 / 北京
原生蓄力 云领未来

腾讯云原生最佳实践

1000 + 企业云原生改造价值体现和最佳实践总结

郭志宏 腾讯云容器解决方案架构师

CAICT 中国信通院

目录

腾讯云

腾讯云

01

云原生价值

云原生价值体现 - 提升资源利用率（降本）

业务云原生化改造，资源利用率提升可达

30%~40%

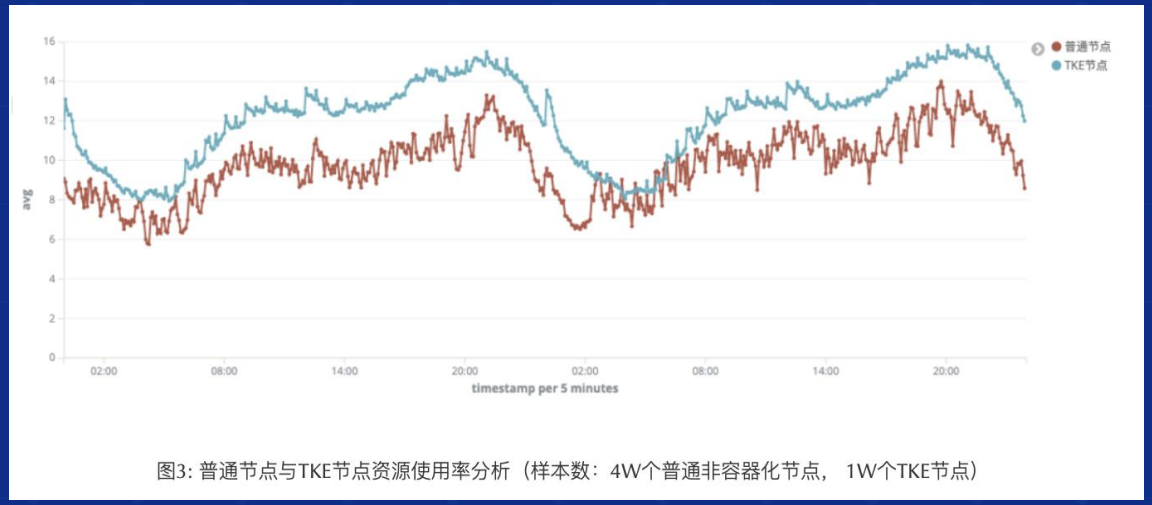


图3: 普通节点与TKE节点资源使用率分析（样本数：4W个普通非容器化节点， 1W个TKE节点）

图4：部分配置了节点池弹性伸缩业务应用启用了HPA平均资源使用率图示（样本1000+TKE节点）

云原生价值体现 - 提升发布效率（增效）

腾讯云

标准化镜像交付

统一业务运行环境

Coding Devops

开发，测试，发布流程统一

滚动更新

更新发布，不影响业务

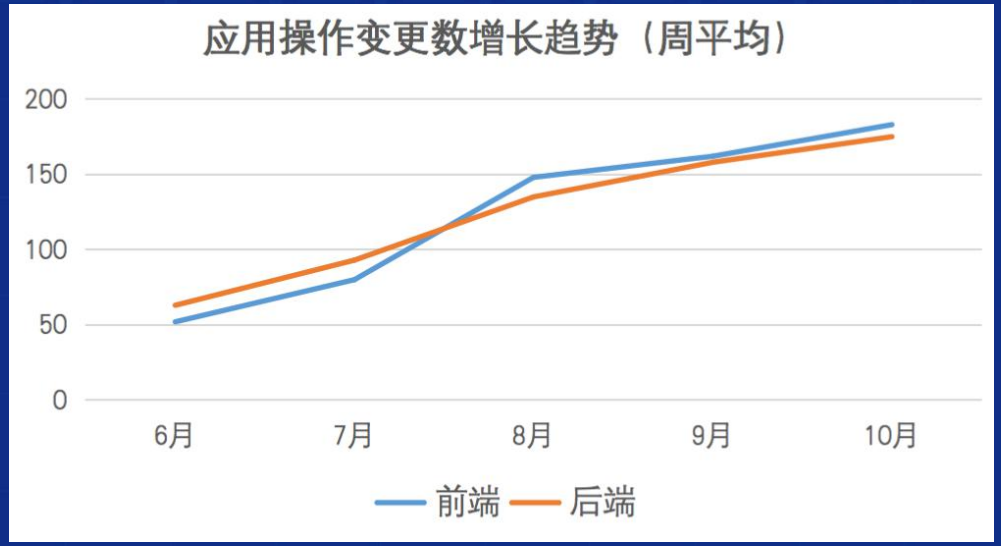
蓝绿/灰度发布

根据业务需求，灵活发布

半夜发布到 随时发布

每周发布几十次 上百次

提高 300%



云原生价值体现 - 支撑业务发展

腾讯云

腾讯会议

2020 年初，8天扩容10w 台云主机

疫情期间，在线办公需求突增，腾讯会迎来爆发式增长，短短一个月的时间，TKE容器服务结合自研AMD 机型，帮助腾讯会议快速扩容10w 台 主机资源，平稳的支撑了业务发展

某在线教育

平稳支撑了开学季拉新活动

2020年开学季，某在线教育投入了大量的资源来做拉新活动，在此期间，TKE 支撑客户业务平稳过渡。

腾讯云

02

腾讯云原生最佳实践

1000 + 客户云原生改造最佳实践总结

云原生改造的问题和挑战

腾讯云

性能

急速弹性

资源利用率

运维管理

易用

效率

稳定

负载不均

安全

内部推广

腾讯云

问题

- 节点crash(OOM)
- 网络毛刺
- Cpu 调度延迟
- 文件系统异常
- 内核死锁
- Docker/kubelet 夯住

云原生内核（Tkernel）

- 针对容器场景的大量优化
- 内核团队维护，热补丁修复策略
- 修复诸多公版内核存在的bug

NPD 增强

- 提供更多kubernetes 事件
- 事件告警，节点自愈
- 动态调度

99.95%

节点稳定性

腾讯云

场景

东西向流量

南北向流向

问题

原生NodePort 损耗

contrack 问题

Bridge 性能损耗

内核协议栈损耗

LB 直通Pod

避免Nat 转发带来的性能损耗

灵活的回话保持

解决ipvs 二次负载不均衡问题

独立网卡

每个pod 独占一张弹性网卡，

不再经过节点网络协议栈，

网络性能接近宿主机

支持Pod 绑定Eip

零损耗网络

多核计算性能

腾讯云

优势

- 高密度部署提升
- 资源利用率
- 控制面，网络中断offload到智能网卡

问题

- 跨Numa 调度问题
- 大量pod 并发启停
- 大内存回收问题
- Pod 间互相干扰

方案

- Numa 亲和性
- 云原生内核（场景优化）
- 弹性容器EKS

TKE + 黑石物理机

腾讯云

安全问题无小事

Runc 逃逸

Apiserver 越权

集群挖矿

全链路保障用户集群安全

从构建、分发、运行，提供全链路的容器安全能力

从仓库、镜像、集群、网络、运行时、合规，提供全栈的容器安全能力

腾讯云容器安全服务 **腾讯容器团队联合云鼎实验室/主机安全团队联合共同研发**， 沉淀腾讯二十年安全能力

腾讯云

大集群节点维护遇到的问题

- 由于资源或多应用混合部署需求，较多集群存在异构节点（cpu, gpu, amd）
- 异构节点扩容麻烦，每次均需把各个参数都设置一遍
- 集群存在异构节点，应用调度规则设置复杂，需要给不同机器设置各种业务标签
- 节点日常管理上如：kubernetes版本升级、docker版本升级、节点镜像管理等功能相对复杂

节点池方案

节点池粒度配置设置（配置模板）

操作系统，机型、规格、启动参数，label， 自定义脚本等

集群自动/快速扩缩容

修改节点池副本数，即可自动创建、销毁节点池下的node 节点。结合伸缩组、节点池，实现集群自动扩容（特定配置的机器）

节点池内节点自愈

通过配置 Npd-plus，oom—gurad等稳定性组件，提高节点稳定性。

如何急速/灵活弹性

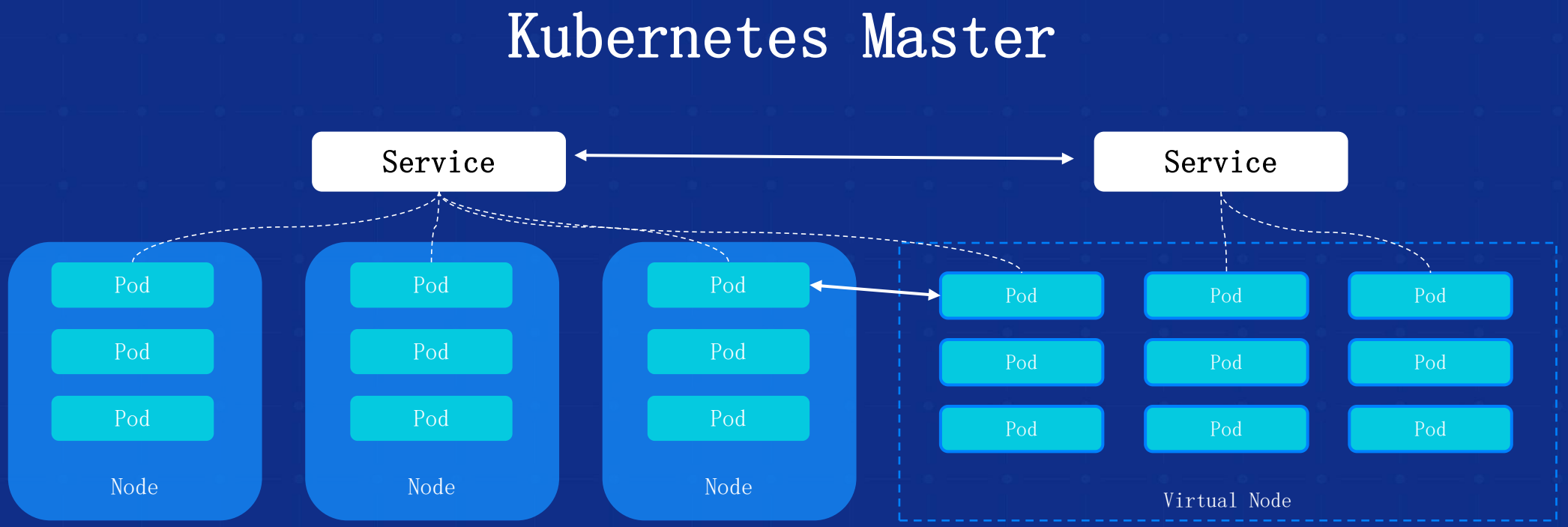
腾讯云

场景

- 大数据批处理
- 业务波峰波谷
- Job/定时任务/视频转码

客户收益

- 资源规划更简单，
- 弹性资源按量计费，成本更低



如何解决集群负载不均衡的问题 - 增强调度

腾讯云

动态调度器 Dynamic Scheduler

场景

集群负载不均衡

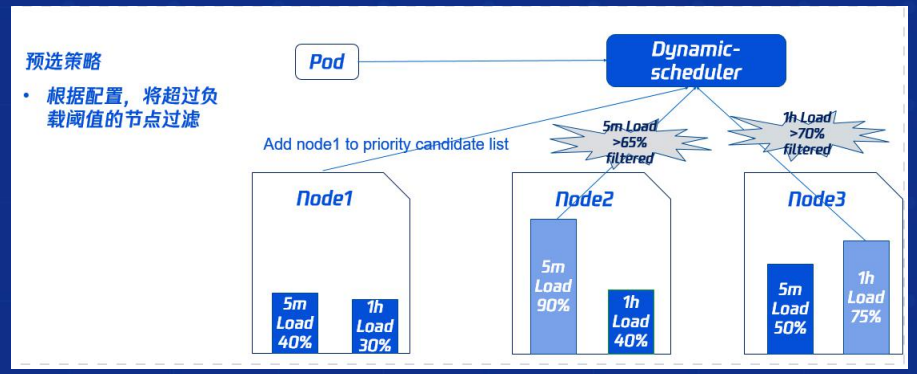
应用调度到热点机器上

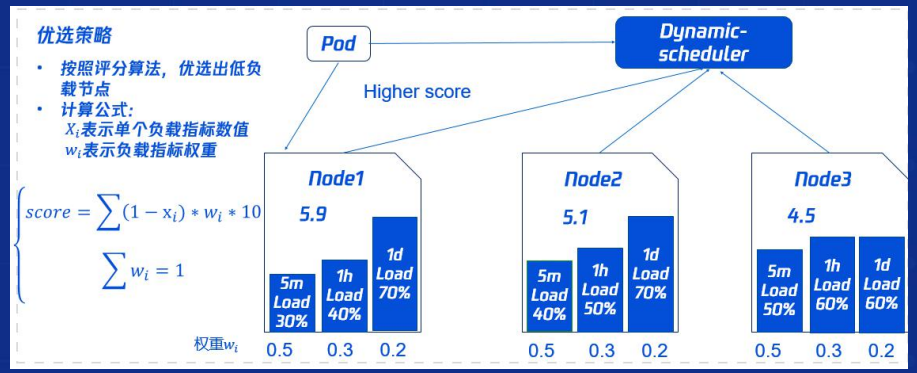
用户收益

根据真实资源使用情况调度，负载更均衡

避免热点调度

Kube Scheduler - Extender机制





重调度器 Descheduler

场景

集群存量节点负载高，

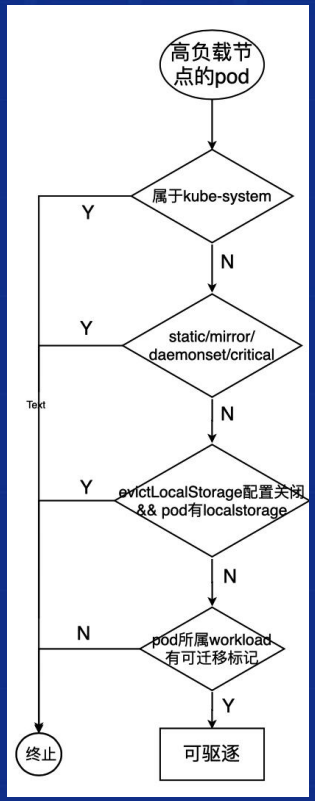
新加的节点负载低，需要均衡。

用户收益

资源再均衡

安全保障

- Pod驱逐时执行筛选逻辑（重要pod不可被驱逐）
- 用户手动标记的业务才会被驱逐
- 只在workload ready的pod比例 > 50%时执行驱逐



进一步提升资源利用率 - 在离线混部

腾讯云

场景：

- 在线任务优先级高、波峰波谷明显
- 很多情况下和在线任务错峰处理
- 整体资源平均利用率低

收益：

- 资源利用率明显提升
- 在离线任务统一调度，运维管理

1. 大数据/离线任务如何容器化？

Yarn on TKE

- 渐进式大数据容器化解决方案
- EMR on TKE

2. 集群中有多少资源给离线使用？

资源预测和回收

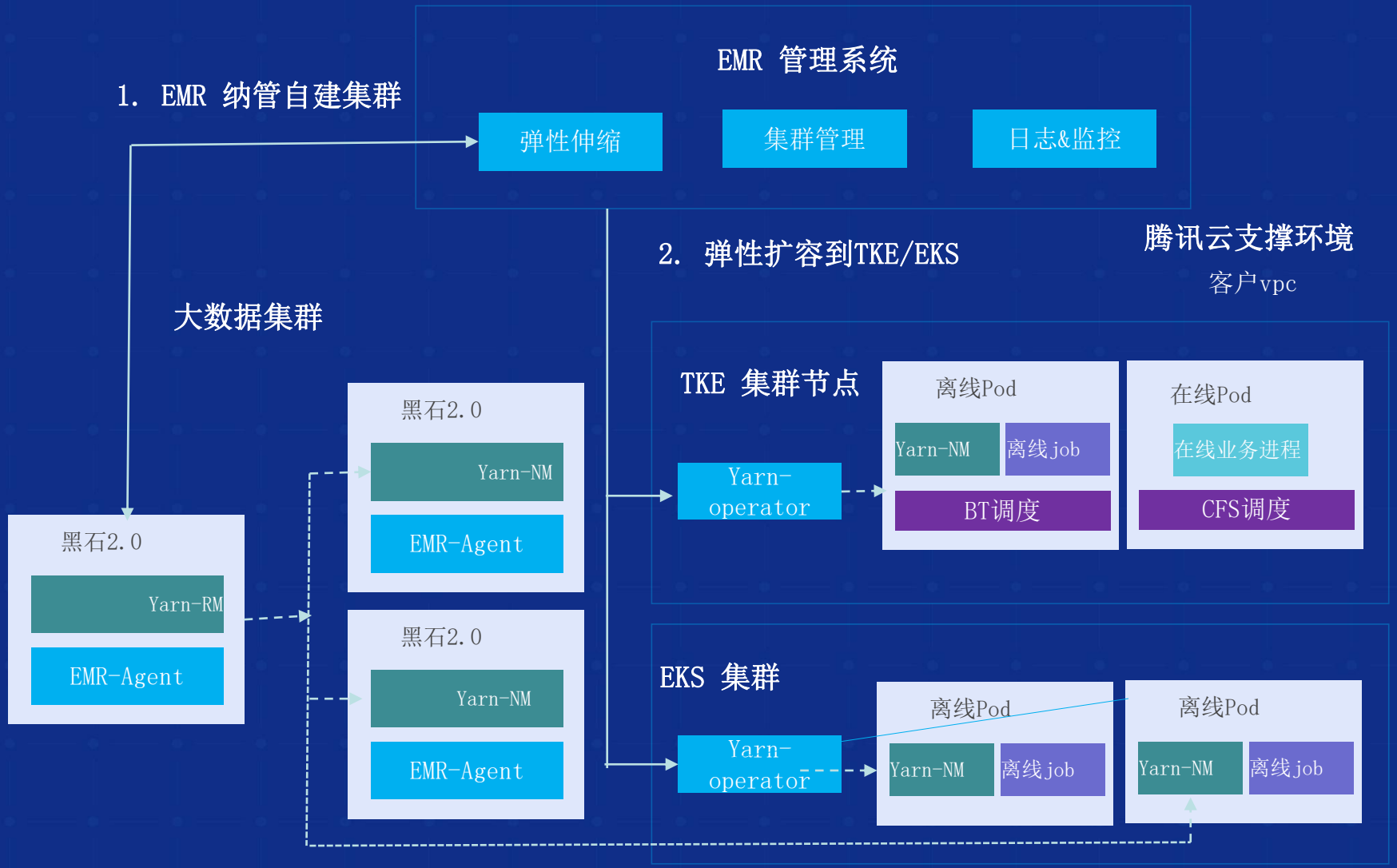
- 扩展离线resources：
- 根据在线服务实例监控数据，预测离线资源，动态更新离线资源可用使用情况。

3. 如何避免在离线任务的互相干扰？

资源Qos 隔离

云原生操作系统 TencentOS， 提供cpu，内存，网络，磁盘Qos

在离线混部 - 典型使用方式



腾讯云

起步最为艰难，短期收益不明显还会遇到各种问题，
决心和方法缺一不可

文化与知识

- 开源协同
- 云原生知识库
- 云原生课堂
- 云原生沙龙

奖励机制

- 云原奖励
- 晋升答辩

质量体系

- 云原生成熟度模型
- 打分机制

更多实践成果

腾讯云

混合云解决方案	AI解决方案	大数据解决方案	在离线混布解决方案
TkeStack	TKE发行版	EKS	TCM
零损耗网络	云原生kernel	GPU虚拟化	集群不停服升级

- 在自研业务落地过程中，打磨出了大量优秀的能力
- 混合云：多云环境下的统一调度与管理
- 降本解决方案：基于成本分析大盘、负载推荐、混合调度、GPU虚拟化和弹性能力的全套云原生降本解决方案
- TKE发行版&TKEStack：海量业务打磨后的k8s runtime和管理系统的开源输出

腾讯云

03

典型客户案例

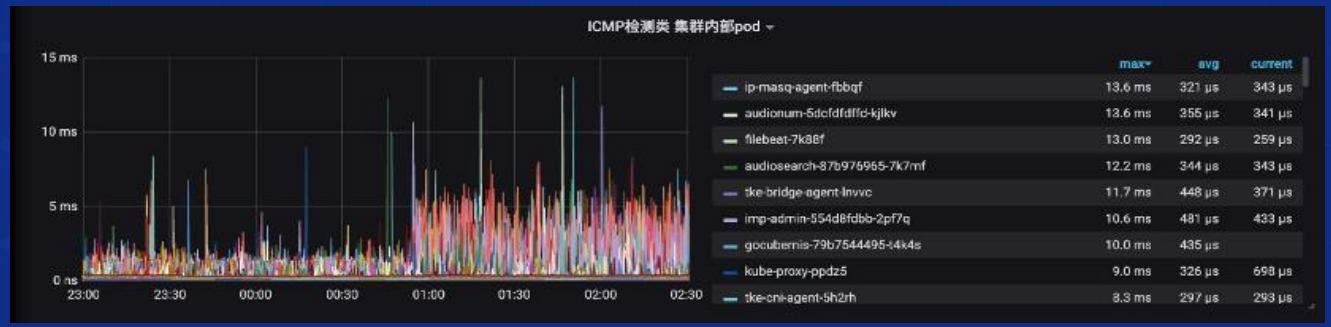
典型案例-某在线教育-客户痛点

时间周期特性明显

- 高峰时段是平峰时段的20倍，低峰时段上百倍
- 流量上升曲线较陡
- 资源使用率低

业务毛刺明显

- 业务间调用响应时间抖动厉害，
- 高峰期更明显



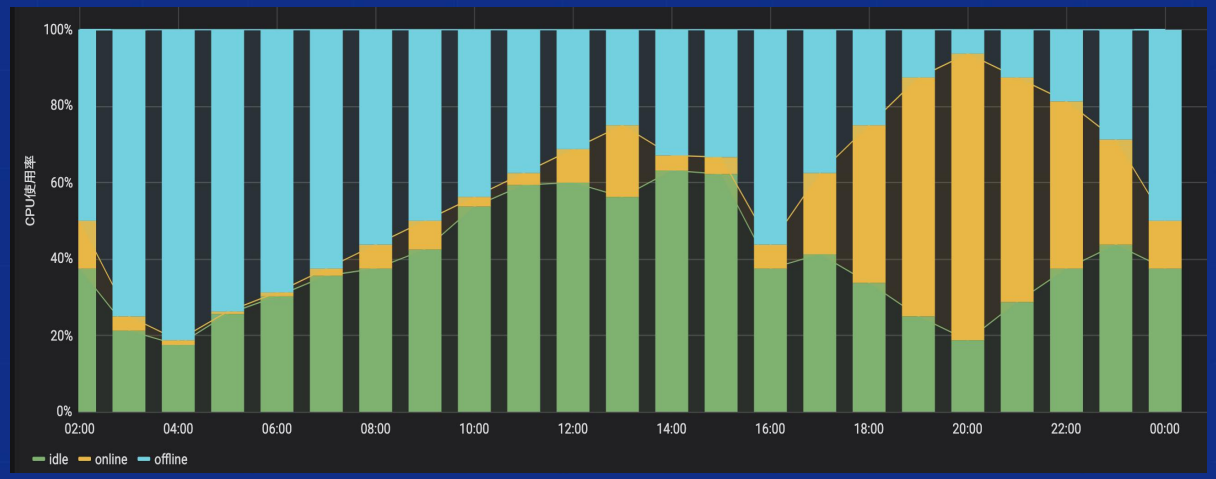
典型案例-某在线教育-解决方案与收益

腾讯云

使用在离线混部技术

在离线任务CPU，内存，网络Qos 隔离

不影响在线业务提前下，提升整体资源利用率。



使用云原生内核

解决大量内核跟容器运行时的bug

优化包括内核调度、ipvs、contrack等问题

使用serverless 容器

解决突发和临时算力不足问题

解决业务高峰弹性需求

解决特务业务隔离性需求

成本下降 **43%**

稳定性提升到 **99.995%**

接口响应提升 **10%**

典型案例-七人普- 客户特征

腾讯云

时间紧

部署，测试，上线两个月时间

高性能

14亿C 端用户，700w B 端用户

高可靠

10年一度，国家级项目，确保万无一失

典型案例-七人普- 解决方案

大量使用云原生技术

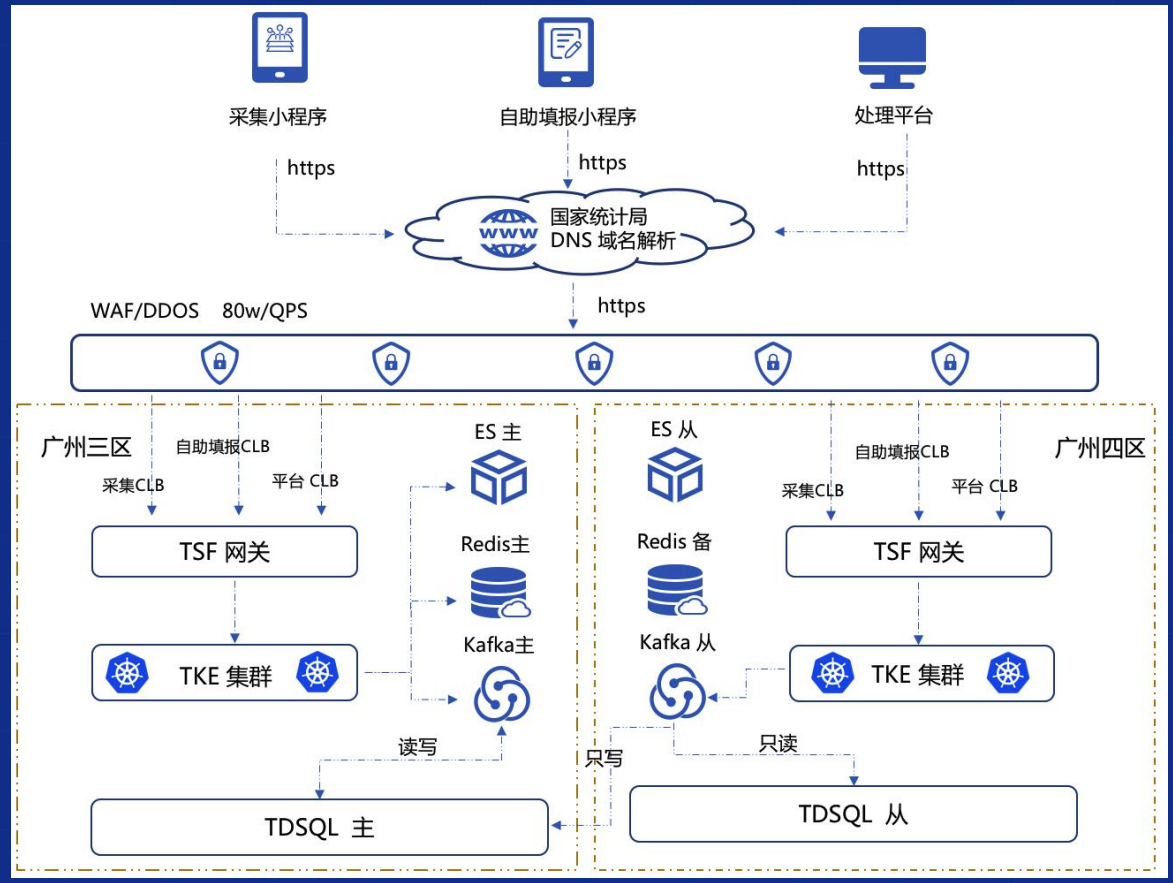
- TKE，TSF，TCB，TDSQL，Ckafka，Redis 等
- 灵活弹性

跨可用区双活架构

- 服务间流量可用区内闭环
- 数据层主从架构
- DNS 快速切流

稳定性和性能保障

云原生内核，节点自愈



云原生
产业大会
2021 CLOUD NATIVE
INDUSTRY CONFERENCE
2021/05/26 × 中国 / 北京
原生蓄力 云领未来

THANKS!

CAICT 中国信通院

腾讯云原生 - 欢迎关注

腾讯云

扫描二维码 关注
我们