Contributing to the Linux Kernel

Omar Sandoval osandov@osandov.com

CSE 451, Autumn 2015

November 23, 2015



Goals

- Introduce you to what it's like to work on a production operating system
- Shed some light on how Linux in particular is developed
- Give you a starting point if you're interested in working on Linux

What is Linux?

- Operating system kernel written by Linus Torvalds, first released in 1991
- Free and open source (GNU GPLv2)
- General-purpose
 - Desktops, servers, smartphones, embedded systems
 - x86, ARM, MIPS, ...



What is Linux? (cont.)

- Huge software engineering endeavor
- For 4.3:
 - Over 50,000 files
 - Almost 20,000,000 lines of code
 - About 1600 developers and 200 companies



Requisites for Working on the Kernel

- Technical knowledge
- 2 Familiarity with the development process
- 3 Motivation

Technical Requirements

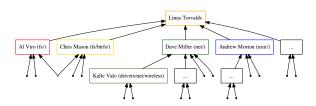
- C
- Unix shell
- Git

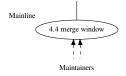
Technical Requirements

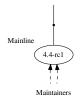
- C
- Unix shell
- Git
- Optional
 - Specific domain knowledge, e.g., filesystems, schedulers, memory management
 - Scripting, working with VMs

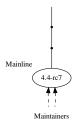
Development Process: Maintainers

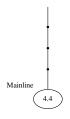
■ Kernel source tree divided into subsystems with different maintainers, submaintainers

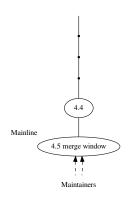


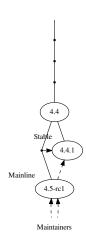


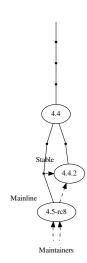


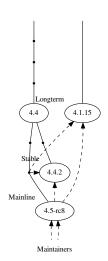


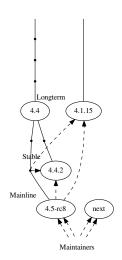












■ Mainline

- Linus Torvalds's linux.git tree
- Merge window
- Weekly release candidates

Mainline

- Linus Torvalds's linux.git tree
- Merge window
- Weekly release candidates

■ Stable

- Greg Kroah-Hartman's linux-stable tree
- Backports of bug fixes for recent releases

■ Mainline

- Linus Torvalds's linux.git tree
- Merge window
- Weekly release candidates

■ Stable

- Greg Kroah-Hartman's linux-stable tree
- Backports of bug fixes for recent releases
- Longterm
 - Greg Kroah-Hartman, others
 - Backports of select versions for years

■ Mainline

- Linus Torvalds's linux.git tree
- Merge window
- Weekly release candidates

■ Stable

- Greg Kroah-Hartman's linux-stable tree
- Backports of bug fixes for recent releases

■ Longterm

- Greg Kroah-Hartman, others
- Backports of select versions for years

Next

- Stephen Rothwell's linux-next tree
- Integration tree combining everything maintainers have in their own trees for the next merge window

Motivation

■ Can't really help you with this one



Finding Something to Work On

- Decide on a few specific subsystems
- Code cleanup (coding style, whitespace)
- Follow the mailing lists
- Try the release candidates
- Reproduce bugs



```
r[osandov@new ~/Dropbox/Homework/CSE551/551ws master]
$ ag sendfile
ERR: error in pthread_setaffinity_np(): Invalid argument [osandov@now ~/Dropbox/Homework/CSES51/551ws master]
```

```
PTHREAD SETAFFINITY NP(3)
                                                   Linux Programmer's Manual
                                                                                                       PTHREAD SETAFFINITY NP(3)
NAME
       pthread setaffinity np. pthread getaffinity np - set/get CPU affinity of a thread
SYNOPSIS
       #define GNU SOURCE
                                      /* See feature test macros(7) */
       #include <pthread.h>
       int pthread_setaffinity_np(pthread_t thread, size_t cpusetsize,
                                  const cou set t *couset):
       int pthread getaffinity np(pthread t thread, size t cpusetsize,
                                 cpu set t *cpuset);
       Compile and link with -pthread.
DESCRIPTION
       The pthread setaffinity np() function sets the CPU affinity mask of the thread thread to the CPU set pointed to by cpuset.
       If the call is successful, and the thread is not currently running on one of the CPUs in couset, then it is migrated to
       one of those CPUs.
       The pthread getaffinity np() function returns the CPU affinity mask of the thread thread in the buffer pointed to by
       cpuset.
       For more details on CPU affinity masks, see sched setaffinity(2). For a description of a set of macros that can be used
       to manipulate and inspect CPU sets, see CPU SET(3).
       The argument <u>cpusetsize</u> is the length (in bytes) of the buffer pointed to by <u>cpuset</u>. Typically, this argument would be
       specified as sizeof(cpu set t). (It may be some other value, if using the macros described in CPU SET(3) for dynamically
       allocating a CPU set.)
RETURN VALUE
       On success, these functions return O: on error, they return a nonzero error number.
FRRORS
       EFAULT A supplied memory address was invalid.
       EINVAL (pthread setaffinity mp()) The affinity bit mask mask contains no processors that are currently physically on the
              system and permitted to the thread according to any restrictions that may be imposed by the 'cpuset' mechanism
             described in cpuset(7).
       EINVAL (pthread setaffinity mp()) couset specified a CPU that was outside the set supported by the kernel. (The kernel
             configuration option CONFIG NR CPUS defines the range of the set supported by the kernel data type used to repre-
              sent (PU sets.)
  Manual page pthread setaffinity np(3) line 1/132 39% (press h for help or g to guit)
```

```
21
               free(opts.query):
20
               opts.guery = word regexp guery:
19
               opts.query len = strlen(opts.query);
18
           compile study(&opts.re, &opts.re extra, opts.guery, pcre opts, study opts):
16
14
       if (opts.search stream) {
13
           search stream(stdin, "");
      } else {
           for (i = 0: i < workers len: i++) {
               workers[i].id = i;
               int rv = pthread create(&(workers[i].thread), NULL, &search file worker, &(workers[i].id));
8
               if (rv != 0) {
                   die("error in pthread create(): %s", strerror(rv));
6
5 #if defined(HAVE PTHREAD SETAFFINITY NP) && defined(USE CPU SET)
               if (opts.use thread affinity) {
 3
                   cpu set t cpu set;
2
                   CPU ZERO(&cpu set);
                   CPU_SET(i % num_cores, &cpu_set);
144
                   rv = Othread setaffinity np(workers[i].thread, sizeof(cpu set), &cpu set);
 2
                       die("error in pthread_setaffinity_np(): %s", strerror(rv));
                   log debug("Thread %i set to CPU %i", i, i):
5
               } else {
6
                   log debug("Thread affinity disabled.");
8 #else
               log debug("No CPU affinity support.");
10 #endif
11
12
           for (i = 0; paths[i] != NULL; i++) {
               log debug('searching path %s for %s", paths[i], opts.guery);
               symhash = NULL;
14
15
               ignores *ig = init ignore(root ignores, "", 0);
               struct stat s = {.st dev = 0 };
17 #ifndef WIN32
18
               /* The device is ignored if opts.one dev is false, so it's fine
19
                * to leave it at the default 0
20
               if (opts.one_dev && lstat(paths[i], &s) == -1) {
21
22
                   log err("Failed to get device information for path %s. Skipping...", paths[i]);
                                                                                                                   144,22
                                                                                                                                 72%
```

```
r[osandov@mew ~/linux/linux.git]
s git bisect start v4.1-rc1 v4.0
Bisecting: 5576 revisions left to test after this (roughly 13 steps)
[6c373ca89399c5a3f7ef210ad8f63dc3437da345] Merge git://git.kernel.org/pub/scm/linux/kernel/git/davem/net-next
-[osandov@mew ~/linux/linux.git]
s git bisect run ~/Dropbox/linux/bugs/affinity-einval/bisect.sh >/dev/null 2>&1
r[osandov@mew ~/linux/linux.git]
s git bisect log
# bad: [b787f68c36d49bb1d9236f40381364lefa74a031] Linux 4.1-rcl
# good: [39a8804455fb23f09157341d3ba7db6d7ae6ee76] Linux 4.0
git bisect start 'v4.1-rc1' 'v4.0'
# bad: [6c373ca89399c5a3f7ef210ad8f63dc3437da345] Merge git://git.kernel.org/pub/scm/linux/kernel/git/davem/net-next
nit bisect bad 6c373ca89399c5a3f7ef210ad8f63dc3437da345
# bad: [e95e7f627062be5e6ce97lce873e6234c9lffc50] Merge branch 'timers-nohz-for-linus' of git://git.kernel.org/pub/scm/linux/kernel/g
it/tip/tip
git bisect bad e95e7f627062be5e6ce971ce873e6234c91ffc50
# bad: [c4be50eee2bd4d50e0f0ca58776f685c08de69c3] Merge tag 'driver-core-4.1-rc1' of git://git.kernel.org/pub/scm/linux/kernel/git/gr
eakh/driver-core
git bisect bad c4be50eee2bd4d50e0f0ca58776f685c08de69c3
# bad: [la370f4cd95e056d55ef5bf1a183880e70195e59] Merge tag 'edac for 4.1' of git;//git,kernel.org/pub/scm/linux/kernel/git/bp/bp
git bisect bad la370f4cd95e056d55ef5bfla183880e70195e59
# bad: [7fd56474db326f7a6df0e2a4e3a9600cc083ab9b] Merge branch 'timers-core-for-linus' of git://git.kernel.org/pub/scm/linux/kernel/g
git bisect bad 7fd56474db326f7a6df0e2a4e3a9600cc083ab9b
# good: [900360131066f192c82311a098d03d6ac6429e20] Merge tag 'for linus' of git://git.kernel.org/pub/scm/virt/kvm/kvm
git bisect good 900360131066f192c82311a098d03d6ac6429e20
# good: [979081e7440056da28b19e57acf20098caf49103] ACPI/PAD: Use explicit broadcast control function
git bisect good 979081e7440056da28b19e57acf20098caf49103
# bad: [62a935b256f68a71697716595347209fb5275426] sched/core: Drop debugging leftover trace printk call
git bisect bad 62a935b256f68a71697716595347209fb5275426
# good: [b5b4860d1d61ddc5308c7d492cbeaa3a6e508d7f] sched: Make scale rt invariant with frequency
git bisect good b5b4860d1d61ddc5308c7d492cbeaa3a6e508d7f
# good: [dfbca41f347997e57048a53755611c8e2d792924] sched: Optimize freg invariant accounting
git bisect good dfbca41f347997e57048a53755611c8e2d792924
# good: [07c54f7a7ff77bb47bae26e566969e9c4b6fb0c6] sched/core: Remove unused argument from init [rt|dl] rg()
git bisect good 07c54f7a7ff77bb47bae26e566969e9c4b6fb0c6
# bad: [3c18d447b3b36a8d3c90dc37dfbd363cdb685d0a] sched/core: Check for available DL bandwidth in cpuset cpu inactive()
git bisect bad 3c18d447b3b36a8d3c90dc37dfbd363cdb685d0a
# good: [4cd57f97l35840f63743lc92380c8da3edbe44ed] sched/deadline: Always engueue on previous rg when dl task timer() fires
git bisect good 4cd57f97135840f637431c92380c8da3edbe44ed
# first bad commit: [3c18d447b3b36a8d3c90dc37dfbd363cdb685d0a] sched/core: Check for available DL bandwidth in cpuset cpu inactive()
-[osandov@mew ~/linux/linux.git]
Ls I
```

```
commit 3c18d447b3b36a8d3c90dc37dfbd363cdb685d0a
Author: Juri Lelli <iuri.lelli@arm.com>
Date: Tue Mar 31 09:53:37 2015 +0100
   sched/core: Check for available DL bandwidth in couset cou inactive()
   Hotplug operations are destructive w.r.t. cpusets. In case such an
    operation is performed on a CPU belonging to an exlusive couset, the
    DL bandwidth information associated with the corresponding root
    domain is gone even if the operation fails (in sched cpu inactive()).
    For this reason we need to move the check we currently have in
    sched cpu inactive() to cpuset cpu inactive() to prevent useless
   cousets reconfiguration in the CPU DOWN FAILED path.
    Signed-off-by: Juri Lelli <juri.lelli@arm.com>
    Signed-off-by: Peter Ziilstra (Intel) <peterz@infradead.org>
    Cc: Juri Lelli <juri.lelli@gmail.com>
   Link: http://lkml.kernel.org/r/1427792017-7356-2-git-send-email-juri.lelli@arm.com
   Signed-off-by: Ingo Molnar <mingo@kernel.org>
diff --git a/kernel/sched/core.c b/kernel/sched/core.c
index 4c49e75ca24d..28b0d75a8273 100644
--- a/kernel/sched/core.c
+++ b/kernel/sched/core.c
00 -5337,36 +5337,13 00 static int sched cpu active(struct notifier block *nfb.
static int sched cou inactive(struct notifier block *nfb.
                                        unsigned long action, void *hcpu)
        unsigned long flags:
        long cpu = (long)hcpu;
        struct dl bw *dl b:
        switch (action & ~CPU TASKS FROZEN) {
        case CPU DOWN PREPARE:
                set cpu active(cpu, false):
                /* explicitly allow suspend */
                if (!(action & CPU TASKS FROZEN)) {
                        bool overflow:
                        int cpus;
                        rcu read lock sched():
                        dl \bar{b} = d\bar{l} bw \bar{o} f(cpu);
```

```
static int __init migration_init(void)
@@ -7006,7 +6983,6 @@ static int cpuset cpu active(struct notifier block *nfb, unsigned long action,
       case CPU ONLINE:
       case CPU DOWN FAILED:
               cpuset_update_active_cpus(true);
                break:
       default:
00 -7018.8 +6994.32 00 static int couset cou active(struct notifier block *nfb. unsigned long action.
static int cpuset cpu inactive(struct notifier block *nfb, unsigned long action,
       switch (action) {
       unsigned long flags;
       long cpu = (long)hcpu;
       struct dl bw *dl b;
       switch (action & ~CPU TASKS FROZEN) {
       case CPU DOWN PREPARE:
               7* explicitly allow suspend */
                if (!(action & CPU TASKS FROZEN)) {
                        bool overflow:
                        int cpus;
                        rcu read_lock_sched();
                        dl \bar{b} = d\bar{l} bw of(cpu);
                        raw spin lock irgsave(&dl b->lock, flags):
                        cpus = dl bw cpus(cpu);
                        overflow = dl overflow(dl b, cpus, 0, 0);
                        raw spin unlock irgrestore(&dl b->lock, flags):
                        rcu read unlock sched();
                        if (overflow) {
                                trace printk("hotplug failed for cpu %lu", cpu);
                                return notifier from errno(-EBUSY):
                cpuset update active cpus(false):
               break:
       case CPU DOWN PREPARE FROZEN:
(END)
```

```
i:Exit -:PrevPg <Space>:NextPg v:View Attachm. r:Replv <Right>:Next ?:Help
Date: Mon. 4 May 2015 03:09:36 -0700
From: Omar Sandoval <osandov@osandov.com>
To: Ingo Molnar ⊲mingo@redhat.com>. Peter Ziilstra <peterz@infradead.org>. Juri Lelli <juri.lelli@arm.com>.
        linux-kernel@vger.kernel.org
Cc: Omar Sandoval <osandov@osandov.com>
Subject: [PATCH] sched/core: fix regression in cpuset cpu inactive for suspend
X-Mailer: git-send-email 2.3.7
Commit 3c18d447b3b3 ("sched/core: Check for available DL bandwidth in
couset cou inactive()"), a SCHED DEADLINE bugfix, had a logic error that
caused a regression in setting a CPU inactive during suspend. I ran into
this when a program was failing pthread setaffinity np() with EINVAL after
a suspend+wake up. A simple reproducer:
$ ./a.out
sched setaffinity: Success
$ systemctl suspend
[wake up from suspend]
$ ./a.out
sched setaffinity: Invalid argument
Where ./a.out is:
        #define GNU SOURCE
       #include <errno.h>
        #include <sched.h>
        #include <stdio.h>
        #include <stdlib b>
       #include <string.h>
        #include <unistd.h>
        int main(void)
                long num cores;
               cpu set t cpu set:
               int ret:
               num cores = sysconf( SC NPROCESSORS ONLN);
                CPU ZERO(&cpu set):
                CPU SET(num cores - 1, &cpu set);
                errno = 0:
                ret = sched setaffinity(getpid(), sizeof(cpu set), &cpu set);
  F- 29/99: Omar Sandoval
                                  [PATCH] sched/core: fix regression in cpuset cpu inactive for suspend
```

```
:Exit -: PrevPg <Space>: NextPg v: View Attachm. r: Replv <Right>: Next ?: Help
       #include <errno.h>
       #include <sched.h>
       #include <stdio.h>
       #include <stdlib.h>
       #include <string.h>
       #include <unistd.h>
       int main(void)
               long num cores:
               cpu set t cpu set;
               int ret:
               num cores = sysconf( SC NPROCESSORS ONLN);
               CPU ZERO(&cpu set);
               CPU SET(num cores - 1, &cpu set);
               errno = 0:
               ret = sched setaffinity(getpid(), sizeof(cpu set), &cpu set);
               perror('sched setaffinity');
               return ret ? EXIT_FAILURE : EXIT_SUCCESS;
The mistake is that suspend is handled in the action ==
CPU DOWN PREPARE FROZEN case of the switch statement in
cpuset cpu inactive. However, the commit in question masked out
CPU TASKS FROZEN from the action, making this case dead. The fix is
straightforward.
Fixes: 3c18d447b3b3 ("sched/core: Check for available DL bandwidth in
cpuset cpu inactive()")
Signed-off-by: Omar Sandoval <osandov@osandov.com>
This applies to v4.1-rc2.
I'm not familiar with this part of the kernel at all, but this Seems
Right (TM). If it isn't, consider this a bug report-plus.
Thanks!
kernel/sched/core.c | 28 +++++++++
 1 file changed, 12 insertions(+), 16 deletions(-)
 F- 29/99: Omar Sandoval
                             [PATCH] sched/core: fix regression in cpuset cpu inactive for suspend
```

```
i:Exit :PrevPg <Space>:NextPg v:View Attachm. r:Replv <Right>:Next ?:Help
diff --git a/kernel/sched/core.c b/kernel/sched/core.c
index fe22f7510bce..ecf05bf39525 100644
--- a/kernel/sched/core.c
+++ b/kernel/sched/core.c
00 -6997,27 +6997,23 00 static int cpuset cpu inactive(struct notifier block *nfb, unsigned long action,
        unsigned long flags;
        long cpu = (long)hcpu;
        struct dl bw *dl b;
        bool overflow;
        int cous:
        switch (action & -CPU TASKS FROZEN) {
        switch (action) {
        case CPU DOWN PREPARE:
                 7* explicitly allow suspend */
                 if (!(action & CPU TASKS FROZEN)) {
                          bool overflow:
                          int cpus;
                          rcu_read_lock_sched();
                          dl \overline{b} = d\overline{l} bw \overline{of(cpu)};
                 rcu read lock sched():
                 dl b = dl bw of (cpu):
                          raw_spin_lock_irqsave(&dl_b->lock, flags);
                          cpus = dl_bw_cpus(cpu);
                          overflow = _dl_overflow(dl_b, cpus, 0, 0);
raw_spin_unlock_irgrestore(&dl_b.>lock, flags);
                 raw spin lock irosave(&dl b->lock, flags):
                 cpus = dl_bw_cpus(cpu);
overflow = __dl_overflow(dl_b, cpus, 0, 0);
                 raw spin unlock irgrestore(&dl b->lock, flags):
                          rcu_read_unlock_sched();
                 rcu read unlock sched():
                          if (overflow)
                                  return notifier from errno(-EBUSY):
                 if (overflow)
                          return notifier from errno(-EBUSY);
                 couset update active cous(false):
                                     [PATCH] sched/core: fix regression in cpuset cpu inactive for suspend
```

```
i:Exit -:PrevPg <Space>:NextPg v:View Attachm. r:Replv <Right>:Next ?:Help
Date: Fri. 15 May 2015 09:19:31 +0200
From: Ingo Molnar amingo@kernel.org>
To: Linus Torvalds <torvalds@linux-foundation.org>
Cc: linux-kernelöyger.kernel.org. Peter Zijlstra <a.p.zijlstraöchello.nl>. Thomas Gleixner <tglxölinutronix.de>. Andrew Morton
       <akpm@linux-foundation.org>
Subject: [GIT PULL] scheduler fixes
User-Agent: Mutt/1.5.23 (2014-03-12)
Linus,
Please pull the latest sched-urgent-for-linus git tree from:
  ait://ait.kernel.org/pub/scm/linux/kernel/ait/tip/tip.ait sched-urgent-for-linus
  # HEAD: 533445c6e53368569e50ab3fb712230c03d523f3 sched/core: Fix regression in cpuset cpu inactive() for suspend
Two fixes: a suspend/resume related regression fix, and an RT priority
boosting fix.
Thanks.
       Indo
Omar Sandoval (1)
     sched/core: Fix regression in couset cou inactive() for suspend
Thomas Gleixner (1).
     sched: Handle priority boosted tasks proper in setscheduler()
 include/linux/sched/rt.h | 7 ++++---
 kernel/locking/rtmutex.c | 12 +++++
 kernel (sched/core c
                        54 +++++++++++++++++++++
 3 files changed, 37 insertions(+), 36 deletions(-)
diff --git a/include/linux/sched/rt.h b/include/linux/sched/rt.h
index 6341f5be6e24..a30b172df6e1 100644
--- a/include/linux/sched/rt.h
+++ b/include/linux/sched/rt.h
00 -18.7 +18.7 00 static inline int rt task(struct task struct *p)
#ifdef CONFIG RT MUTEXES
  I- 11/33: Indo Molnar
                                  [GIT PULL] scheduler fixes
                                                                                                                          -- (19%)
```

Where to Go From Here

- Look at Documentation/HOWTO
- Build and install a kernel from scratch
- Watch "Write and Submit your first Linux kernel Patch" on YouTube
- Take the Eudyptula Challenge
- Read LWN.net
- Email me at osandov@osandov.com