国产操作系统自主核心能力打磨实践

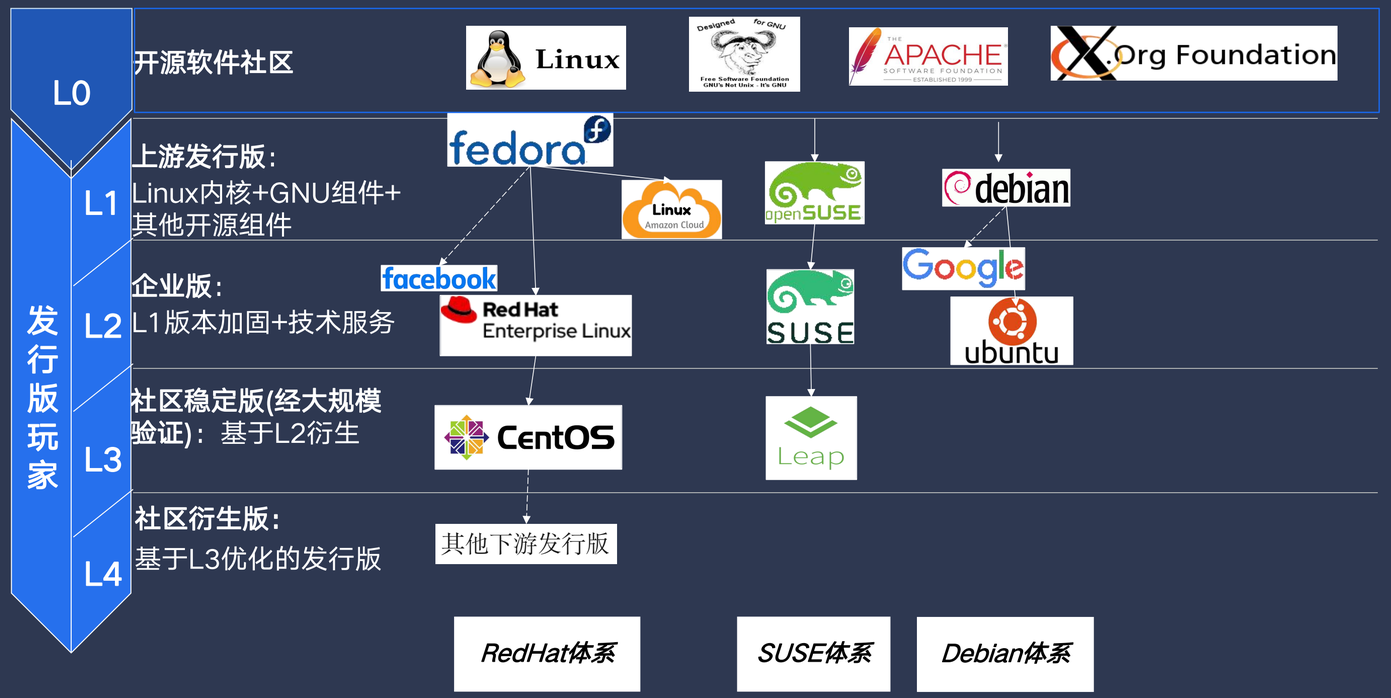
–TencentOS Server进击之路

蒋彪–腾讯云操作系统研发负责人/OpenCloudOS社区ToC委员

目录

- Linux行业背景
- TencentOS Server简介
- 经济操作系统打磨实践
- 绿色操作系统打磨实践

Linux行业背景-前



L1国产发行版不足

L1上游发行版需聚焦创新，投入大，社区版本未经过大规模生产环境验证，**非稳定版本，无法直接用于生产环境**

L2国产商业版不足

L2国产商业版本稀缺。主要原因是上游社区维护**能力与投入不足**

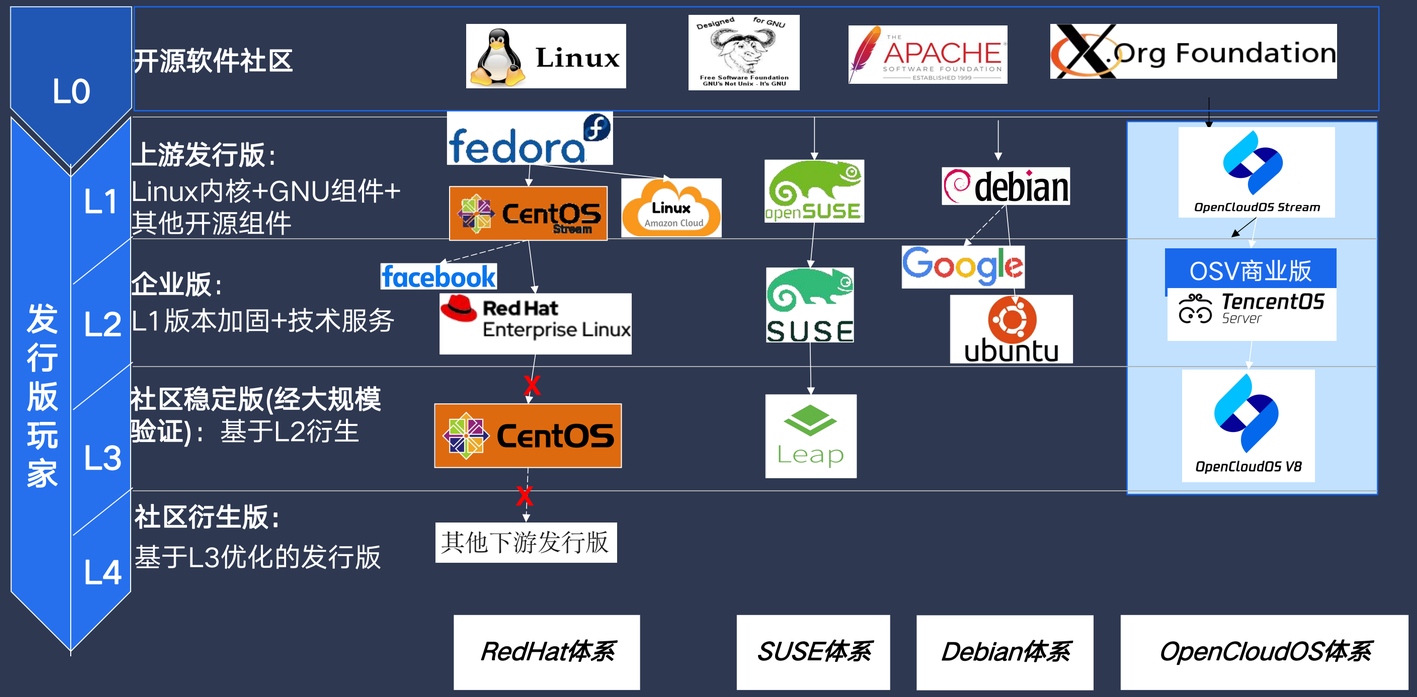
L3/L4国产发行版不足

L3/L4社区聚焦版本的稳定和生产价值，但需要依赖可靠上游版本（商业版本）；

供应链风险暴露、核心能力不足、国产OS亟待自主

Linux行业背景-后

OpenCloudOS覆盖L1\L2\L3全链路，实现全链路国产化，输出生产级可用版本



行业问题:开源供应链安全风险

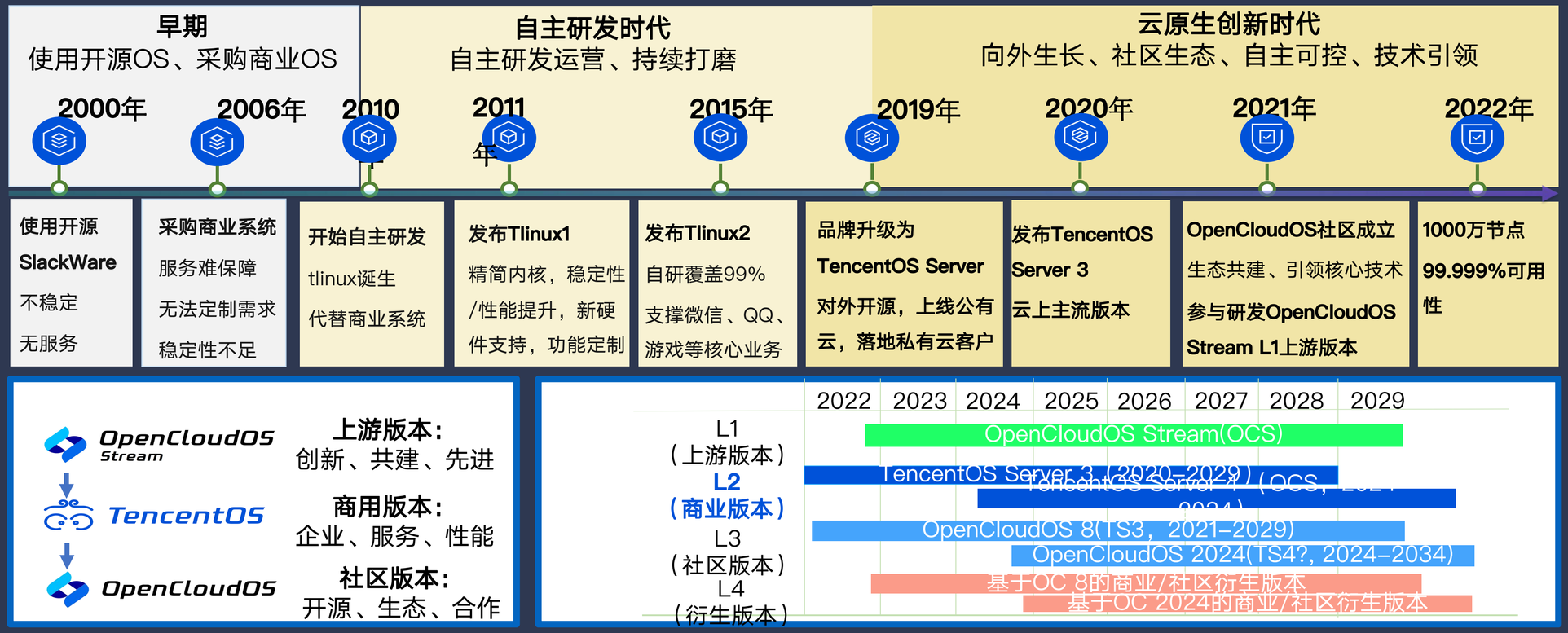
红帽不再维护CentOS8；国产OS对其强依赖，影响较大；充分暴露**开源软件供应链安全风险**

目录

- Linux行业背景
- TencentOS Server简介
- 经济操作系统打磨实践
- 绿色操作系统打磨实践

TencentOS Server简介

三个时代、十年积累、千万节点

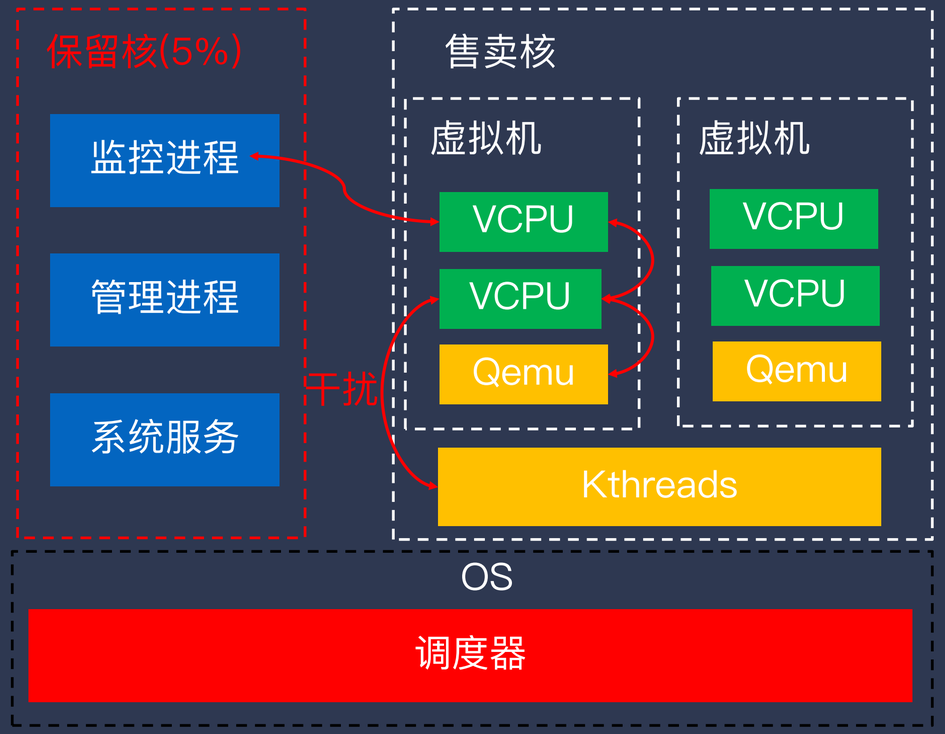


目录

- Linux行业背景
- TencentOS Server简介
- 经济操作系统打磨实践–降本增效
- 绿色操作系统打磨实践

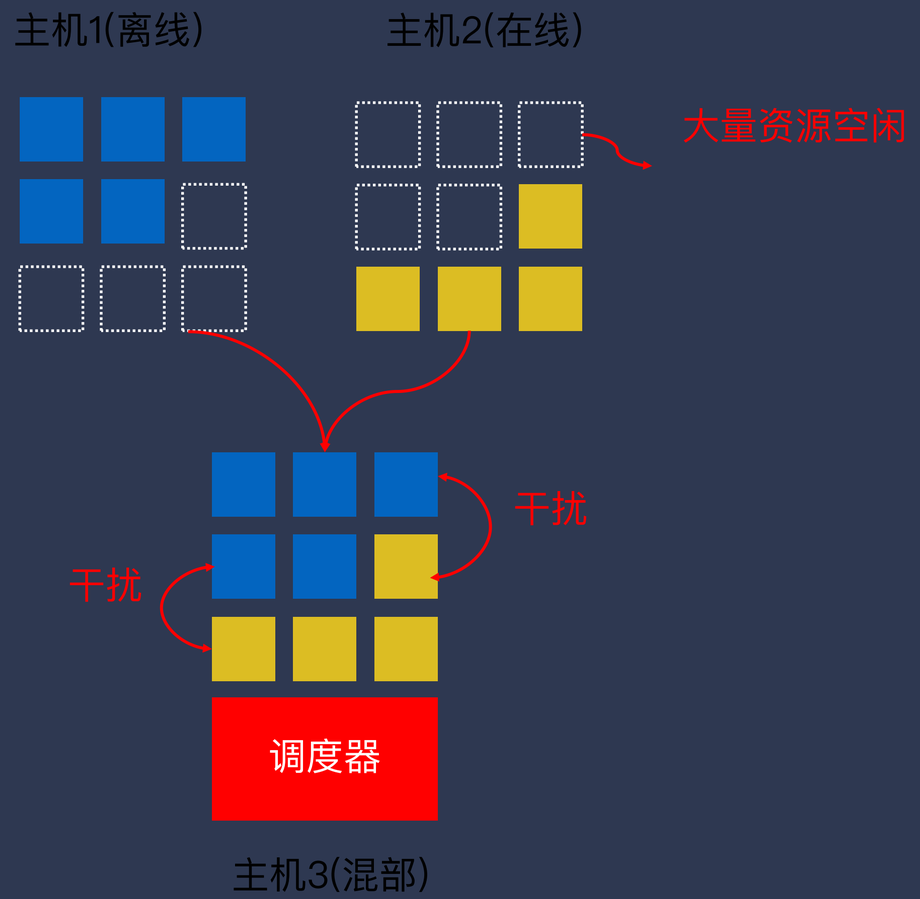
VMF(VM First)调度器–背景

(CPU全售卖场景)



(CPU全售卖场景)：5%保留核，干扰严重，实时性差
核心目标：CPU全售卖，微妙级延迟

(混部场景)



(混部场景)：大盘资源利用率低(15%)，离线干扰
核心目标：绝对压制离线，业务无感知

核心：OS内核调度器

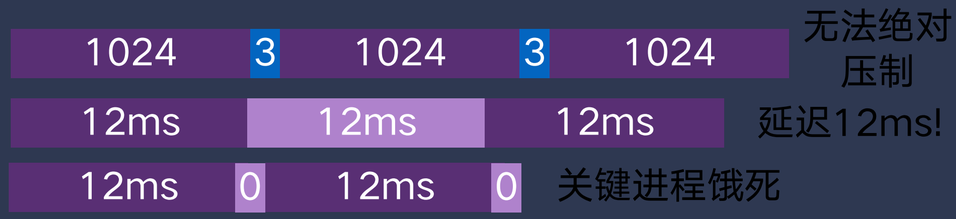
VMF内核调度器–设计

核心挑战
- CFS无法满足要求，需要重写
- 原因：公平性设计

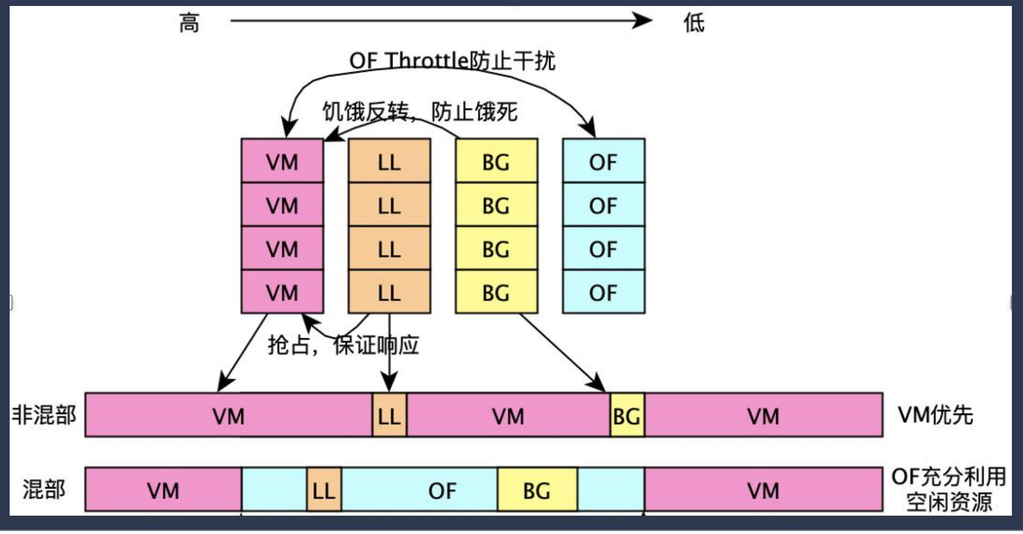
核心设计（基于任务类型的非公平调度器）
- **特征画像–>任务分类**
- **离线任务绝对低优先级**

效果
- VM优先(VM First)，更实时
- 对离线绝对压制，完美隔离

高级特性：
- **超线程干扰隔离**
- **BG饥饿保护**
- **超线程协同调度**
- 动态MWait
- ...



任务	VCPU	内核线程	普通进程	离线任务
优先级	4	4/3	1	0
运行时长	长	短	长	长
延迟敏感	是	是	否	否
容忍饥饿	否	否	否	是
抽象建模	VM	LL (Lowlatency)	BG (BackGround)	OF (Offline)



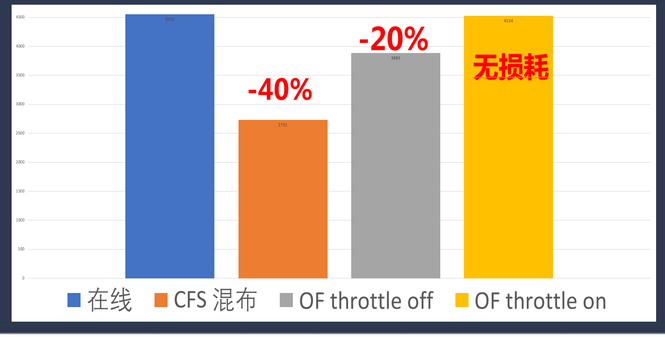
VMF内核调度器–效果

(全售卖) 实时性（测试工具：cyclictest）

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

- 延迟微妙级，提升1个数量级
- CPU全售卖

(混部) 吞吐性能（测试工具：sysbench）



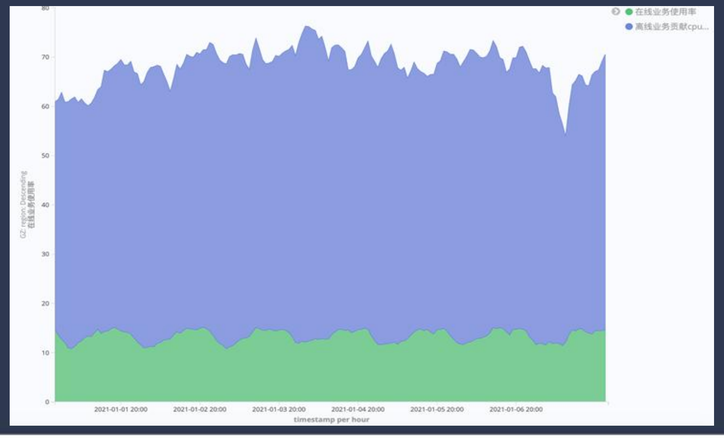
(混部) 现网业务(AMS)压测表现

- 在线业务对离线无感知
- 通过真实业务压测(敏感程度不同)

4月28日	朋友圈–mixer					
					0428 18:30–22:00	
业务子机	状态	母机	离线子机	离线子机	cpu使用率	失败率
9.142.4.32	混合	9.178.80.239	9.142.132.85	9.142.132.81	63%	0.04%
9.142.4.33	对比	9.178.81.154			63%	0.04%

(混部) 资源利用率

- 样板集群CPU达65%，行业标杆
- 大盘CPU利用率翻倍



如意(RUE)-容器混部-架构

1 场景进阶

虚拟机混部->容器混部->多优先级混部

2 架构进阶（三层架构）

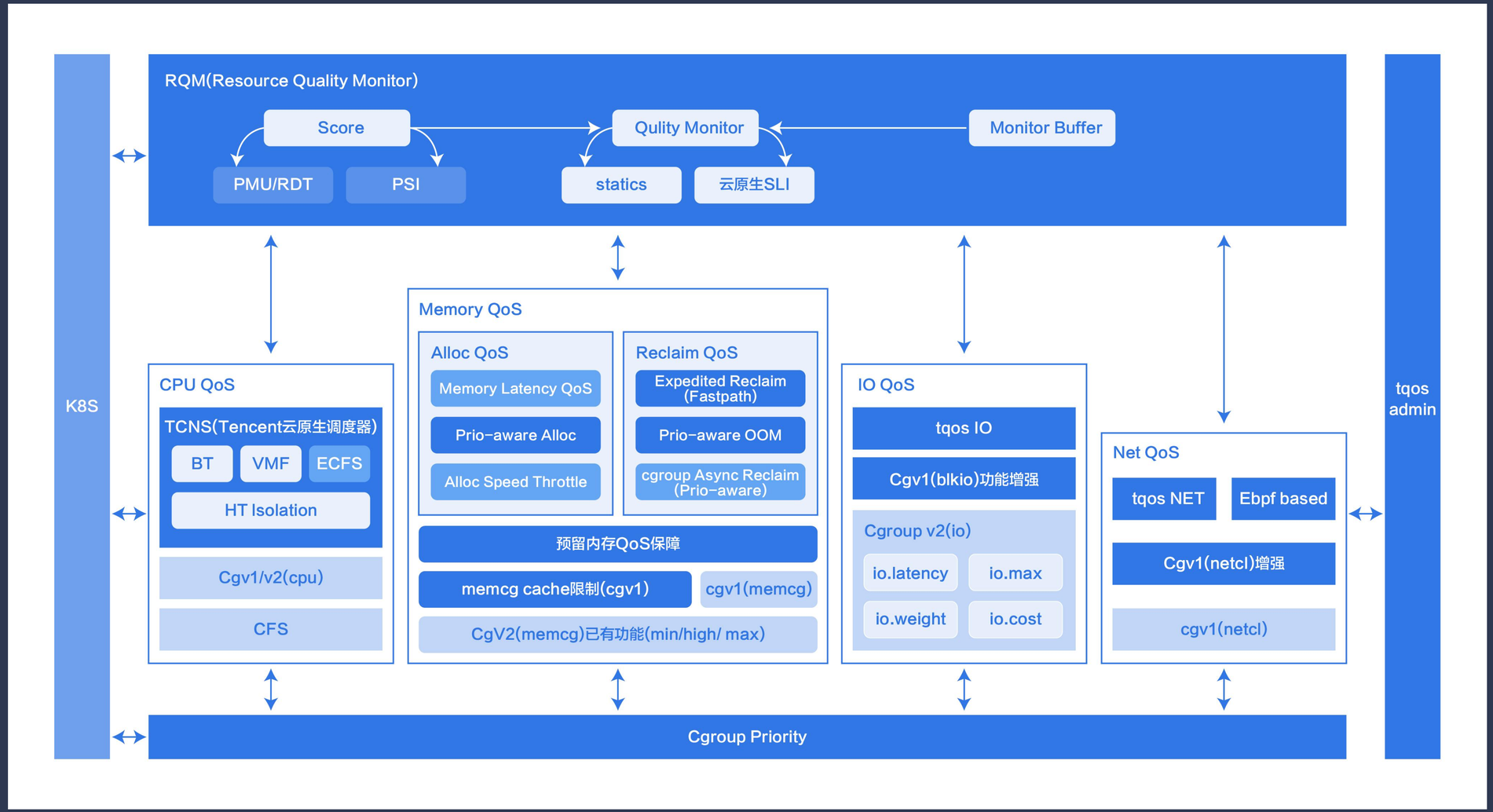
- 统一优先级
- 资源全隔离
- 服务质量监控框架

3 资源隔离进阶

CPU->内存、IO、网络(全覆盖)

4 影响力进阶

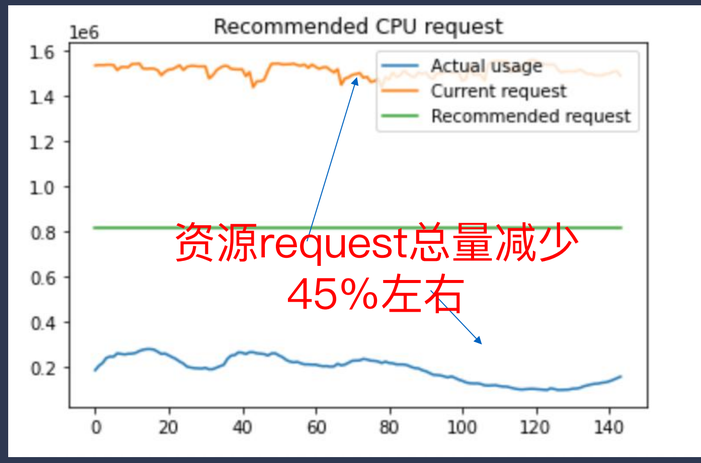
技术品牌打造：如意(RUE)



如意(RUE)–多优先级混部–经济操作系统

AI模型+如意(RUE)，提升资源利用率，降低成本

基于FinOps理念，基于RUE实现多优先级混部，



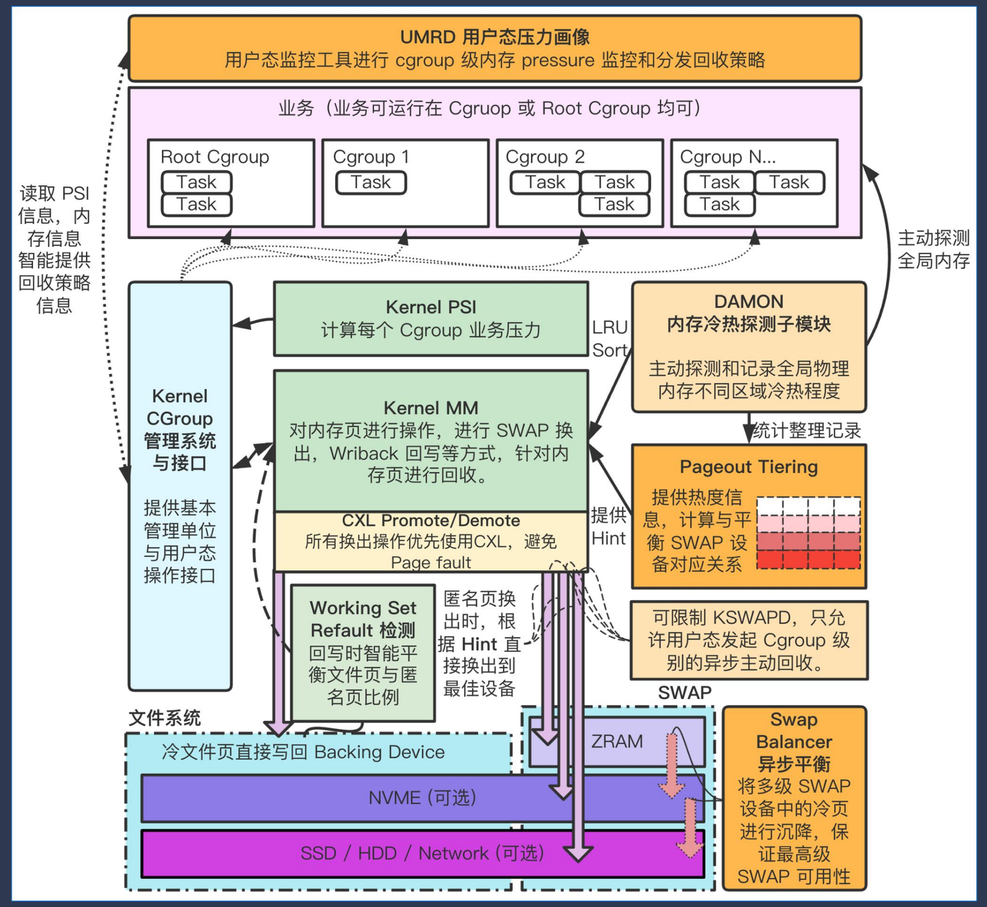
Pod规格从8–16核减少为1–4核

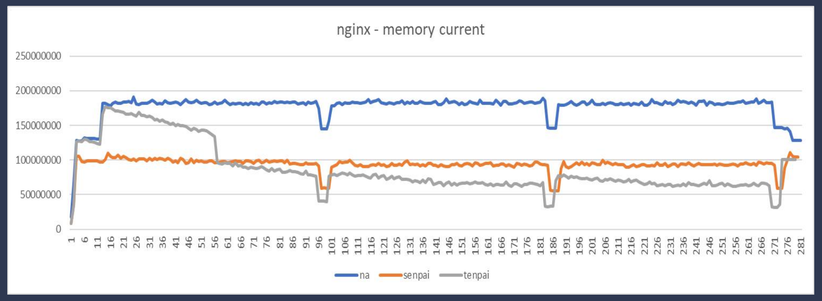
内存分级卸载–悟净

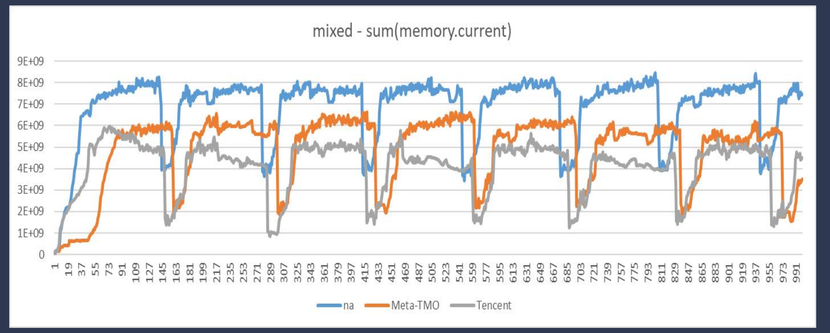
•UMRD模块：根据PSI模块提供的cgroup内存访问延迟敏感性，决策出对应cgroup中能够回收的页面量。

•Pageout Tiering模块：结合社区DAMON物理地址监控功能，在待回收的页面链表中，根据页面冷热频率（DAMON动态迭代的采样频率）换出到不同速度的后备设备上。

•SWAP BALANCER模块：每个SWAP后备设备维护一个LRU链表，当本SWAP设备快满时，demote冷页到速度更慢的设备。





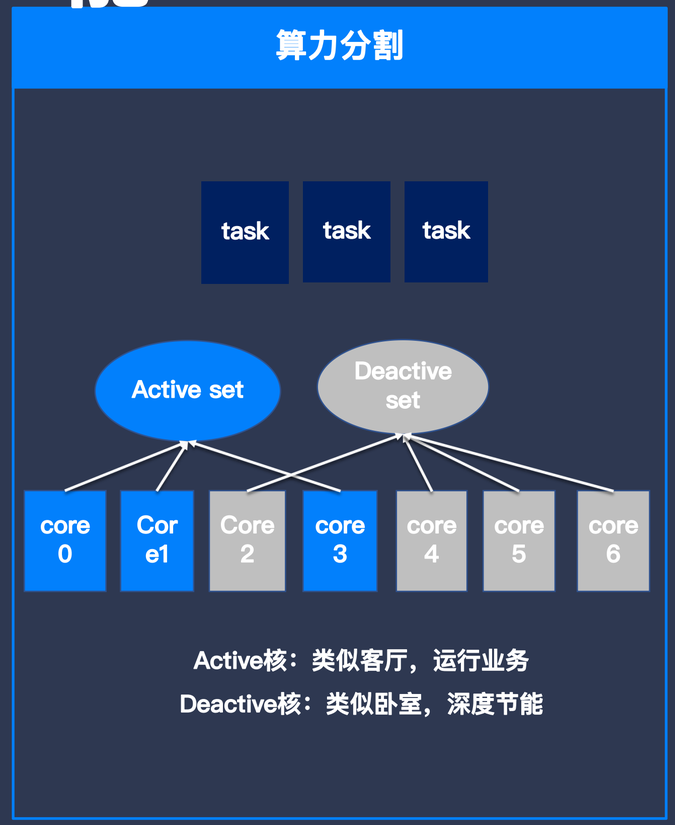


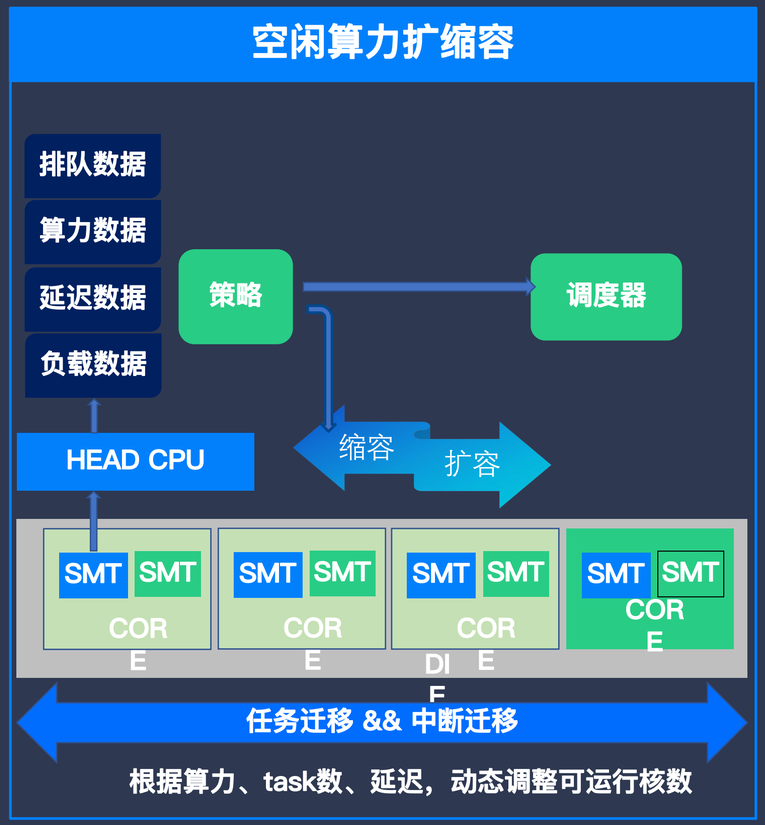
内存节省30%+

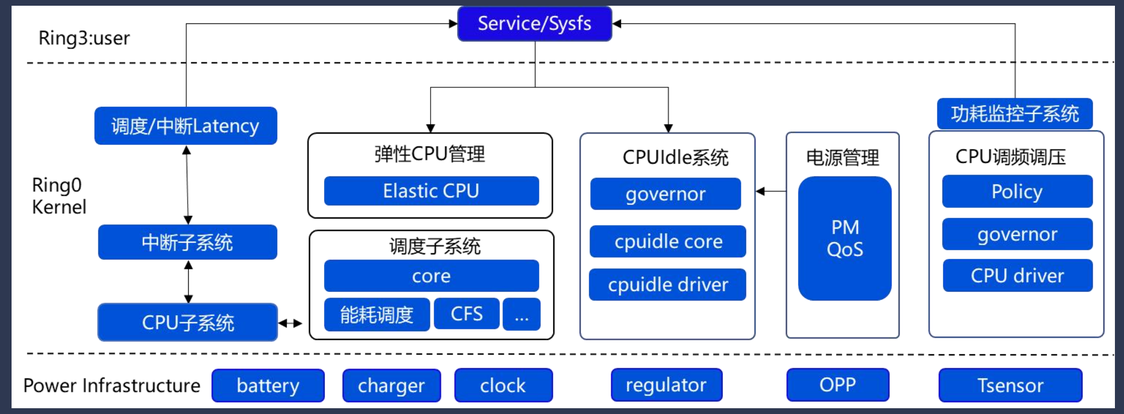
目录

- Linux行业背景
- TencentOS Server简介
- 经济操作系统打磨实践
- 绿色操作系统打磨实践–节能低碳

绿色操作系统–系统级能耗优化–悟能







- 空闲算力感知设计，完全自适应业务负载，业务侧透明，性能影响小于1%
- 根据业务实时负载动态调整cpu core深度睡眠状态、 cpu core频率以及调整 uncore频率
- 自动退出机制，轻松应对请求突发高负载场景
- 提供多个配置接口，用户可结合业务场景敏感度调节节能选项
- 双平面功耗监控

想一想，我该如何把这些技术应用在工作实践中？

THANKS