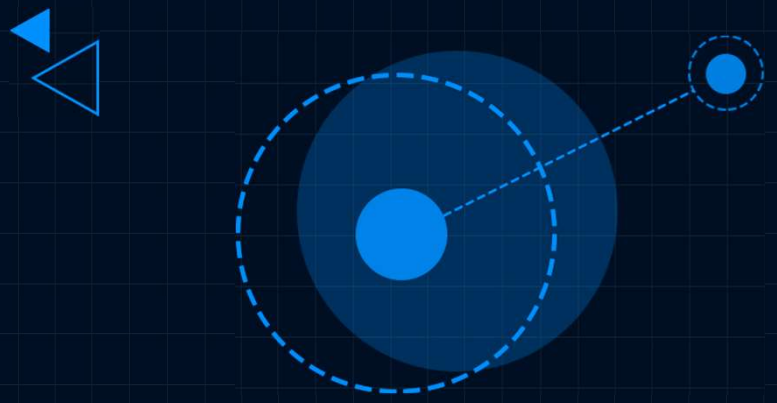
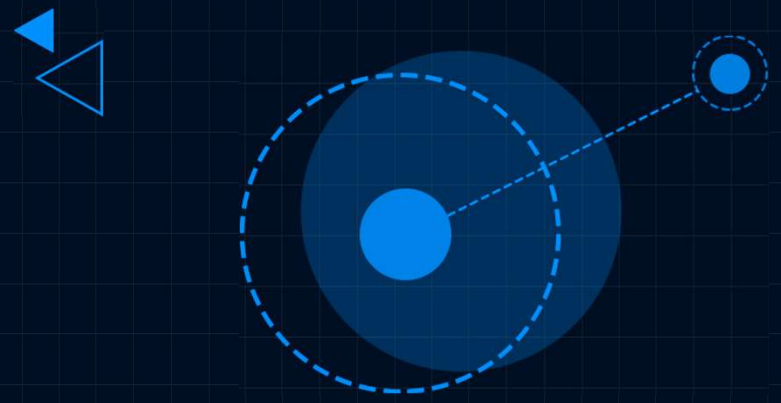
TencentOS Server云原生特性优化实践

蒋彪 @ CID2021
腾讯云虚拟化团队



坐标成都，专注内核/虚拟化，虚位以待

benbjiang@benbjiang.com



目录

TencentOS产品布局

覆盖云、边、端全场景

云

云服务器操作系统

云场景深度定制优化，高性能，安全可靠

TencentOS Server

边

边缘计算操作系统

云端计算能力边缘化

TencentOS Edge

端

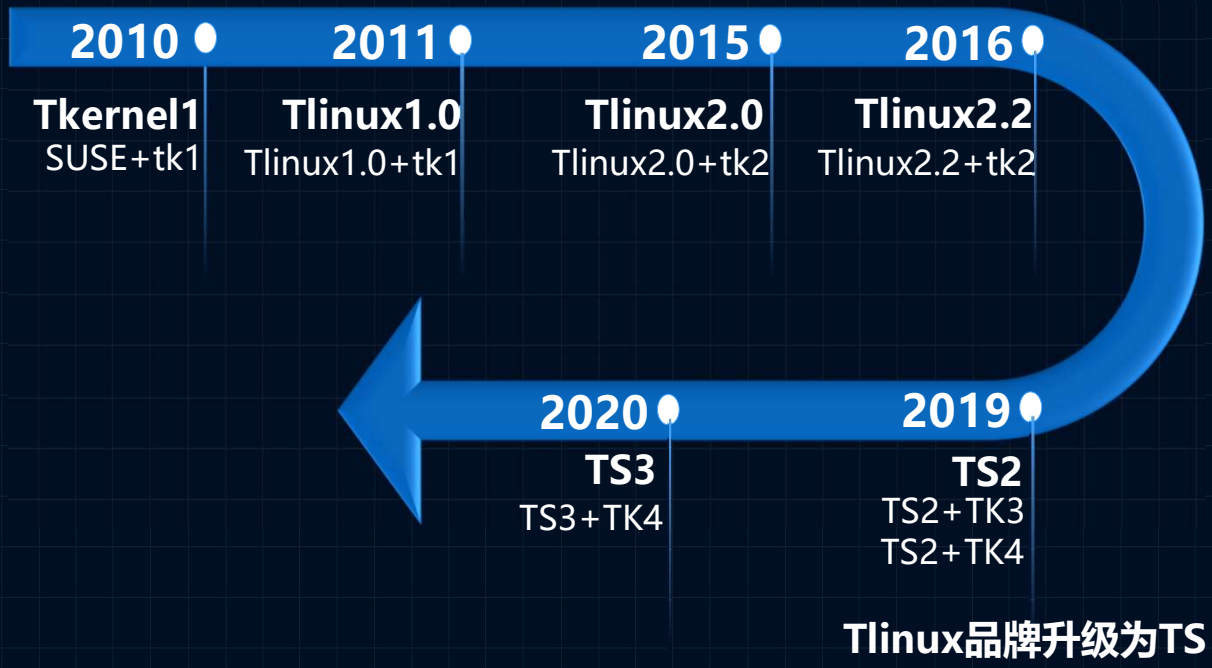
物联网操作系统

低功耗，低资源占用，模块化，安全可靠，端云协同

TencentOS Tiny

TencentOS Server历程

10年+征程



- 10+ 年技术迭代和积累
- 400万+ 商用节点
- 100% 覆盖Tencent业务场景，海量场景考验

内核解耦、独立演进(L1)

TencentOS产品特性

为云而生

为云开发的OS，适用于各种工作负载，是全面开放的 Linux 操作系统，含有最新的、基于开放标准的虚拟化和云原生工具

全面优化

经过全面优化，高度定制的高性能 OS，针对系统内的各类负载都进行了深度优化，让用户获得更高性能。典型场景性能提升50%+

极致稳定

10+年专业打磨，99.999%系统可用性，400w+商用节点海量应用考验，95%宕机自动分析成功率

生态共建

完全开源，社区协同共建。致力于将腾讯开源生态中的优秀成果引入其中，如Tencent Kona，TKE Kubernetes Distro等。

专业服务

国内顶尖的技术团队提供服务，每版本提供长达5年的支持。提供丰富的系统性能和故障定位工具能提升服务效率；拥有宕机自动化分析能力、零停机内核修复能力，保障业务连续性

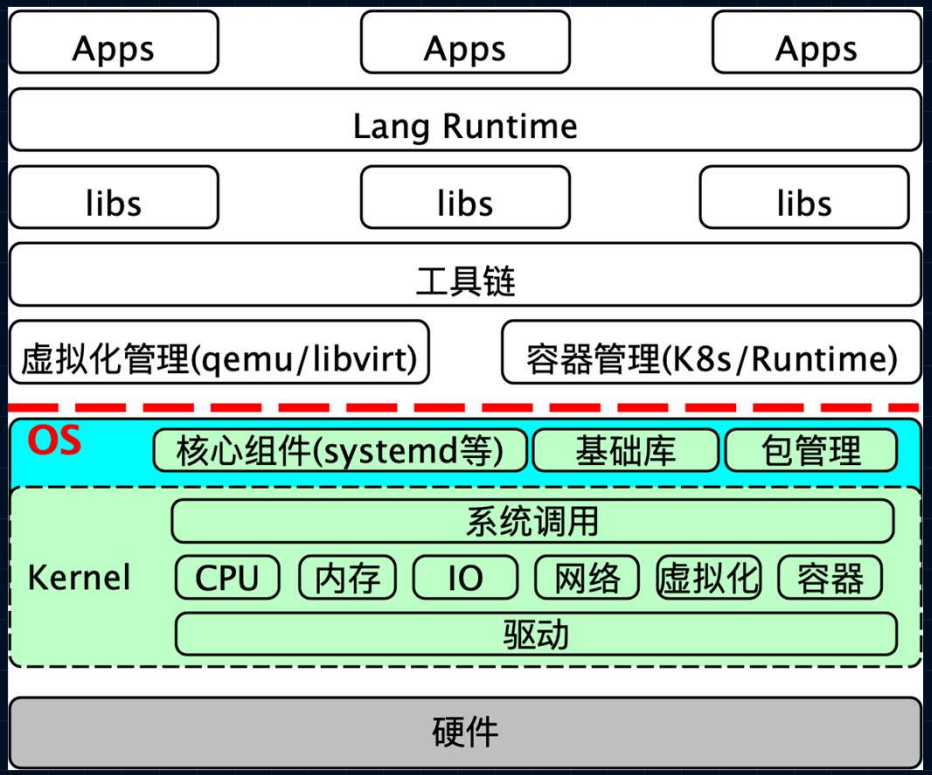
服务集成

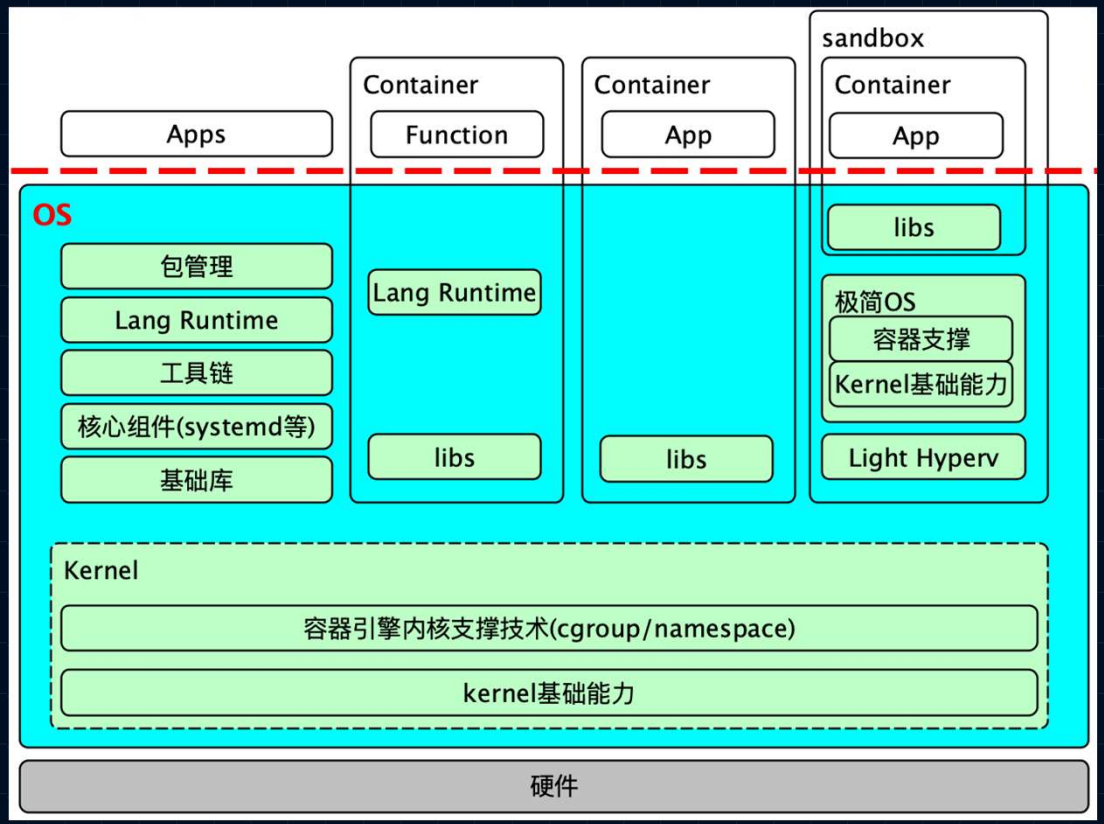
完全融入腾讯云产品体系，为您提供满足不同工作负载的完整的解决方案。 腾讯云安全专家服务为企业提供安全咨询、渗透测试、应急响应、等保合规等服务

目录

云原生 For OS

应用之外皆为OS





传统场景

- OS专注于核心能力(机制)，业务提供策略
- OS应用层较薄，业务臃肿

云原生场景

- 云原生场景业务专注于业务逻辑，其他交给OS
- 边界上移，应用之外皆为OS

目录

TencentOS For 云原生

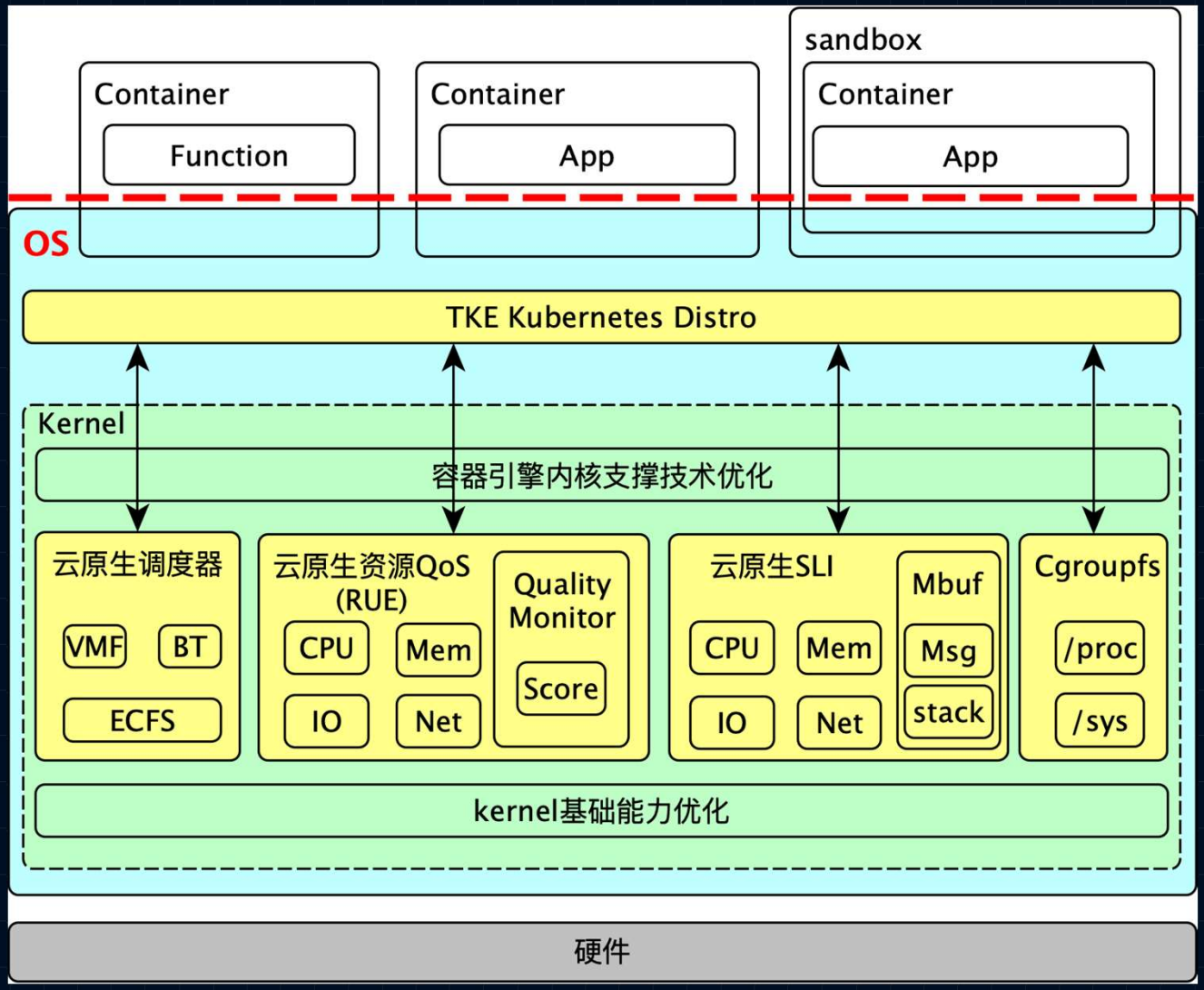
整体架构

Features

- 云原生调度器(TCNS)
- 云原生资源隔离(RUE)
- Quality monitor(服务质量监控)
- 云原生SLI(容器视图的专业指标)
- Mbuf(常态化内核关键监控)
- Cgroupfs(容器资源视图隔离)

整体解决方案

- 结合腾讯云原生场景
- 与TKE深度融合(TKS k8s发行版)



云原生调度器

内核调度器的两次重构

背景

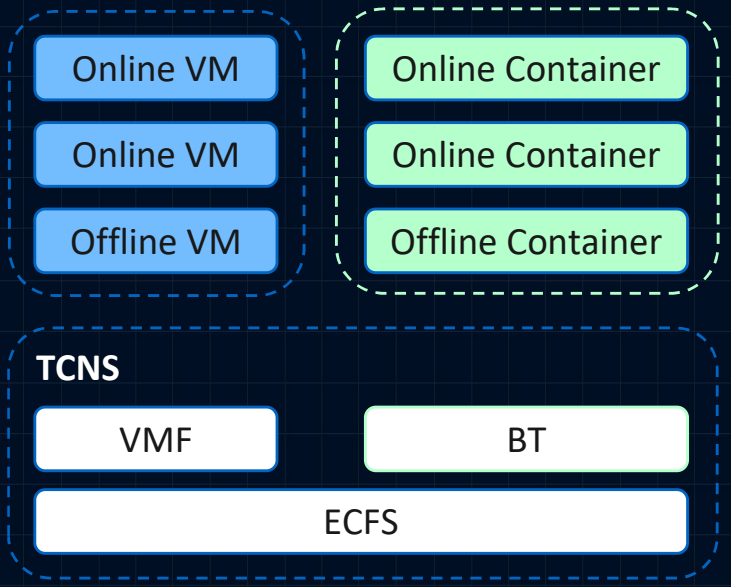
- 硬件Offload，虚拟机运行环境越来越简单
- 云原生化趋势，容器承载更多、更复杂的业务
- 虚拟机和容器场景对内核调度器有完全不同的诉求

痛点

- 虚拟机实时性差，干扰大，性能差/稳定性差
- CPU全售卖
- 混部CPU干扰（绝对压制、超线程干扰隔离）

解决

- VMF(VM First)- 虚拟机调度完整解决方案
- BT(Batch)- 离线任务专用调度器
- ECFS(Enhanced CFS)- 增强CFS



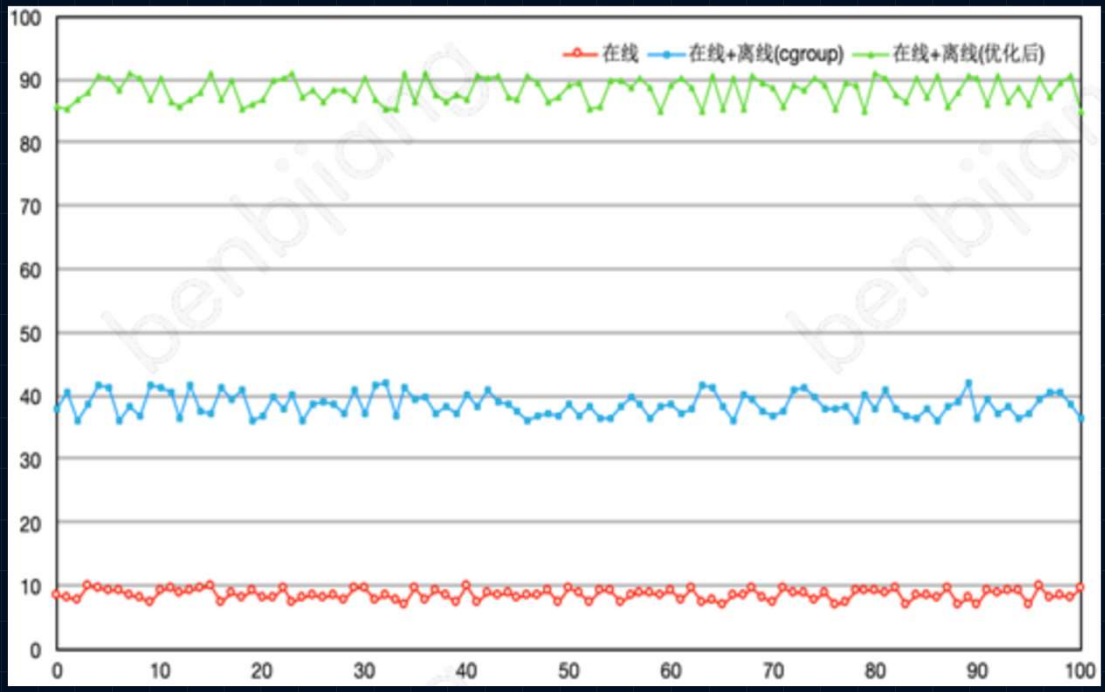
效果

- VMF(VM First)- 微秒级延迟、零干扰
- BT(Batch)- 全混部场景支持，零干扰
- ECFS(Enhanced CFS)- 社区路线

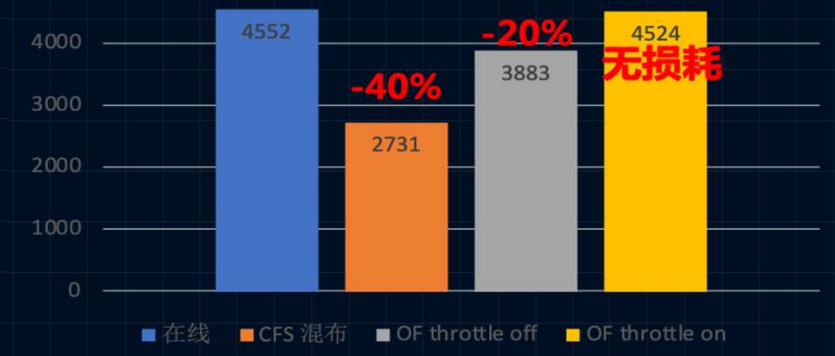
效果-云原生调度器

测试延迟（测试工具：cyclictest）

	类型	S5 VMF	S5 CFS
时延 Idle (ns)	Max(us)	116	4689
	Overflow（100us）	0.28	0.82
时延 Busy (ns)	Max(us)	452	19969
	Overflow（100us）	2.2	20

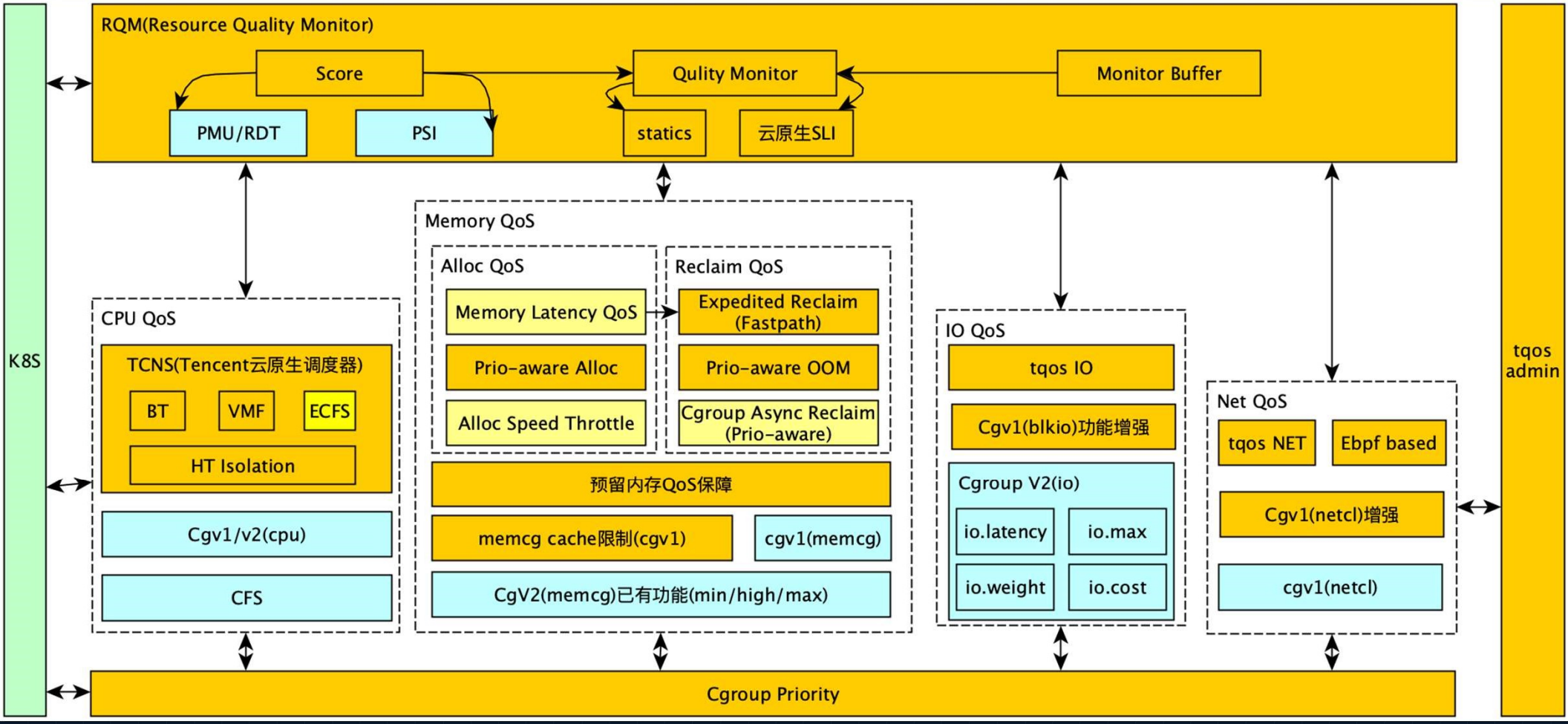


业务延迟（测试工具：codis(真实业务)）



如意(RUE)

统一资源隔离(混部底层隔离完整解决方案)



效果-如意(RUE)

痛点:

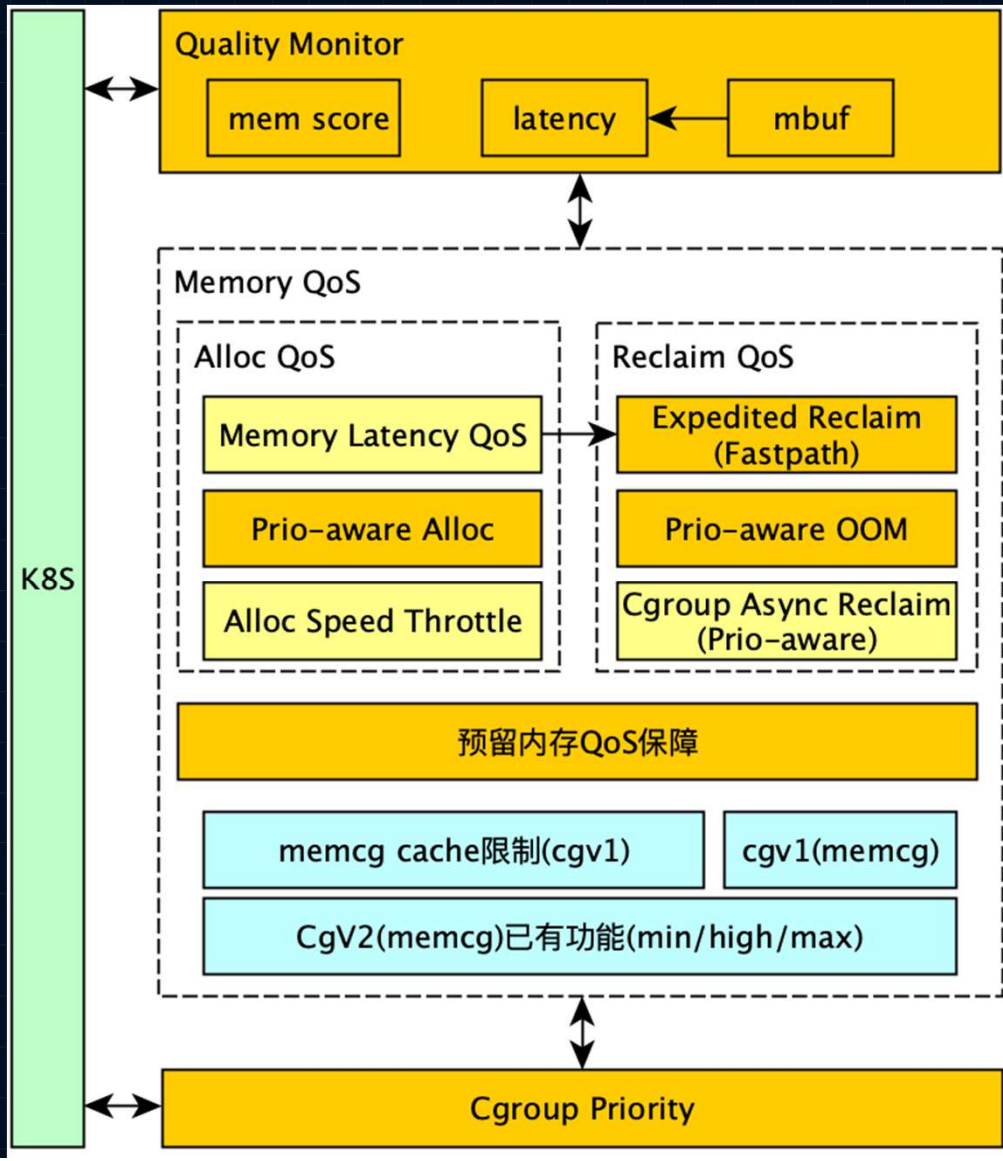
- 内存不可压缩，复用难度大，混部深水区
- 内存(分配)延迟无保障，经常抖动(us->ms->s级)
- 内存回收慢、不区分优先级，不可控
- 内存分配不区分优先级，速度不可控，相互干扰

解决:

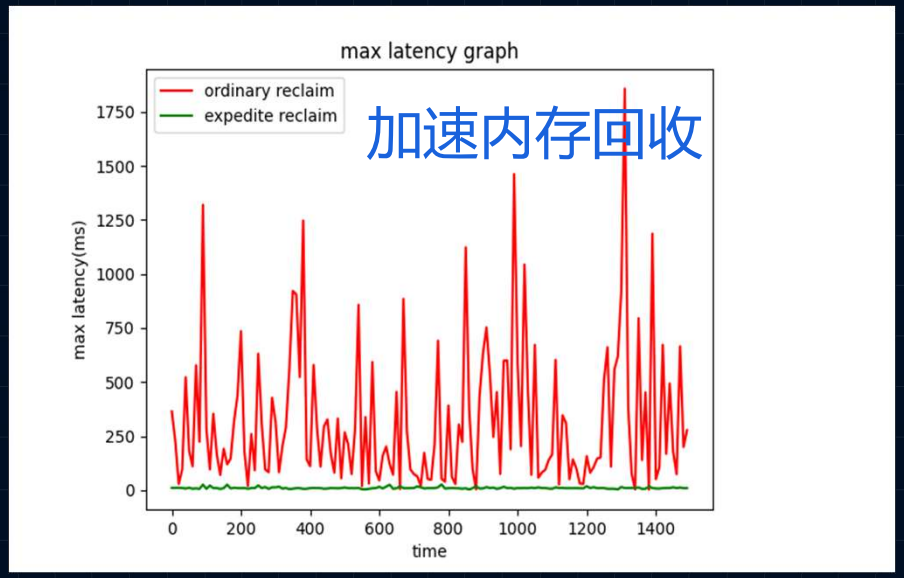
- 优先级分配(分级水线)
- 优先级回收(优先级OOM/Reclaim/异步回收)
- 加速回收
- 内存分配限速(MST)
- Latency QoS保障
- 保留内存

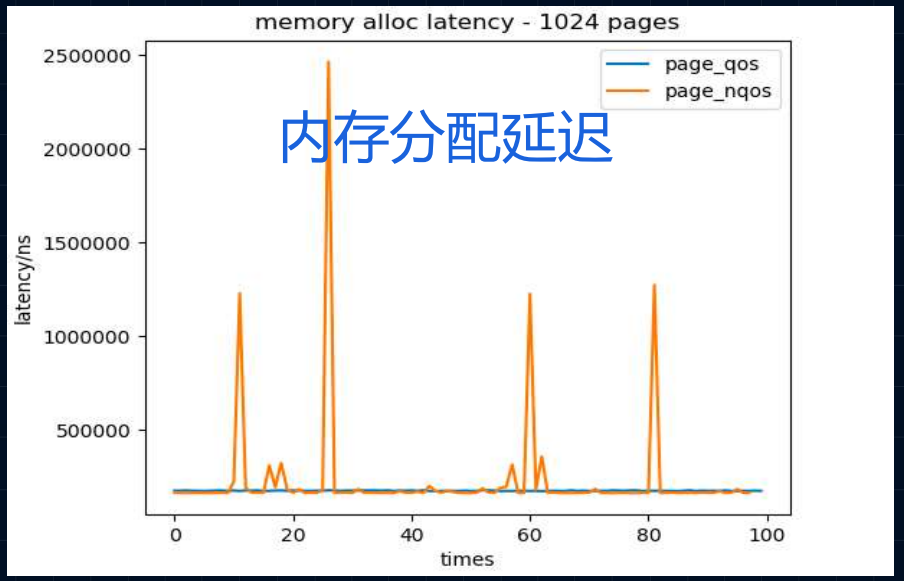
效果:

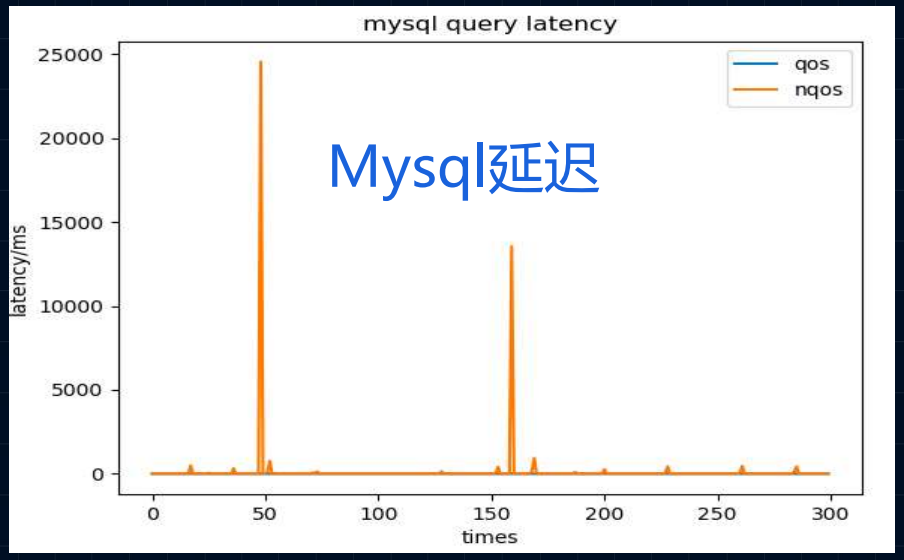
- 保障高优内存分配Latency

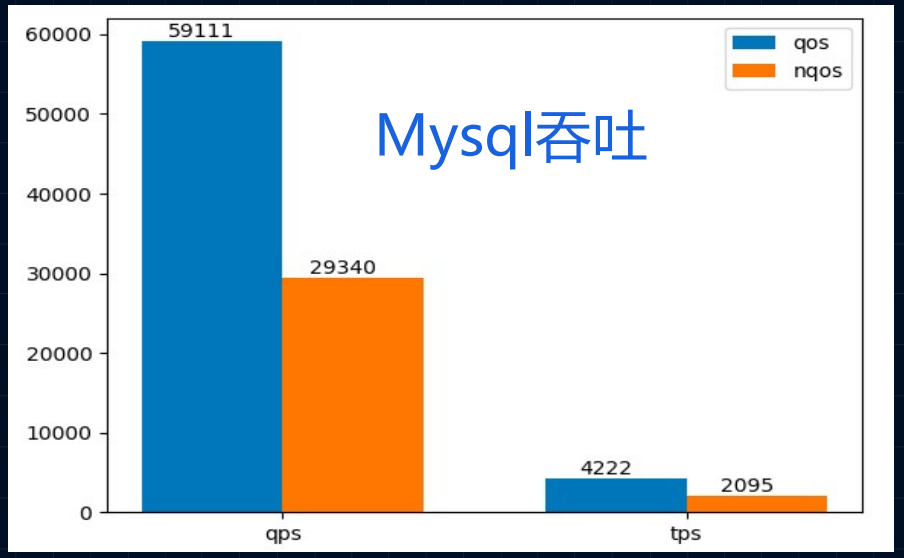


效果-RUE


max latency graph
ordinary reclaim
expedite reclaim
加速内存回收
max latency(ms)
time


memory alloc latency - 1024 pages
page_qos
page_nqos
内存分配延迟
latency/ns
times


mysql query latency
qos
nqos
Mysql延迟
latency/ms
times


59111
29340
4222
2095
qos
nqos
Mysql吞吐
qps
tps

云原生SLI

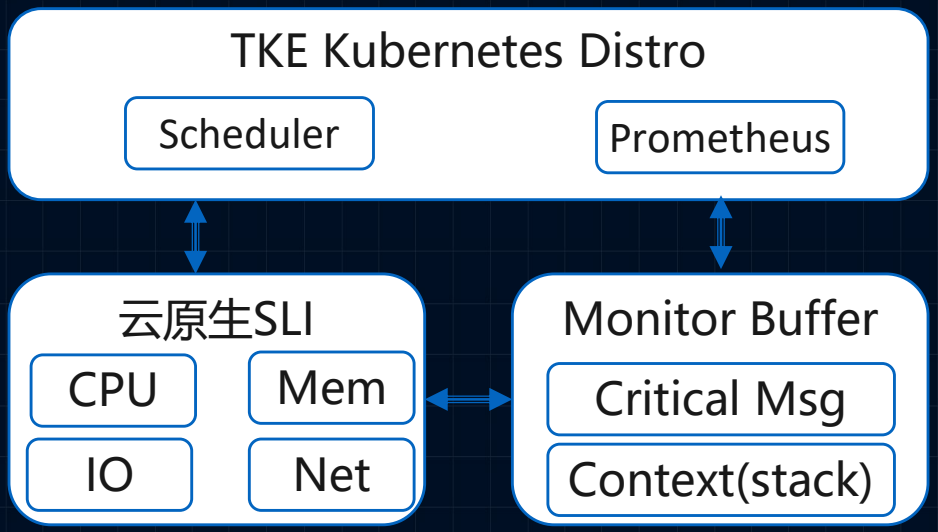
容器专属指标和监控，助力业务稳定性提升

背景

- 容器业务多样而复杂
- 容器隔离性天然较弱
- 容器级别指标缺乏

容器平台痛点

- 业务稳定性感知难、保障难
- 资源利用率普遍偏低，提升难
- 敏感业务随机抖动难题



解决(容器平台与OS深入融合)

- 云原生SLI
- Monitor Buffer
- TKE k8s发行版

效果

- 平台可视化能力提升，助力业务稳定性提升
- 基于SLI的调度优化，业务资源利用率提升
- 常态化内核实时监控，捕捉每一次随机抖动

Cgroupfs

容器资源视图隔离

背景

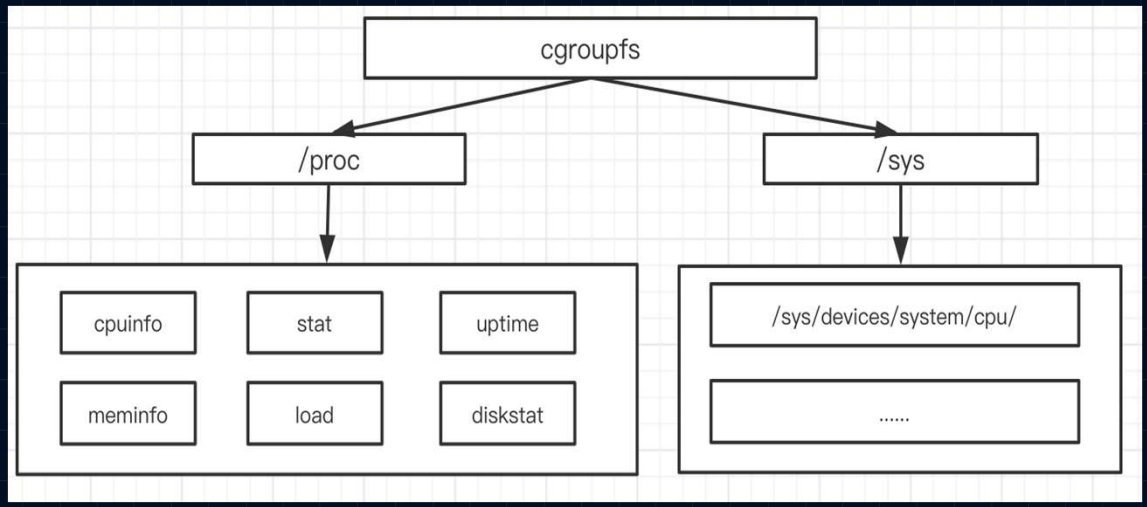
- 容器隔离性不完整
- 资源视图(/proc、/sys)未容器化

Lxcfs痛点

- 用户态实现，稳定性差(hung,crash无法恢复)，开销大
- 用户态信息获取，可扩展性、定制性差
- 额外依赖Lxcfs组件
- 富容器信息不准确

Cgroupfs

- 独立文件系统，与Lxcfs兼容(k8s适配)
- 全内核实现，可靠，开销可控
- 可深度定制，更专业
- 容器平台深入融合，识别容器，完善解决方案



目录

未来

邀你一起共建

全面开源

- https://github.com/Tencent/TencentOS-kernel

OS社区筹建中，敬请期待，邀你共建

Thanks_