Tuning the Linux CFS Scheduler

Stuart Bowman

December 15, 2015

1 Introduction

Nearly any modern computing device today has a well balanced and finely tuned scheduling system. As a result, operating system Kernels must be widely configurable and flexible among a large array of hardware configurations and applications. Additionally, while it may be easy to make a scheduler that performs with very high throughput, it is a much larger task to build a scheduling system that caters to all types of scheduling criteria (turnaround time, responsiveness, wait time) without causing process starvation or instabilities in system operation.

This paper presents an overview of a select group of the main components of cpu scheduling under the Linux Kernel as well as a set of tests and benchmarks detailing the observed performance of various scheduler modifications. Additionally, it aims to provide an overview of the Linux CFS scheduling operation as well as discuss several other alternative real-world scheduling algorithms.

2 Brief History

In 20+ years of Linux development, the Kernel scheduler has undergone several major revisions. Among these revisions are three primary scheduling systems.

O(N) scheduler This scheduling system ran through every task in the run queue during each scheduling event, picking and choosing which processes to run. Despite the simplicity of rating and picking processes sequentially at each scheduling routine, this method was not suitable for a wide variety of hardware and not ready for multi-core systems. [1]

Linux 2.6.x Saw the change of several different scheduling systems as it served the Linux community from 2003 until 2016 through it's long term support variant. Kernel iterations up to 2.6.22 included the **O(1) scheduler** by Ingo Molnar, an effectively constant time scheduling algorithm designed to replace to formerly costly O(N) scheduler. O(1) made use of a scheduling priority run queue, which allowed it to simply pick the top-most process in constant time. As the Kernel matured however, the heuristics of the O(1) run queue became difficult to manage and beginning in 2.6.23 the **Completely Fair Scheduler** (CFS) by Ingo Molnar was introduced. [1]

With CFS came a large change in process ordering and organization. CFS operates through a modified variant of round robin scheduling and utilizes a self-balancing red-black tree system. All operations in a red-black tree can be performed in O(log n), including adding nodes, deleting nodes and finding nodes. The CFS red-black tree organizes processes that need the cpu the most on the left of the tree and process that need the cpu the least on the right. A process's need for the cpu is inversely proportional to it's "virtual runtime". The virtual runtime for a given process is calculated as the amount of time assigned to a given process. [1]

Additionally, CFS provides a method of dynamic process prioritization through the use of "nice" values. This allows the system a secondary way to adjust a process's need for the cpu and allows users to arbitrarily give certain processes higher priority over others. CFS also has support for various other methods of load balancing between cores, sharing runtime quotas between runqueues and various other rules and features.

Other schedulers such as Con Kolivas' **Rotating Staircase Deadline Scheduler** and his **BFS scheduler** make use of alternative methods for organizing and scheduling processes in the Kernel. These are frequently offered in the form of a Kernel patch or sometimes as a completely packaged Kernel, where the Kernel variant is represented by the author's initials, such as **linux-ck**. [1]

3 Scheduler Flow Control

As defined by the Kernel's documentation on CFS,

"80% of CFS's design can be summed up in a single sentence: CFS basically models an 'ideal, precise multi-tasking CPU' on real hardware." [2]

Looking inside the Kernel source, there are several files of interest that hold particular importance to the development and tuning of the scheduler.

Kernel/sched/sched.h contains numerous preprocessor definitions, function prototypes, constants for use in weighting and organizing processes.

Kernel/sched/core.c handles the primary scheduling overhead including the main schedule() function, and manages multiple scheduling policies and run queues.

Kernel/sched/fair.c defines nearly all CFS behaviors and rules (such as picking the next task to run), defines constants specific to CFS, and includes structs such as fair_sched_class.

lib/rbtree.c contains the class definition for the red-black tree system used by CFS, and includes functions for quickly navigating between adjacent nodes in the tree as well as finding the leftmost node (which is typically the most critical process to run).

include/linux/sched/prio.h holds several priority constants including minimum and maximum niceness values, and the default values for newly created processes

include/linux/sched/rt.h contains, among other things, the Round Robin timeslice multiplier for SCHED_RR which directly impacts the system's final default Round Robin timeslice.

Digging deeper into each of these files revealed some particular areas of interest for modifying scheduling behavior. **Core.c** holds the primary overhead control on scheduling in the Kernel which makes it a prime candidate to start exploring. Core.c's primary function is `Schedule()` which handles scheduling flow control between runqueues such as the CFS runqueue. `Schedule()` calls can be triggered by blocking from Mutexes, Semephores and several other elements and it is capable of handling both realtime scheduling policies (Round Robin, FIFO) and non-preemptable policies (Batch).

Schedule() makes a call to `pick_next_task(rq, prev)`, where rq represents the active runqueue and prev represents the current task_struct being executed. `Pick_next_task()` then passes control to the appropriate scheduler class (such as CFS) where the scheduler class will run its own implementation of `pick_next_task()`. When CFS is the current scheduling class, `fair_sched_class.pick_next_task(rq, prev)` is called, sending control to fair.c's function `pick_next_task_fair(rq,prev)`.

`pick_next_task_fair()` contains a variety of additional rules, regulations and exceptions for choosing tasks. `pick_next_task_fair()` will return the task_struct of the process it selects to run next to core.c and, in the event that the CFS chooses to run the idle process, it will return `NULL`.

Under the "Simple" clause of this function, `pick_next_entity(cfs_rq, sched_entity)` is called, where the leftmost node of the red-black tree of tasks is chosen unless overridden to re-run the last process or run the skip buddy. This is effectively the method where processes are actually chosen in the CFS scheduler.

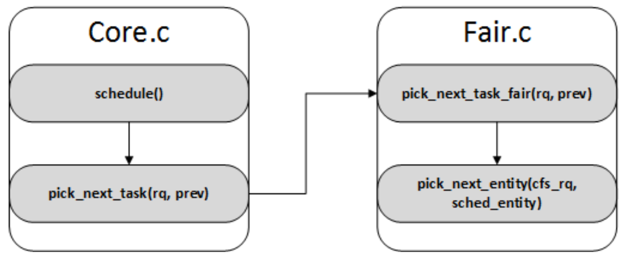

Figure 1: Simplified version of the scheduler's flow control

4 Scheduler Modifications

With the above scheduling flow control in mind, we sought out to change and tune the CFS system in some meaningful way. Among the many traits a scheduler is desired to exhibit, we decided to modify the scheduler to measure and tweak its throughput, scheduling period, and individual process runtime. These tests were not necessarily designed to improve the Kernel, but rather to explicitly test how various properties affect its performance. There were three main source modifications we attempted:

1. Modify the logic behind which process is chosen in `pick_next_entity()`
2. Change the Kernel parameters at runtime to cause frequent context switching
3. Modify CFS bandwidth and quota to cause artificial throttling

1 Picking Next Entity Changes

The first, and most simplistic change involved modifying the operation of `pick_next_entity()` in fair.c. As mentioned above, this function is the last authority queried in the scheduling flow control of CFS, which means process traffic can be changed here. A simplified psudocode version of `pick_next_entity()` is as follows:

```
1  pick_next_entity(cfs_rq, currentProcess)
2
3  SchedulerEntity left = PickFirstEntity()
4  SchedulerEntity se
5
6  if (currentProcess is left of leftmost entity)
7          left = currentProcess
8
9  se = left
10
11 if (cfs_rq->skip should skip se){
12         //skip buddy
13 }
14
15 if (cfs_rq->last and WakeupPreemptEntity())
16         se = cfs_rq->last
17
18 if (cfs_rq->next and WakeupPreemptEntity())
19         se = cfs_rq->next
20
21 return se
```

By removing the additional logic from lines 11-19, we can force the cfs scheduler to strictly run the leftmost (most needed) entity from the red-black tree, and bypass the additional logic used to make exceptions. The resulting psudocode is as follows:

```
pick_next_entity(cfs_rq, currentProcess)

SchedulerEntity left = PickFirstEntity()
SchedulerEntity se

if (currentProcess is left of leftmost entity)
        left = currentProcess

se = left

return se
```

2 Changing Kernel Parameters

Changing Kernel parameters is not terribly difficult and can be done fairly easily from several different source files before compilation. Areas of focus include the targeted preemption latency and minimum preemption granularity in fair.c, min/max priority niceness values in prio.h, `SCHED_LOAD_SCALE` (which adjusts the number of shares available for the root group) and `RUNTIME_INF` for cfs_rq quota in sched.h. [10]

Some Kernel parameters have the benefit of being publically readable and writable during runtime. The utility `sysctl` can be used to read and write to Kernel parameters at runtime, changing the behavior of everything from average Round Robin time slicing to CFS bandwidth.

Additionally, many Kernel parameters can be read from the /proc folder, providing a secondary resource for details on cgroups, hardware info and process specific information.

```
[root@HAL9000 ~]# uname -mrs
Linux 4.2.5-1-ARCH x86_64
[root@HAL9000 ~]# sysctl -A | grep "sched" | grep -v "domain"
sysctl: reading key "net.ipv6.conf.all.stable_secret"
sysctl: reading key "net.ipv6.conf.default.stable_secret"
sysctl: reading key "net.ipv6.conf.enp0s3.stable_secret"
sysctl: reading key "net.ipv6.conf.lo.stable_secret"
kernel.sched_autogroup_enabled = 1
kernel.sched_cfs_bandwidth_slice_us = 5000
kernel.sched_child_runs_first = 0
kernel.sched_latency_ns = 12000000
kernel.sched_migration_cost_ns = 500000
kernel.sched_min_granularity_ns = 1500000
kernel.sched_nr_migrate = 32
kernel.sched_rr_timeslice_ms = 30
kernel.sched_rt_period_us = 1000000
kernel.sched_rt_runtime_us = 950000
kernel.sched_shares_window_ns = 10000000
kernel.sched_time_avg_ms = 1000
kernel.sched_tunable_scaling = 1
kernel.sched_wakeup_granularity_ns = 2000000
[root@HAL9000 ~]# █
```

Figure 2: Displaying the default values for the stock Arch Linux Kernel

3 CFS Bandwidth and Quota

As it turns out, the Linux scheduling system makes use of time in any way it possibly can. Additionally, it can also be made to limit or even restrict certain groups of processes from using 100% of cpu resources at all times. These features are made possible through several parameters:

rt_period_us represents the bandwidth enforcement interval in microseconds (default is 100000000us)

rt_runtime_us represents the permitted amount of rt_period_us that may be used during a given period (default is 95000000us)

runtime_inf changes the default quota assigned to CFS (default value is disabled, meaning that CFS has unconstrained use of the cpu)

[7][8][10]

5 Benchmarks

The following benchmarks were performed individually for each of the previously mentioned Kernel modifications on an Arch Linux Virtual Machine. To empirically measure the throughput and turnaround time of various workloads, Prime95 was used to stress test the various Kernel builds. In addition, a custom built C++ program was used to run a synthetic workload on 16 concurrent child processes, calculated the average time required to finish each process.

The benchmarks used were as follows:

- Prime 95 Benchmark, sampled when running 25 iterations of 2048K FFT on 1, 2 and 4 cores.

- SyntheticProcess, our custom C++ synthetic load program, running 16 child processes each running a synthetic load and then averaging the completion time of all children (with a 1 sec delay between iterations to ensure the children were dequeued completely).

- A composite test where Prime 95 and SyntheticProcess are run in parallel to evaluate CFS's ability to share between entirely separate pid parents.

The Kernels used in our benchmarks were as follows:

STOCK A stock build of Linux 4.2.6

Test1 A modified version of the stock Kernel, implementing the pick_next_task psudocode changes detailed above.

Test2 A stock build of Linux 4.2.6 with the following Kernel parameters configured using `sysctl`:

sched_rt_period_us = 10,

sched_rt_runtime_us = 10,

sched_rr_timeslice_ms = 1,

sched_cfs_bandwidth_slice_us = 1

Test3 A modified version of the stock Kernel, implementing the CFS Bandwidth and Quota changes detailed above. Specifically, RUNTIME_INF was set to 10% of the CFS bandwidth period (previously it was unrestricted), refill_cfs_bandwidth_runtime() was short circuited to disable refilling alloted bandwidth, and the sched.h preprocessor constant MAX_SHARES was changed from 2^{18} to 2^{9}.

Benchmark Results

As can be seen from the benchmark results, Prime 95 performed nearly identical on all variants of the Kernel, with very little variance. Likewise, our Synthetic load program also saw nearly identical results for all Kernel variants, with only a 0.001 second variance between each test.

Interestingly, Prime 95 saw a significant amount of variance during the composite test during it's 2 core run on all Kernels. This may be a result of CFS struggling to decide which Kernel threads to map with process threads, resulting in frequent process migrations between cores. This hypothesis can be backed up by the fact that the results were very consistent when Prime 95 is run on all 4 cores, when all Kernel threads must be utilized all the time. Additionally, the Synthetic load also saw a considerable degree of variance between tests (range of 0.06 sec) during the same composite test with Prime 95, meaning that CFS's scheduling decisions effected both processes.

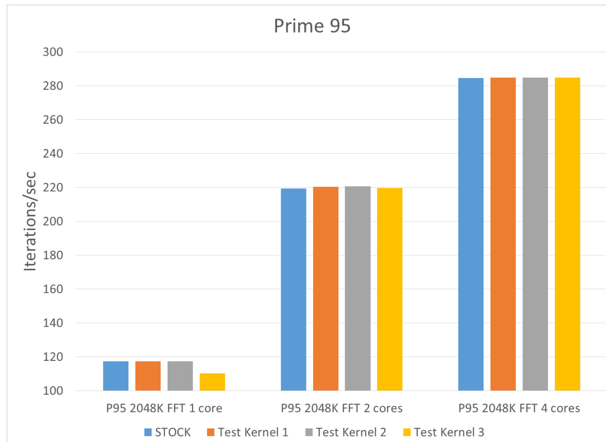

Figure 3: Prime 95 Benchmark sampled at 2048K FFT on 1, 2, and 4 cores

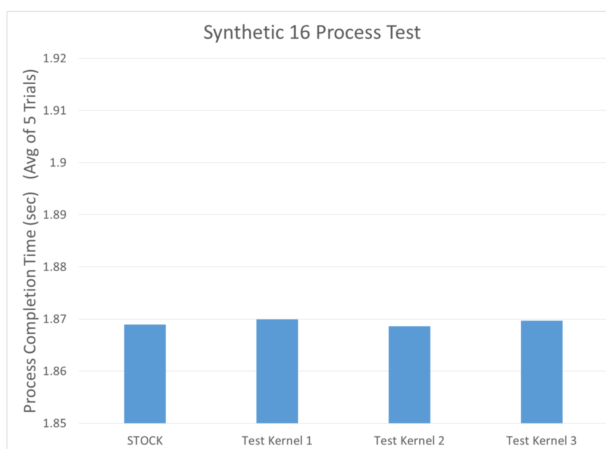

Figure 4: Synthetic load, completion time of each forked process, averaged over 5 trials of batches of 16 concurrent forks.

6 Summary

In summary, the Linux CFS scheduling system is a vastly robust and highly configurable system for managing and running processes in a real-time environment. It is designed for a wide array of hardware, ranging from smartphones to desktop computers to corporate servers.

One of the biggest discoveries of this study has been that, although the CFS scheduler can be highly sensitive to particular Kernel source code changes, it is also robust enough to adapt and change it's behavior so that consistent behavior can be expected from it even when individual components are missing or compromised. And that is likely one of the biggest hurtles in CPU scheduling today, to make Kernels that are adaptable enough for everyone's needs, but also robust enough to adapt under a variety of constraints.

7 Citations

[1] M. Tim Jones, IBM. (2009, Dec). "Inside the Linux 2.6 Completely Fair Scheduler" . [Online] http://www.ibm.com/developerworks/library/l-completely-fair-scheduler/

[2] Red Hat. "Scheduler Design CFS" . [Online] http://people.redhat.com/mingo/cfs-scheduler/sched-design-CFS.txt

[3] Kernel.org. "CFS Bandwidth Control" . [Online] https://www.Kernel.org/doc/Documentation/scheduler/sched-bwc.txt

[4] Kernel.org. "Red-black Trees" . [Online] https://www.Kernel.org/doc/Documentation/rbtree.txt

[5] tbazant, Open Suse (2013, Nov). "Tuning the Task Scheduler" . [Online] https://activedoc.opensuse.org/book/opensuse-system-analysis-and-tuning-guide/chapter-14-tuning-the-task-scheduler

[6] Guy Rutenberg (2007, Sept). "Profiling code using clock_gettime" . [Online] http://www.guyrutenberg.com/2007/09/22/profiling-code-using-clock_gettime/

[7] Christophe Blaess (2012, Jan). "Linux 3.2 - CFS CPU Bandwidth" . [Online] http://www.blaess.fr/christophe/2012/01/07/linux-3-2-cfs-cpu-bandwidth-english-version/

[8] Red Hat. "CPU Scheduling" . [Online] https://access.redhat.com/documentation/en-US/Red_Hat_Enterprise_Linux/6/html/Performance_Tuning_Guide/s-cpu-scheduler.html

[9] Martin B. J. (2013, Jul) "How to Pick Next Task" . [Online] http://martinbj2008.github.io/blog/2013/07/26/how-to-pick-next-task/

[10] Paul Turner, Bharata B Rao, Nikhil Rao (2010) "CPU bandwidth control for CFS" . [Online] http://landley.net/kdocs/ols/2010/ols2010-pages-245-254.pdf

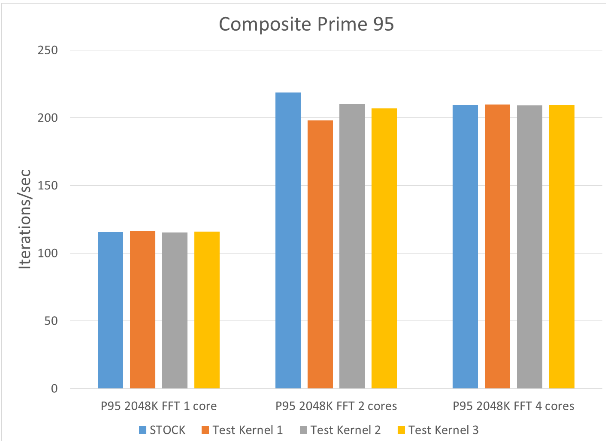

Figure 5: Composite test running both Prime 95 and Synthetic load at the same time. A high amount of variance can be observed while Prime 95 was running on only 2 cores.

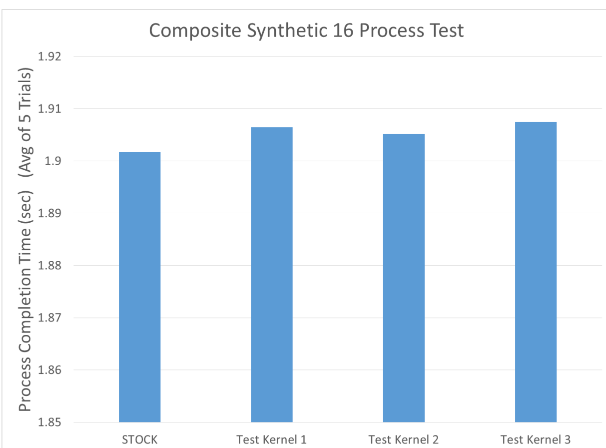

Figure 6: The same composite test, running both Prime 95 and Synthetic load at the same time. Each sample represents the average of the last 5 iterations of Synthetic load while Prime 95 was completing it's 2048K FFT run.