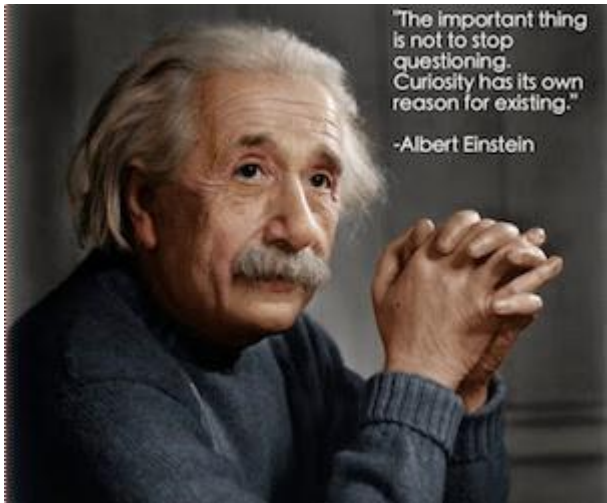


Linux 内核调度

Version: 3.0
Author: 孙垒
Email:sunlei0625@163.com
Tel:18221493945



摘要：本文详细介绍了 Linux 内核调度架构的实现细节。本文基于 Linux kernel4.9.76 版本介绍。

版本	时间	基于内核	作者	主要修改
V1.0	2016 年 2 月	4.4.2	孙垒	
V2.0	2017 年 1 月 12	4.4.41	孙垒	
V3.0	2017 年 8 月 12	4.9.76	孙垒	

目录

第 1 章 基础概念	2
1.1 进程	2
1.2 线程	4
1.3 就绪队列	5
1.4 内核空间	5
1.5 进程上下文	6
1.6 抢占	6
1.7 系统调用	8
1.8 同步和并发	9
1.9 中断和中断处理	10
1.10 延时任务	11
1.11 时间	12
1.12 小结	13
第 2 章 基本数据结构	14
1.1 运行队列 <code>rq</code>	14
1.2 完全公平调度 <code>cfs</code>	27
1.3 实时进程调度 <code>rt</code>	35
1.4 最终期限调度 <code>dl</code>	41
1.5 调度类 <code>sched_class</code>	44
1.6 调度实体 <code>sched_entity</code>	48
1.7 调度域 <code>sched_domain</code>	55
1.8 任务组 <code>task_group</code>	71
1.9 进程描述符 <code>task_struct</code>	81
1.10 控制组 <code>Cgroup</code>	82
1.11 小结	114
第 3 章 调度架构	115
1.12 简介	115
1.13 <code>rq</code> 就绪队列	115
1.14 优先级计算	119

1.15	负荷权重计算	122
1.16	进程时间度量	123
1.16.1	进程时间统计	124
1.16.2	虚拟时间计算	126
1.16.3	进程时间更新	127
1.16.4	运行时间控制	129
1.16.5	延迟计算	130
1.17	进程创建时调度设置	133
1.18	进程加载时调度设置	137
1.19	核心调度器	138
1.19.1	就绪队列时间更新	142
1.19.2	查找 next 进程	146
1.19.3	上下文切换	147
1.19.4	平衡处理	154
1.20	周期性调度器	156
1.20.1	CPU 本地时间更新	157
1.20.2	调度类 task_tick().....	159
1.20.3	更新 CPU 负载.....	160
1.20.4	更新可运行任务数	161
1.20.5	性能事件处理	162
1.20.6	负载均衡	164
1.21	小结	188
第 4 章	调度核心层	189
1.22	简介	189
1.23	调度架构层函数	189
1.23.1	update_rq_clock.....	190
1.23.2	enqueue_task	190
1.23.3	dequeue_task	191
1.23.4	activate_task.....	195
1.23.5	deactivate_task.....	196

1.23.6	check_preempt_curr	196
1.23.7	set_cpus_allowed_ptr	197
1.23.8	set_task_cpu.....	201
1.23.9	select_task_rq.....	202
1.23.10	wake_up_if_idle.....	203
1.23.11	migration_cpu_stop.....	204
1.23.12	wait_task_inactive	207
1.23.13	migrate_swap	209
1.23.14	resched_cpu.....	210
1.23.15	resched_curr	210
1.23.16	task_can_attach.....	211
1.23.17	nr_iowait.....	213
1.24	小结	213
第 5 章 CFS 调度类		215
1.25	简介	215
1.26	CFS 类	215
1.26.1	enqueue_task_fair()	219
1.26.2	dequeue_task_fair()	223
1.26.3	yield_task_fair()	226
1.26.4	yield_to_task_fair().....	227
1.26.5	check_preempt_wakeup().....	227
1.26.6	pick_next_task_fair()	230
1.26.7	put_prev_task_fair()	234
1.26.8	select_task_rq_fair().....	236
1.26.9	migrate_task_rq_fair().....	238
1.26.10	rq_online_fair()	239
1.26.11	rq_offline_fair().....	241
1.26.12	task_dead_fair	242
1.26.13	set_cpus_allowed_common	242
1.26.14	set_curr_task_fair()	242

1.26.15	task_tick_fair()	243
1.26.16	task_fork_fair()	244
1.26.17	prio_changed_fair()	245
1.26.18	switched_from_fair()	246
1.26.19	switched_to_fair()	246
1.26.20	get_rr_interval_fair()	247
1.26.21	update_curr_fair()	247
1.26.22	task_change_group_fair	248
1.27	其他函数	248
1.28	小结	249
第 6 章	RT 调度类	250
1.29	RT 类	250
1.29.1	enqueue_task_rt	253
1.29.2	dequeue_task_rt	257
1.29.3	yield_task_rt	263
1.29.4	check_preempt_curr_rt	264
1.29.5	pick_next_task_rt	266
1.29.6	put_prev_task_rt	268
1.29.7	select_task_rq_rt	268
1.29.8	set_cpus_allowed_common	270
1.29.9	rq_online_rt	270
1.29.10	rq_offline_rt	271
1.29.11	task_woken_rt	271
1.29.12	switched_from_rt	271
1.29.13	set_curr_task_rt	272
1.29.14	task_tick_rt	272
1.29.15	get_rr_interval_rt	273
1.29.16	prio_changed_rt	274
1.29.17	switched_to_rt	275
1.29.18	update_curr_rt	276

1.30	小结	278
第 7 章	DL 调度类	279
1.31	简介	279
1.32	DL 调度类	279
1.32.1	enqueue_task_dl	280
1.32.2	dequeue_task_dl	282
1.32.3	yield_task_dl	284
1.32.4	check_preempt_curr_dl	285
1.32.5	pick_next_task_dl	285
1.32.6	put_prev_task_dl	287
1.32.7	select_task_rq_dl	288
1.32.8	set_cpus_allowed_dl	289
1.32.9	rq_online_dl	289
1.32.10	rq_offline_dl	290
1.32.11	task_woken_dl	290
1.32.12	set_curr_task_dl	290
1.32.13	task_tick_dl	291
1.32.14	task_fork_dl	291
1.32.15	task_dead_dl	292
1.32.16	prio_changed_dl	292
1.32.17	switched_from_dl	293
1.32.18	switched_to_dl	294
1.32.19	update_curr_dl	294
1.33	小结	295
第 8 章	IDLE 调度类	296
1.34	简介	296
1.35	IDLE 类	296
1.35.1	check_preempt_curr_idle	297
1.35.2	pick_next_task_idle	298
1.35.3	select_task_rq_idle	298

1.36	小结	298
第 9 章	调度 API 函数.....	299
1.37	简介	299
1.38	接口函数.....	299
1.38.1	wake_up_process	299
1.38.2	try_to_wake_up_local.....	305
1.38.3	wake_up_new_task.....	307
1.38.4	preempt	308
1.38.5	sched_setscheduler	317
1.38.6	_cond_resched	322
1.38.7	__might_sleep	324
1.38.8	yield	327
1.38.9	schedule_timeout.....	329
1.38.10	io_schedule_timeout.....	330
1.38.11	scheduler_ipi.....	331
1.38.12	sched_setattr	333
1.38.13	set_user_nice.....	334
1.38.14	kick_process.....	336
1.38.15	sched_setaffinity.....	341
1.38.16	__migrate_task	343
1.38.17	sched_move_task	344
1.38.18	dump_cpu_task	345
1.39	小结	346
第 10 章	completion.....	347
1.40	简介	347
1.41	主要数据结构	347
1.42	使用方式.....	348
1.43	等待处理	349
1.44	唤醒处理	352
1.45	小结	353

第 11 章 wait 队列	354
1.46 简介	354
1.47 主要数据结构	354
1.48 wait_queue_t	362
1.49 wait_bit_queue	365
1.50 bit_wait_table	368
1.51 小结	370
第 12 章 内核 workqueue	371
1.52 简介	371
1.53 主要数据结构	372
1.53.1 struct workqueue_struct	372
1.53.2 struct worker	381
1.53.3 struct work_struct	392
1.54 workqueue 初始化	394
1.55 工作队列创建	397
1.56 入队列处理	404
1.56.1 queue_work_on	404
1.56.2 queue_delayed_work_on	409
1.57 任务处理	412
1.57.1 基本函数	412
1.57.2 工作者创建与销毁	421
1.57.3 工作者状态	422
1.57.4 工作者线程	422
1.58 rescuer_thread	424
1.59 小结	428
第 14 章 CPU 时间计量	429
1.60 简介	429
1.61 主要数据结构	429
1.62 初始化	433
1.63 cpucct 文件接口	434

1.64	时间计量	436
1.64.1	cpustat 更新	436
1.64.2	cpuusage 更新	436
1.65	irq 时间度量	437
1.66	软中断时间度量	440
1.67	小结	440
第 15 章	Cgroup 实现	441
1.68	简介	441
1.68.1	管理控制组	441
1.68.2	主要子系统	441
1.69	数据结构关系	443
1.70	Cgroup 使用	444
1.71	Cgroup 初始化	445
1.71.1	cgroup_init_early	445
1.71.2	cgroup_init	451
1.72	mount 处理	461
1.73	文件操作	471
1.73.1	cgroup_remount	471
1.73.2	cgroup_show_options	473
1.73.3	cgroup_mkdir	474
1.73.4	cgroup_rmdir	477
1.74	Cgroup 与进程创建	480
1.75	Cgroup 子系统	483
1.75.1	CPU 子系统	484
1.75.2	cpuset 子系统	486
1.75.3	memory 子系统	490
1.75.4	blkio 子系统	494
1.76	小结	497
第 16 章	进程 taskstats	498
1.77	简介	498

1.78	主要数据结构	498
1.79	模块初始化	507
1.80	监听注册	509
1.81	信息收集	512
1.82	小结	517
第 17 章 内核 profile		518
1.83	介绍	518
1.84	数据描述	518
1.85	处理过程	519
1.85.1	初始化	519
1.85.2	proc 文件处理	522
1.85.3	动态配置 profile	524
1.85.4	profile 事件处理	525
1.85.5	profile 处理	526
1.86	perf 使用	530
1.86.1	perf list	531
1.86.2	perf top	532
1.86.3	perf stat	533
1.86.4	perf record	534
1.86.5	perf report	534
1.87	小结	535
第 18 章 IDLE 进程		536
1.88	介绍	536
1.89	idle 进程创建	536
1.90	idle 运行时机	539
1.91	idle 进程主体	540
1.92	小结	545
第 19 章 调度统计		546
1.93	基本介绍	546
1.94	基本数据	546

1.95	处理过程	550
1.96	sched 域 debug	563
1.97	小结	566
第 20 章 系统调用		567
1.98	nice()	567
1.99	sched_setscheduler()	567
1.100	sched_setparam()	568
1.101	sched_getscheduler()	568
1.102	sched_setaffinity()	569
1.103	sched_yield()	569
1.104	sched_get_priority_max()	570
1.105	sched_rr_get_interval()	570
1.106	setpriority()	570
1.107	getpriority()	571
1.108	小结	571