Linux Kernel: monitoring the scheduler by trace_sched* events

Marco Perronet (Università degli Studi di Torino)

Thesis advisor: Enrico Bini

July 2019

Objectives

1. Illustrate how scheduling works in a real operating system
   - Implementation of the current scheduler (CFS, the Completely Fair Scheduler)
   - Comparison with the previous schedulers
2. Write documentation for the scheduler events
   - Function and event tracing
   - `ftrace` usage

GNU/Linux

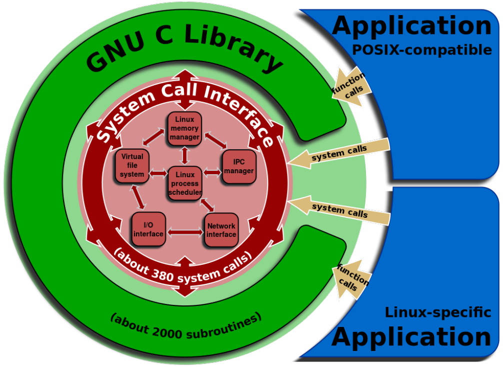


Figure 1: Userspace and Kernelspace

Why "GNU/Linux"?

- Linux is the kernel
- GNU is the application software, running on top of the kernel

Linux Kernel

The kernel is the core of the operating system. It is the software intended to manage the hardware resources. Some of its roles are:

- Responding to I/O requests
- Managing memory allocation
- Deciding how CPU time is shared among the demanding processes (*scheduling*)

Scheduler

- Scheduling classes are an extensible hierarchy of scheduler modules
- A task from a scheduling class can be chosen to run only if there are no runnable tasks in classes higher in the hierarchy
- Each process has a scheduling policy associated

Scheduling classes and policies

Scheduling classes	**Scheduling policies**
`stop_sched_class`	
`dl_sched_class`	`SCHED_DEADLINE`
`rt_sched_class`	`SCHED_FIFO` `SCHED_RR`

`fair_sched_class`	`SCHED_NORMAL`
	`SCHED_BATCH`
	`SCHED_IDLE`

`idle_sched_class`	

Tuning and extending the scheduler

The workload on servers is different from the workload on desktops (CPU bound vs I/O bound). The scheduler must be changed accordingly:

- Implementing a new scheduling policy. Scheduling classes are made to be extensible, so scheduling policies are handled by the scheduler without the core code assuming about them too much.
- Changing the existing scheduler's behavior with tunable values

Source files

- `dl_sched_class` – `kernel/sched/deadline.c`
- `rt_sched_class` – `kernel/sched/rt.c`
- `fair_sched_class` – `kernel/sched/fair.c`
- Core code shared by all classes – `kernel/sched/core.c`

Tuning and extending the scheduler

```c
// Code from ./kernel/sched/fair.c
/* The idea is to set a period in which each task runs once.
 *
 * When there are too many tasks (sched_nr_latency)
 * we have to stretch this period
 * because otherwise the slices get too small.
 */
static u64 __sched_period(unsigned long nr_running) {
    if (unlikely(nr_running > sched_nr_latency))
        return nr_running * sysctl_sched_min_granularity;
    else
        return sysctl_sched_latency;
}
```

- `sysctl_sched_min_granularity` – The minimum time a task is allowed to run on a CPU before being preempted
- `sysctl_sched_latency` – The default scheduler period

What if we tweak the granularity?

Completely Fair Scheduler (CFS)

CFS tries to model an ideal multi-tasking CPU where each task runs for the same amount of time. Total fairness would mean that with n tasks, every task receives $\frac{1}{n}$ of the processor's time.

CFS is a Dynamic Priority scheduling policy. In order to know which task deserves to run next, every task keeps track of the total amount of time that it has spent running (*runtime*). This value is used as priority.

Picking the next task

Being as fair as possible with all the task means keeping all the tasks' runtimes as close as possible to each other.
Following this logic, the task that deserves more than anyone to be executed next is the one with the smallest runtime.

Completely Fair Scheduler (CFS)

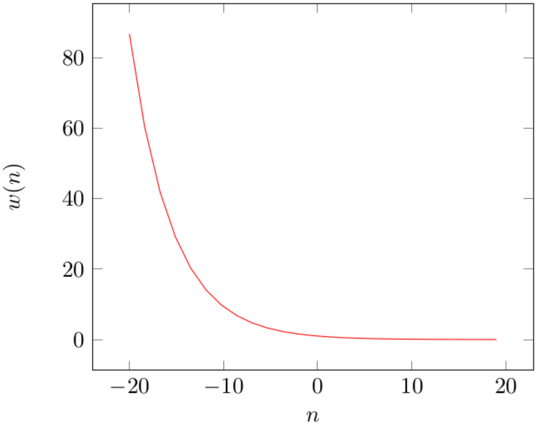
Virtual runtime

Vruntime is the runtime of the task, weighted based on *nice*. For high priority tasks (low *nice* value), vruntime is less than the real time spent on the CPU. In this case, vruntime grows slower than real time. The opposite is true for low priority tasks (high *nice* value).

The runqueue is implemented through a red-black tree, which is ordered with the virtual runtime.

Completely Fair Scheduler (CFS)

$$weight = \frac{1024}{(1.25)^{nice}}. \tag{1}$$



Completely Fair Scheduler (CFS)

Both the timeslice and the runtime must be weighted.

$$assigned\_time = target\_latency \frac{task\_weight}{total\_weight} \tag{2}$$

$$vruntime = runtime \frac{weight\_of\_nice\_0}{task\_weight} \tag{3}$$

- *Equation 2*: the task's timeslice is proportional to its weight.
- *Equation 3*: tasks with a higher priority modifier will have a low vruntime, and will be picked more often by the scheduler.

`ftrace` (Function tracer)

`ftrace` is a debug tool embedded in the kernel. It is also useful to approach and understand the code.

`ftrace` can perform function and event tracing.

Function tracing

Traces the path taken by kernel functions. Entry and exit point of the functions are traced, so the total duration of the function can be calculated. This allows latencies to be easily detected.

Event tracing

Based on static tracepoints in the code, which are called just like functions. Unlike function tracing, tracepoints are static, so they cannot be toggled at runtime.

Function tracing

- Thanks to code instrumentation, function tracing creates zero overhead when it's not used.
- It's possible to filter what is being traced by dynamically activating tracing only on functions from a single subsystem, or on one function alone.

Code instrumentation

At compile time, extra assembly instructions are generated to help debuggers and analysis tools. `ftrace` exploits `gcc`'s code instrumentation feature, and uses *runtime injection* for dynamic toggling.

Function tracing

```
 0)               |  scheduler_tick() {
 0)   0.094 us    |    _raw_spin_lock();
 0)   0.116 us    |    update_rq_clock.part.84();
 0)               |    task_tick_fair() {
 0)               |      update_curr() {
 0)   0.086 us    |        update_min_vruntime();
 0)   0.093 us    |        cpuacct_charge();
 0)   1.631 us    |      } /* update_curr */
 0)   0.074 us    |      update_cfs_shares();
 0)   0.124 us    |      hrtimer_active();
 0)   4.320 us    |    } /* task_tick_fair */
 0)               |    cpu_load_update_active() {
 0)   0.069 us    |      tick_nohz_tick_stopped();
 0)               |      cpu_load_update() {
 0)   0.088 us    |        sched_avg_update();
 0)   0.940 us    |      } /* cpu_load_update */
 0)   2.419 us    |    } /* cpu_load_update_active */
 0)   0.102 us    |    calc_global_load_tick();
 0)               |    trigger_load_balance() {
 0)   0.094 us    |      raise_softirq();
 0)   0.183 us    |      nohz_balance_exit_idle.part.85();
 0)   1.890 us    |    } /* trigger_load_balance */
 0) + 13.238 us   |  } /* scheduler_tick */
```

Event tracing

Event tracing is performed at specific points in the code known as *tracepoints*. It's less efficient than function tracing because it doesn't use runtime injection.

Tracepoint functions are generated by the `TRACE_EVENT(...)` macro, which allows developers to quickly declare their own events to trace from outside the kernel.

Why not just use `printk()` ?

- Tracepoints are more efficient: it is faster to write in `ftrace`'s ring buffer than in standard output.
- When debugging the scheduler or other high-volume areas, `printk()`'s overhead can introduce heisenbugs or even create a live lock.
- Output from tracepoints can be quickly filtered by selectively toggling them from userspace.

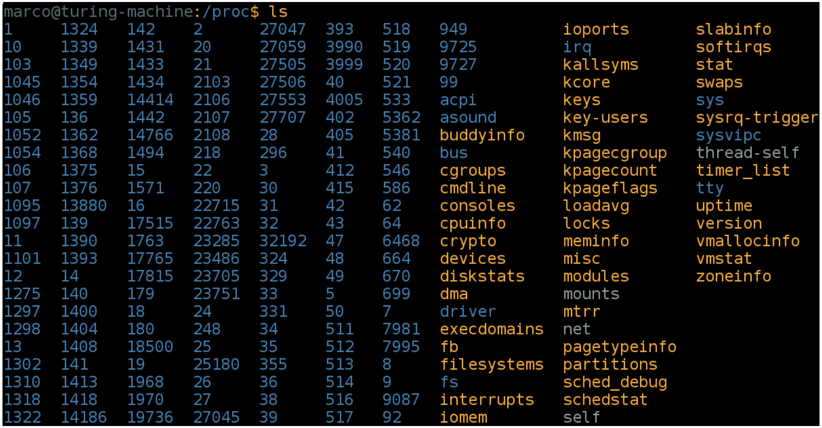
Event tracing

```c
static void update_curr(struct cfs_rq *cfs_rq) {
    struct sched_entity *curr = cfs_rq->curr;
    u64 now = rq_clock_task(rq_of(cfs_rq));
    u64 delta_exec = now - curr->exec_start;

    if (unlikely((s64)delta_exec <= 0))
        return;

    curr->exec_start = now; // Reset exec_start
    schedstat_set(curr->statistics.exec_max, max(delta_exec,
 ↪  curr->statistics.exec_max));
    curr->sum_exec_runtime += delta_exec; // Non-weighted runtime
    schedstat_add(cfs_rq->exec_clock, delta_exec);
    curr->vruntime += calc_delta_fair(delta_exec, curr); // Applies
 ↪  vruntime equation
    update_min_vruntime(cfs_rq);

    if (entity_is_task(curr)) {
        struct task_struct *curtask = task_of(curr);
        trace_sched_stat_runtime(curtask, delta_exec,
 ↪  curr->vruntime); // Tracepoint
    }
}
```

Event tracing

```
# tracer: nop
#
# entries-in-buffer/entries-written: 116546/459475   #P:4
#
#                              _-----=> irqs-off
#                             / _----=> need-resched
#                            | / _---=> hardirq/softirq
#                            || / _--=> preempt-depth
#                            ||| /     delay
#   TASK-PID   CPU#  ||||    TIMESTAMP  FUNCTION
#      | |       |   ||||       |         |
  <idle>-0     [000] d...   611.283814: sched_switch:
 ↪   prev_comm=swapper/0 prev_pid=0 prev_prio=120 prev_state=R ==>
 ↪   next_comm=Xorg next_pid=1450 next_prio=120
    Xorg-1450  [000] d...   611.283921: sched_stat_runtime: comm=Xorg
 ↪   pid=1450 runtime=117083 [ns] vruntime=17539094302 [ns]
    Xorg-1450  [000] d...   611.283924: sched_switch: prev_comm=Xorg
 ↪   prev_pid=1450 prev_prio=120 prev_state=S ==> next_comm=swapper/0
 ↪   next_pid=0 next_prio=120
  <idle>-0     [000] d...   611.283957: sched_switch:
 ↪   prev_comm=swapper/0 prev_pid=0 prev_prio=120 prev_state=R ==>
 ↪   next_comm=Xorg next_pid=1450 next_prio=120
```

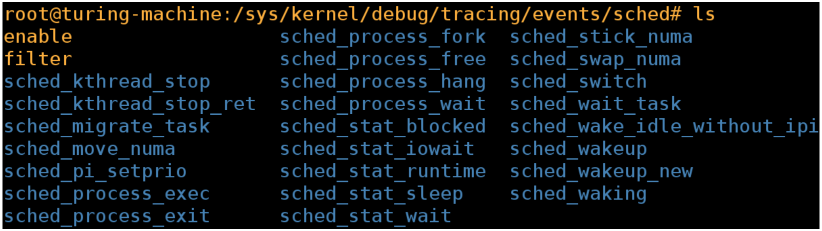
Interfacing with ftrace



```
marco@turing-machine:/proc$ ls
1     1324  142   2     27047 393  518  949         ioports       slabinfo
10    1339  1431  20    27059 3990 519  9725        irq           softirqs
103   1349  1433  21    27505 3999 520  9727        kallsyms      stat
1045  1354  1434  2103  27506 40   521  99          kcore         swaps
1046  1359  14414 2106  27553 4005 533  acpi        keys          sys
105   136   1442  2107  27707 402  5362 asound      key-users     sysrq-trigger
1052  1362  14766 2108  28    405  5381 buddyinfo   kmsg          sysvipc
1054  1368  1494  218   296   41   540  bus         kpagecgroup   thread-self
106   1375  15    22    3     412  546  cgroups     kpagecount    timer_list
107   1376  1571  220   30    415  586  cmdline     kpageflags    tty
1095  13880 16    22715 31    42   62   consoles    loadavg       uptime
1097  139   17515 22763 32    43   64   cpuinfo     locks         version
11    1390  1763  23285 32192 47   6468 crypto      meminfo       vmallocinfo
1101  1393  17765 23486 324   48   664  devices     misc          vmstat
12    14    17815 23705 329   49   670  diskstats   modules       zoneinfo
1275  140   179   23751 33    5    699  dma         mounts
1297  1400  18    24    331   50   7    driver      mtrr
1298  1404  180   248   34    511  7981 execdomains net
13    1408  18500 25    35    512  7995 fb          pagetypeinfo
1302  141   19    25180 355   513  8    filesystems partitions
1310  1413  1968  26    36    514  9    fs          sched_debug
1318  1418  1970  27    38    516  9087 interrupts  schedstat
1322  14186 19736 27045 39    517  92   iomem       self
```

Figure 2: The procfs special filesystem

Interfacing with ftrace



```
root@turing-machine:/sys/kernel/debug/tracing/events/sched# ls
enable                  sched_process_fork  sched_stick_numa
filter                  sched_process_free  sched_swap_numa
sched_kthread_stop      sched_process_hang  sched_switch
sched_kthread_stop_ret  sched_process_wait  sched_wait_task
sched_migrate_task      sched_stat_blocked  sched_wake_idle_without_ipi
sched_move_numa         sched_stat_iowait   sched_wakeup
sched_pi_setprio        sched_stat_runtime  sched_wakeup_new
sched_process_exec      sched_stat_sleep    sched_waking
sched_process_exit      sched_stat_wait
```

Figure 3: Scheduler events in the `tracefs` special filesystem

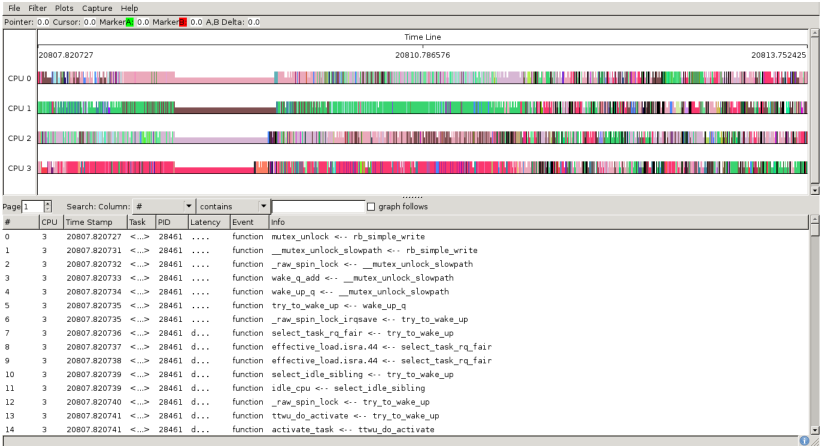
- Enable scheduler events:

```
echo 1 > enable
```

- Enable only the `sched_stat_runtime` event:

```
echo 1 > sched_stat_runtime/enable
```

KernelShark



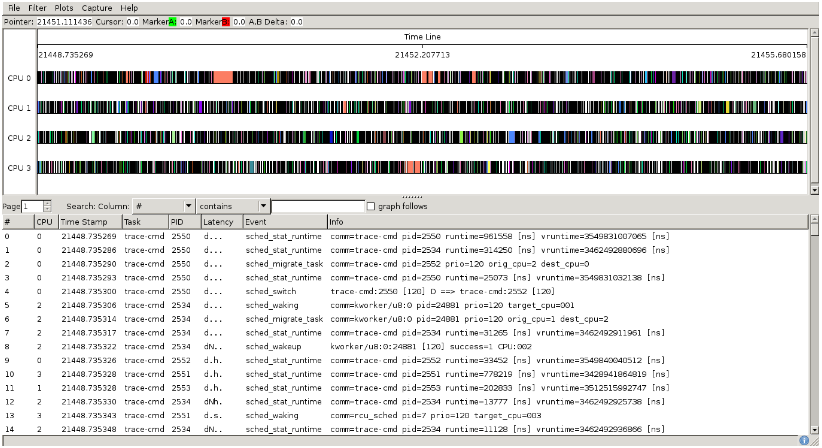
Pointer: 0.0 Cursor: 0.0 MarkerA: 0.0 MarkerB: 0.0 A,B Delta: 0.0

Page 1 Search: Column: # contains ☐ graph follows

#	CPU	Time Stamp	Task	PID	Latency	Event	Info
0	3	20807.820727	<...>	28461	....	function	mutex_unlock <-- rb_simple_write
1	3	20807.820731	<...>	28461	....	function	__mutex_unlock_slowpath <-- rb_simple_write
2	3	20807.820732	<...>	28461	....	function	_raw_spin_lock <-- __mutex_unlock_slowpath
3	3	20807.820733	<...>	28461	....	function	wake_q_add <-- __mutex_unlock_slowpath
4	3	20807.820734	<...>	28461	....	function	wake_up_q <-- __mutex_unlock_slowpath
5	3	20807.820735	<...>	28461	....	function	try_to_wake_up <-- wake_up_q
6	3	20807.820735	<...>	28461	....	function	_raw_spin_lock_irqsave <-- try_to_wake_up
7	3	20807.820736	<...>	28461	d...	function	select_task_rq_fair <-- try_to_wake_up
8	3	20807.820737	<...>	28461	d...	function	effective_load.isra.44 <-- select_task_rq_fair
9	3	20807.820738	<...>	28461	d...	function	effective_load.isra.44 <-- select_task_rq_fair
10	3	20807.820739	<...>	28461	d...	function	select_idle_sibling <-- try_to_wake_up
11	3	20807.820739	<...>	28461	d...	function	idle_cpu <-- select_idle_sibling
12	3	20807.820740	<...>	28461	d...	function	_raw_spin_lock <-- try_to_wake_up
13	3	20807.820741	<...>	28461	d...	function	ttwu_do_activate <-- try_to_wake_up
14	3	20807.820741	<...>	28461	d...	function	activate_task <-- ttwu_do_activate

Figure 4: Function tracing with KernelShark

KernelShark



Pointer: 21451.111436 Cursor: 0.0 MarkerA: 0.0 MarkerB: 0.0 A,B Delta: 0.0

Page 1 Search: Column: # contains

#	CPU	Time Stamp	Task	PID	Latency	Event	Info
0	0	21448.735269	trace-cmd	2550	d...	sched_stat_runtime	comm=trace-cmd pid=2550 runtime=961558 [ns] vruntime=3549831007065 [ns]
1	0	21448.735286	trace-cmd	2550	d...	sched_stat_runtime	comm=trace-cmd pid=2534 runtime=314250 [ns] vruntime=3462492880696 [ns]
2	0	21448.735290	trace-cmd	2550	d...	sched_migrate_task	comm=trace-cmd pid=2552 prio=120 orig_cpu=2 dest_cpu=0
3	0	21448.735293	trace-cmd	2550	d...	sched_stat_runtime	comm=trace-cmd pid=2550 runtime=25073 [ns] vruntime=3549831032138 [ns]
4	0	21448.735300	trace-cmd	2550	d...	sched_switch	trace-cmd:2550 [120] D ==> trace-cmd:2552 [120]
5	2	21448.735306	trace-cmd	2534	d...	sched_waking	comm=kworker/u8:0 pid=24881 prio=120 target_cpu=001
6	2	21448.735314	trace-cmd	2534	d...	sched_migrate_task	comm=kworker/u8:0 pid=24881 prio=120 orig_cpu=1 dest_cpu=2
7	2	21448.735317	trace-cmd	2534	d...	sched_stat_runtime	comm=trace-cmd pid=2534 runtime=31265 [ns] vruntime=3462492911961 [ns]
8	2	21448.735322	trace-cmd	2534	dN..	sched_wakeup	kworker/u8:0:24881 [120] success=1 CPU:002
9	0	21448.735326	trace-cmd	2552	d.h.	sched_stat_runtime	comm=trace-cmd pid=2552 runtime=33452 [ns] vruntime=3549840040512 [ns]
10	3	21448.735328	trace-cmd	2551	d.h.	sched_stat_runtime	comm=trace-cmd pid=2551 runtime=778219 [ns] vruntime=3428941864819 [ns]
11	1	21448.735328	trace-cmd	2553	d.h.	sched_stat_runtime	comm=trace-cmd pid=2553 runtime=202833 [ns] vruntime=3512515992747 [ns]
12	2	21448.735330	trace-cmd	2534	dNh.	sched_stat_runtime	comm=trace-cmd pid=2534 runtime=13777 [ns] vruntime=3462492925738 [ns]
13	3	21448.735343	trace-cmd	2551	d.s.	sched_waking	comm=rcu_sched pid=7 prio=120 target_cpu=003
14	2	21448.735348	trace-cmd	2534	dN..	sched_stat_runtime	comm=trace-cmd pid=2534 runtime=11128 [ns] vruntime=3462492936866 [ns]

Figure 5: Event tracing with KernelShark

Thank you!