OpenEuler

调度系列分享之调度器简史

目录

CONTENT

OpenEuler

目标

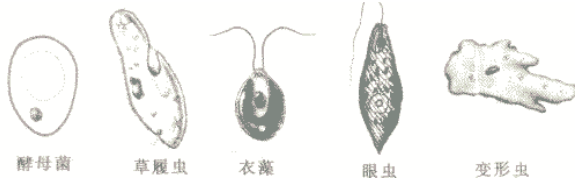
回答两个问题：

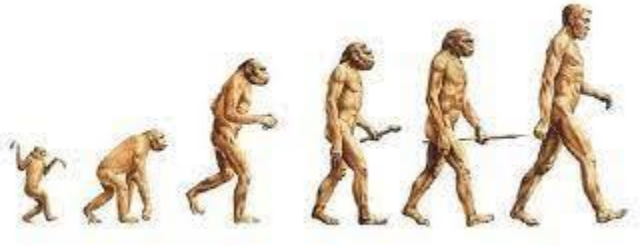
推动调度器发展的动力是什么？

Linux调度器都包含哪些关键子模块？

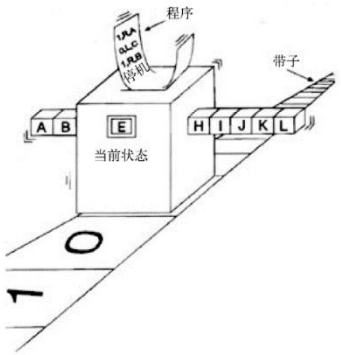
演进

> 调度是为了解决资源和需求之间的不匹配问题，现实往往是资源少&需求多，计算机领域也是如此。





回答更复杂的问题



Before 1955: Human operators
1955: 1st OS with batch scheduler (GM-NAA)
1967: Multiprogramming (IBM OS/360 MFT/MVT)
1968: Multiprocessors (IBM OS/360 M65MP)
1971: Time sharing (IBM OS/360)

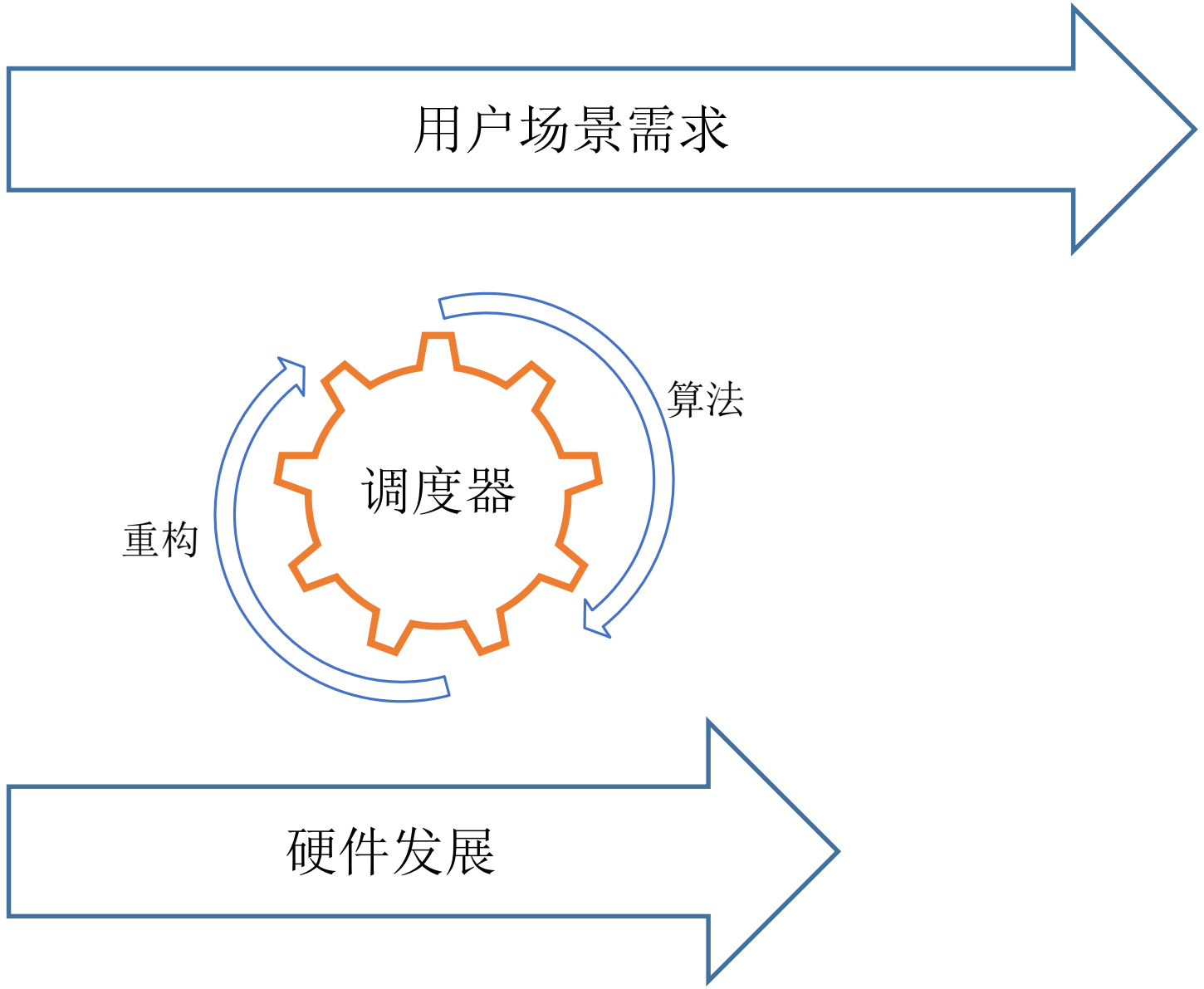
e.g. linux
1991: RR
2001: O(n)
2003: O(1)
2007: CFS

e.g. CFS+
NUMA
SMT
CPUFREQ
big.LITTLE

e.g. CFS+
嵌入式
云&计算
终端

OpenEuler

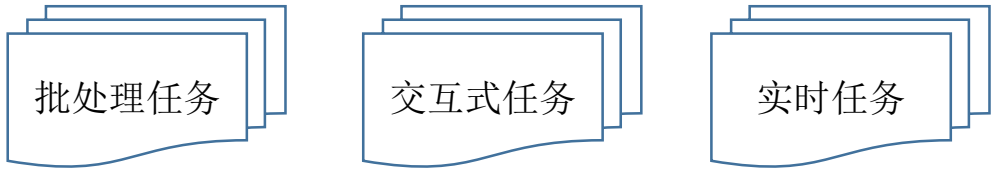
调度器发展的动力



调度器组件

OpenEuler

早期调度器

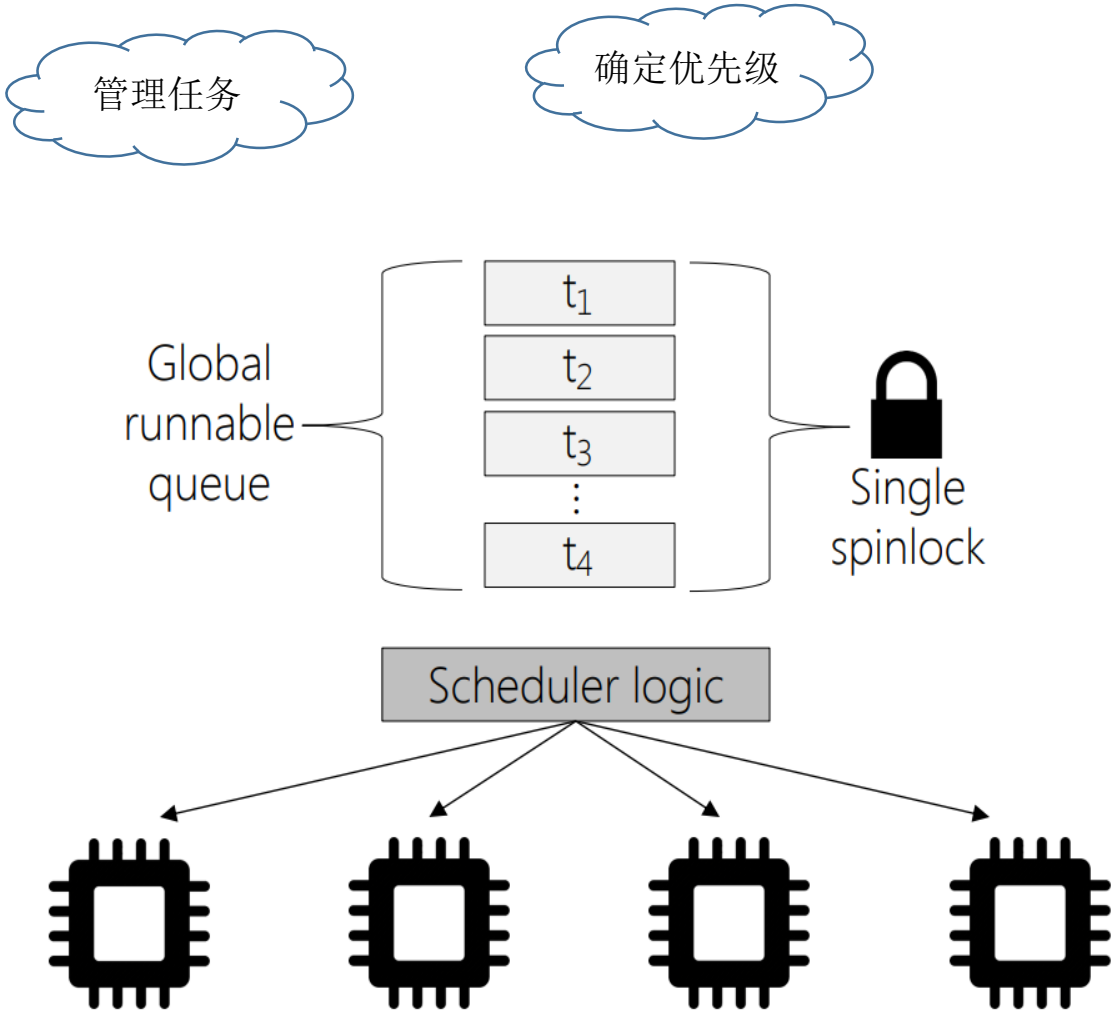


Before 1955: Human operators
1955: 1st OS with batch scheduler (GM-NAA)
1967: Multiprogramming (IBM OS/360 MFT/MVT)
1968: Multiprocessors (IBM OS/360 M65MP)
1971: Time sharing (IBM OS/360)
1969: Real Time (RTOS/360)

公平优先	BVT			
多核	负载分担	两级调度		
实时	RM	EDF		
公平	彩票调度	步幅调度		
优先级	MLQ	MLFQ		
经典	FCFS	SJF	PSJF	RR

OpenEuler

Linux O(n)调度器



实时任务：weight = 1000 + p->rt_priority;
普通任务：p->counter = NICE_TO_TICKS(prev->nice)
weight = p->counter + (20 - p->nice)

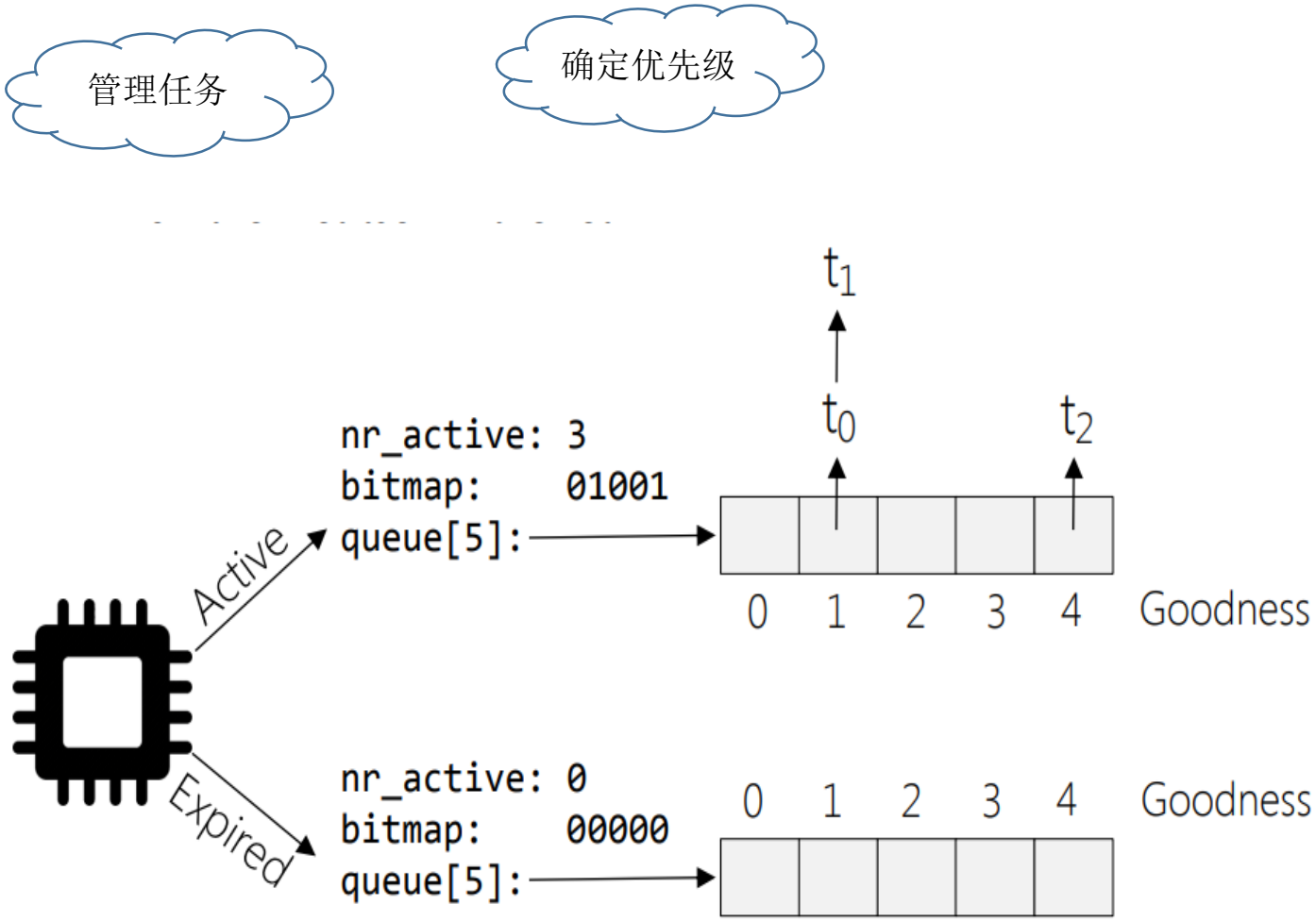
- 以weight作为动态优先级来选任务，每次选择遍历整个runqueue链表计算weight
- Tick周期性减掉普通和RR任务运行时间片，所有任务weight为0后统一充值
- 全局队列大锁

问题：

- O(n)的算法复杂度
- 大锁导致SMP扩展性差
- 重新计算时间片和active task数量<CPU核数时CPU空转
- 任务多核迁移频繁
- 只通过奖励睡眠任务来奖励交互式任务，IO-bound的任务会对其造成干扰
- Epoch时间片粒度跟系统负载相关，会导致响应延迟的尾时延过大

OpenEuler

Linux O(1)调度器



- 以per-cpu runqueue
- runqueue多链表，支持140个优先级
- Bitmap标识链表是否为空
- active和expired两个队列
- 动态优先级与静态优先级一致
- 判断交互式进程的因子更多，更准但复杂

问题：

- 对交互式进程复杂的启发式判断不完全准确，导致桌面用户卡顿

OpenEuler

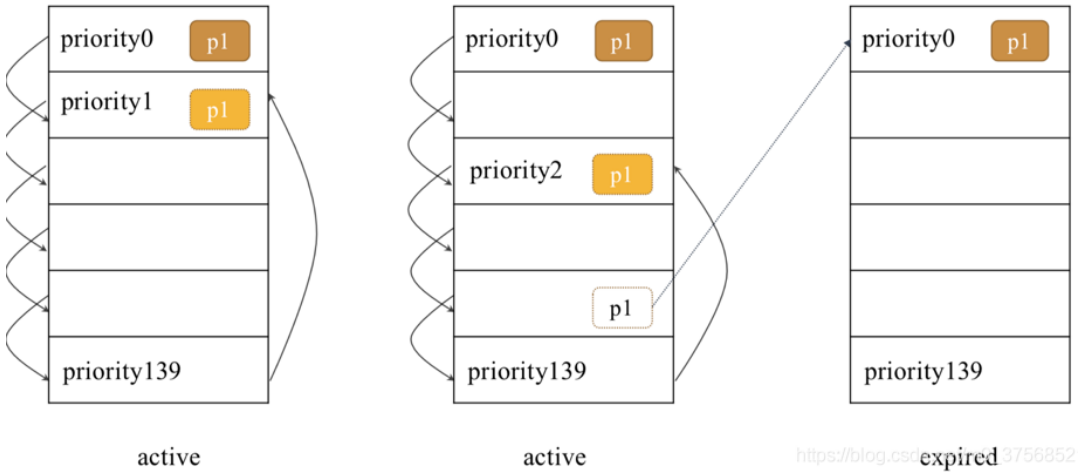
Linux CFS调度器

管理任务

确定优先级

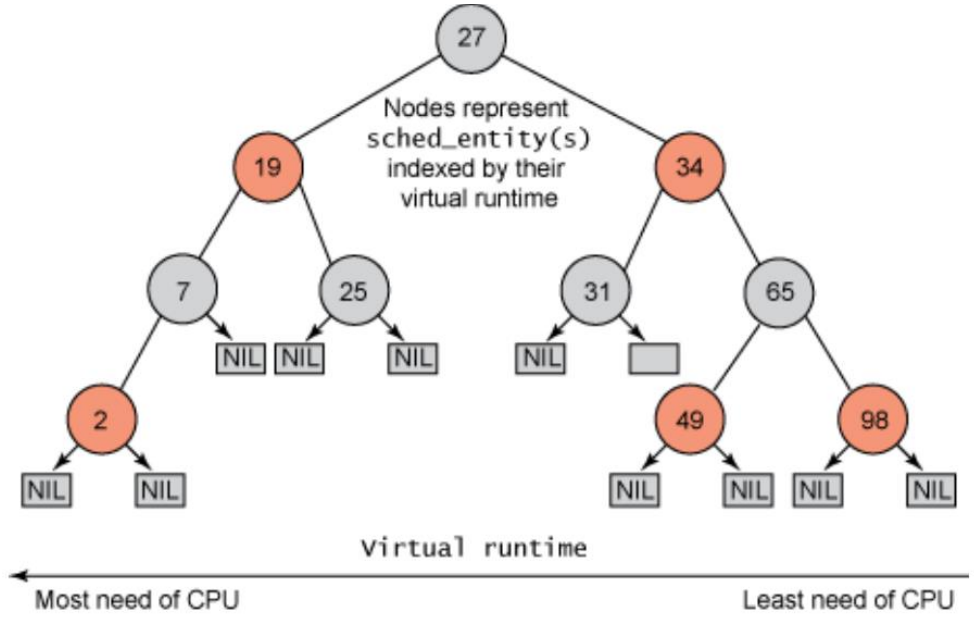
SD/RSDL引入公平调度的思想

- 猜不准就定义一个人人平等的规则，核心是规则要足够好
- 每个优先级有一个tg时间片，每个任务在各优先级有固定的tp时间片



CFS改进：

- 红黑树代替优先级队列
- 使用Virtual runtime = （physical runtime） X （nice value 0的权重）/进程的权重



OpenEuler

硬件发展带来的需求

Multi-processors

Multi-core

NUMA

big.LITTLE

Dynamic frequency

- Load Balance(sched_domain/cputopo)
- Load Tracking(per-runqueue/PELT/WALT)
- CPUfreq(schedutil)
- CPUidle(TEO/menu/ladder)
- System suspend/runtime PM
- HMP/EAS

OpenEuler

用户场景带来的需求

real time

Multi-user

- RT scheduler/DL scheduler
- CFS scheduler
- group scheduler

调度器底噪高	资源隔离	调度底噪高/能效
CT	计算&云	终端

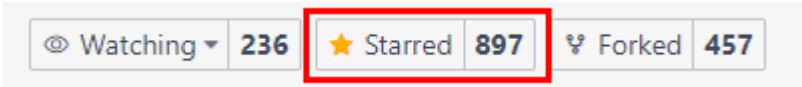
OpenEuler

后续分享

调度分享议题
进程创建与切换
Linux调度器框架
RT&DL调度器
CFS调度器
组调度
负载追踪
负载均衡
CPU调频
CPU IDLE
System suspend/runtime PM
EAS
性能分析工具
调度器性能评估测试
业界技术分享
openEuler开源特性系列

OpenEuler

✓ openEuler kernel gitee 仓库



源代码仓库

https://gitee.com/openeuler/kernel

欢迎大家多多 Star，多多参与社区开发，多多贡献补丁。

✓ maillist、issue、bugzilla

可以通过邮件列表、issue、bugzilla 参与社区讨论

欢迎大家多多讨论问题，发现问题多提 issue、bugzilla

https://gitee.com/openeuler/kernel/issues

https://bugzilla.openeuler.org

kernel@openeuler.org

✓ openEuler kernel SIG 微信技术交流群

请扫描右方二维码添加小助手微信

或者直接添加小助手微信（微信号：openeuler-kernel）

备注“交流群”或“技术交流”

加入 openEuler kernel SIG 技术交流群

技术交流

OpenEuler

Thank you

OpenEuler