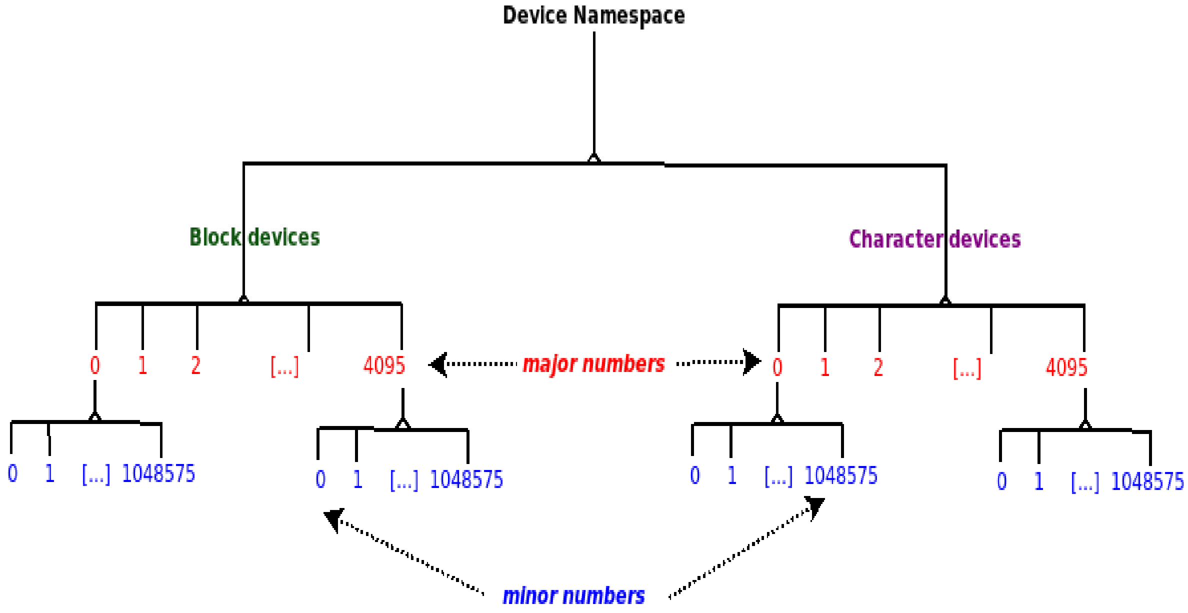
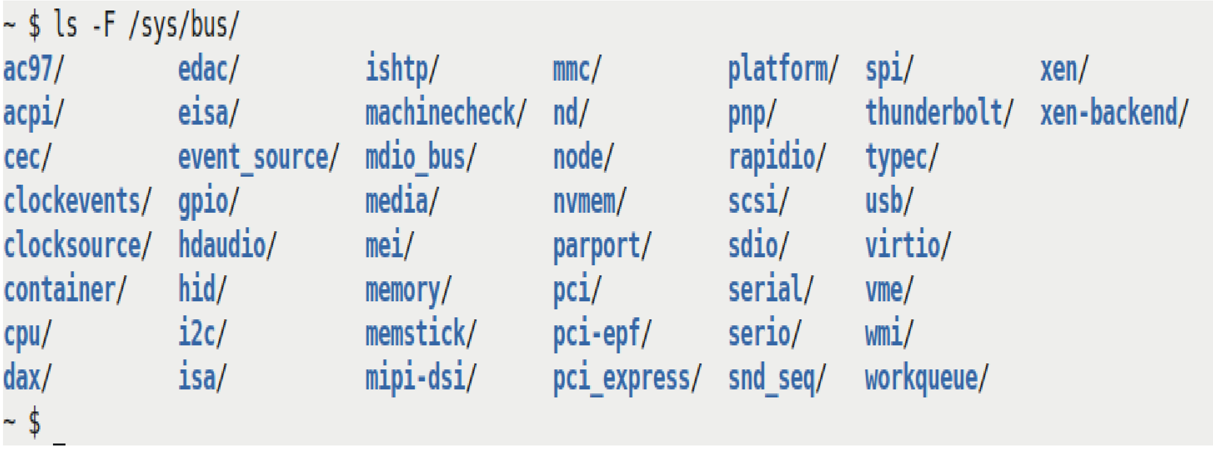
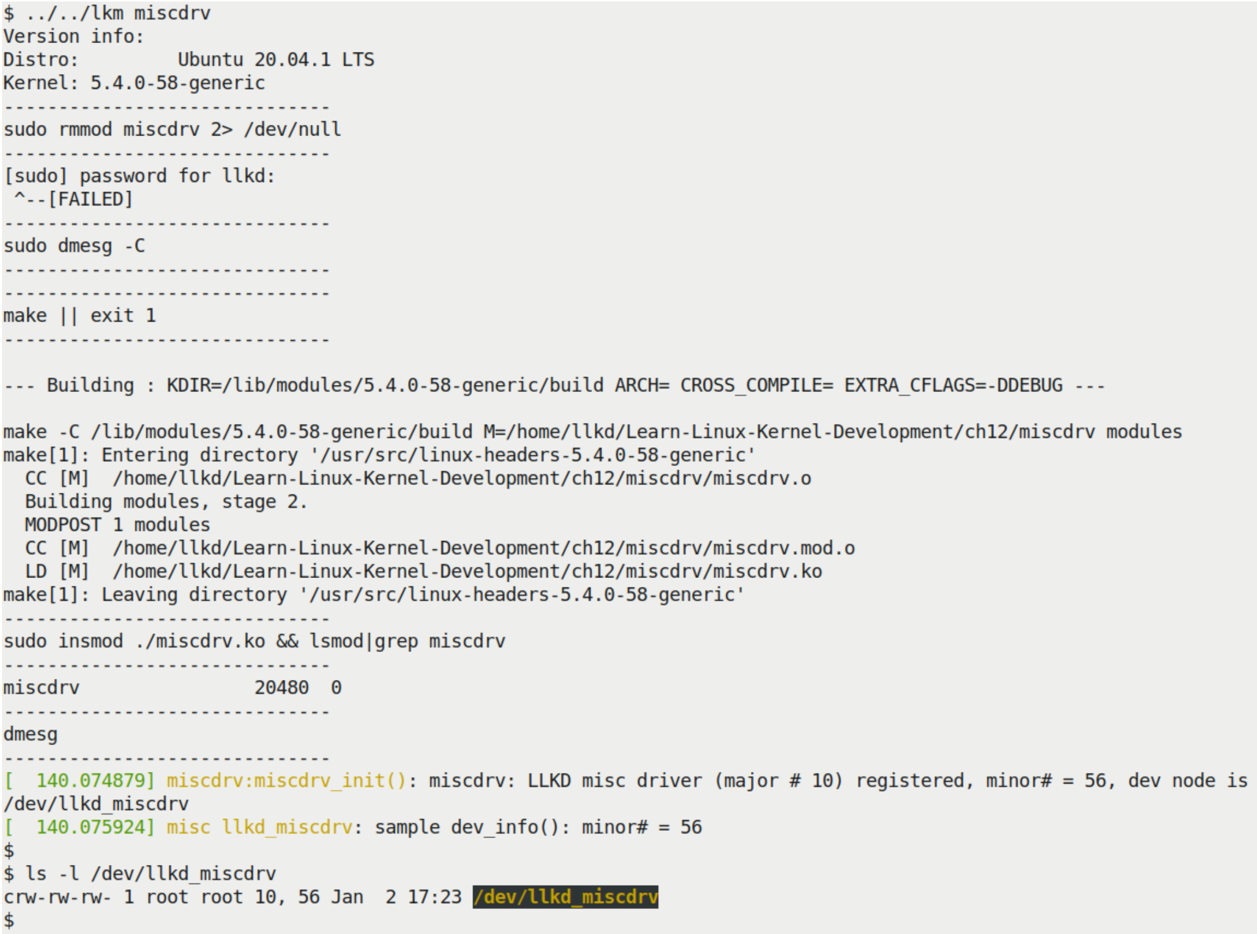
Chapter 1: Writing a Simple misc Character Device Driver



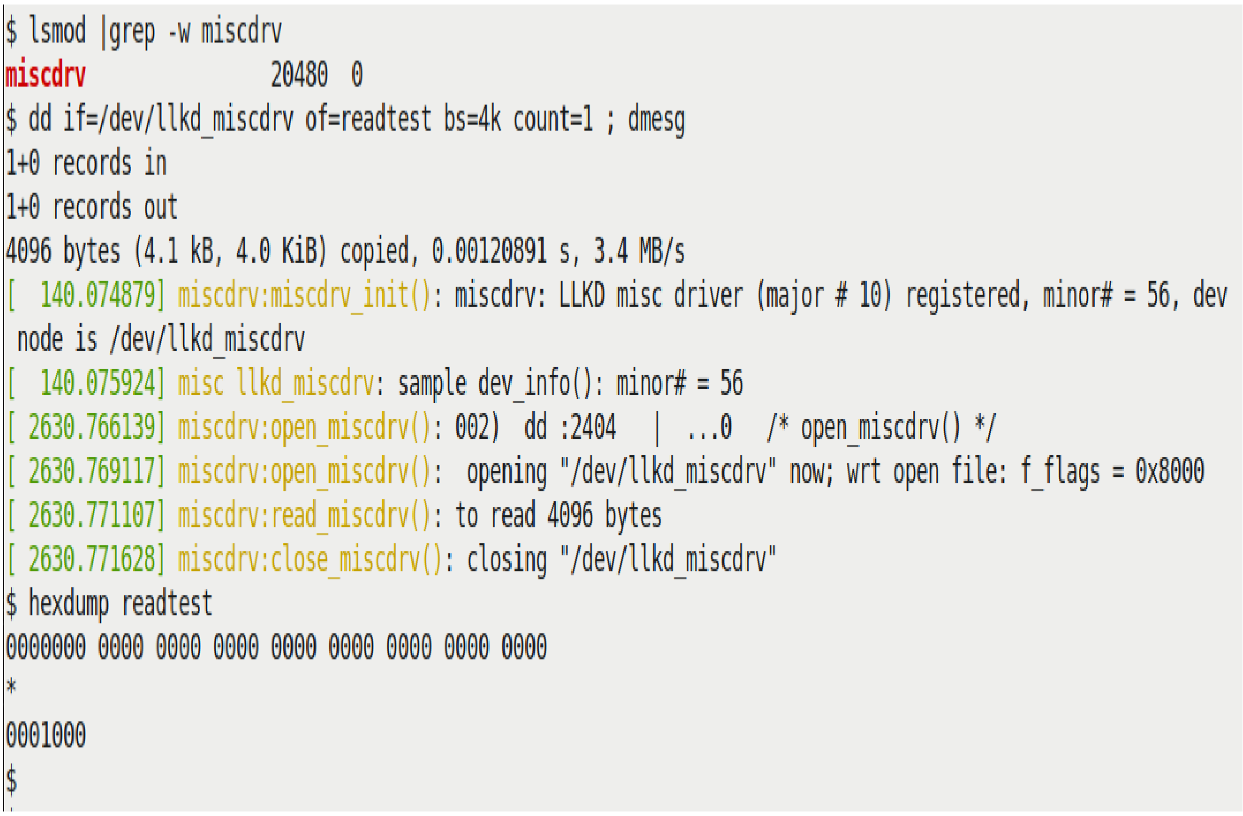
```
10 char       Non-serial mice, misc features
                0 = /dev/logibm       Logitech bus mouse
                1 = /dev/psaux        PS/2-style mouse port
                2 = /dev/inportbm     Microsoft Inport bus mouse
                3 = /dev/atibm        ATI XL bus mouse
                4 = /dev/jbm          J-mouse
                4 = /dev/amigamouse   Amiga mouse (68k/Amiga)
                5 = /dev/atarimouse   Atari mouse
                6 = /dev/sunmouse     Sun mouse
                7 = /dev/amigamouse1  Second Amiga mouse
                8 = /dev/smouse       Simple serial mouse driver
                9 = /dev/pc110pad     IBM PC-110 digitizer pad
               10 = /dev/adbmouse     Apple Desktop Bus mouse
               11 = /dev/vrtpanel     Vr41xx embedded touch panel
               13 = /dev/vpcmouse     Connectix Virtual PC Mouse
               14 = /dev/touchscreen/ucb1x00  UCB 1x00 touchscreen
               15 = /dev/touchscreen/mk712    MK712 touchscreen
              128 = /dev/beep         Fancy beep device
              129 =
              130 = /dev/watchdog     Watchdog timer port
              131 = /dev/temperature  Machine internal temperature
              132 = /dev/hwtrap       Hardware fault trap
              133 = /dev/exttrp       External device trap
              134 = /dev/apm_bios     Advanced Power Management BIOS
              135 = /dev/rtc          Real Time Clock
              137 = /dev/vhci         Bluetooth virtual HCI driver
              139 = /dev/openprom     SPARC OpenBoot PROM
              140 = /dev/relay8       Berkshire Products Octal relay card
              141 = /dev/relay16      Berkshire Products ISO-16 relay card
              142 =
              143 = /dev/pciconf      PCI configuration space
              144 = /dev/nvram        Non-volatile configuration RAM
```



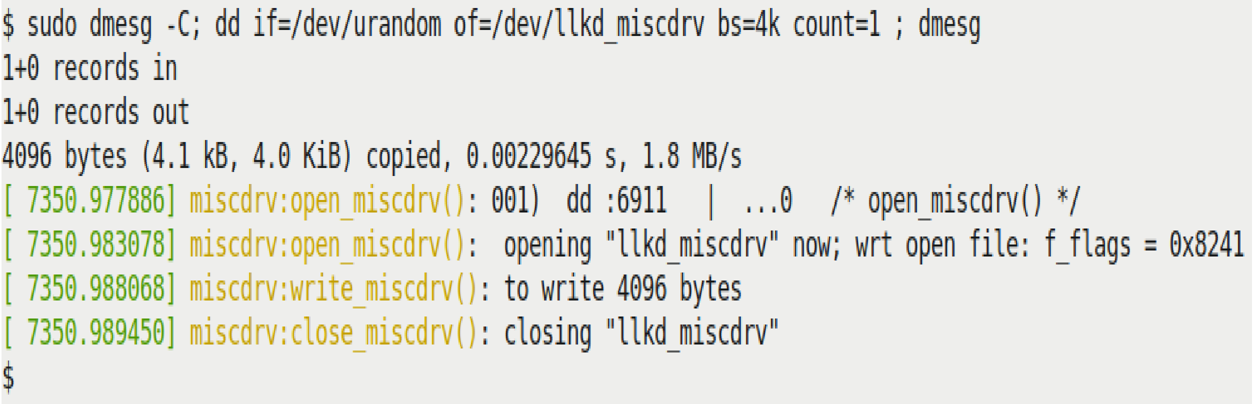
```
~ $ ls -F /sys/bus/
ac97/         edac/          ishtp/         mmc/          platform/  spi/          xen/
acpi/         eisa/          machinecheck/  nd/           pnp/       thunderbolt/  xen-backend/
cec/          event_source/  mdio_bus/      node/         rapidio/   typec/
clockevents/  gpio/          media/         nvmem/        scsi/      usb/
clocksource/  hdaudio/       mei/           parport/      sdio/      virtio/
container/    hid/           memory/        pci/          serial/    vme/
cpu/          i2c/           memstick/      pci-epf/      serio/     wmi/
dax/          isa/           mipi-dsi/      pci_express/  snd_seq/   workqueue/
~ $ _
```



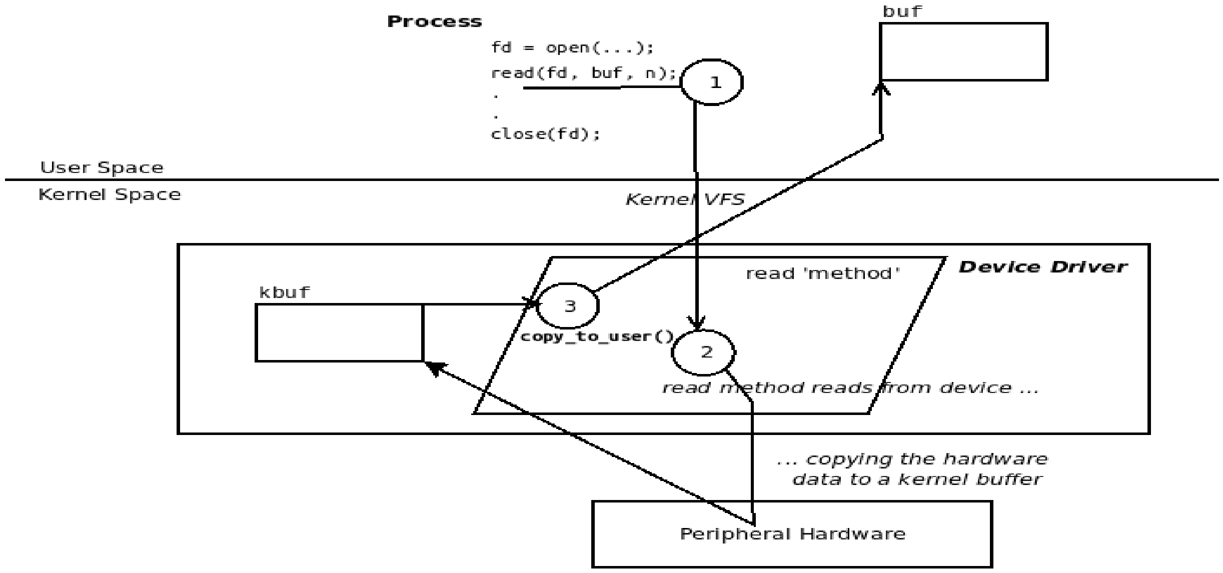
```
$ ../../lkm miscdrv
Version info:
Distro:         Ubuntu 20.04.1 LTS
Kernel: 5.4.0-58-generic
------------------------------
sudo rmmod miscdrv 2> /dev/null
------------------------------
[sudo] password for llkd:
 ^--[FAILED]
------------------------------
sudo dmesg -C
------------------------------
------------------------------
make || exit 1
------------------------------

--- Building : KDIR=/lib/modules/5.4.0-58-generic/build ARCH= CROSS_COMPILE= EXTRA_CFLAGS=-DDEBUG ---

make -C /lib/modules/5.4.0-58-generic/build M=/home/llkd/Learn-Linux-Kernel-Development/ch12/miscdrv modules
make[1]: Entering directory '/usr/src/linux-headers-5.4.0-58-generic'
  CC [M]  /home/llkd/Learn-Linux-Kernel-Development/ch12/miscdrv/miscdrv.o
  Building modules, stage 2.
  MODPOST 1 modules
  CC [M]  /home/llkd/Learn-Linux-Kernel-Development/ch12/miscdrv/miscdrv.mod.o
  LD [M]  /home/llkd/Learn-Linux-Kernel-Development/ch12/miscdrv/miscdrv.ko
make[1]: Leaving directory '/usr/src/linux-headers-5.4.0-58-generic'
------------------------------
sudo insmod ./miscdrv.ko && lsmod|grep miscdrv
------------------------------
miscdrv                20480  0
------------------------------
dmesg
------------------------------
[  140.074879] miscdrv:miscdrv_init(): miscdrv: LLKD misc driver (major # 10) registered, minor# = 56, dev node is
/dev/llkd_miscdrv
[  140.075924] misc llkd_miscdrv: sample dev_info(): minor# = 56
$
$ ls -l /dev/llkd_miscdrv
crw-rw-rw- 1 root root 10, 56 Jan  2 17:23 /dev/llkd_miscdrv
$
```

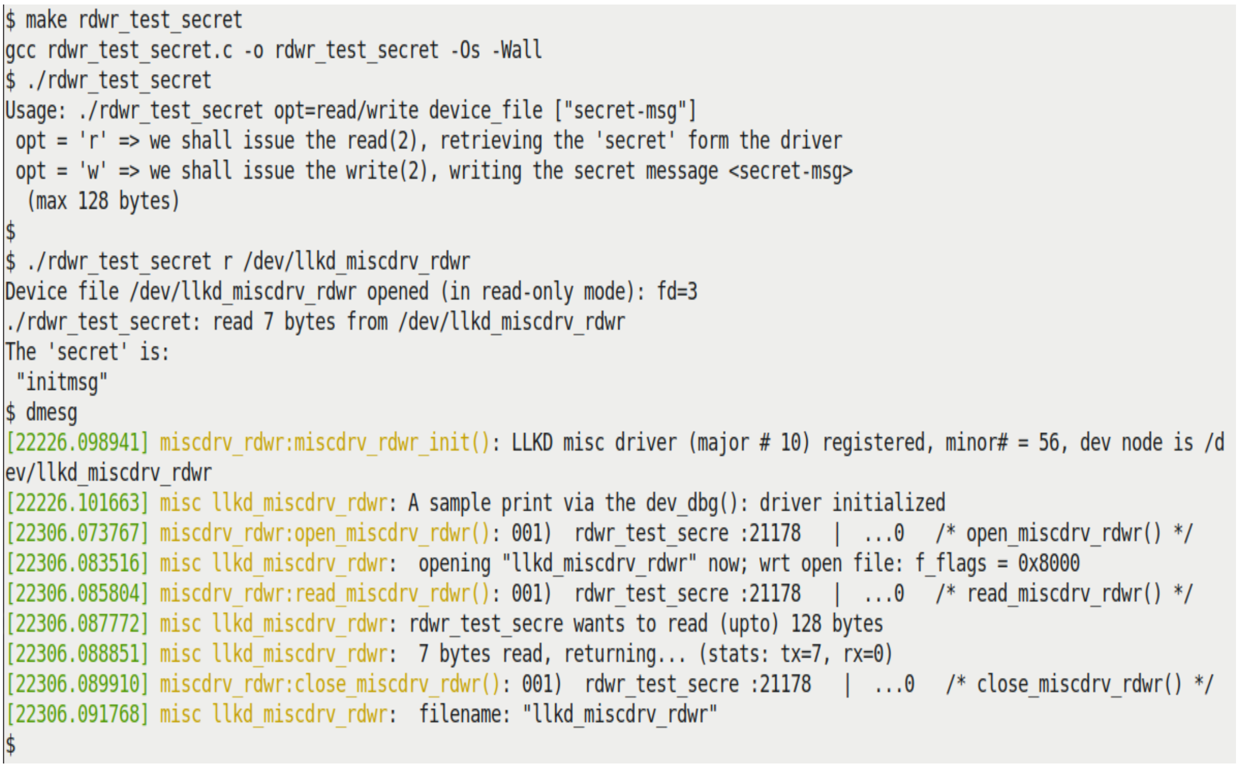


```
$ lsmod |grep -w miscdrv
miscdrv                20480  0
$ dd if=/dev/llkd_miscdrv of=readtest bs=4k count=1 ; dmesg
1+0 records in
1+0 records out
4096 bytes (4.1 kB, 4.0 KiB) copied, 0.00120891 s, 3.4 MB/s
[  140.074879] miscdrv:miscdrv_init(): miscdrv: LLKD misc driver (major # 10) registered, minor# = 56, dev
 node is /dev/llkd_miscdrv
[  140.075924] misc llkd_miscdrv: sample dev_info(): minor# = 56
[ 2630.766139] miscdrv:open_miscdrv(): 002)  dd :2404   |  ...0   /* open_miscdrv() */
[ 2630.769117] miscdrv:open_miscdrv():  opening "/dev/llkd_miscdrv" now; wrt open file: f_flags = 0x8000
[ 2630.771107] miscdrv:read_miscdrv(): to read 4096 bytes
[ 2630.771628] miscdrv:close_miscdrv(): closing "/dev/llkd_miscdrv"
$ hexdump readtest
0000000 0000 0000 0000 0000 0000 0000 0000 0000
*
0001000
$
```

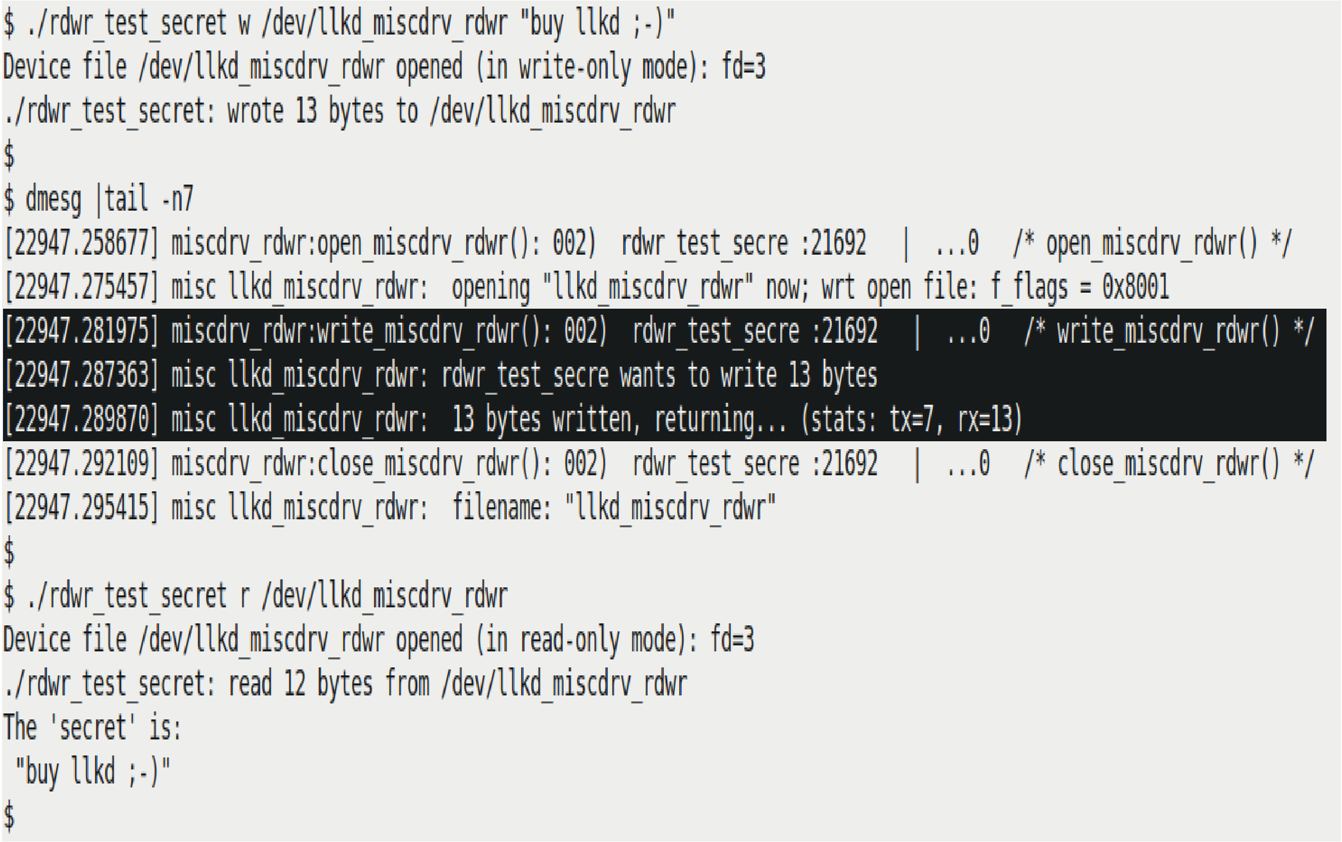


```
$ sudo dmesg -C; dd if=/dev/urandom of=/dev/llkd_miscdrv bs=4k count=1 ; dmesg
1+0 records in
1+0 records out
4096 bytes (4.1 kB, 4.0 KiB) copied, 0.00229645 s, 1.8 MB/s
[ 7350.977886] miscdrv:open_miscdrv(): 001)  dd :6911   |  ...0   /* open_miscdrv() */
[ 7350.983078] miscdrv:open_miscdrv():  opening "llkd_miscdrv" now; wrt open file: f_flags = 0x8241
[ 7350.988068] miscdrv:write_miscdrv(): to write 4096 bytes
[ 7350.989450] miscdrv:close_miscdrv(): closing "llkd_miscdrv"
$
```

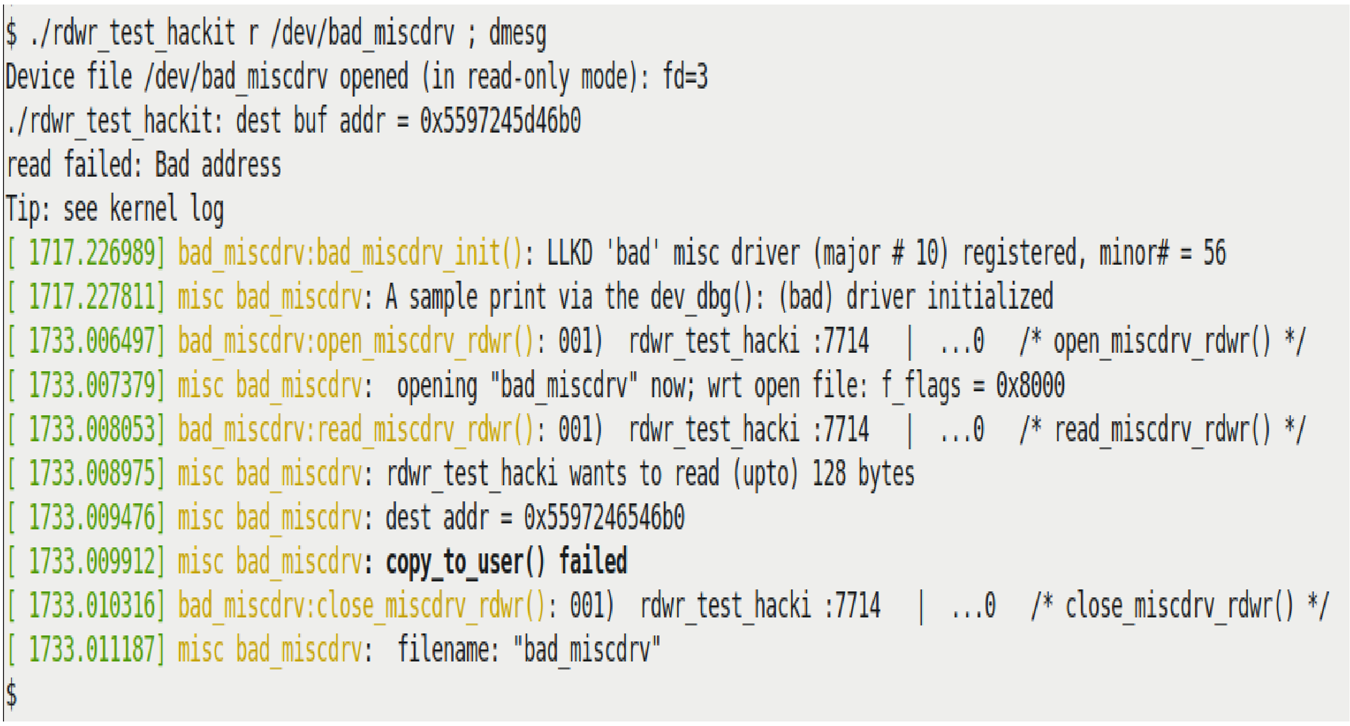




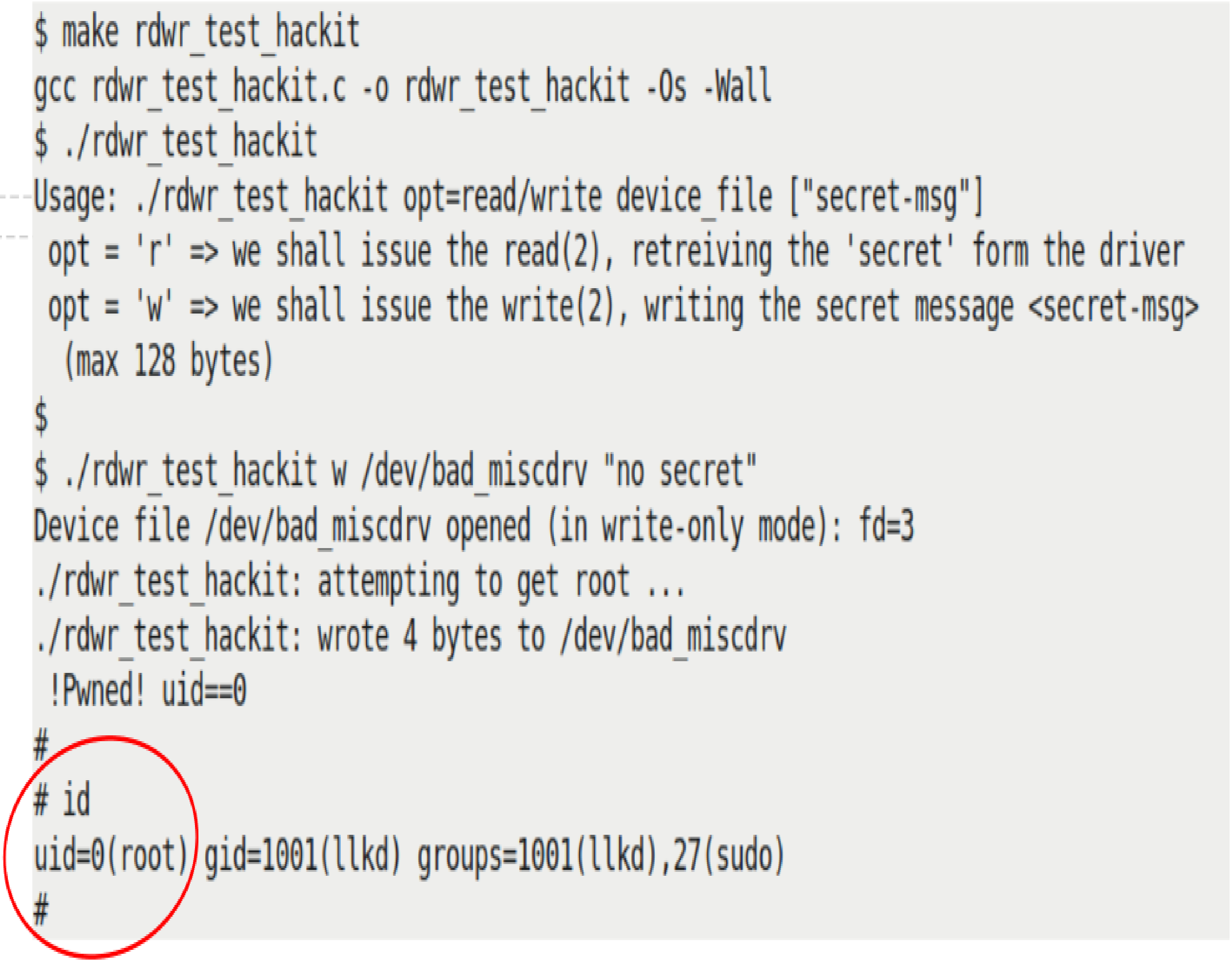
```
$ make rdwr_test_secret
gcc rdwr_test_secret.c -o rdwr_test_secret -Os -Wall
$ ./rdwr_test_secret
Usage: ./rdwr_test_secret opt=read/write device_file ["secret-msg"]
 opt = 'r' => we shall issue the read(2), retrieving the 'secret' form the driver
 opt = 'w' => we shall issue the write(2), writing the secret message <secret-msg>
  (max 128 bytes)
$
$ ./rdwr_test_secret r /dev/llkd_miscdrv_rdwr
Device file /dev/llkd_miscdrv_rdwr opened (in read-only mode): fd=3
./rdwr_test_secret: read 7 bytes from /dev/llkd_miscdrv_rdwr
The 'secret' is:
 "initmsg"
$ dmesg
[22226.098941] miscdrv_rdwr:miscdrv_rdwr_init(): LLKD misc driver (major # 10) registered, minor# = 56, dev node is /d
ev/llkd_miscdrv_rdwr
[22226.101663] misc llkd_miscdrv_rdwr: A sample print via the dev_dbg(): driver initialized
[22306.073767] miscdrv_rdwr:open_miscdrv_rdwr(): 001)  rdwr_test_secre :21178   |  ...0   /* open_miscdrv_rdwr() */
[22306.083516] misc llkd_miscdrv_rdwr:  opening "llkd_miscdrv_rdwr" now; wrt open file: f_flags = 0x8000
[22306.085804] miscdrv_rdwr:read_miscdrv_rdwr(): 001)  rdwr_test_secre :21178   |  ...0   /* read_miscdrv_rdwr() */
[22306.087772] misc llkd_miscdrv_rdwr: rdwr_test_secre wants to read (upto) 128 bytes
[22306.088851] misc llkd_miscdrv_rdwr:  7 bytes read, returning... (stats: tx=7, rx=0)
[22306.089910] miscdrv_rdwr:close_miscdrv_rdwr(): 001)  rdwr_test_secre :21178   |  ...0   /* close_miscdrv_rdwr() */
[22306.091768] misc llkd_miscdrv_rdwr:  filename: "llkd_miscdrv_rdwr"
$
```



```
$ ./rdwr_test_secret w /dev/llkd_miscdrv_rdwr "buy llkd ;-)"
Device file /dev/llkd_miscdrv_rdwr opened (in write-only mode): fd=3
./rdwr_test_secret: wrote 13 bytes to /dev/llkd_miscdrv_rdwr
$
$ dmesg |tail -n7
[22947.258677] miscdrv_rdwr:open_miscdrv_rdwr(): 002)  rdwr_test_secre :21692   |  ...0   /* open_miscdrv_rdwr() */
[22947.275457] misc llkd_miscdrv_rdwr:  opening "llkd_miscdrv_rdwr" now; wrt open file: f_flags = 0x8001
[22947.281975] miscdrv_rdwr:write_miscdrv_rdwr(): 002)  rdwr_test_secre :21692   |  ...0   /* write_miscdrv_rdwr() */
[22947.287363] misc llkd_miscdrv_rdwr: rdwr_test_secre wants to write 13 bytes
[22947.289870] misc llkd_miscdrv_rdwr:  13 bytes written, returning... (stats: tx=7, rx=13)
[22947.292109] miscdrv_rdwr:close_miscdrv_rdwr(): 002)  rdwr_test_secre :21692   |  ...0   /* close_miscdrv_rdwr() */
[22947.295415] misc llkd_miscdrv_rdwr:  filename: "llkd_miscdrv_rdwr"
$
$ ./rdwr_test_secret r /dev/llkd_miscdrv_rdwr
Device file /dev/llkd_miscdrv_rdwr opened (in read-only mode): fd=3
./rdwr_test_secret: read 12 bytes from /dev/llkd_miscdrv_rdwr
The 'secret' is:
 "buy llkd ;-)"
$
```

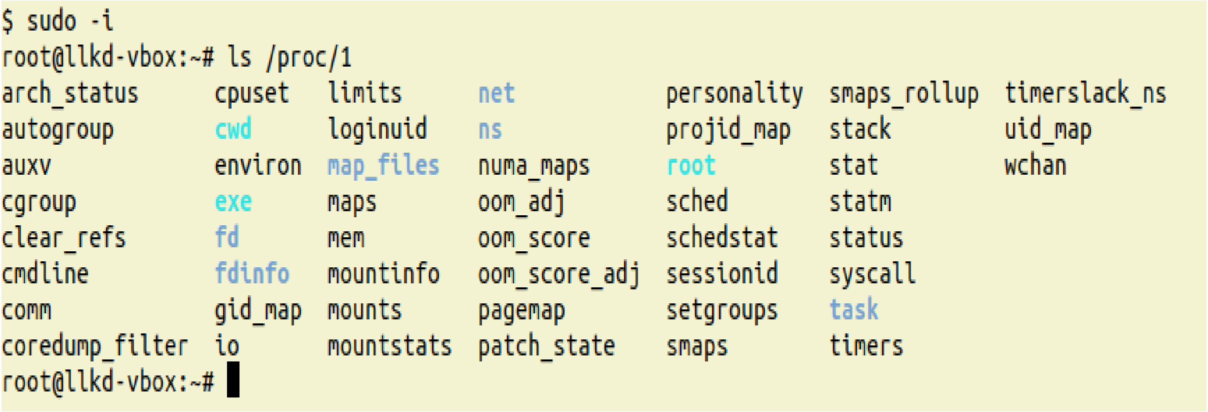


```
$ ./rdwr_test_hackit r /dev/bad_miscdrv ; dmesg
Device file /dev/bad_miscdrv opened (in read-only mode): fd=3
./rdwr_test_hackit: dest buf addr = 0x5597245d46b0
read failed: Bad address
Tip: see kernel log
[ 1717.226989] bad_miscdrv:bad_miscdrv_init(): LLKD 'bad' misc driver (major # 10) registered, minor# = 56
[ 1717.227811] misc bad_miscdrv: A sample print via the dev_dbg(): (bad) driver initialized
[ 1733.006497] bad_miscdrv:open_miscdrv_rdwr(): 001)  rdwr_test_hacki :7714   |  ...0   /* open_miscdrv_rdwr() */
[ 1733.007379] misc bad_miscdrv:  opening "bad_miscdrv" now; wrt open file: f_flags = 0x8000
[ 1733.008053] bad_miscdrv:read_miscdrv_rdwr(): 001)  rdwr_test_hacki :7714   |  ...0   /* read_miscdrv_rdwr() */
[ 1733.008975] misc bad_miscdrv: rdwr_test_hacki wants to read (upto) 128 bytes
[ 1733.009476] misc bad_miscdrv: dest addr = 0x5597246546b0
[ 1733.009912] misc bad_miscdrv: copy_to_user() failed
[ 1733.010316] bad_miscdrv:close_miscdrv_rdwr(): 001)  rdwr_test_hacki :7714   |  ...0   /* close_miscdrv_rdwr() */
[ 1733.011187] misc bad_miscdrv:  filename: "bad_miscdrv"
$
```

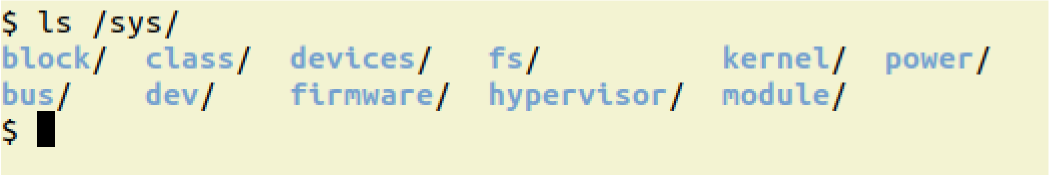


```
$ make rdwr_test_hackit
gcc rdwr_test_hackit.c -o rdwr_test_hackit -Os -Wall
$ ./rdwr_test_hackit
Usage: ./rdwr_test_hackit opt=read/write device_file ["secret-msg"]
 opt = 'r' => we shall issue the read(2), retreiving the 'secret' form the driver
 opt = 'w' => we shall issue the write(2), writing the secret message <secret-msg>
  (max 128 bytes)
$
$ ./rdwr_test_hackit w /dev/bad_miscdrv "no secret"
Device file /dev/bad_miscdrv opened (in write-only mode): fd=3
./rdwr_test_hackit: attempting to get root ...
./rdwr_test_hackit: wrote 4 bytes to /dev/bad_miscdrv
 !Pwned! uid==0
#
# id
uid=0(root) gid=1001(llkd) groups=1001(llkd),27(sudo)
#
```

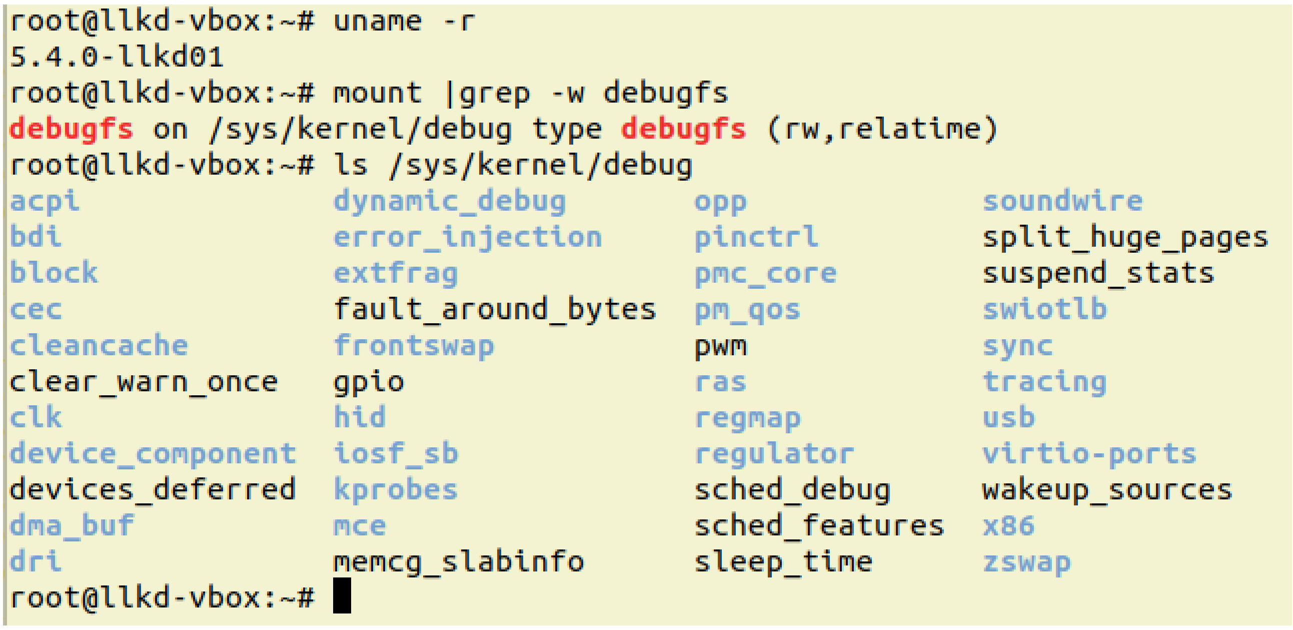
Chapter 2: User-Kernel Communication Pathways



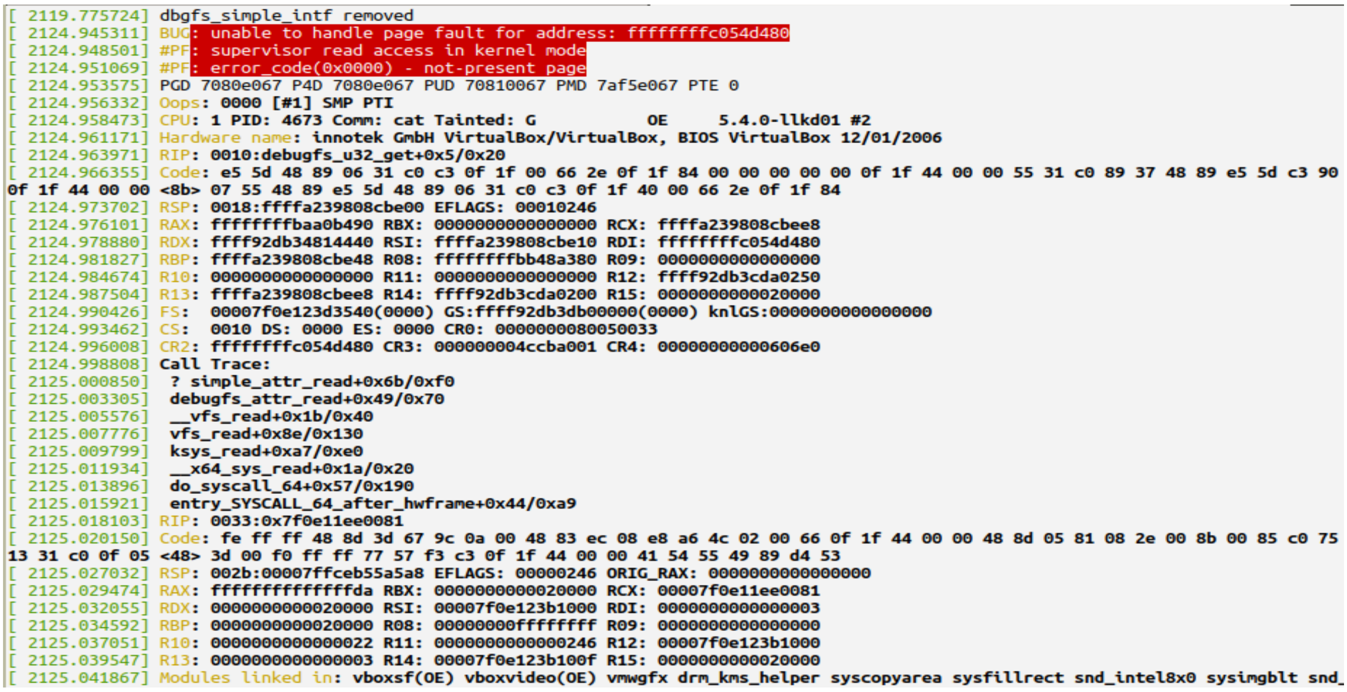
```
$ sudo -i
root@llkd-vbox:~# ls /proc/1
arch_status      cpuset   limits      net            personality  smaps_rollup  timerslack_ns
autogroup        cwd      loginuid    ns             projid_map   stack         uid_map
auxv             environ  map_files   numa_maps      root         stat          wchan
cgroup           exe      maps        oom_adj        sched        statm
clear_refs       fd       mem         oom_score      schedstat    status
cmdline          fdinfo   mountinfo   oom_score_adj  sessionid    syscall
comm             gid_map  mounts      pagemap        setgroups    task
coredump_filter  io       mountstats  patch_state    smaps        timers
root@llkd-vbox:~# █
```



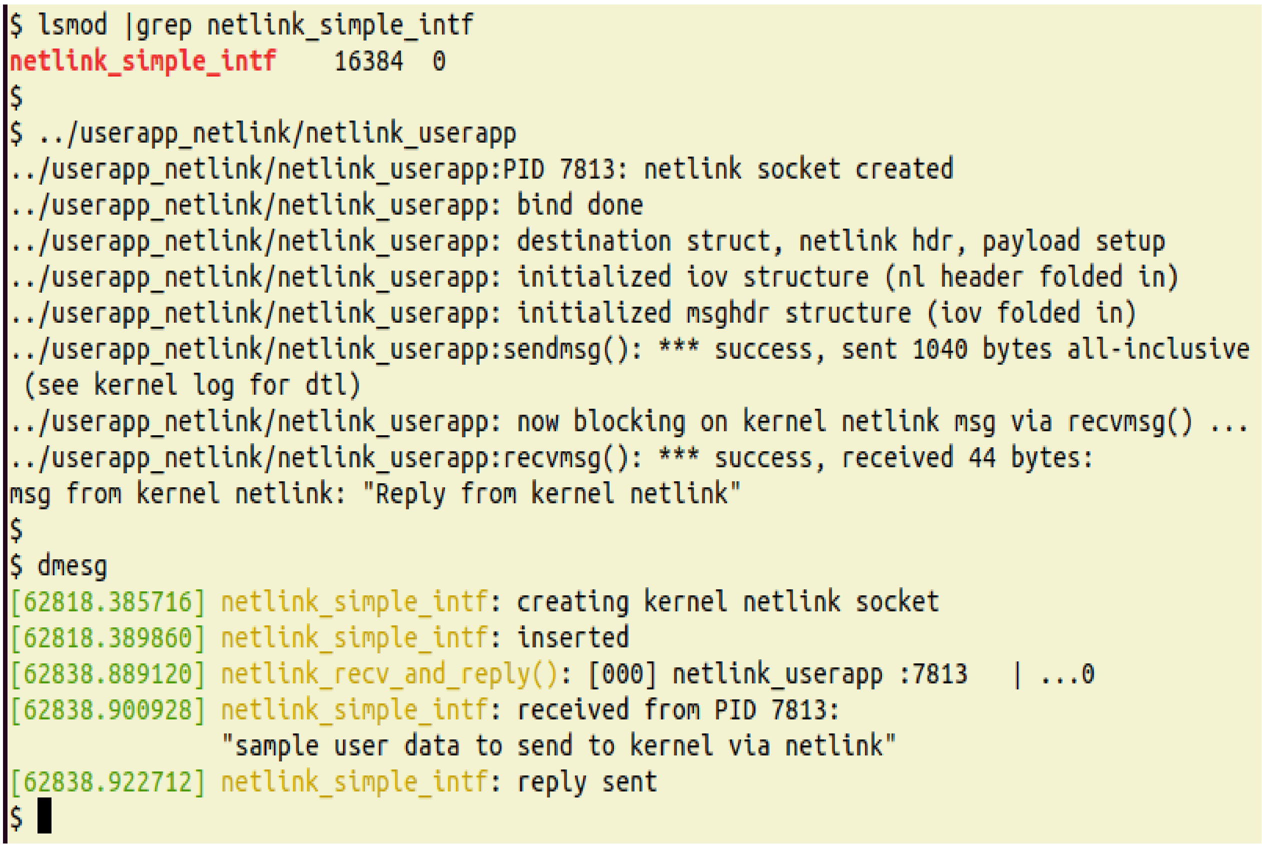
```
$ ls /sys/
block/  class/  devices/   fs/          kernel/  power/
bus/    dev/    firmware/  hypervisor/  module/
$ █
```



```
root@llkd-vbox:~# uname -r
5.4.0-llkd01
root@llkd-vbox:~# mount |grep -w debugfs
debugfs on /sys/kernel/debug type debugfs (rw,relatime)
root@llkd-vbox:~# ls /sys/kernel/debug
acpi              dynamic_debug       opp             soundwire
bdi               error_injection     pinctrl         split_huge_pages
block             extfrag             pmc_core        suspend_stats
cec               fault_around_bytes  pm_qos          swiotlb
cleancache        frontswap           pwm             sync
clear_warn_once   gpio                ras             tracing
clk               hid                 regmap          usb
device_component  iosf_sb             regulator       virtio-ports
devices_deferred  kprobes             sched_debug     wakeup_sources
dma_buf           mce                 sched_features  x86
dri               memcg_slabinfo      sleep_time      zswap
root@llkd-vbox:~# █
```



```
[ 2119.775724] dbgfs_simple_intf removed
[ 2124.945311] BUG: unable to handle page fault for address: ffffffffc054d480
[ 2124.948501] #PF: supervisor read access in kernel mode
[ 2124.951069] #PF: error_code(0x0000) - not-present page
[ 2124.953575] PGD 7080e067 P4D 7080e067 PUD 70810067 PMD 7af5e067 PTE 0
[ 2124.956332] Oops: 0000 [#1] SMP PTI
[ 2124.958473] CPU: 1 PID: 4673 Comm: cat Tainted: G           OE     5.4.0-llkd01 #2
[ 2124.961171] Hardware name: innotek GmbH VirtualBox/VirtualBox, BIOS VirtualBox 12/01/2006
[ 2124.963971] RIP: 0010:debugfs_u32_get+0x5/0x20
[ 2124.966355] Code: e5 5d 48 89 06 31 c0 c3 0f 1f 00 66 2e 0f 1f 84 00 00 00 00 00 0f 1f 44 00 00 55 31 c0 89 37 48 89 e5 5d c3 90
0f 1f 44 00 00 <8b> 07 55 48 89 e5 5d 48 89 06 31 c0 c3 0f 1f 40 00 66 2e 0f 1f 84
[ 2124.973702] RSP: 0018:ffffa239808cbe00 EFLAGS: 00010246
[ 2124.976101] RAX: ffffffffbaa0b490 RBX: 0000000000000000 RCX: ffffa239808cbee8
[ 2124.978880] RDX: ffff92db34814440 RSI: ffffa239808cbe10 RDI: ffffffffc054d480
[ 2124.981827] RBP: ffffa239808cbe48 R08: ffffffffbb48a380 R09: 0000000000000000
[ 2124.984674] R10: 0000000000000000 R11: 0000000000000000 R12: ffff92db3cda0250
[ 2124.987504] R13: ffffa239808cbee8 R14: ffff92db3cda0200 R15: 0000000000020000
[ 2124.990426] FS:  00007f0e123d3540(0000) GS:ffff92db3db00000(0000) knlGS:0000000000000000
[ 2124.993462] CS:  0010 DS: 0000 ES: 0000 CR0: 0000000080050033
[ 2124.996008] CR2: ffffffffc054d480 CR3: 000000004ccba001 CR4: 00000000000606e0
[ 2124.998808] Call Trace:
[ 2125.000850]  ? simple_attr_read+0x6b/0xf0
[ 2125.003305]  debugfs_attr_read+0x49/0x70
[ 2125.005576]  __vfs_read+0x1b/0x40
[ 2125.007776]  vfs_read+0x8e/0x130
[ 2125.009799]  ksys_read+0xa7/0xe0
[ 2125.011934]  __x64_sys_read+0x1a/0x20
[ 2125.013896]  do_syscall_64+0x57/0x190
[ 2125.015921]  entry_SYSCALL_64_after_hwframe+0x44/0xa9
[ 2125.018103] RIP: 0033:0x7f0e11ee0081
[ 2125.020150] Code: fe ff ff 48 8d 3d 67 9c 0a 00 48 83 ec 08 e8 a6 4c 02 00 66 0f 1f 44 00 00 48 8d 05 81 08 2e 00 8b 00 85 c0 75
13 31 c0 0f 05 <48> 3d 00 f0 ff ff 77 57 f3 c3 0f 1f 44 00 00 41 54 55 49 89 d4 53
[ 2125.027032] RSP: 002b:00007ffceb55a5a8 EFLAGS: 00000246 ORIG_RAX: 0000000000000000
[ 2125.029474] RAX: ffffffffffffffda RBX: 0000000000020000 RCX: 00007f0e11ee0081
[ 2125.032055] RDX: 0000000000020000 RSI: 00007f0e123b1000 RDI: 0000000000000003
[ 2125.034592] RBP: 0000000000020000 R08: 00000000ffffffff R09: 0000000000000000
[ 2125.037051] R10: 0000000000000022 R11: 0000000000000246 R12: 00007f0e123b1000
[ 2125.039547] R13: 0000000000000003 R14: 00007f0e123b100f R15: 0000000000020000
[ 2125.041867] Modules linked in: vboxsf(OE) vboxvideo(OE) vmwgfx drm_kms_helper syscopyarea sysfillrect snd_intel8x0 sysimgblt snd_
```



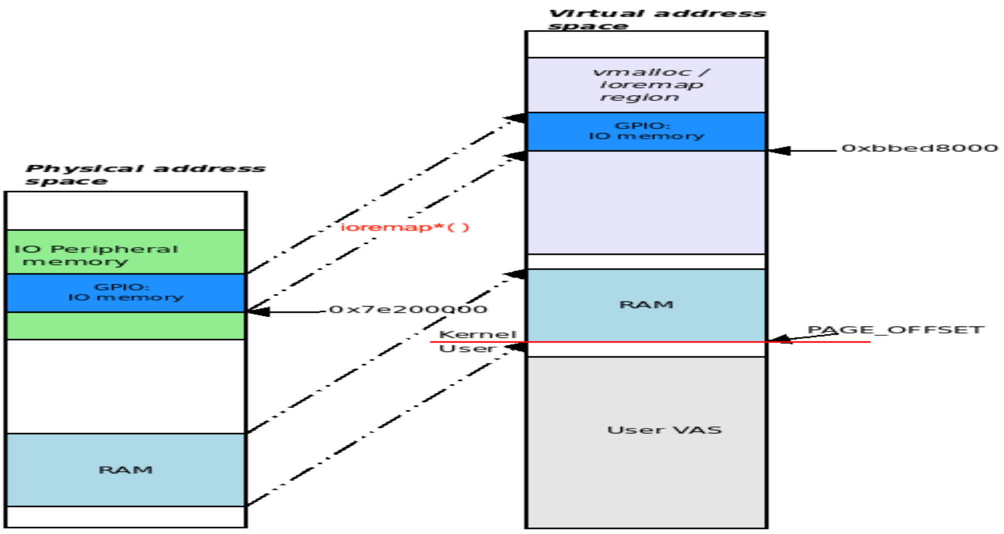
```
$ lsmod |grep netlink_simple_intf
netlink_simple_intf    16384  0
$
$ ../userapp_netlink/netlink_userapp
../userapp_netlink/netlink_userapp:PID 7813: netlink socket created
../userapp_netlink/netlink_userapp: bind done
../userapp_netlink/netlink_userapp: destination struct, netlink hdr, payload setup
../userapp_netlink/netlink_userapp: initialized iov structure (nl header folded in)
../userapp_netlink/netlink_userapp: initialized msghdr structure (iov folded in)
../userapp_netlink/netlink_userapp:sendmsg(): *** success, sent 1040 bytes all-inclusive
 (see kernel log for dtl)
../userapp_netlink/netlink_userapp: now blocking on kernel netlink msg via recvmsg() ...
../userapp_netlink/netlink_userapp:recvmsg(): *** success, received 44 bytes:
msg from kernel netlink: "Reply from kernel netlink"
$
$ dmesg
[62818.385716] netlink_simple_intf: creating kernel netlink socket
[62818.389860] netlink_simple_intf: inserted
[62838.889120] netlink_recv_and_reply(): [000] netlink_userapp :7813   | ...0
[62838.900928] netlink_simple_intf: received from PID 7813:
               "sample user data to send to kernel via netlink"
[62838.922712] netlink_simple_intf: reply sent
$ █
```

Chapter 3: Working with Hardware I/O Memory

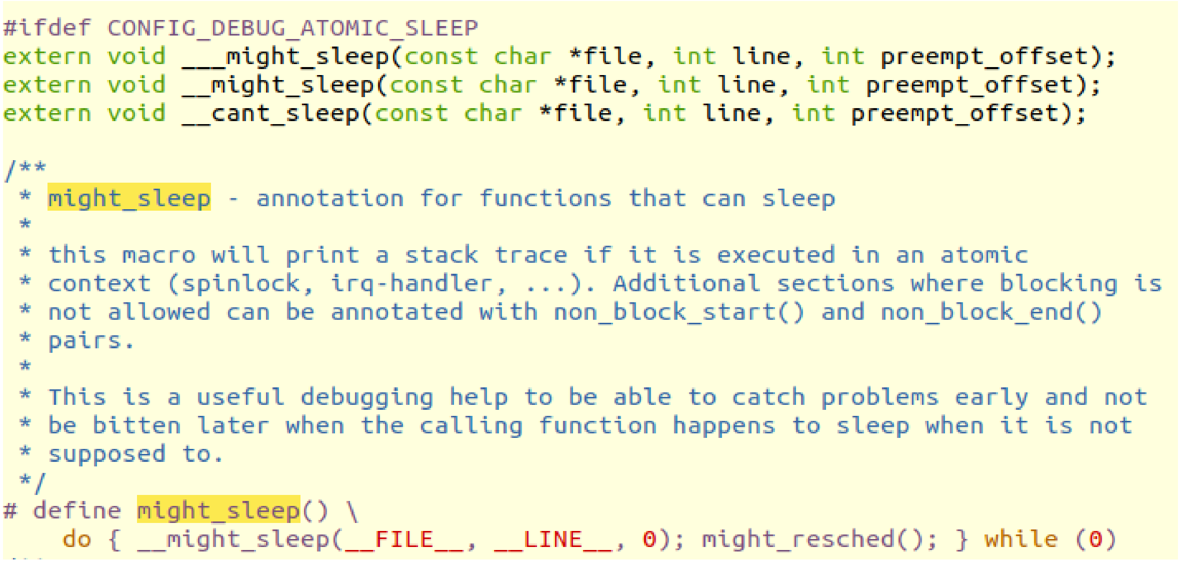
6.1 Register View

The GPIO has 41 registers. All accesses are assumed to be 32-bit.

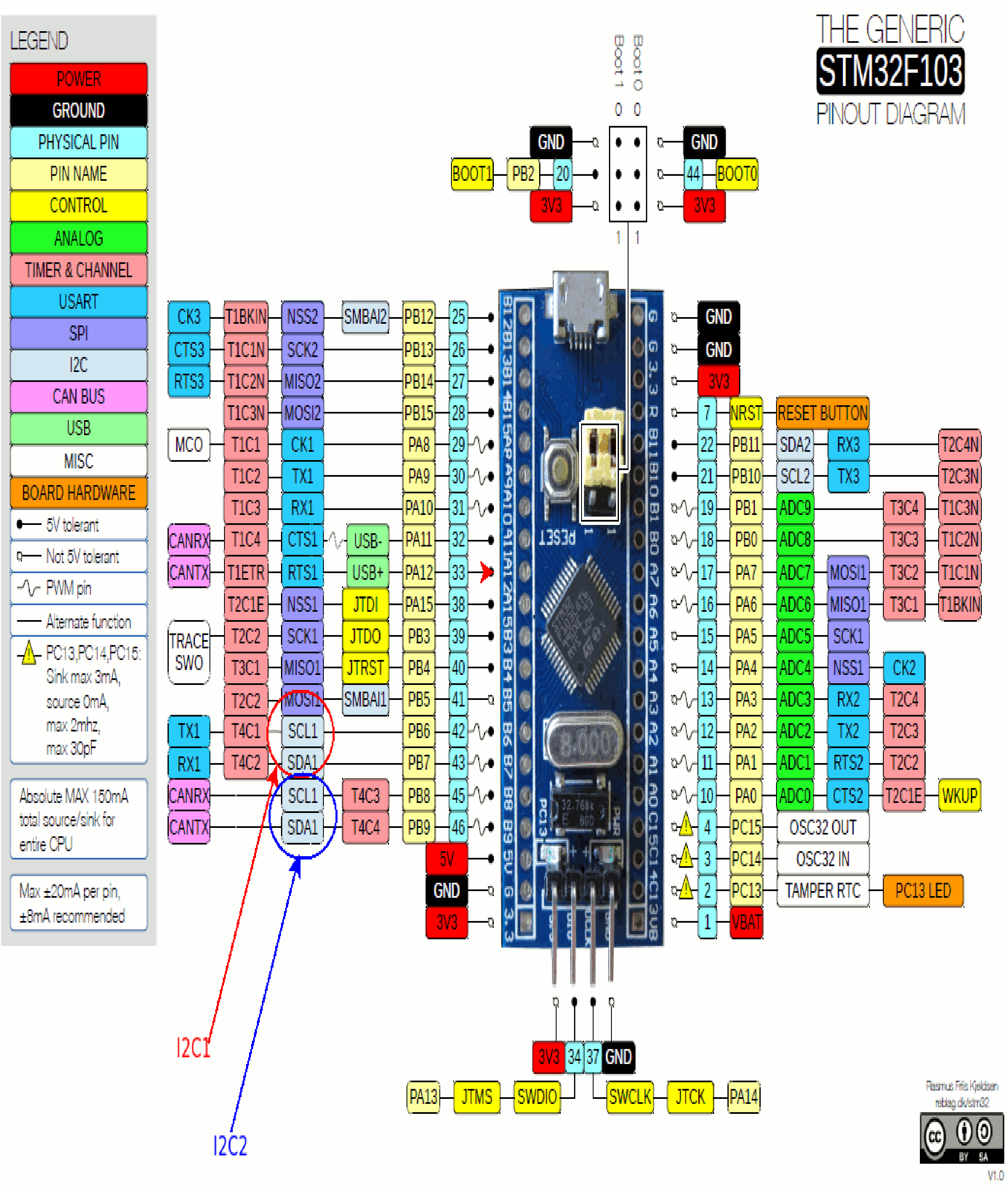
Address	Field Name	Description	Size	Read/ Write
0x 7E20 0000	GPFSEL0	GPIO Function Select 0	32	R/W
0x 7E20 0000	GPFSEL0	GPIO Function Select 0	32	R/W
0x 7E20 0004	GPFSEL1	GPIO Function Select 1	32	R/W
0x 7E20 0008	GPFSEL2	GPIO Function Select 2	32	R/W
0x 7E20 000C	GPFSEL3	GPIO Function Select 3	32	R/W
0x 7E20 0010	GPFSEL4	GPIO Function Select 4	32	R/W
0x 7E20 0014	GPFSEL5	GPIO Function Select 5	32	R/W
0x 7E20 0018	-	Reserved	-	-
0x 7E20 001C	GPSET0	GPIO Pin Output Set 0	32	W
0x 7E20 0020	GPSET1	GPIO Pin Output Set 1	32	W

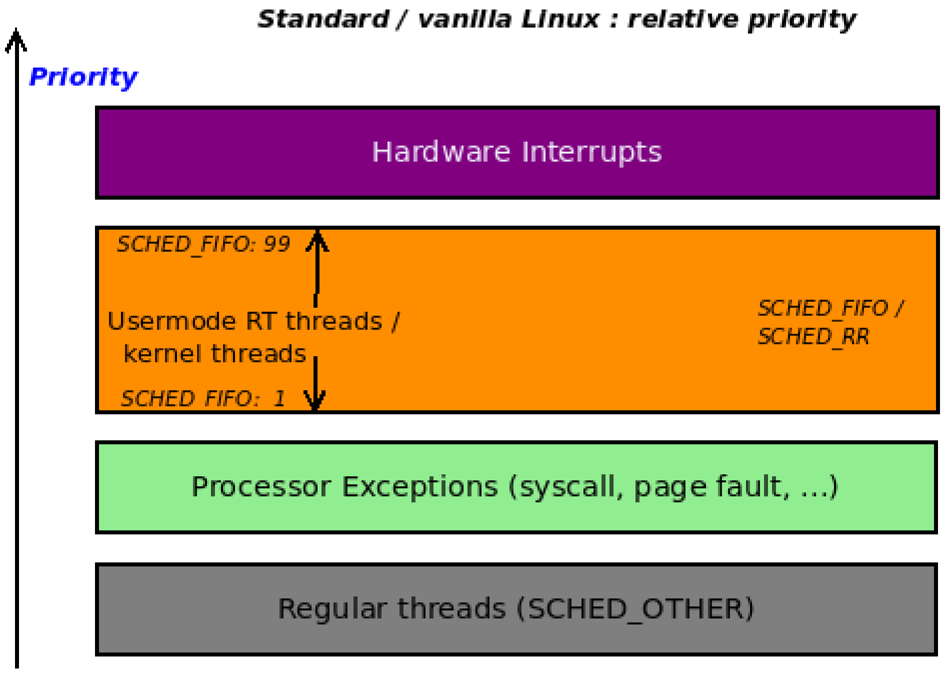


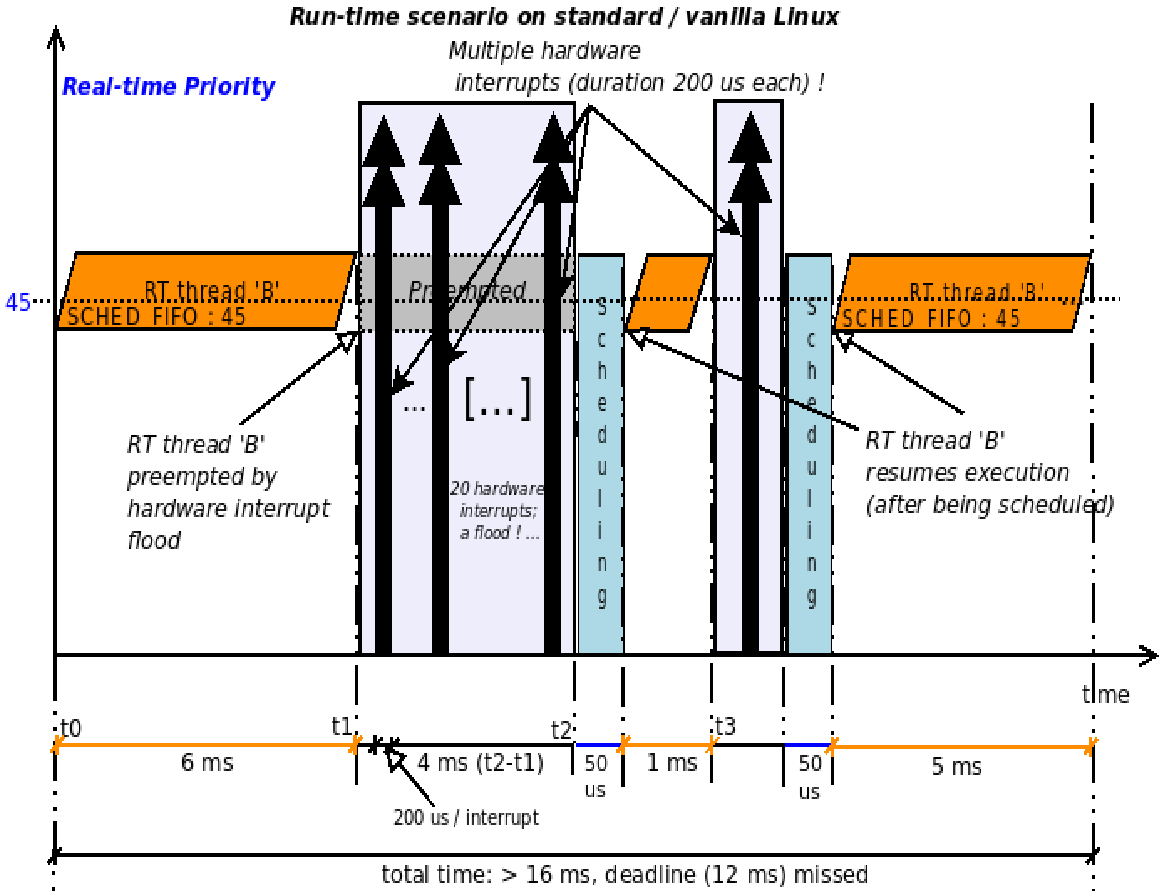
Chapter 4: Handling Hardware Interrupts



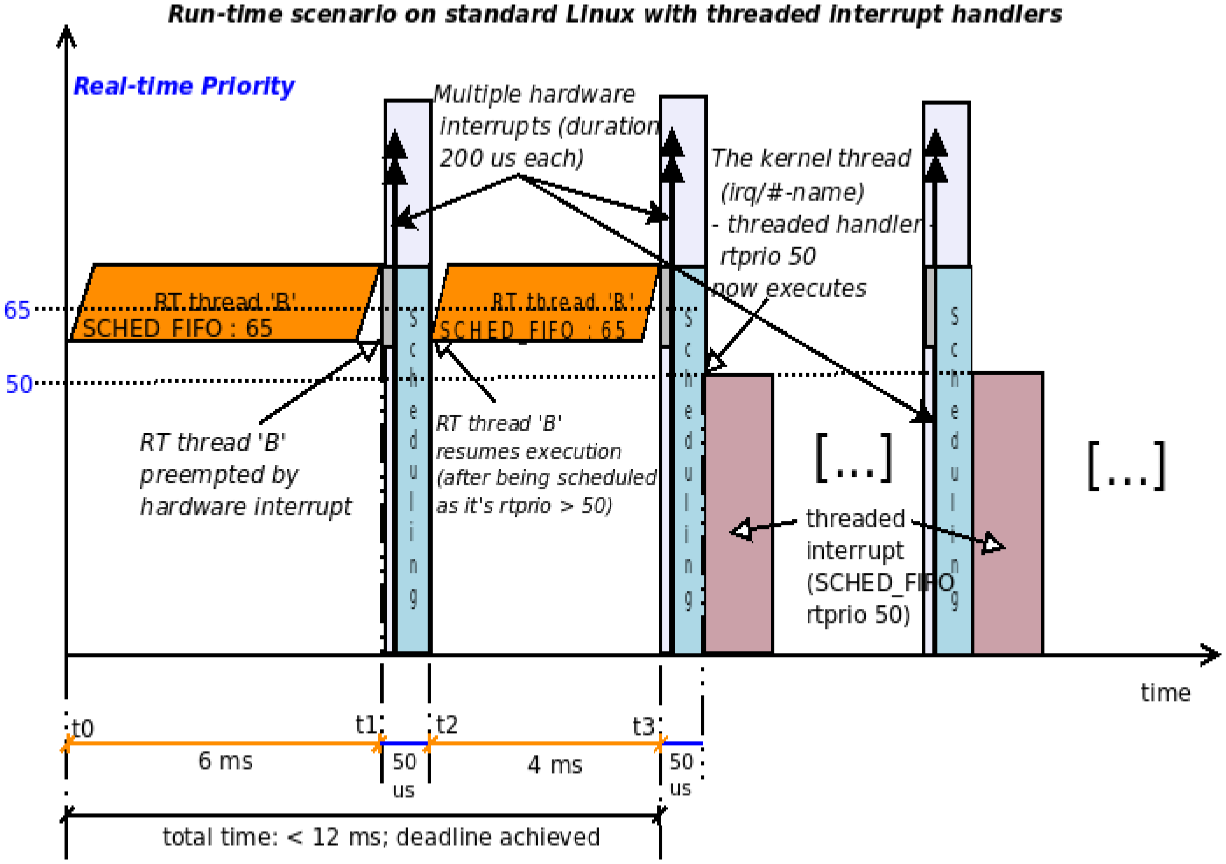
```
#ifdef CONFIG_DEBUG_ATOMIC_SLEEP
extern void ___might_sleep(const char *file, int line, int preempt_offset);
extern void __might_sleep(const char *file, int line, int preempt_offset);
extern void __cant_sleep(const char *file, int line, int preempt_offset);

/**
 * might_sleep - annotation for functions that can sleep
 *
 * this macro will print a stack trace if it is executed in an atomic
 * context (spinlock, irq-handler, ...). Additional sections where blocking is
 * not allowed can be annotated with non_block_start() and non_block_end()
 * pairs.
 *
 * This is a useful debugging help to be able to catch problems early and not
 * be bitten later when the calling function happens to sleep when it is not
 * supposed to.
 */
# define might_sleep() \
	do { __might_sleep(__FILE__, __LINE__, 0); might_resched(); } while (0)
```

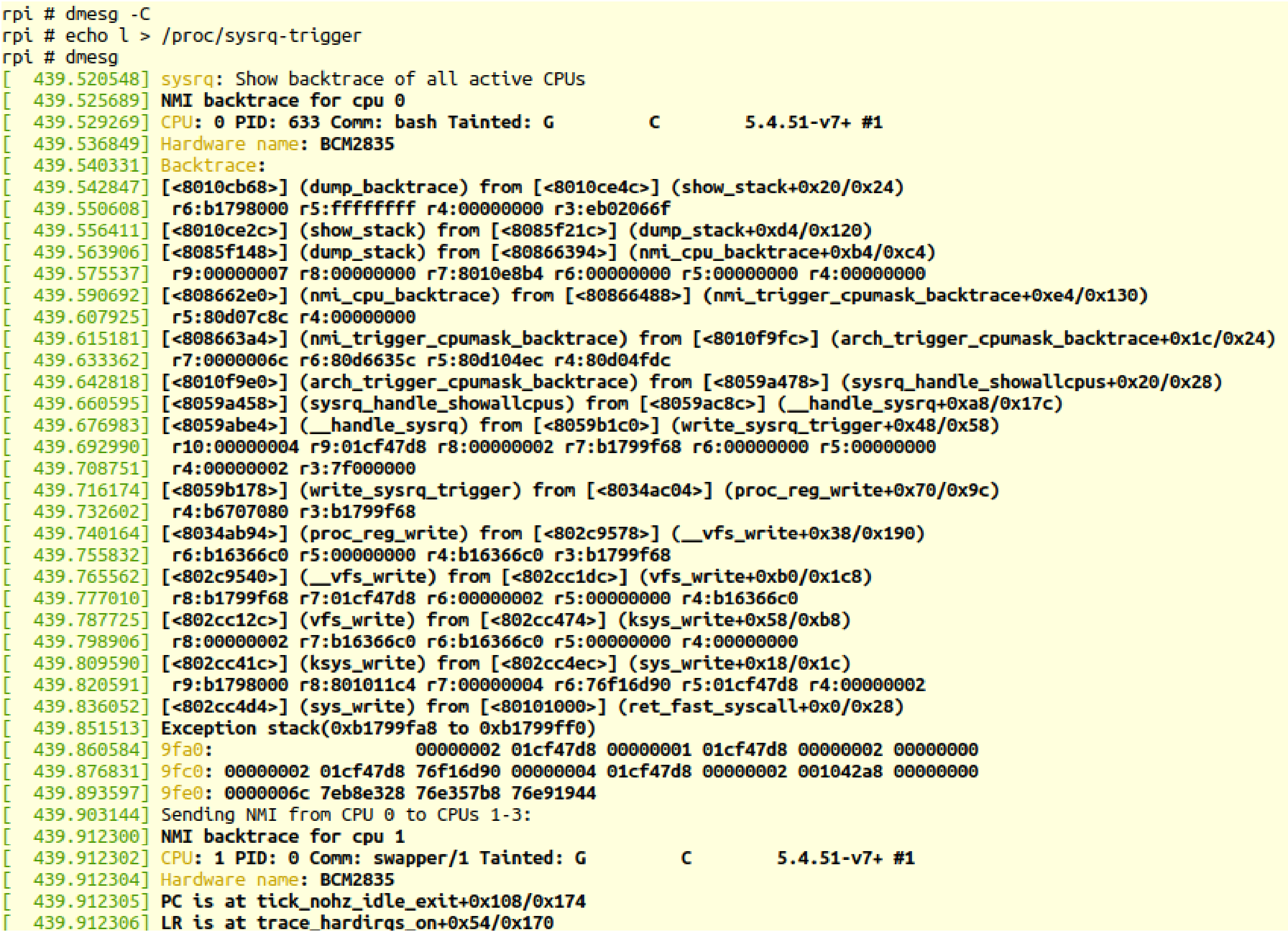

THE GENERIC
STM32F103
PINOUT DIAGRAM
LEGEND
POWER
GROUND
PHYSICAL PIN
PIN NAME
CONTROL
ANALOG
TIMER & CHANNEL
USART
SPI
I2C
CAN BUS
USB
MISC
BOARD HARDWARE
5V tolerant
Not 5V tolerant
PWM pin
Alternate function
PC13,PC14,PC15: Sink max 3mA, source 0mA, max 2mhz, max 30pF
Absolute MAX 150mA total source/sink for entire CPU
Max ±20mA per pin, ±8mA recommended
Boot 1
Boot 0
0 0
1 1
GND
GND
BOOT1 PB2 20
44 BOOT0
3V3
3V3
CK3 T1BKIN NSS2 SMBAI2 PB12 25
CTS3 T1C1N SCK2 PB13 26
RTS3 T1C2N MISO2 PB14 27
T1C3N MOSI2 PB15 28
MCO T1C1 CK1 PA8 29
T1C2 TX1 PA9 30
T1C3 RX1 PA10 31
CANRX T1C4 CTS1 USB- PA11 32
CANTX T1ETR RTS1 USB+ PA12 33
T2C1E NSS1 JTDI PA15 38
TRACE SWO T2C2 SCK1 JTDO PB3 39
T3C1 MISO1 JTRST PB4 40
T2C2 MOSI1 SMBAI1 PB5 41
TX1 T4C1 SCL1 PB6 42
RX1 T4C2 SDA1 PB7 43
CANRX SCL1 T4C3 PB8 45
CANTX SDA1 T4C4 PB9 46
5V
GND
3V3
GND
GND
3V3
7 NRST RESET BUTTON
22 PB11 SDA2 RX3 T2C4N
21 PB10 SCL2 TX3 T2C3N
19 PB1 ADC9 T3C4 T1C3N
18 PB0 ADC8 T3C3 T1C2N
17 PA7 ADC7 MOSI1 T3C2 T1C1N
16 PA6 ADC6 MISO1 T3C1 T1BKIN
15 PA5 ADC5 SCK1
14 PA4 ADC4 NSS1 CK2
13 PA3 ADC3 RX2 T2C4
12 PA2 ADC2 TX2 T2C3
11 PA1 ADC1 RTS2 T2C2
10 PA0 ADC0 CTS2 T2C1E WKUP
4 PC15 OSC32 OUT
3 PC14 OSC32 IN
2 PC13 TAMPER RTC PC13 LED
1 VBAT
3V3 34 37 GND
PA13 JTMS SWDIO SWCLK JTCK PA14
B12 B13 B14 B15 A8 A9 A10 A11 A12 A15 B3 B4 B5 B6 B7 B8 B9 5V G 3.3
G G 3.3 R B11 B10 B1 B0 A7 A6 A5 A4 A3 A2 A1 A0 C15 C14 C13 VB
RESET
8.000
32.768k
I2C1
I2C2
Rasmus Friis Kjeldsen
reblag.dk/stm32
cc BY SA
V1.0


Standard / vanilla Linux : relative priority
Priority
Hardware Interrupts
SCHED_FIFO: 99
Usermode RT threads /
kernel threads
SCHED_FIFO: 1
SCHED_FIFO /
SCHED_RR
Processor Exceptions (syscall, page fault, ...)
Regular threads (SCHED_OTHER)

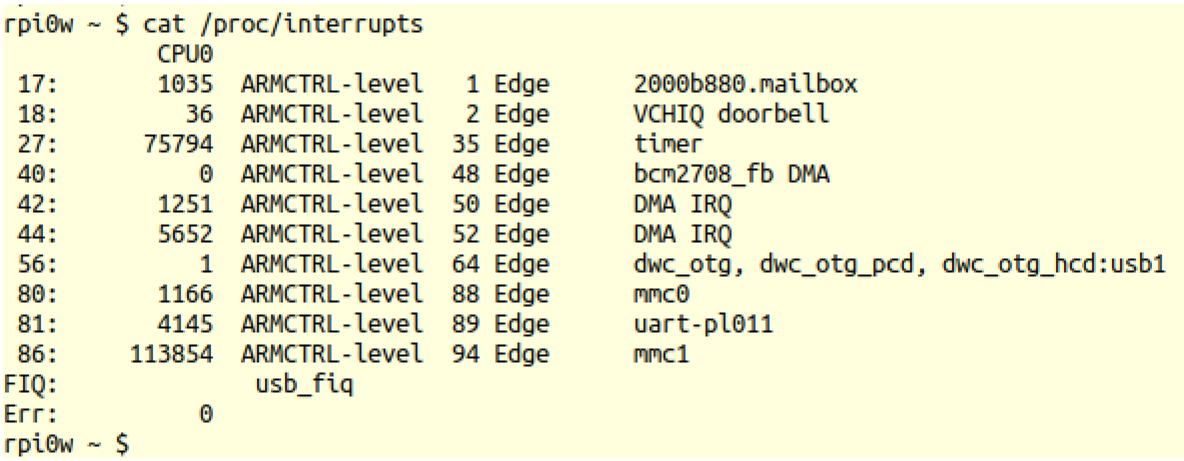

Run-time scenario on standard / vanilla Linux
Multiple hardware
interrupts (duration 200 us each) !
Real-time Priority
45
RT thread 'B'
SCHED_FIFO : 45
Preempted
...
[...]
20 hardware
interrupts;
a flood ! ...
Scheduling
Scheduling
RT thread 'B'
SCHED_FIFO : 45
RT thread 'B'
preempted by
hardware interrupt
flood
RT thread 'B'
resumes execution
(after being scheduled)
time
t0
t1
t2
t3
6 ms
4 ms (t2-t1)
200 us / interrupt
50
us
1 ms
50
us
5 ms
total time: > 16 ms, deadline (12 ms) missed

Run-time scenario on standard Linux with threaded interrupt handlers

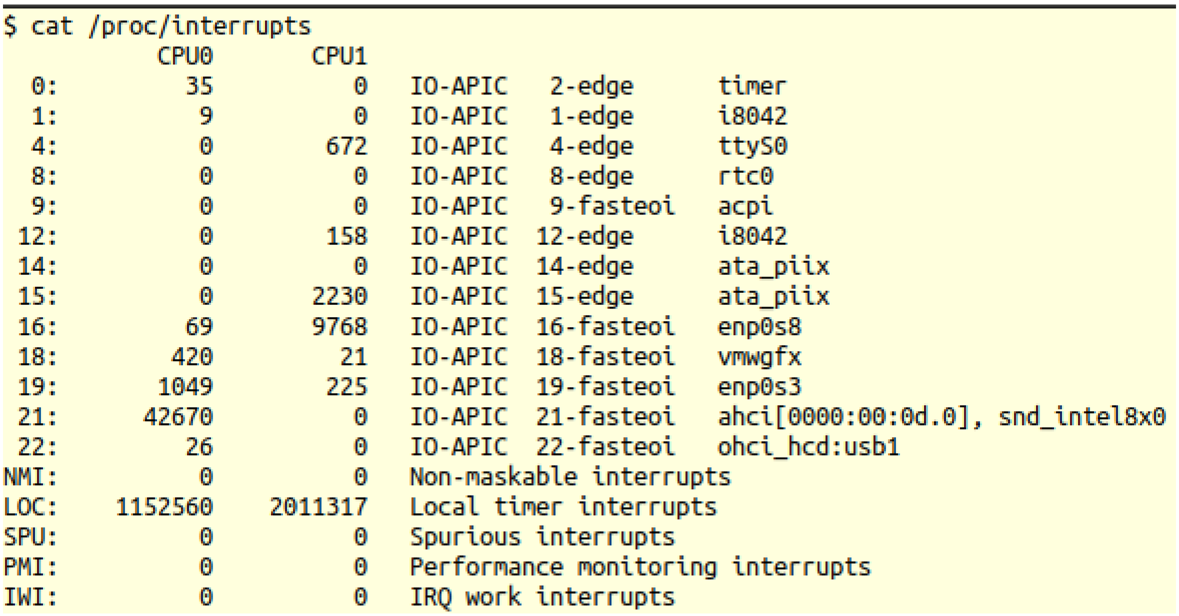

Real-time Priority
65
50
RT thread 'B'
SCHED_FIFO : 65
RT thread 'B'
SCHED_FIFO : 65
Multiple hardware interrupts (duration 200 us each)
The kernel thread (irq/#-name) - threaded handler - rtprio 50 now executes
Scheduling
Scheduling
Scheduling
RT thread 'B' preempted by hardware interrupt
RT thread 'B' resumes execution (after being scheduled as it's rtprio > 50)
[...]
[...]
threaded interrupt (SCHED_FIFO rtprio 50)
time
t0
t1
t2
t3
6 ms
50 us
4 ms
50 us
total time: < 12 ms; deadline achieved



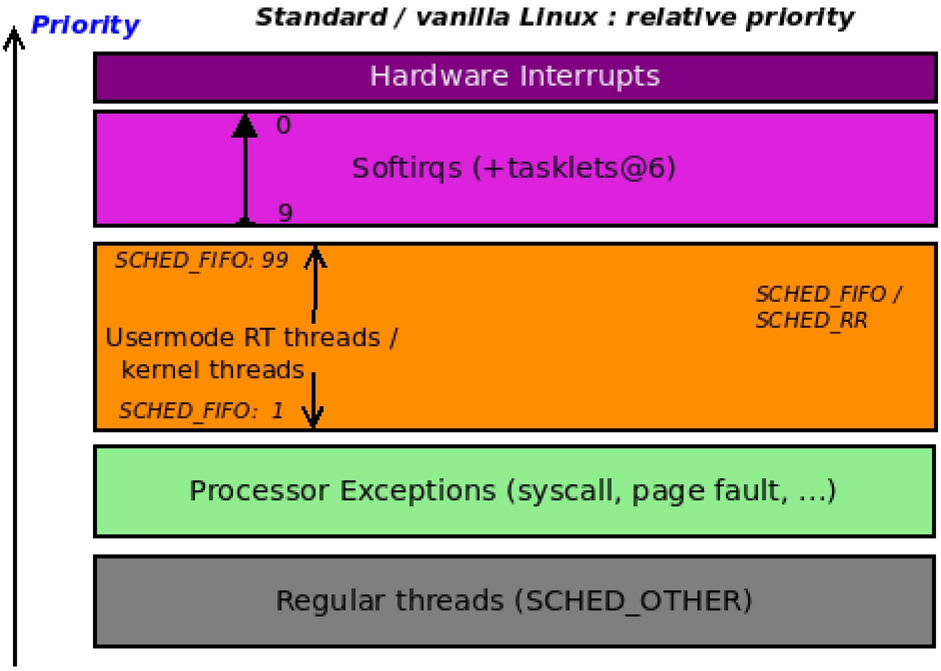
```
rpi # dmesg -C
rpi # echo l > /proc/sysrq-trigger
rpi # dmesg
[  439.520548] sysrq: Show backtrace of all active CPUs
[  439.525689] NMI backtrace for cpu 0
[  439.529269] CPU: 0 PID: 633 Comm: bash Tainted: G         C        5.4.51-v7+ #1
[  439.536849] Hardware name: BCM2835
[  439.540331] Backtrace:
[  439.542847] [<8010cb68>] (dump_backtrace) from [<8010ce4c>] (show_stack+0x20/0x24)
[  439.550608]  r6:b1798000 r5:ffffffff r4:00000000 r3:eb02066f
[  439.556411] [<8010ce2c>] (show_stack) from [<8085f21c>] (dump_stack+0xd4/0x120)
[  439.563906] [<8085f148>] (dump_stack) from [<80866394>] (nmi_cpu_backtrace+0xb4/0xc4)
[  439.575537]  r9:00000007 r8:00000000 r7:8010e8b4 r6:00000000 r5:00000000 r4:00000000
[  439.590692] [<808662e0>] (nmi_cpu_backtrace) from [<80866488>] (nmi_trigger_cpumask_backtrace+0xe4/0x130)
[  439.607925]  r5:80d07c8c r4:00000000
[  439.615181] [<808663a4>] (nmi_trigger_cpumask_backtrace) from [<8010f9fc>] (arch_trigger_cpumask_backtrace+0x1c/0x24)
[  439.633362]  r7:0000006c r6:80d6635c r5:80d104ec r4:80d04fdc
[  439.642818] [<8010f9e0>] (arch_trigger_cpumask_backtrace) from [<8059a478>] (sysrq_handle_showallcpus+0x20/0x28)
[  439.660595] [<8059a458>] (sysrq_handle_showallcpus) from [<8059ac8c>] (__handle_sysrq+0xa8/0x17c)
[  439.676983] [<8059abe4>] (__handle_sysrq) from [<8059b1c0>] (write_sysrq_trigger+0x48/0x58)
[  439.692990]  r10:00000004 r9:01cf47d8 r8:00000002 r7:b1799f68 r6:00000000 r5:00000000
[  439.708751]  r4:00000002 r3:7f000000
[  439.716174] [<8059b178>] (write_sysrq_trigger) from [<8034ac04>] (proc_reg_write+0x70/0x9c)
[  439.732602]  r4:b6707080 r3:b1799f68
[  439.740164] [<8034ab94>] (proc_reg_write) from [<802c9578>] (__vfs_write+0x38/0x190)
[  439.755832]  r6:b16366c0 r5:00000000 r4:b16366c0 r3:b1799f68
[  439.765562] [<802c9540>] (__vfs_write) from [<802cc1dc>] (vfs_write+0xb0/0x1c8)
[  439.777010]  r8:b1799f68 r7:01cf47d8 r6:00000002 r5:00000000 r4:b16366c0
[  439.787725] [<802cc12c>] (vfs_write) from [<802cc474>] (ksys_write+0x58/0xb8)
[  439.798906]  r8:00000002 r7:b16366c0 r6:b16366c0 r5:00000000 r4:00000000
[  439.809590] [<802cc41c>] (ksys_write) from [<802cc4ec>] (sys_write+0x18/0x1c)
[  439.820591]  r9:b1798000 r8:801011c4 r7:00000004 r6:76f16d90 r5:01cf47d8 r4:00000002
[  439.836052] [<802cc4d4>] (sys_write) from [<80101000>] (ret_fast_syscall+0x0/0x28)
[  439.851513] Exception stack(0xb1799fa8 to 0xb1799ff0)
[  439.860584] 9fa0:                   00000002 01cf47d8 00000001 01cf47d8 00000002 00000000
[  439.876831] 9fc0: 00000002 01cf47d8 76f16d90 00000004 01cf47d8 00000002 001042a8 00000000
[  439.893597] 9fe0: 0000006c 7eb8e328 76e357b8 76e91944
[  439.903144] Sending NMI from CPU 0 to CPUs 1-3:
[  439.912300] NMI backtrace for cpu 1
[  439.912302] CPU: 1 PID: 0 Comm: swapper/1 Tainted: G         C        5.4.51-v7+ #1
[  439.912304] Hardware name: BCM2835
[  439.912305] PC is at tick_nohz_idle_exit+0x108/0x174
[  439.912306] LR is at trace_hardirqs_on+0x54/0x170
```

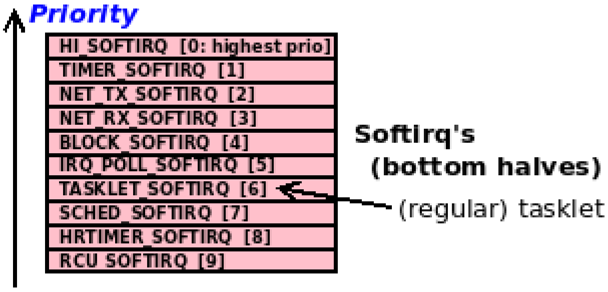


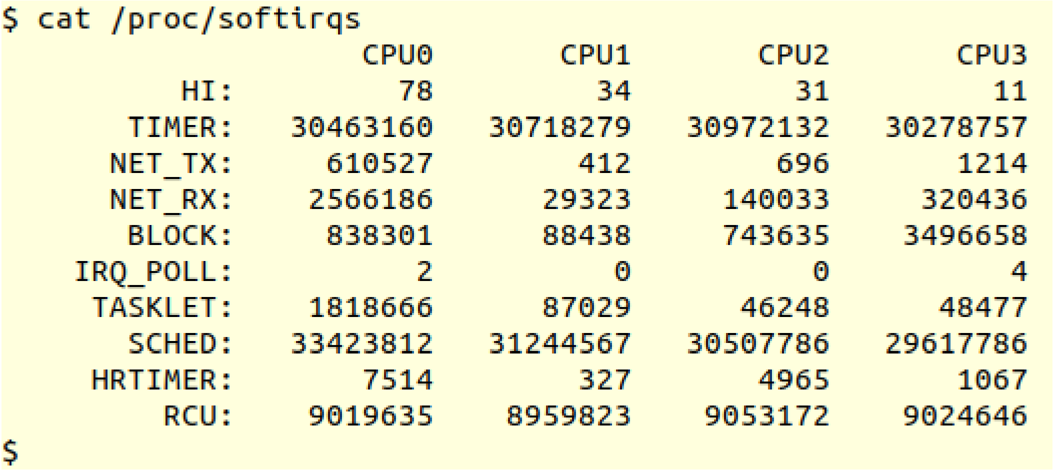
```
rpi0w ~ $ cat /proc/interrupts
           CPU0
 17:       1035  ARMCTRL-level   1 Edge      2000b880.mailbox
 18:         36  ARMCTRL-level   2 Edge      VCHIQ doorbell
 27:      75794  ARMCTRL-level  35 Edge      timer
 40:          0  ARMCTRL-level  48 Edge      bcm2708_fb DMA
 42:       1251  ARMCTRL-level  50 Edge      DMA IRQ
 44:       5652  ARMCTRL-level  52 Edge      DMA IRQ
 56:          1  ARMCTRL-level  64 Edge      dwc_otg, dwc_otg_pcd, dwc_otg_hcd:usb1
 80:       1166  ARMCTRL-level  88 Edge      mmc0
 81:       4145  ARMCTRL-level  89 Edge      uart-pl011
 86:     113854  ARMCTRL-level  94 Edge      mmc1
FIQ:              usb_fiq
Err:          0
rpi0w ~ $
```



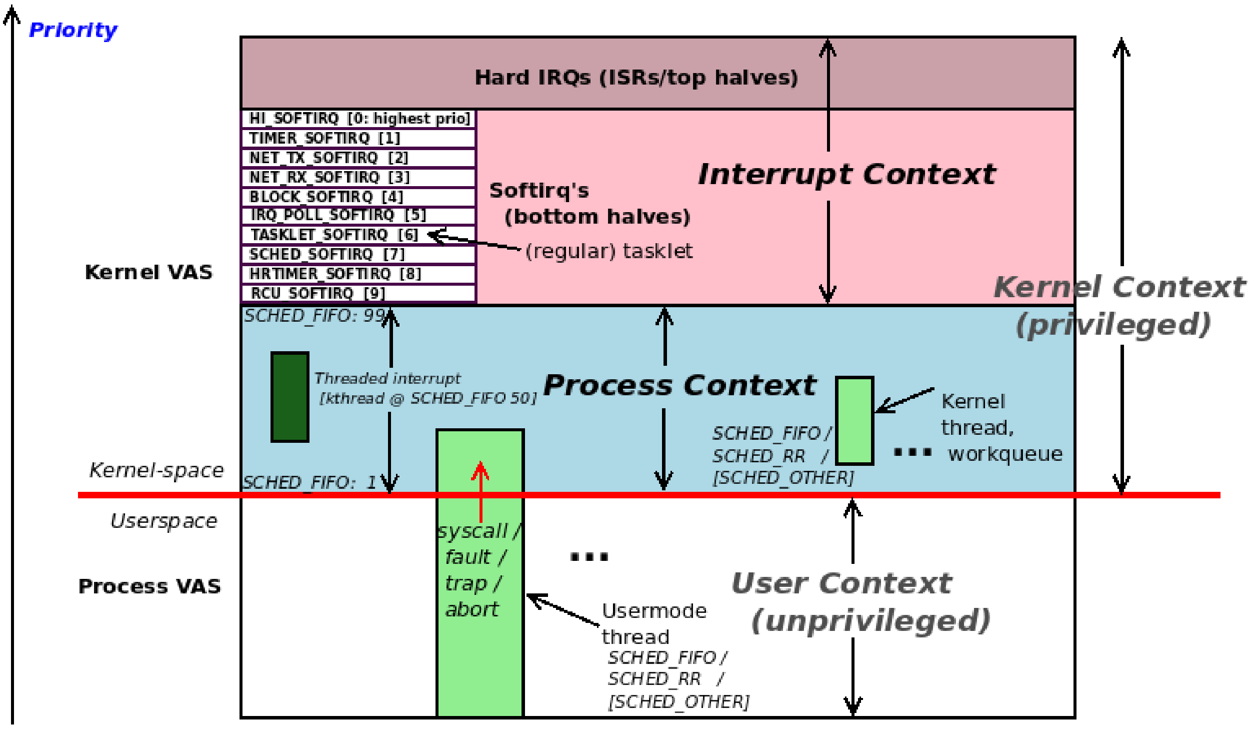
```
$ cat /proc/interrupts
           CPU0       CPU1
  0:         35          0   IO-APIC   2-edge      timer
  1:          9          0   IO-APIC   1-edge      i8042
  4:          0        672   IO-APIC   4-edge      ttyS0
  8:          0          0   IO-APIC   8-edge      rtc0
  9:          0          0   IO-APIC   9-fasteoi   acpi
 12:          0        158   IO-APIC  12-edge      i8042
 14:          0          0   IO-APIC  14-edge      ata_piix
 15:          0       2230   IO-APIC  15-edge      ata_piix
 16:         69       9768   IO-APIC  16-fasteoi   enp0s8
 18:        420         21   IO-APIC  18-fasteoi   vmwgfx
 19:       1049        225   IO-APIC  19-fasteoi   enp0s3
 21:      42670          0   IO-APIC  21-fasteoi   ahci[0000:00:0d.0], snd_intel8x0
 22:         26          0   IO-APIC  22-fasteoi   ohci_hcd:usb1
NMI:          0          0   Non-maskable interrupts
LOC:    1152560    2011317   Local timer interrupts
SPU:          0          0   Spurious interrupts
PMI:          0          0   Performance monitoring interrupts
IWI:          0          0   IRQ work interrupts
```

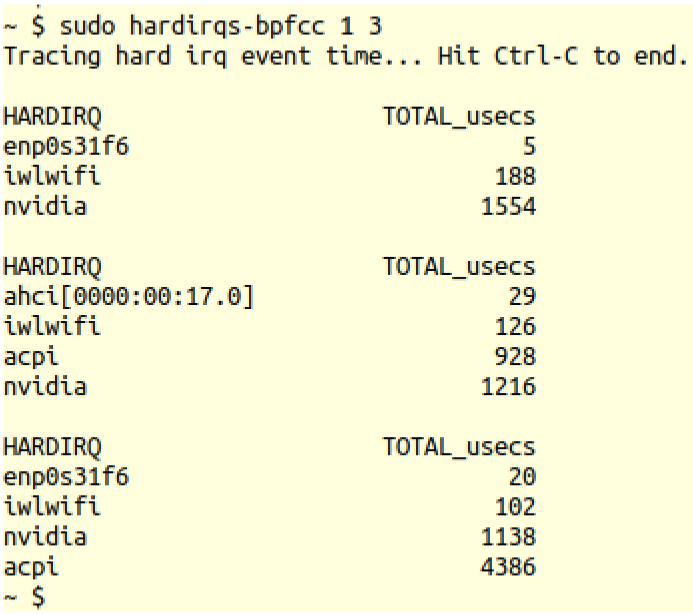




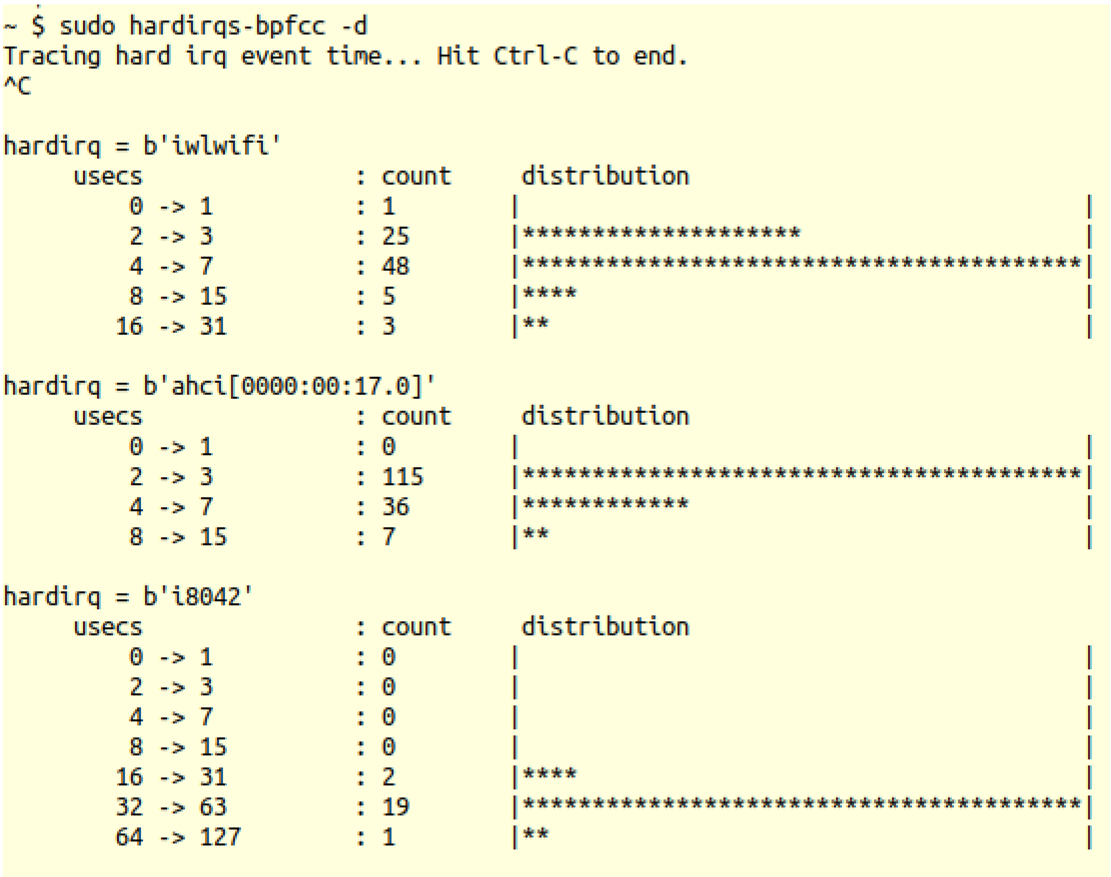


```
$ cat /proc/softirqs
                    CPU0       CPU1       CPU2       CPU3
          HI:         78         34         31         11
       TIMER:   30463160   30718279   30972132   30278757
      NET_TX:     610527        412        696       1214
      NET_RX:    2566186      29323     140033     320436
       BLOCK:     838301      88438     743635    3496658
    IRQ_POLL:          2          0          0          4
     TASKLET:    1818666      87029      46248      48477
       SCHED:   33423812   31244567   30507786   29617786
     HRTIMER:       7514        327       4965       1067
         RCU:    9019635    8959823    9053172    9024646
$
```

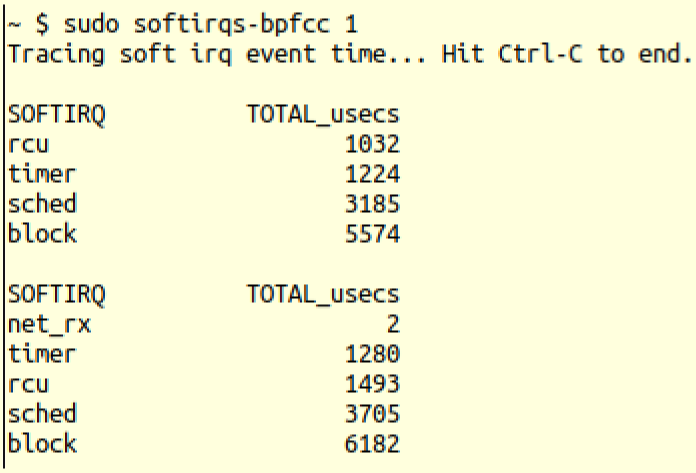




```
~ $ sudo hardirqs-bpfcc 1 3
Tracing hard irq event time... Hit Ctrl-C to end.

HARDIRQ                    TOTAL_usecs
enp0s31f6                            5
iwlwifi                            188
nvidia                            1554

HARDIRQ                    TOTAL_usecs
ahci[0000:00:17.0]                  29
iwlwifi                            126
acpi                               928
nvidia                            1216

HARDIRQ                    TOTAL_usecs
enp0s31f6                           20
iwlwifi                            102
nvidia                            1138
acpi                              4386
~ $
```

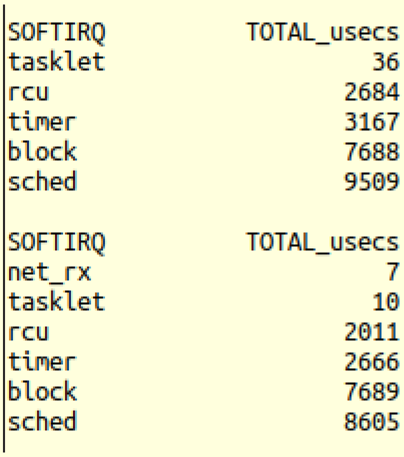


```
~ $ sudo hardirqs-bpfcc -d
Tracing hard irq event time... Hit Ctrl-C to end.
^C

hardirq = b'iwlwifi'
     usecs               : count     distribution
         0 -> 1          : 1        |                                        |
         2 -> 3          : 25       |********************                    |
         4 -> 7          : 48       |****************************************|
         8 -> 15         : 5        |****                                    |
        16 -> 31         : 3        |**                                      |

hardirq = b'ahci[0000:00:17.0]'
     usecs               : count     distribution
         0 -> 1          : 0        |                                        |
         2 -> 3          : 115      |****************************************|
         4 -> 7          : 36       |************                            |
         8 -> 15         : 7        |**                                      |

hardirq = b'i8042'
     usecs               : count     distribution
         0 -> 1          : 0        |                                        |
         2 -> 3          : 0        |                                        |
         4 -> 7          : 0        |                                        |
         8 -> 15         : 0        |                                        |
        16 -> 31         : 2        |****                                    |
        32 -> 63         : 19       |****************************************|
        64 -> 127        : 1        |**                                      |
```

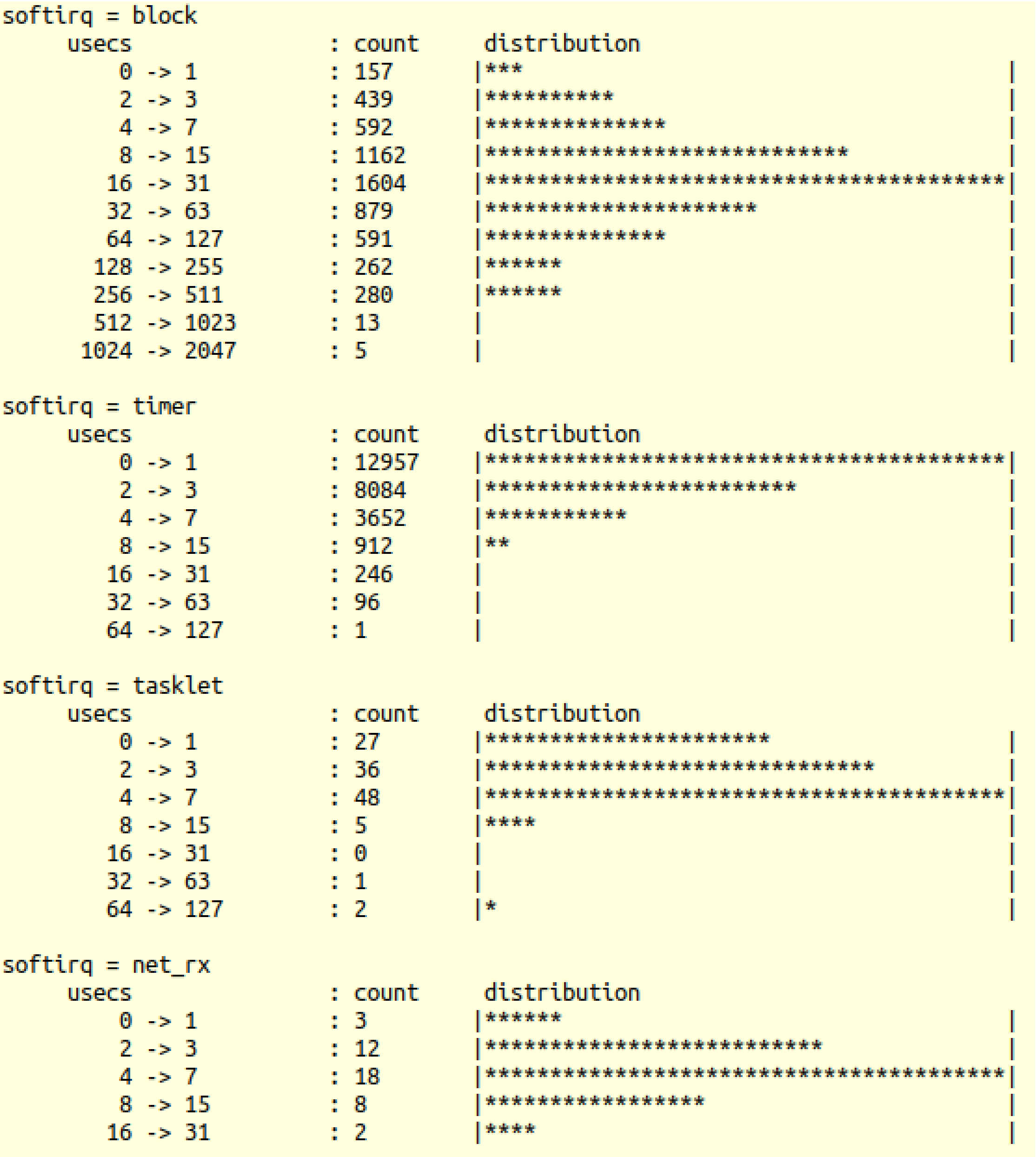


```
~ $ sudo softirqs-bpfcc 1
Tracing soft irq event time... Hit Ctrl-C to end.

SOFTIRQ          TOTAL_usecs
rcu                     1032
timer                   1224
sched                   3185
block                   5574

SOFTIRQ          TOTAL_usecs
net_rx                     2
timer                   1280
rcu                     1493
sched                   3705
block                   6182
```

[...]

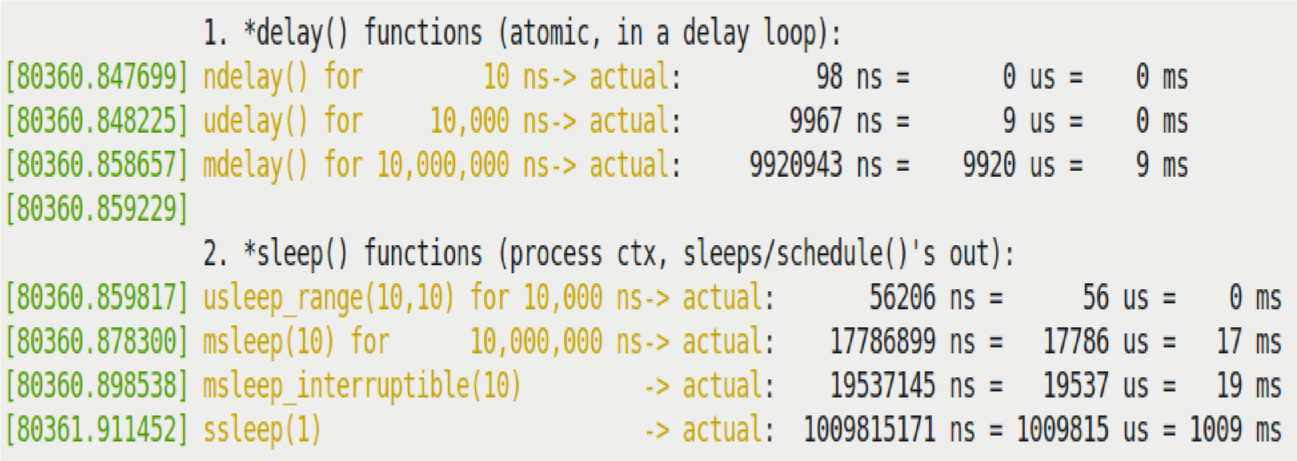


```
SOFTIRQ          TOTAL_usecs
tasklet                   36
rcu                     2684
timer                   3167
block                   7688
sched                   9509

SOFTIRQ          TOTAL_usecs
net_rx                     7
tasklet                   10
rcu                     2011
timer                   2666
block                   7689
sched                   8605
```

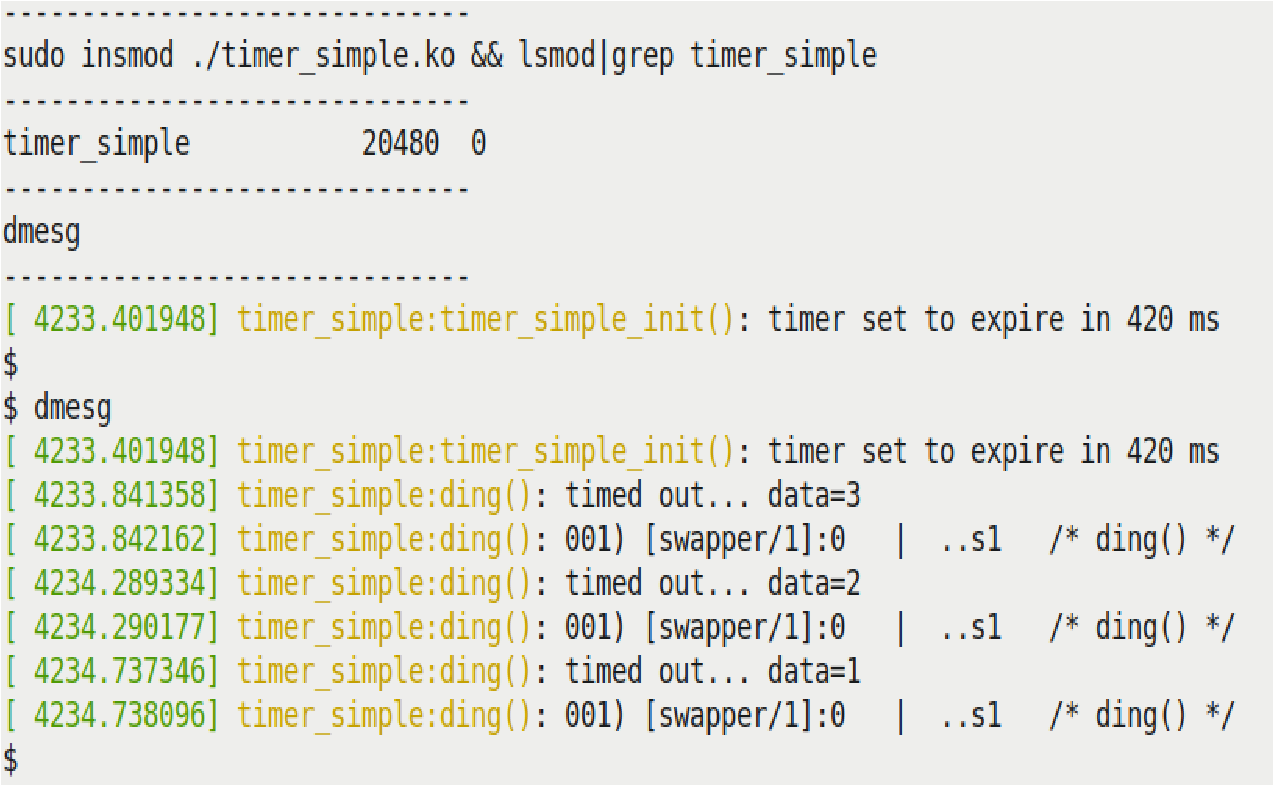


```
softirq = block
     usecs               : count     distribution
         0 -> 1          : 157      |***                                     |
         2 -> 3          : 439      |**********                              |
         4 -> 7          : 592      |**************                          |
         8 -> 15         : 1162     |****************************            |
        16 -> 31         : 1604     |****************************************|
        32 -> 63         : 879      |*********************                   |
        64 -> 127        : 591      |**************                          |
       128 -> 255        : 262      |******                                  |
       256 -> 511        : 280      |******                                  |
       512 -> 1023       : 13       |                                        |
      1024 -> 2047       : 5        |                                        |

softirq = timer
     usecs               : count     distribution
         0 -> 1          : 12957    |****************************************|
         2 -> 3          : 8084     |************************                |
         4 -> 7          : 3652     |***********                             |
         8 -> 15         : 912      |**                                      |
        16 -> 31         : 246      |                                        |
        32 -> 63         : 96       |                                        |
        64 -> 127        : 1        |                                        |

softirq = tasklet
     usecs               : count     distribution
         0 -> 1          : 27       |**********************                  |
         2 -> 3          : 36       |******************************          |
         4 -> 7          : 48       |****************************************|
         8 -> 15         : 5        |****                                    |
        16 -> 31         : 0        |                                        |
        32 -> 63         : 1        |                                        |
        64 -> 127        : 2        |*                                       |

softirq = net_rx
     usecs               : count     distribution
         0 -> 1          : 3        |******                                  |
         2 -> 3          : 12       |**************************              |
         4 -> 7          : 18       |****************************************|
         8 -> 15         : 8        |*****************                       |
        16 -> 31         : 2        |****                                    |
```

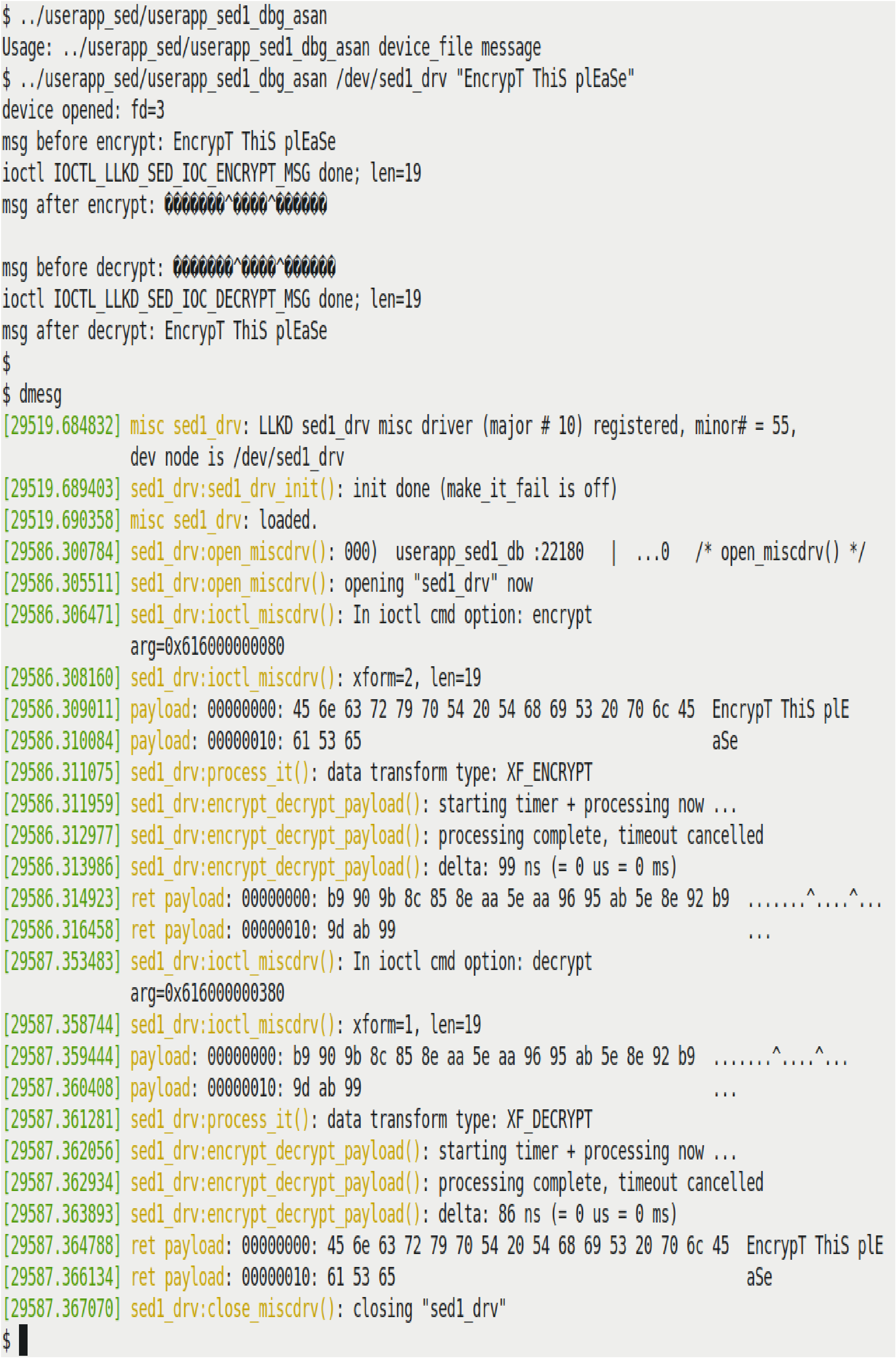
Chapter 5: Working with Kernel Timers, Threads, and Workqueues



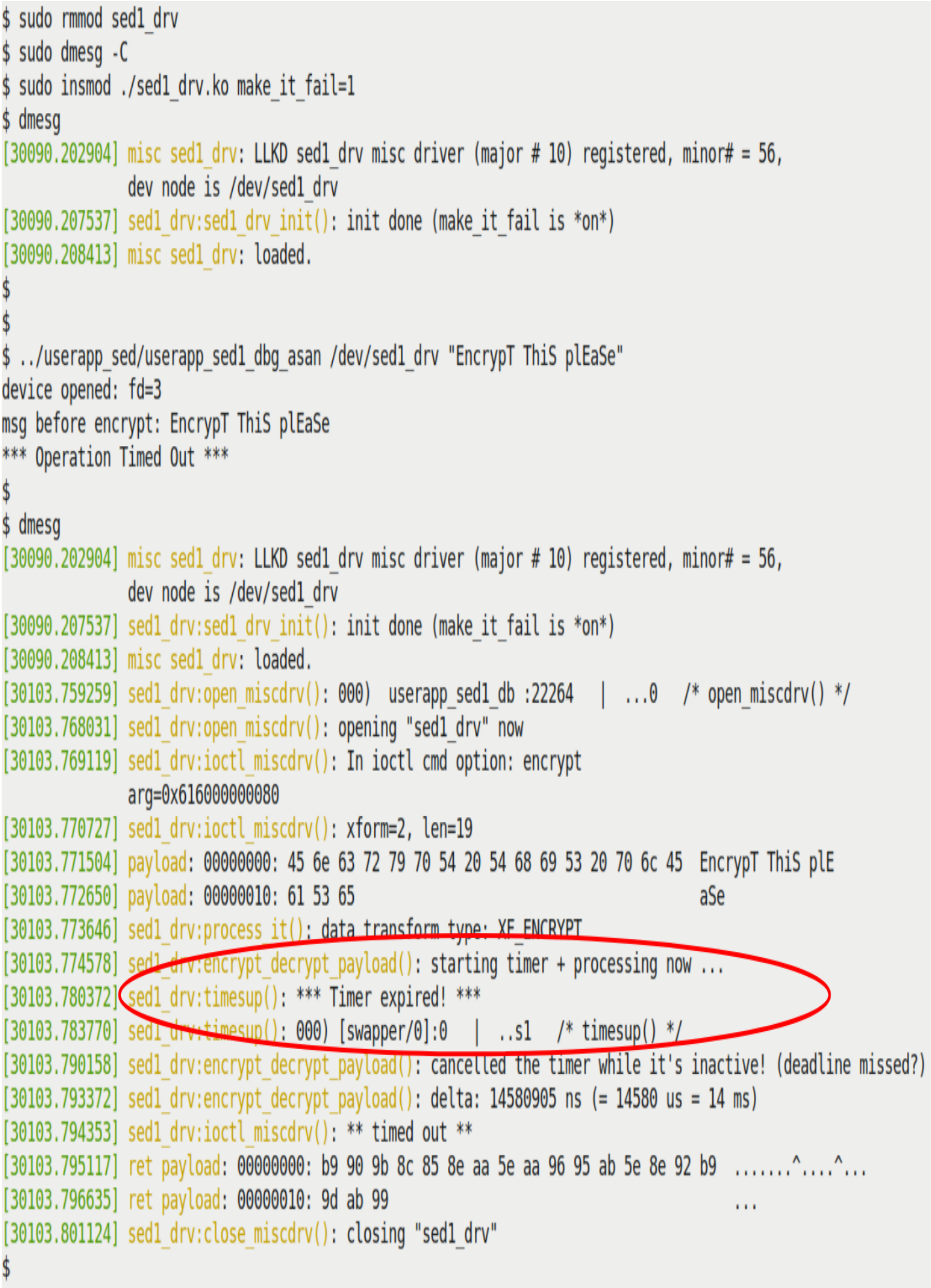
```
              1. *delay() functions (atomic, in a delay loop):
[80360.847699] ndelay() for         10 ns-> actual:         98 ns =       0 us =    0 ms
[80360.848225] udelay() for     10,000 ns-> actual:       9967 ns =       9 us =    0 ms
[80360.858657] mdelay() for 10,000,000 ns-> actual:    9920943 ns =    9920 us =    9 ms
[80360.859229]
              2. *sleep() functions (process ctx, sleeps/schedule()'s out):
[80360.859817] usleep_range(10,10) for 10,000 ns-> actual:      56206 ns =      56 us =    0 ms
[80360.878300] msleep(10) for      10,000,000 ns-> actual:   17786899 ns =   17786 us =   17 ms
[80360.898538] msleep_interruptible(10)        -> actual:   19537145 ns =   19537 us =   19 ms
[80361.911452] ssleep(1)                       -> actual: 1009815171 ns = 1009815 us = 1009 ms
```



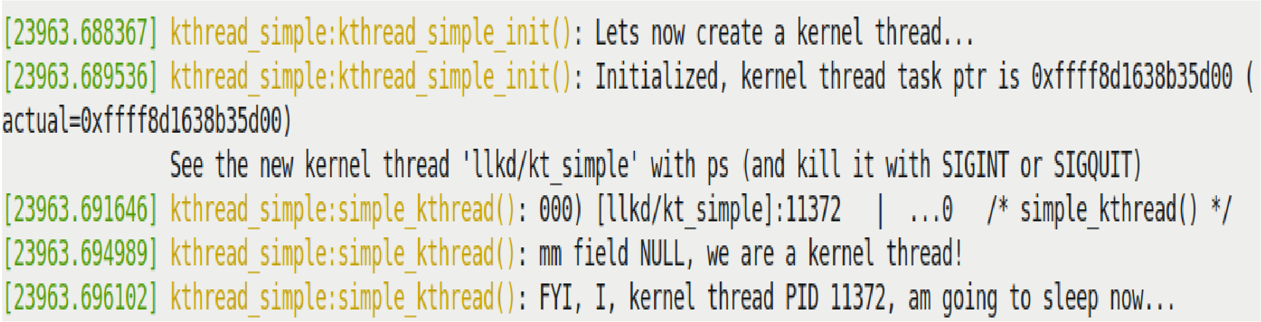
```
------------------------------
sudo insmod ./timer_simple.ko && lsmod|grep timer_simple
------------------------------
timer_simple           20480  0
------------------------------
dmesg
------------------------------
[ 4233.401948] timer_simple:timer_simple_init(): timer set to expire in 420 ms
$
$ dmesg
[ 4233.401948] timer_simple:timer_simple_init(): timer set to expire in 420 ms
[ 4233.841358] timer_simple:ding(): timed out... data=3
[ 4233.842162] timer_simple:ding(): 001) [swapper/1]:0   |  ..s1   /* ding() */
[ 4234.289334] timer_simple:ding(): timed out... data=2
[ 4234.290177] timer_simple:ding(): 001) [swapper/1]:0   |  ..s1   /* ding() */
[ 4234.737346] timer_simple:ding(): timed out... data=1
[ 4234.738096] timer_simple:ding(): 001) [swapper/1]:0   |  ..s1   /* ding() */
$
```



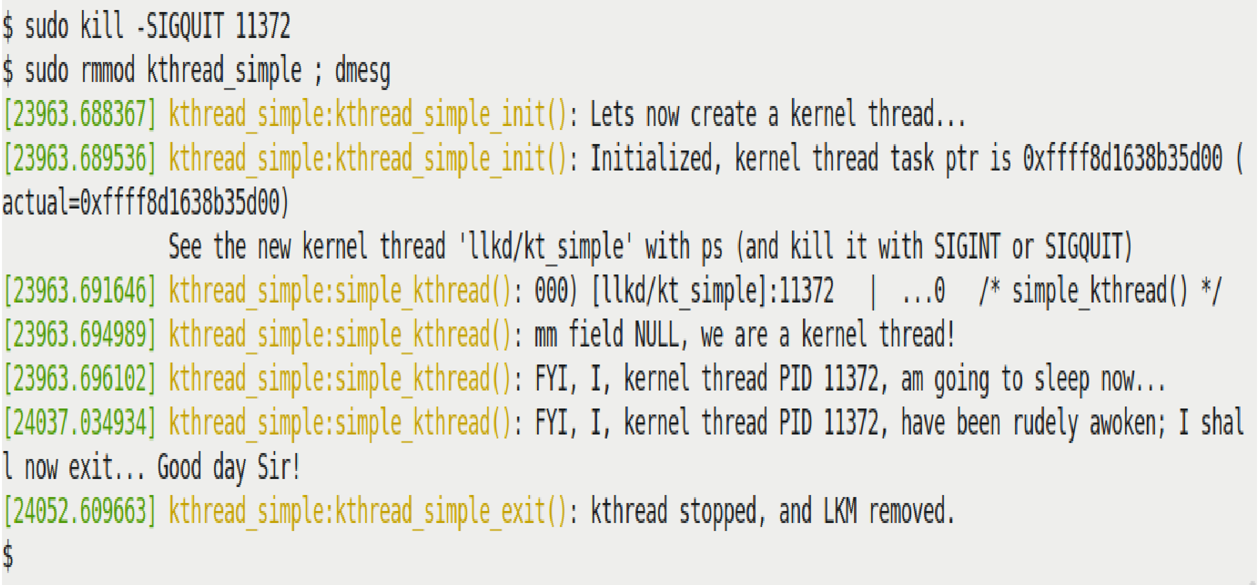
```
$ ../userapp_sed/userapp_sed1_dbg_asan
Usage: ../userapp_sed/userapp_sed1_dbg_asan device_file message
$ ../userapp_sed/userapp_sed1_dbg_asan /dev/sed1_drv "EncrypT ThiS plEaSe"
device opened: fd=3
msg before encrypt: EncrypT ThiS plEaSe
ioctl IOCTL_LLKD_SED_IOC_ENCRYPT_MSG done; len=19
msg after encrypt: �������^����^������

msg before decrypt: �������^����^������
ioctl IOCTL_LLKD_SED_IOC_DECRYPT_MSG done; len=19
msg after decrypt: EncrypT ThiS plEaSe
$
$ dmesg
[29519.684832] misc sed1_drv: LLKD sed1_drv misc driver (major # 10) registered, minor# = 55,
               dev node is /dev/sed1_drv
[29519.689403] sed1_drv:sed1_drv_init(): init done (make_it_fail is off)
[29519.690358] misc sed1_drv: loaded.
[29586.300784] sed1_drv:open_miscdrv(): 000)  userapp_sed1_db :22180   |  ...0   /* open_miscdrv() */
[29586.305511] sed1_drv:open_miscdrv(): opening "sed1_drv" now
[29586.306471] sed1_drv:ioctl_miscdrv(): In ioctl cmd option: encrypt
               arg=0x616000000080
[29586.308160] sed1_drv:ioctl_miscdrv(): xform=2, len=19
[29586.309011] payload: 00000000: 45 6e 63 72 79 70 54 20 54 68 69 53 20 70 6c 45  EncrypT ThiS plE
[29586.310084] payload: 00000010: 61 53 65                                         aSe
[29586.311075] sed1_drv:process_it(): data transform type: XF_ENCRYPT
[29586.311959] sed1_drv:encrypt_decrypt_payload(): starting timer + processing now ...
[29586.312977] sed1_drv:encrypt_decrypt_payload(): processing complete, timeout cancelled
[29586.313986] sed1_drv:encrypt_decrypt_payload(): delta: 99 ns (= 0 us = 0 ms)
[29586.314923] ret payload: 00000000: b9 90 9b 8c 85 8e aa 5e aa 96 95 ab 5e 8e 92 b9  .......^....^...
[29586.316458] ret payload: 00000010: 9d ab 99                                         ...
[29587.353483] sed1_drv:ioctl_miscdrv(): In ioctl cmd option: decrypt
               arg=0x616000000380
[29587.358744] sed1_drv:ioctl_miscdrv(): xform=1, len=19
[29587.359444] payload: 00000000: b9 90 9b 8c 85 8e aa 5e aa 96 95 ab 5e 8e 92 b9  .......^....^...
[29587.360408] payload: 00000010: 9d ab 99                                         ...
[29587.361281] sed1_drv:process_it(): data transform type: XF_DECRYPT
[29587.362056] sed1_drv:encrypt_decrypt_payload(): starting timer + processing now ...
[29587.362934] sed1_drv:encrypt_decrypt_payload(): processing complete, timeout cancelled
[29587.363893] sed1_drv:encrypt_decrypt_payload(): delta: 86 ns (= 0 us = 0 ms)
[29587.364788] ret payload: 00000000: 45 6e 63 72 79 70 54 20 54 68 69 53 20 70 6c 45  EncrypT ThiS plE
[29587.366134] ret payload: 00000010: 61 53 65                                         aSe
[29587.367070] sed1_drv:close_miscdrv(): closing "sed1_drv"
$ 
```



```
$ sudo rmmod sed1_drv
$ sudo dmesg -C
$ sudo insmod ./sed1_drv.ko make_it_fail=1
$ dmesg
[30090.202904] misc sed1_drv: LLKD sed1_drv misc driver (major # 10) registered, minor# = 56,
               dev node is /dev/sed1_drv
[30090.207537] sed1_drv:sed1_drv_init(): init done (make_it_fail is *on*)
[30090.208413] misc sed1_drv: loaded.
$
$
$ ../userapp_sed/userapp_sed1_dbg_asan /dev/sed1_drv "EncrypT ThiS plEaSe"
device opened: fd=3
msg before encrypt: EncrypT ThiS plEaSe
*** Operation Timed Out ***
$
$ dmesg
[30090.202904] misc sed1_drv: LLKD sed1_drv misc driver (major # 10) registered, minor# = 56,
               dev node is /dev/sed1_drv
[30090.207537] sed1_drv:sed1_drv_init(): init done (make_it_fail is *on*)
[30090.208413] misc sed1_drv: loaded.
[30103.759259] sed1_drv:open_miscdrv(): 000)  userapp_sed1_db :22264   |  ...0   /* open_miscdrv() */
[30103.768031] sed1_drv:open_miscdrv(): opening "sed1_drv" now
[30103.769119] sed1_drv:ioctl_miscdrv(): In ioctl cmd option: encrypt
               arg=0x616000000080
[30103.770727] sed1_drv:ioctl_miscdrv(): xform=2, len=19
[30103.771504] payload: 00000000: 45 6e 63 72 79 70 54 20 54 68 69 53 20 70 6c 45  EncrypT ThiS plE
[30103.772650] payload: 00000010: 61 53 65                                         aSe
[30103.773646] sed1_drv:process_it(): data transform type: XF_ENCRYPT
[30103.774578] sed1_drv:encrypt_decrypt_payload(): starting timer + processing now ...
[30103.780372] sed1_drv:timesup(): *** Timer expired! ***
[30103.783770] sed1_drv:timesup(): 000) [swapper/0]:0   |  ..s1   /* timesup() */
[30103.790158] sed1_drv:encrypt_decrypt_payload(): cancelled the timer while it's inactive! (deadline missed?)
[30103.793372] sed1_drv:encrypt_decrypt_payload(): delta: 14580905 ns (= 14580 us = 14 ms)
[30103.794353] sed1_drv:ioctl_miscdrv(): ** timed out **
[30103.795117] ret payload: 00000000: b9 90 9b 8c 85 8e aa 5e aa 96 95 ab 5e 8e 92 b9  .......^....^...
[30103.796635] ret payload: 00000010: 9d ab 99                                         ...
[30103.801124] sed1_drv:close_miscdrv(): closing "sed1_drv"
$
```

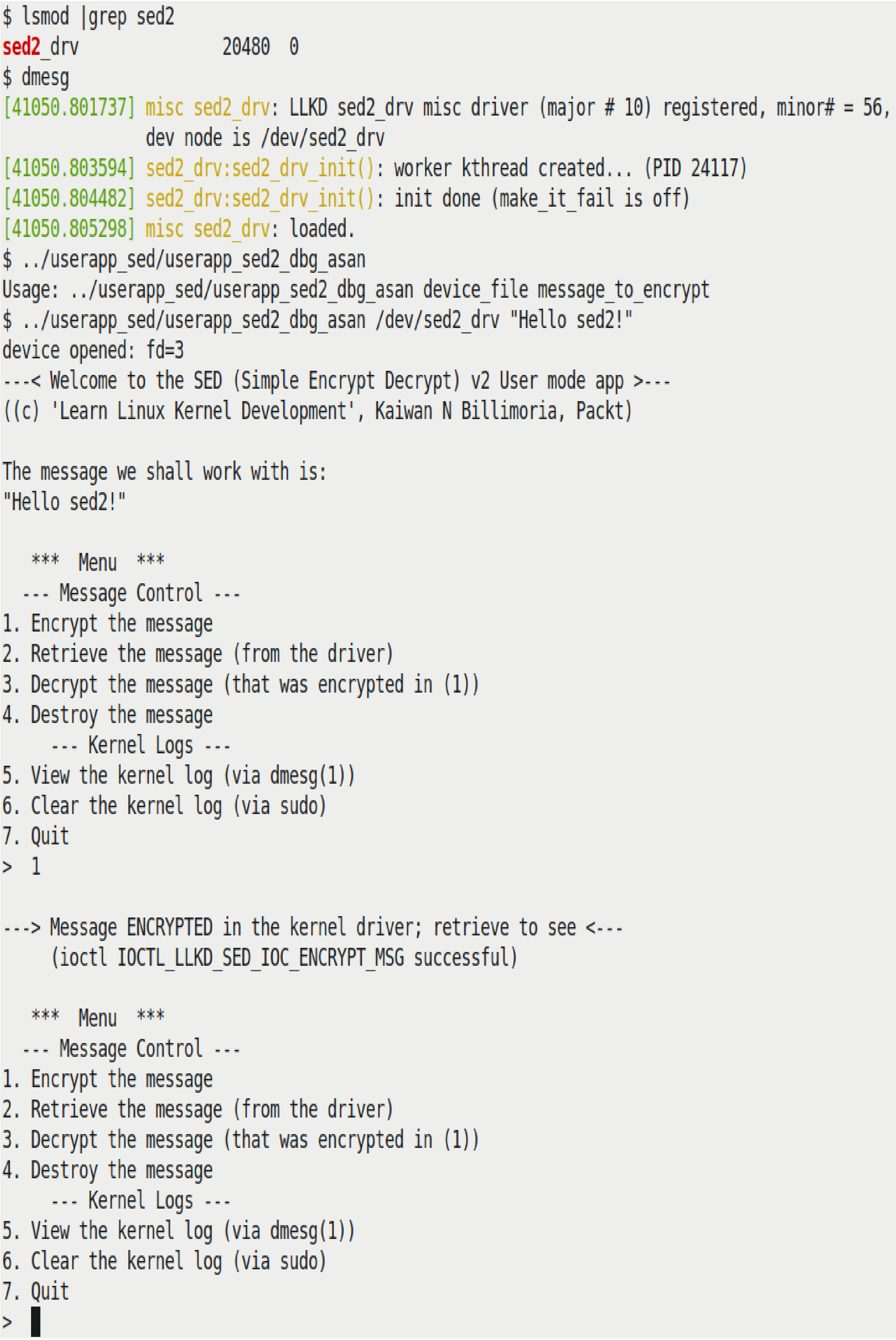


```
[23963.688367] kthread_simple:kthread_simple_init(): Lets now create a kernel thread...
[23963.689536] kthread_simple:kthread_simple_init(): Initialized, kernel thread task ptr is 0xffff8d1638b35d00 (
actual=0xffff8d1638b35d00)
               See the new kernel thread 'llkd/kt_simple' with ps (and kill it with SIGINT or SIGQUIT)
[23963.691646] kthread_simple:simple_kthread(): 000) [llkd/kt_simple]:11372   |  ...0   /* simple_kthread() */
[23963.694989] kthread_simple:simple_kthread(): mm field NULL, we are a kernel thread!
[23963.696102] kthread_simple:simple_kthread(): FYI, I, kernel thread PID 11372, am going to sleep now...
```

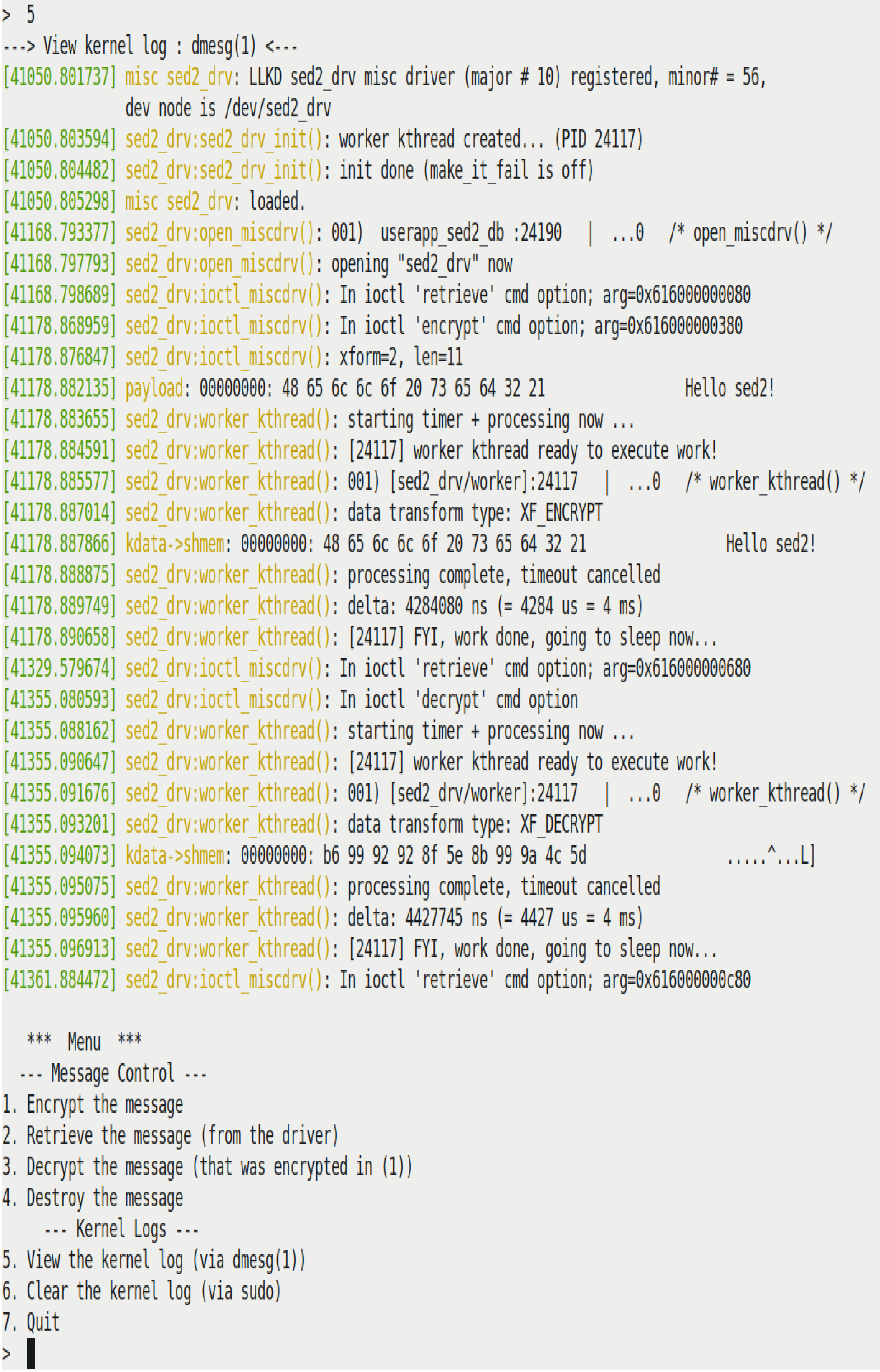


```
$ sudo kill -SIGQUIT 11372
$ sudo rmmod kthread_simple ; dmesg
[23963.688367] kthread_simple:kthread_simple_init(): Lets now create a kernel thread...
[23963.689536] kthread_simple:kthread_simple_init(): Initialized, kernel thread task ptr is 0xffff8d1638b35d00 (
actual=0xffff8d1638b35d00)
               See the new kernel thread 'llkd/kt_simple' with ps (and kill it with SIGINT or SIGQUIT)
[23963.691646] kthread_simple:simple_kthread(): 000) [llkd/kt_simple]:11372   |  ...0   /* simple_kthread() */
[23963.694989] kthread_simple:simple_kthread(): mm field NULL, we are a kernel thread!
[23963.696102] kthread_simple:simple_kthread(): FYI, I, kernel thread PID 11372, am going to sleep now...
[24037.034934] kthread_simple:simple_kthread(): FYI, I, kernel thread PID 11372, have been rudely awoken; I shal
l now exit... Good day Sir!
[24052.609663] kthread_simple:kthread_simple_exit(): kthread stopped, and LKM removed.
$
```

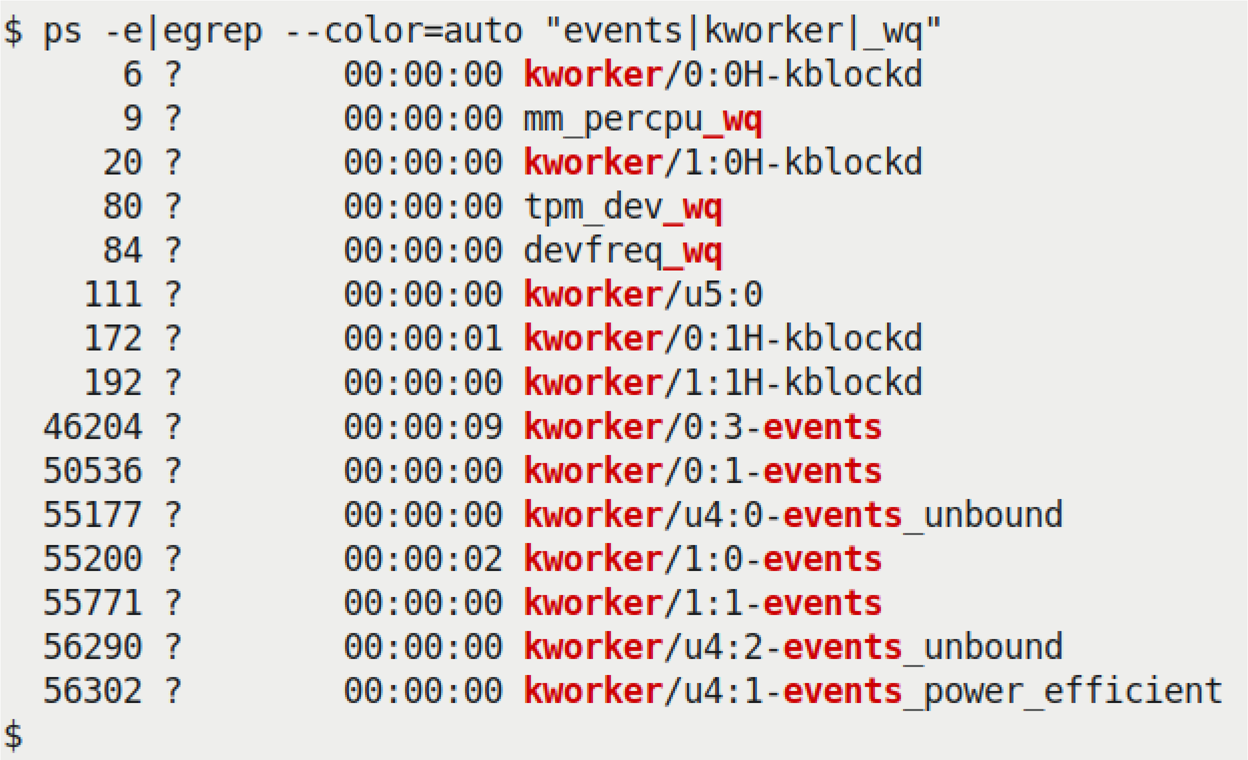
```
                    User-space application
                     /     |        |     \
             |-------      |        |      -------|
    op:   encrypt      retrieve  decrypt       destroy
                        [ioctl]                [ioctl]

        <---------------- sed2 driver ---------------->
    op:   encrypt      retrieve    decrypt      destroy
    by:  [kthread]      [ioctl]   [kthread]     [ioctl]
             ^            ^          ^            ^
             |            |          |            |
             v            v          v            v
           {--------- shared memory region ---------}
```



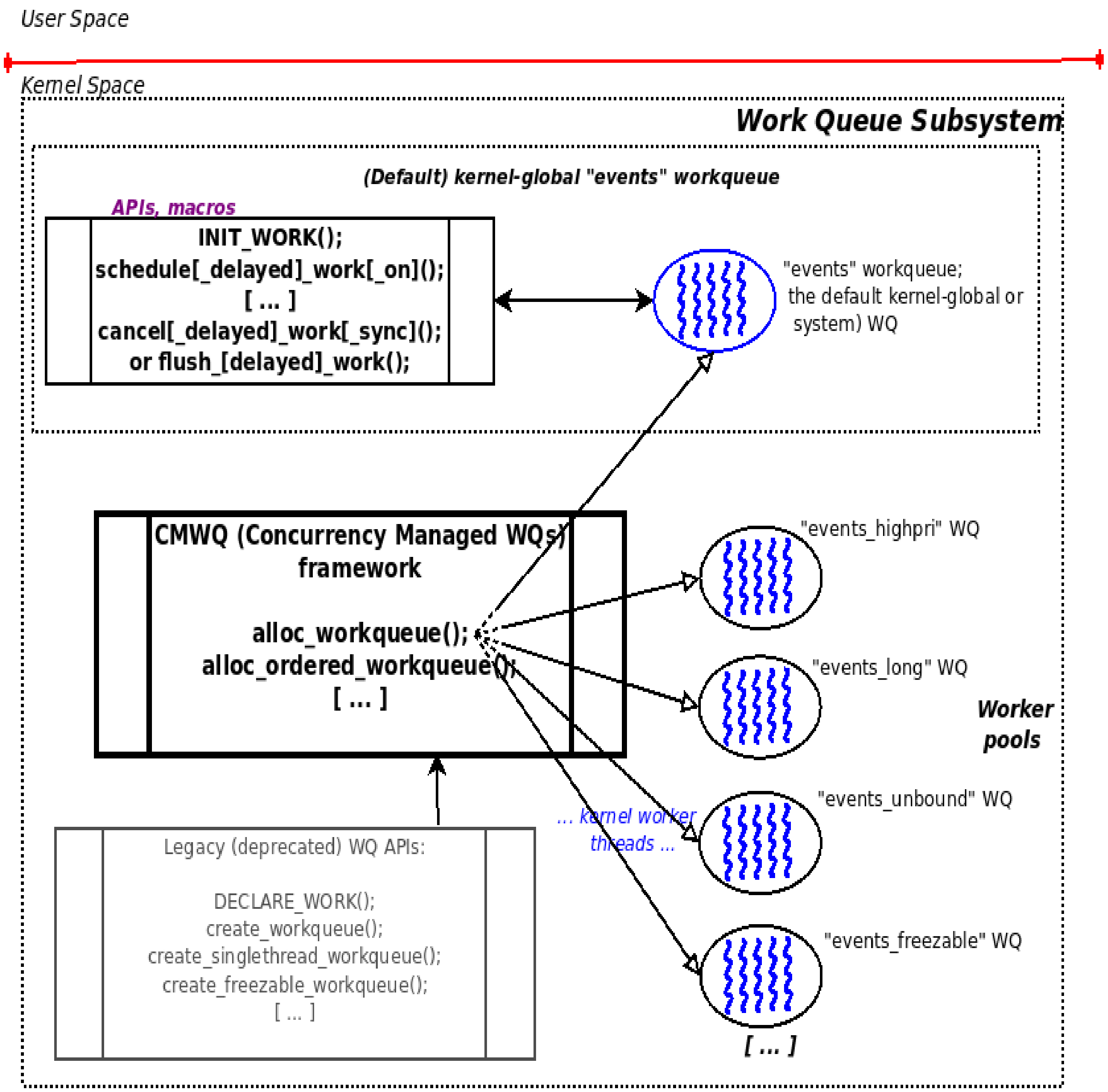
```
$ lsmod |grep sed2
sed2_drv               20480  0
$ dmesg
[41050.801737] misc sed2_drv: LLKD sed2_drv misc driver (major # 10) registered, minor# = 56,
               dev node is /dev/sed2_drv
[41050.803594] sed2_drv:sed2_drv_init(): worker kthread created... (PID 24117)
[41050.804482] sed2_drv:sed2_drv_init(): init done (make_it_fail is off)
[41050.805298] misc sed2_drv: loaded.
$ ../userapp_sed/userapp_sed2_dbg_asan
Usage: ../userapp_sed/userapp_sed2_dbg_asan device_file message_to_encrypt
$ ../userapp_sed/userapp_sed2_dbg_asan /dev/sed2_drv "Hello sed2!"
device opened: fd=3
---< Welcome to the SED (Simple Encrypt Decrypt) v2 User mode app >---
((c) 'Learn Linux Kernel Development', Kaiwan N Billimoria, Packt)

The message we shall work with is:
"Hello sed2!"

   ***  Menu  ***
  --- Message Control ---
1. Encrypt the message
2. Retrieve the message (from the driver)
3. Decrypt the message (that was encrypted in (1))
4. Destroy the message
     --- Kernel Logs ---
5. View the kernel log (via dmesg(1))
6. Clear the kernel log (via sudo)
7. Quit
>  1

---> Message ENCRYPTED in the kernel driver; retrieve to see <---
     (ioctl IOCTL_LLKD_SED_IOC_ENCRYPT_MSG successful)

   ***  Menu  ***
  --- Message Control ---
1. Encrypt the message
2. Retrieve the message (from the driver)
3. Decrypt the message (that was encrypted in (1))
4. Destroy the message
     --- Kernel Logs ---
5. View the kernel log (via dmesg(1))
6. Clear the kernel log (via sudo)
7. Quit
>  █
```

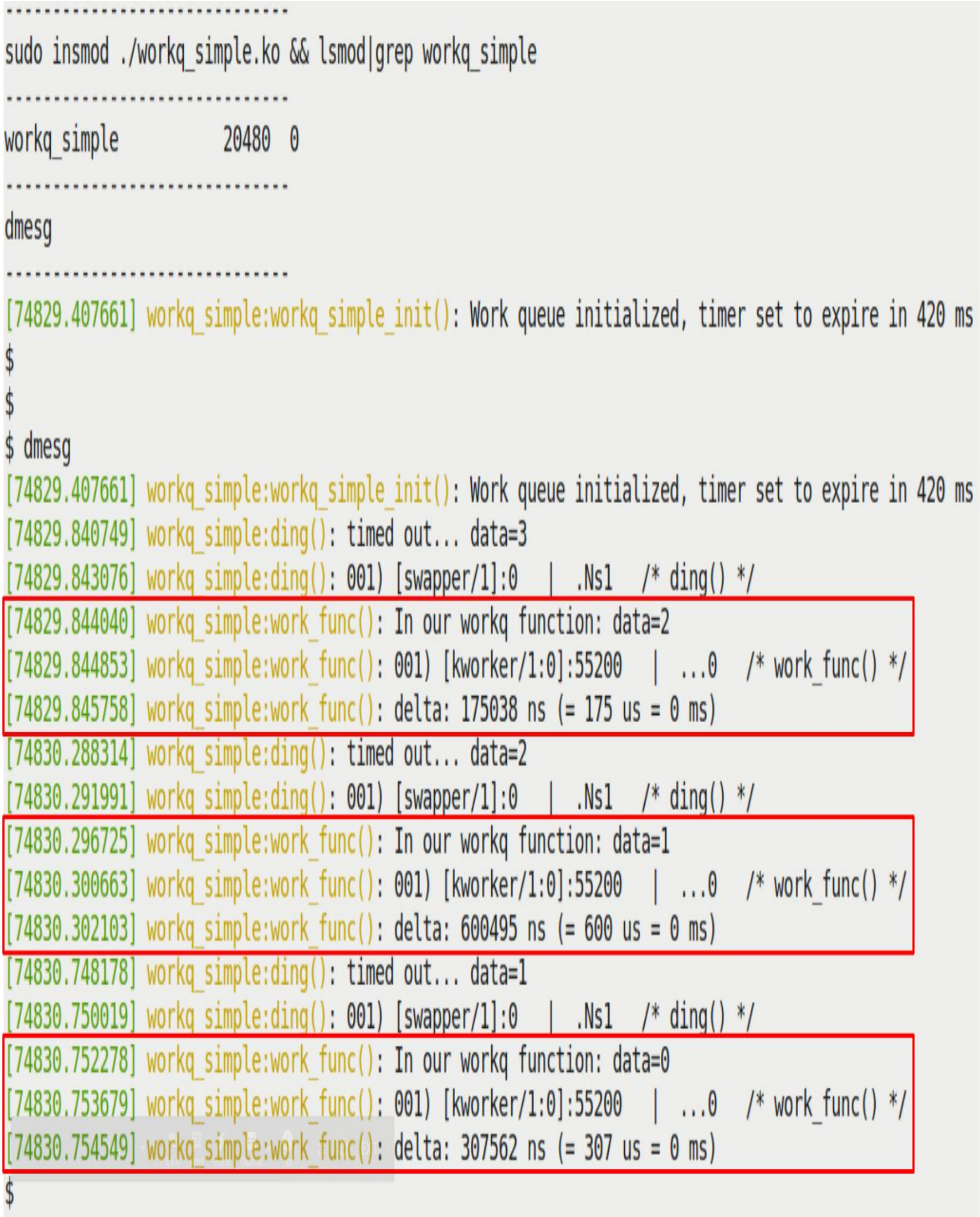


```
>  5
---> View kernel log : dmesg(1) <---
[41050.801737] misc sed2_drv: LLKD sed2_drv misc driver (major # 10) registered, minor# = 56,
               dev node is /dev/sed2_drv
[41050.803594] sed2_drv:sed2_drv_init(): worker kthread created... (PID 24117)
[41050.804482] sed2_drv:sed2_drv_init(): init done (make_it_fail is off)
[41050.805298] misc sed2_drv: loaded.
[41168.793377] sed2_drv:open_miscdrv(): 001)  userapp_sed2_db :24190   |  ...0   /* open_miscdrv() */
[41168.797793] sed2_drv:open_miscdrv(): opening "sed2_drv" now
[41168.798689] sed2_drv:ioctl_miscdrv(): In ioctl 'retrieve' cmd option; arg=0x616000000080
[41178.868959] sed2_drv:ioctl_miscdrv(): In ioctl 'encrypt' cmd option; arg=0x616000000380
[41178.876847] sed2_drv:ioctl_miscdrv(): xform=2, len=11
[41178.882135] payload: 00000000: 48 65 6c 6c 6f 20 73 65 64 32 21                Hello sed2!
[41178.883655] sed2_drv:worker_kthread(): starting timer + processing now ...
[41178.884591] sed2_drv:worker_kthread(): [24117] worker kthread ready to execute work!
[41178.885577] sed2_drv:worker_kthread(): 001) [sed2_drv/worker]:24117   |  ...0   /* worker_kthread() */
[41178.887014] sed2_drv:worker_kthread(): data transform type: XF_ENCRYPT
[41178.887866] kdata->shmem: 00000000: 48 65 6c 6c 6f 20 73 65 64 32 21                Hello sed2!
[41178.888875] sed2_drv:worker_kthread(): processing complete, timeout cancelled
[41178.889749] sed2_drv:worker_kthread(): delta: 4284080 ns (= 4284 us = 4 ms)
[41178.890658] sed2_drv:worker_kthread(): [24117] FYI, work done, going to sleep now...
[41329.579674] sed2_drv:ioctl_miscdrv(): In ioctl 'retrieve' cmd option; arg=0x616000000680
[41355.080593] sed2_drv:ioctl_miscdrv(): In ioctl 'decrypt' cmd option
[41355.088162] sed2_drv:worker_kthread(): starting timer + processing now ...
[41355.090647] sed2_drv:worker_kthread(): [24117] worker kthread ready to execute work!
[41355.091676] sed2_drv:worker_kthread(): 001) [sed2_drv/worker]:24117   |  ...0   /* worker_kthread() */
[41355.093201] sed2_drv:worker_kthread(): data transform type: XF_DECRYPT
[41355.094073] kdata->shmem: 00000000: b6 99 92 92 8f 5e 8b 99 9a 4c 5d                .....^...L]
[41355.095075] sed2_drv:worker_kthread(): processing complete, timeout cancelled
[41355.095960] sed2_drv:worker_kthread(): delta: 4427745 ns (= 4427 us = 4 ms)
[41355.096913] sed2_drv:worker_kthread(): [24117] FYI, work done, going to sleep now...
[41361.884472] sed2_drv:ioctl_miscdrv(): In ioctl 'retrieve' cmd option; arg=0x616000000c80

   ***  Menu  ***
  --- Message Control ---
1. Encrypt the message
2. Retrieve the message (from the driver)
3. Decrypt the message (that was encrypted in (1))
4. Destroy the message
     --- Kernel Logs ---
5. View the kernel log (via dmesg(1))
6. Clear the kernel log (via sudo)
7. Quit
>
```

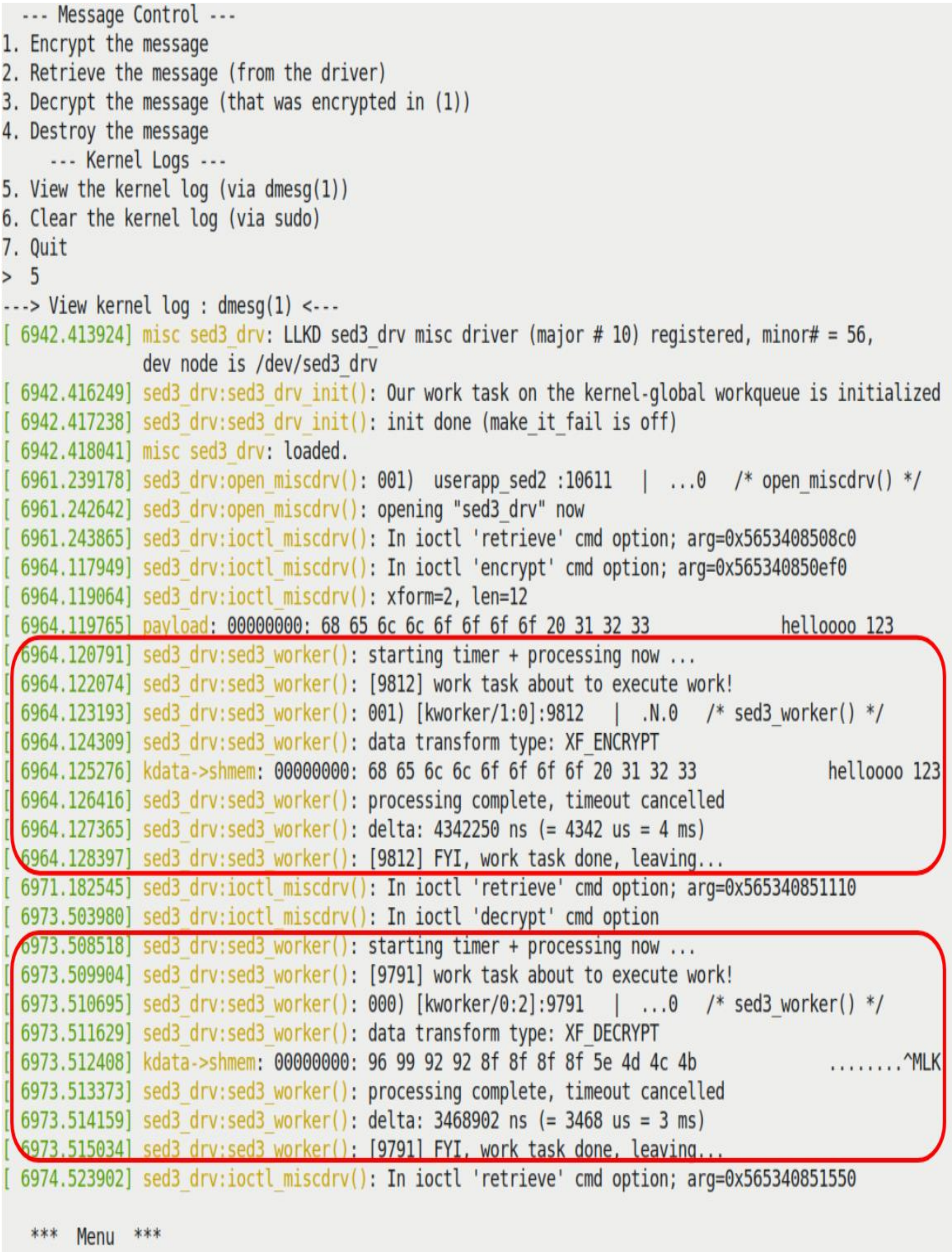


```
$ ps -e|egrep --color=auto "events|kworker|_wq"
      6 ?        00:00:00 kworker/0:0H-kblockd
      9 ?        00:00:00 mm_percpu_wq
     20 ?        00:00:00 kworker/1:0H-kblockd
     80 ?        00:00:00 tpm_dev_wq
     84 ?        00:00:00 devfreq_wq
    111 ?        00:00:00 kworker/u5:0
    172 ?        00:00:01 kworker/0:1H-kblockd
    192 ?        00:00:00 kworker/1:1H-kblockd
  46204 ?        00:00:09 kworker/0:3-events
  50536 ?        00:00:00 kworker/0:1-events
  55177 ?        00:00:00 kworker/u4:0-events_unbound
  55200 ?        00:00:02 kworker/1:0-events
  55771 ?        00:00:00 kworker/1:1-events
  56290 ?        00:00:00 kworker/u4:2-events_unbound
  56302 ?        00:00:00 kworker/u4:1-events_power_efficient
$
```


User Space
Kernel Space
Work Queue Subsystem
(Default) kernel-global "events" workqueue
APIs, macros
INIT_WORK();
schedule[_delayed]_work[_on]();
[...]
cancel[_delayed]_work[_sync]();
or flush_[delayed]_work();
"events" workqueue;
the default kernel-global or
system) WQ
CMWQ (Concurrency Managed WQs)
framework
alloc_workqueue();
alloc_ordered_workqueue();
[...]
"events_highpri" WQ
"events_long" WQ
Worker
pools
"events_unbound" WQ
... kernel worker
threads ...
"events_freezable" WQ
[...]
Legacy (deprecated) WQ APIs:
DECLARE_WORK();
create_workqueue();
create_singlethread_workqueue();
create_freezable_workqueue();
[...]

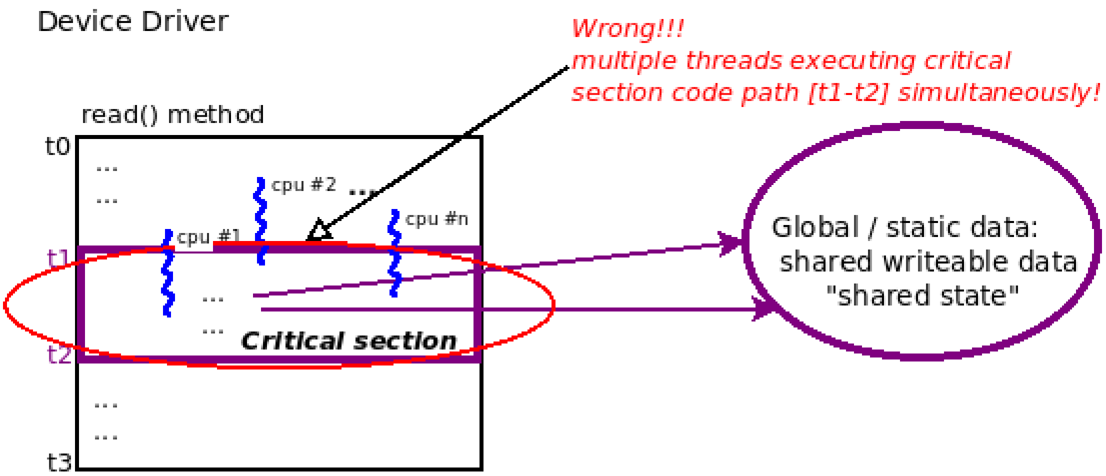


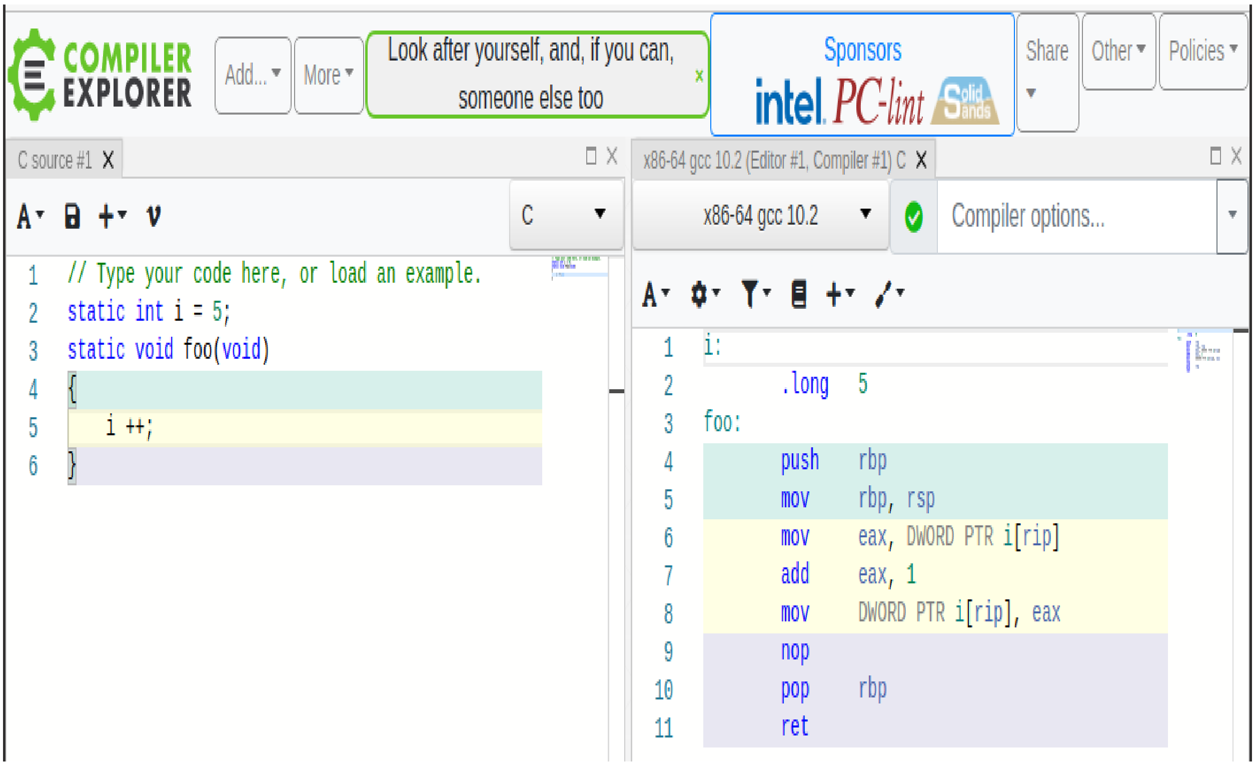
```
------------------------------
sudo insmod ./workq_simple.ko && lsmod|grep workq_simple
------------------------------
workq_simple          20480  0
------------------------------
dmesg
------------------------------
[74829.407661] workq_simple:workq_simple_init(): Work queue initialized, timer set to expire in 420 ms
$
$
$ dmesg
[74829.407661] workq_simple:workq_simple_init(): Work queue initialized, timer set to expire in 420 ms
[74829.840749] workq_simple:ding(): timed out... data=3
[74829.843076] workq_simple:ding(): 001) [swapper/1]:0   |  .Ns1   /* ding() */
[74829.844040] workq_simple:work_func(): In our workq function: data=2
[74829.844853] workq_simple:work_func(): 001) [kworker/1:0]:55200   |  ...0   /* work_func() */
[74829.845758] workq_simple:work_func(): delta: 175038 ns (= 175 us = 0 ms)
[74830.288314] workq_simple:ding(): timed out... data=2
[74830.291991] workq_simple:ding(): 001) [swapper/1]:0   |  .Ns1   /* ding() */
[74830.296725] workq_simple:work_func(): In our workq function: data=1
[74830.300663] workq_simple:work_func(): 001) [kworker/1:0]:55200   |  ...0   /* work_func() */
[74830.302103] workq_simple:work_func(): delta: 600495 ns (= 600 us = 0 ms)
[74830.748178] workq_simple:ding(): timed out... data=1
[74830.750019] workq_simple:ding(): 001) [swapper/1]:0   |  .Ns1   /* ding() */
[74830.752278] workq_simple:work_func(): In our workq function: data=0
[74830.753679] workq_simple:work_func(): 001) [kworker/1:0]:55200   |  ...0   /* work_func() */
[74830.754549] workq_simple:work_func(): delta: 307562 ns (= 307 us = 0 ms)
$
```



```
  --- Message Control ---
1. Encrypt the message
2. Retrieve the message (from the driver)
3. Decrypt the message (that was encrypted in (1))
4. Destroy the message
     --- Kernel Logs ---
5. View the kernel log (via dmesg(1))
6. Clear the kernel log (via sudo)
7. Quit
> 5
---> View kernel log : dmesg(1) <---
[ 6942.413924] misc sed3_drv: LLKD sed3_drv misc driver (major # 10) registered, minor# = 56,
               dev node is /dev/sed3_drv
[ 6942.416249] sed3_drv:sed3_drv_init(): Our work task on the kernel-global workqueue is initialized
[ 6942.417238] sed3_drv:sed3_drv_init(): init done (make_it_fail is off)
[ 6942.418041] misc sed3_drv: loaded.
[ 6961.239178] sed3_drv:open_miscdrv(): 001)  userapp_sed2 :10611   |  ...0   /* open_miscdrv() */
[ 6961.242642] sed3_drv:open_miscdrv(): opening "sed3_drv" now
[ 6961.243865] sed3_drv:ioctl_miscdrv(): In ioctl 'retrieve' cmd option; arg=0x5653408508c0
[ 6964.117949] sed3_drv:ioctl_miscdrv(): In ioctl 'encrypt' cmd option; arg=0x565340850ef0
[ 6964.119064] sed3_drv:ioctl_miscdrv(): xform=2, len=12
[ 6964.119765] payload: 00000000: 68 65 6c 6c 6f 6f 6f 6f 20 31 32 33              helloooo 123
[ 6964.120791] sed3_drv:sed3_worker(): starting timer + processing now ...
[ 6964.122074] sed3_drv:sed3_worker(): [9812] work task about to execute work!
[ 6964.123193] sed3_drv:sed3_worker(): 001) [kworker/1:0]:9812   |  .N.0   /* sed3_worker() */
[ 6964.124309] sed3_drv:sed3_worker(): data transform type: XF_ENCRYPT
[ 6964.125276] kdata->shmem: 00000000: 68 65 6c 6c 6f 6f 6f 6f 20 31 32 33              helloooo 123
[ 6964.126416] sed3_drv:sed3_worker(): processing complete, timeout cancelled
[ 6964.127365] sed3_drv:sed3_worker(): delta: 4342250 ns (= 4342 us = 4 ms)
[ 6964.128397] sed3_drv:sed3_worker(): [9812] FYI, work task done, leaving...
[ 6971.182545] sed3_drv:ioctl_miscdrv(): In ioctl 'retrieve' cmd option; arg=0x565340851110
[ 6973.503980] sed3_drv:ioctl_miscdrv(): In ioctl 'decrypt' cmd option
[ 6973.508518] sed3_drv:sed3_worker(): starting timer + processing now ...
[ 6973.509904] sed3_drv:sed3_worker(): [9791] work task about to execute work!
[ 6973.510695] sed3_drv:sed3_worker(): 000) [kworker/0:2]:9791   |  ...0   /* sed3_worker() */
[ 6973.511629] sed3_drv:sed3_worker(): data transform type: XF_DECRYPT
[ 6973.512408] kdata->shmem: 00000000: 96 99 92 92 8f 8f 8f 8f 5e 4d 4c 4b              ........^MLK
[ 6973.513373] sed3_drv:sed3_worker(): processing complete, timeout cancelled
[ 6973.514159] sed3_drv:sed3_worker(): delta: 3468902 ns (= 3468 us = 3 ms)
[ 6973.515034] sed3_drv:sed3_worker(): [9791] FYI, work task done, leaving...
[ 6974.523902] sed3_drv:ioctl_miscdrv(): In ioctl 'retrieve' cmd option; arg=0x565340851550

  ***  Menu  ***
```

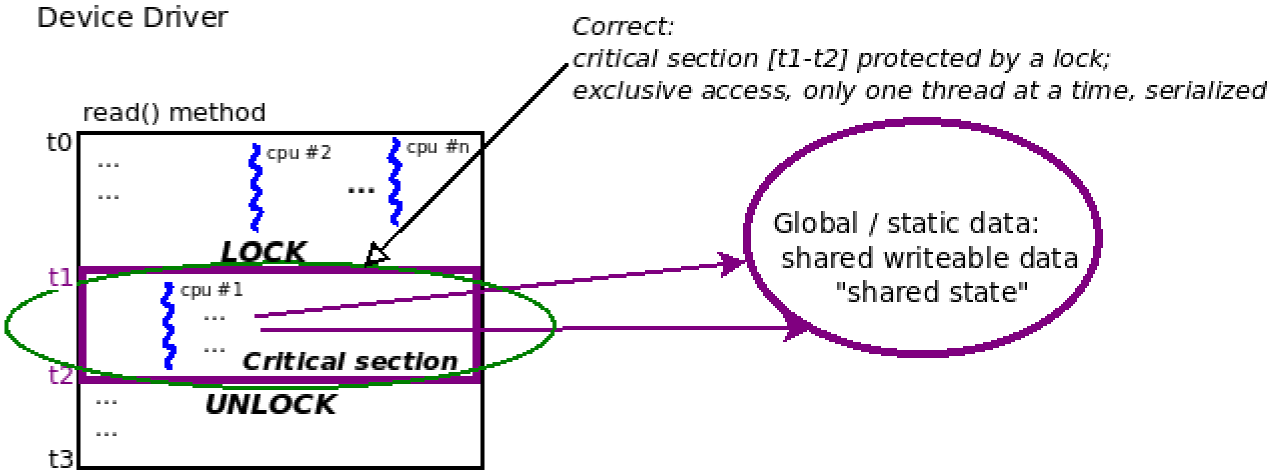
Chapter 6: Kernel Synchronization - Part 1

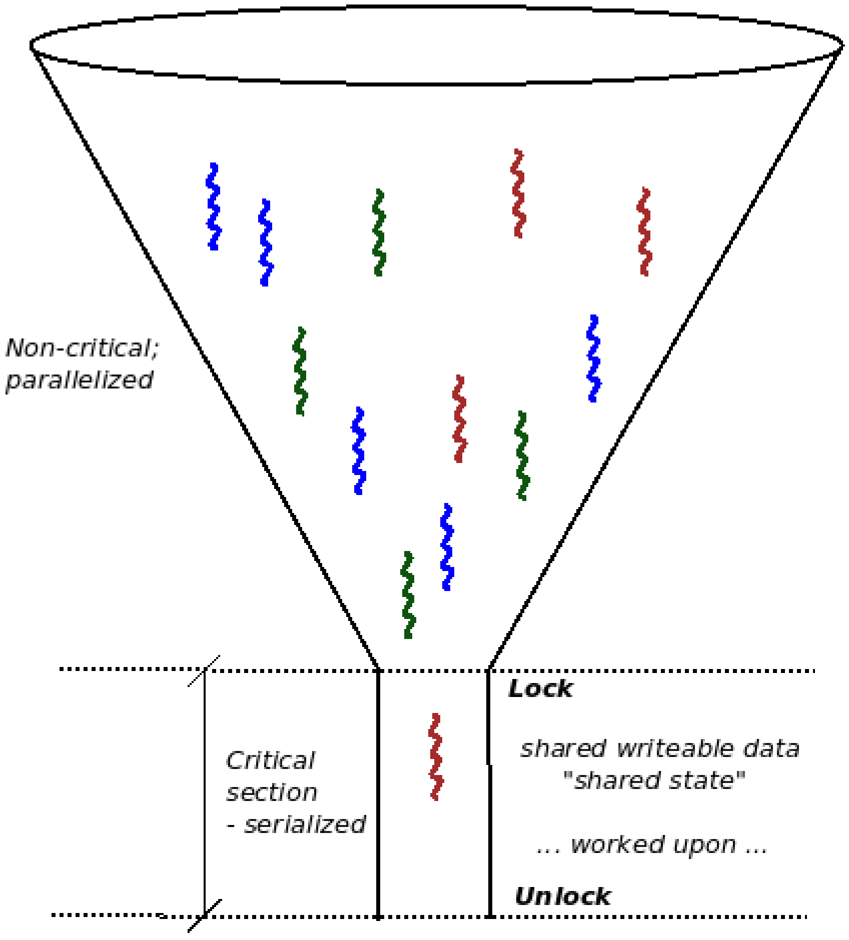


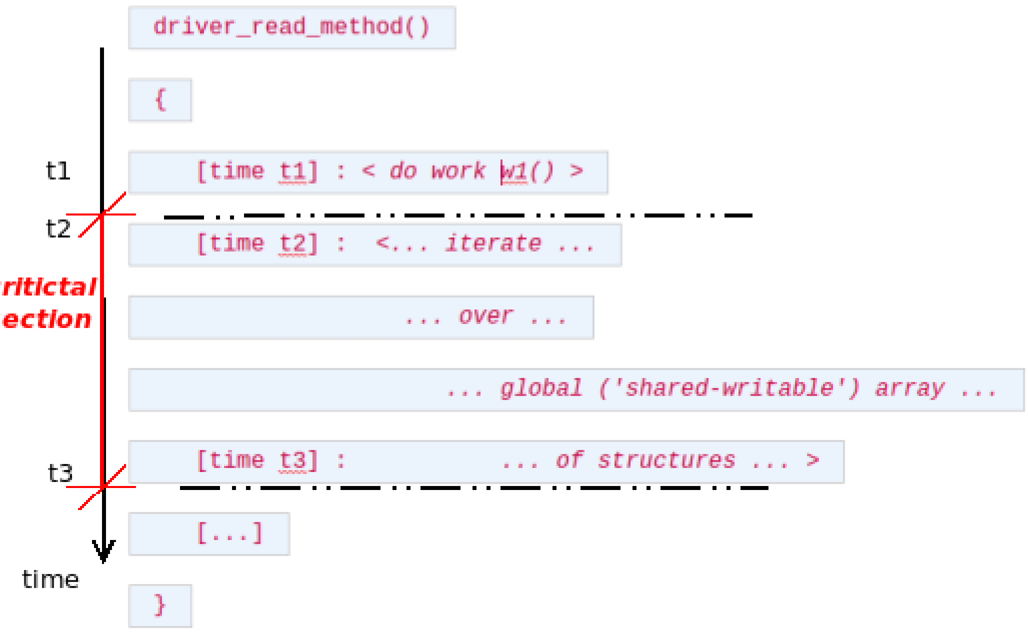

COMPILER EXPLORER
Add...
More
Look after yourself, and, if you can, someone else too
Sponsors
intel PC-lint Solid Sands
Share
Other
Policies
C source #1
C
x86-64 gcc 10.2 (Editor #1, Compiler #1) C
x86-64 gcc 10.2
Compiler options...

```c
// Type your code here, or load an example.
static int i = 5;
static void foo(void)
{
    i ++;
}
```

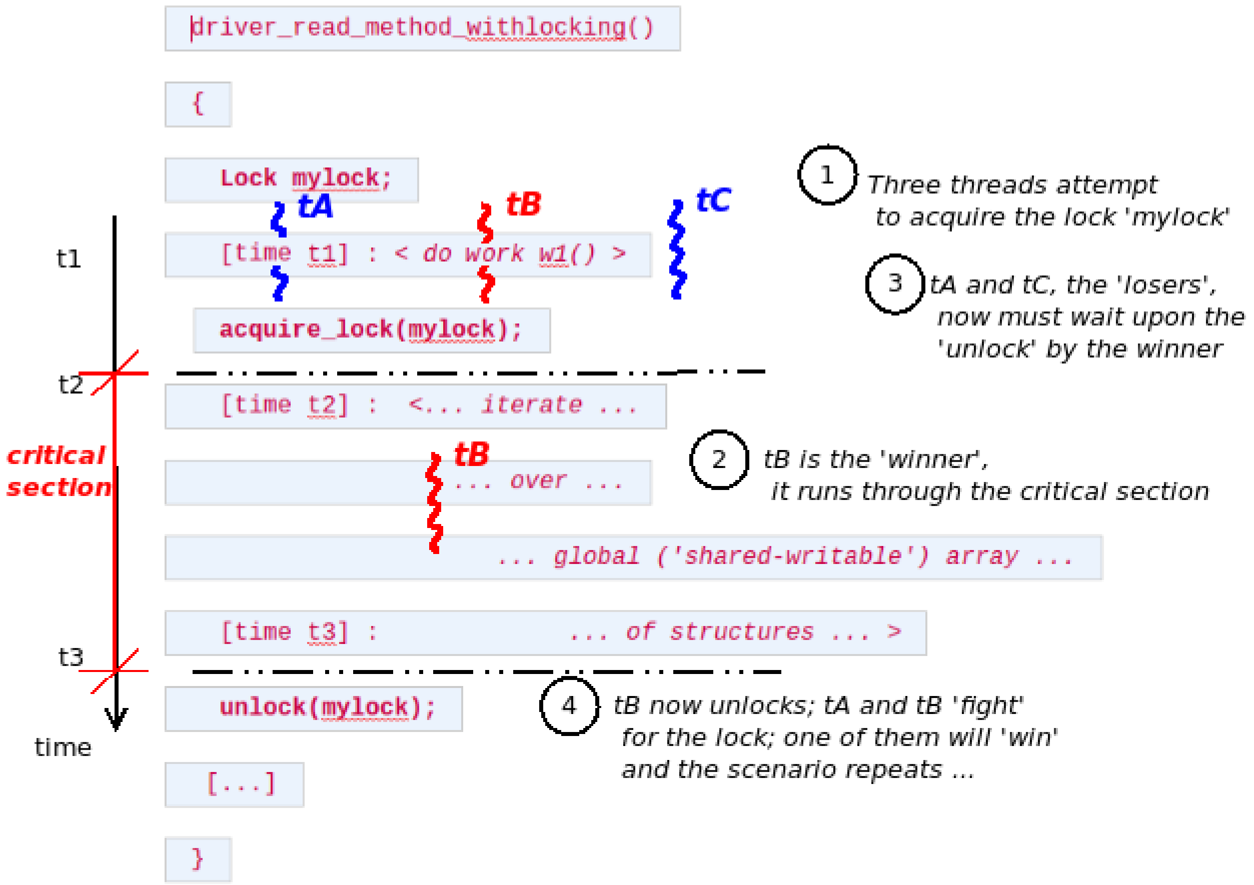
```
i:
        .long   5
foo:
        push    rbp
        mov     rbp, rsp
        mov     eax, DWORD PTR i[rip]
        add     eax, 1
        mov     DWORD PTR i[rip], eax
        nop
        pop     rbp
        ret
```


Device Driver
Correct:
critical section [t1-t2] protected by a lock;
exclusive access, only one thread at a time, serialized
read() method
t0
...
...
cpu #2
cpu #n
...
LOCK
t1
cpu #1
...
...
Critical section
t2
...
...
UNLOCK
t3
Global / static data:
shared writeable data
"shared state"


Non-critical;
parallelized
Lock
Critical
section
- serialized
shared writeable data
"shared state"
... worked upon ...
Unlock



```
driver_read_method()
{
    [time t1] : < do work w1() >
    [time t2] :  <... iterate ...
                         ... over ...
                            ... global ('shared-writable') array ...
    [time t3] :                ... of structures ... >
    [...]
}
```



```
driver_read_method_withlocking()
{
    Lock mylock;
    [time t1] : < do work w1() >
    acquire_lock(mylock);
    [time t2] :  <... iterate ...
                         ... over ...
                            ... global ('shared-writable') array ...
    [time t3] :                ... of structures ... >
    unlock(mylock);
    [...]
}
```

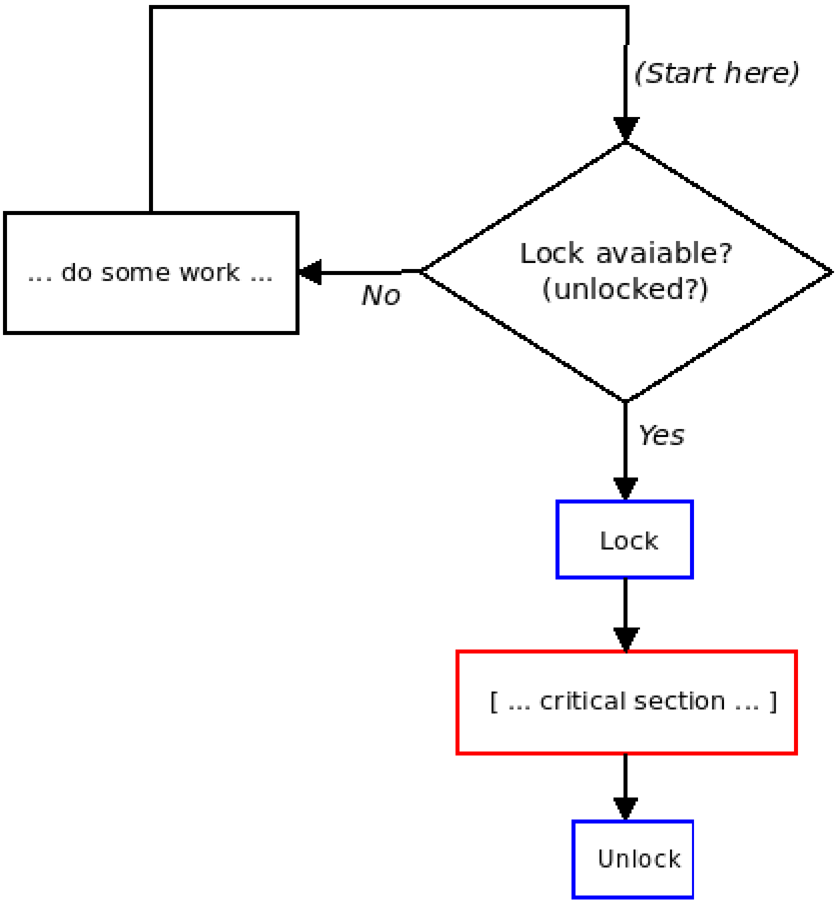
1 Three threads attempt to acquire the lock 'mylock'

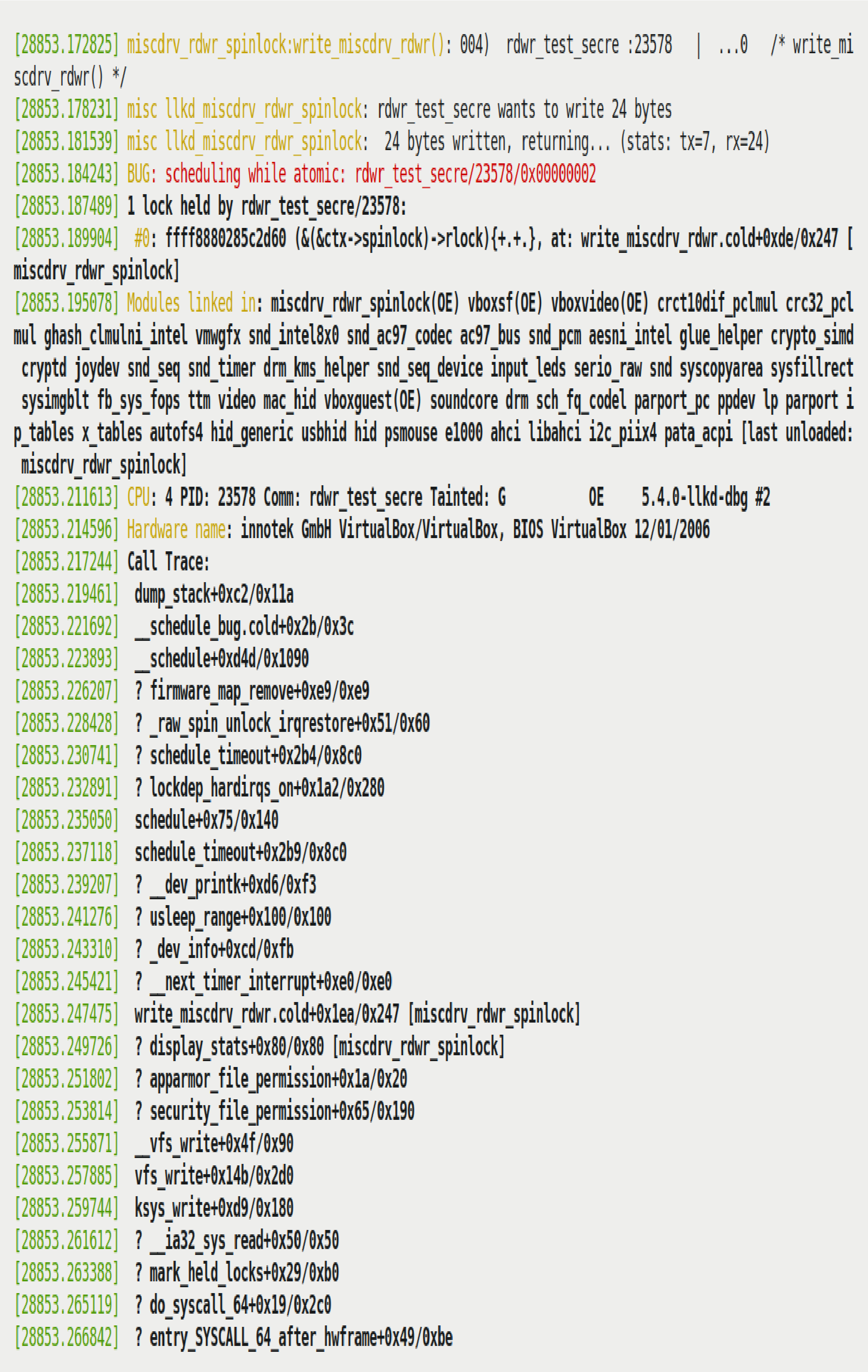
3 tA and tC, the 'losers', now must wait upon the 'unlock' by the winner

2 tB is the 'winner', it runs through the critical section

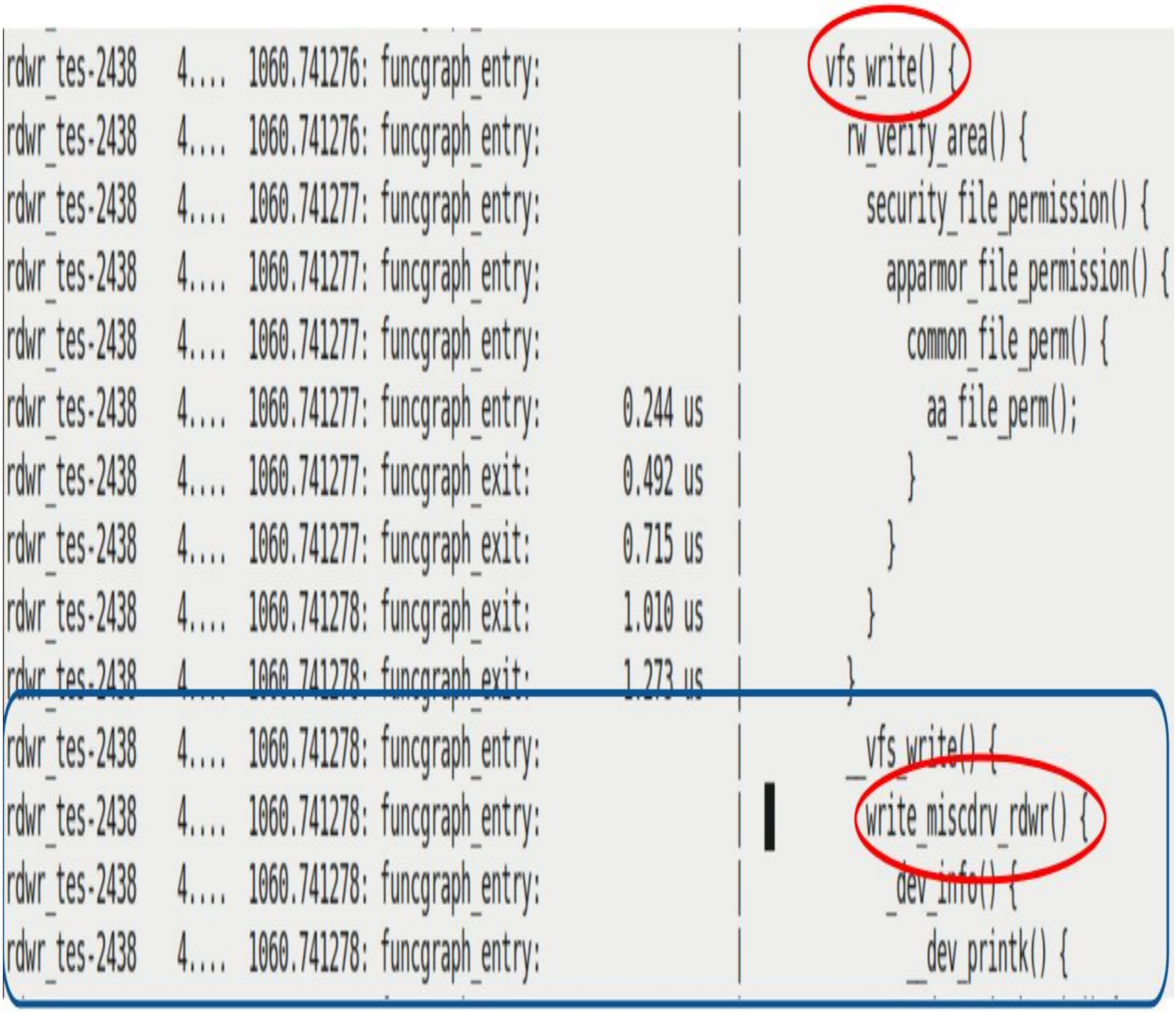
4 tB now unlocks; tA and tB 'fight' for the lock; one of them will 'win' and the scenario repeats ...

```
 static ssize_t read_miscdrv_rdwr(struct file *filp, char __user *ubuf,
-                size_t count, loff_t *off)
+               size_t count, loff_t *off)
 {
-   int ret = count, secret_len = strnlen(ctx->oursecret, MAXBYTES);
+   int ret = count, secret_len;
    struct device *dev = ctx->dev;

+   mutex_lock(&ctx->lock);
+   secret_len = strlen(ctx->oursecret);
+   mutex_unlock(&ctx->lock);
+
    PRINT_CTX();
    dev_info(dev, "%s wants to read (upto) %zd bytes\n", current->comm, count);

@@ -134,17 +140,20 @@
     * member to userspace.
     */
    ret = -EFAULT;
+   mutex_lock(&ctx->lock);
    if (copy_to_user(ubuf, ctx->oursecret, secret_len)) {
        dev_warn(dev, "copy_to_user() failed\n");
-       goto out_notok;
+       goto out_ctu;
    }
    ret = secret_len;

    // Update stats
-   ctx->tx += secret_len;  // our 'transmit' is wrt this driver
+   ctx->tx += secret_len;  // our 'transmit' is wrt this driver
    dev_info(dev, " %d bytes read, returning... (stats: tx=%d, rx=%d)\n",
-       secret_len, ctx->tx, ctx->rx);
- out_notok:
+           secret_len, ctx->tx, ctx->rx);
+out_ctu:
+   mutex_unlock(&ctx->lock);
+out_notok:
    return ret;
```

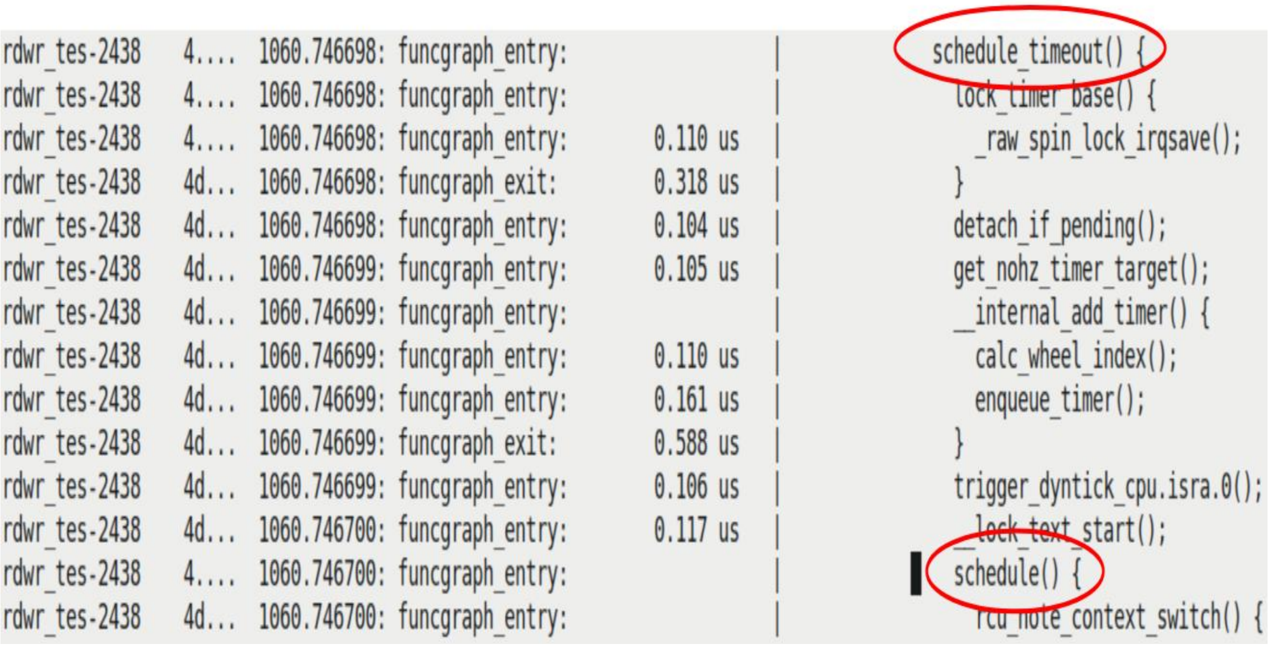

(Start here)
Lock avaiable?
(unlocked?)
... do some work ...
No
Yes
Lock
[... critical section ...]
Unlock



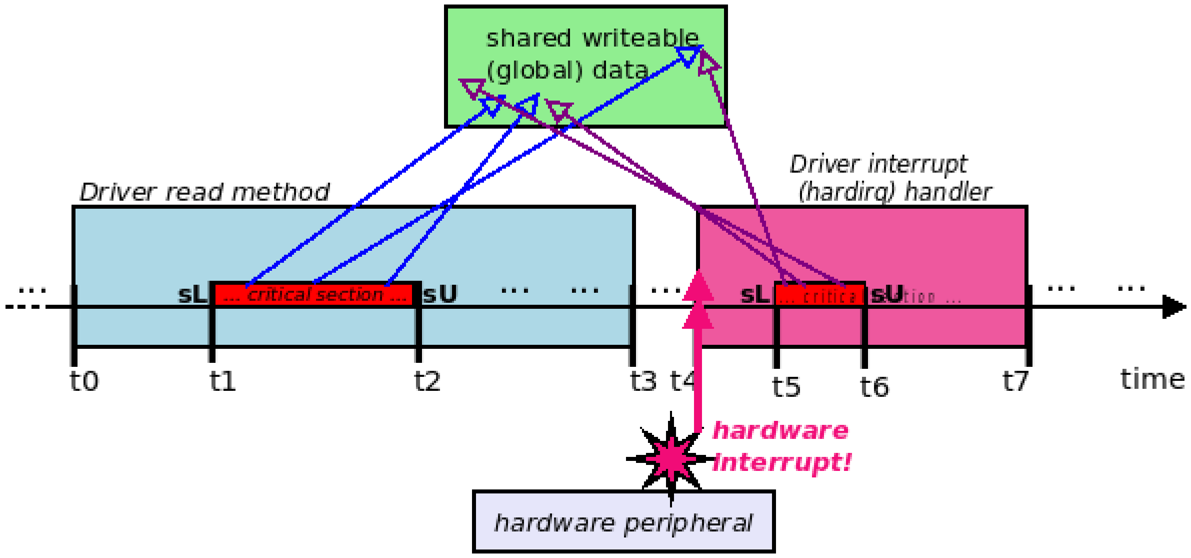
```
[28853.172825] miscdrv_rdwr_spinlock:write_miscdrv_rdwr(): 004)  rdwr_test_secre :23578   |  ...0   /* write_mi
scdrv_rdwr() */
[28853.178231] misc llkd_miscdrv_rdwr_spinlock: rdwr_test_secre wants to write 24 bytes
[28853.181539] misc llkd_miscdrv_rdwr_spinlock:  24 bytes written, returning... (stats: tx=7, rx=24)
[28853.184243] BUG: scheduling while atomic: rdwr_test_secre/23578/0x00000002
[28853.187489] 1 lock held by rdwr_test_secre/23578:
[28853.189904]  #0: ffff8880285c2d60 (&(&ctx->spinlock)->rlock){+.+.}, at: write_miscdrv_rdwr.cold+0xde/0x247 [
miscdrv_rdwr_spinlock]
[28853.195078] Modules linked in: miscdrv_rdwr_spinlock(OE) vboxsf(OE) vboxvideo(OE) crct10dif_pclmul crc32_pcl
mul ghash_clmulni_intel vmwgfx snd_intel8x0 snd_ac97_codec ac97_bus snd_pcm aesni_intel glue_helper crypto_simd
 cryptd joydev snd_seq snd_timer drm_kms_helper snd_seq_device input_leds serio_raw snd syscopyarea sysfillrect
 sysimgblt fb_sys_fops ttm video mac_hid vboxguest(OE) soundcore drm sch_fq_codel parport_pc ppdev lp parport i
p_tables x_tables autofs4 hid_generic usbhid hid psmouse e1000 ahci libahci i2c_piix4 pata_acpi [last unloaded:
 miscdrv_rdwr_spinlock]
[28853.211613] CPU: 4 PID: 23578 Comm: rdwr_test_secre Tainted: G           OE     5.4.0-llkd-dbg #2
[28853.214596] Hardware name: innotek GmbH VirtualBox/VirtualBox, BIOS VirtualBox 12/01/2006
[28853.217244] Call Trace:
[28853.219461]  dump_stack+0xc2/0x11a
[28853.221692]  __schedule_bug.cold+0x2b/0x3c
[28853.223893]  __schedule+0xd4d/0x1090
[28853.226207]  ? firmware_map_remove+0xe9/0xe9
[28853.228428]  ? _raw_spin_unlock_irqrestore+0x51/0x60
[28853.230741]  ? schedule_timeout+0x2b4/0x8c0
[28853.232891]  ? lockdep_hardirqs_on+0x1a2/0x280
[28853.235050]  schedule+0x75/0x140
[28853.237118]  schedule_timeout+0x2b9/0x8c0
[28853.239207]  ? __dev_printk+0xd6/0xf3
[28853.241276]  ? usleep_range+0x100/0x100
[28853.243310]  ? _dev_info+0xcd/0xfb
[28853.245421]  ? __next_timer_interrupt+0xe0/0xe0
[28853.247475]  write_miscdrv_rdwr.cold+0x1ea/0x247 [miscdrv_rdwr_spinlock]
[28853.249726]  ? display_stats+0x80/0x80 [miscdrv_rdwr_spinlock]
[28853.251802]  ? apparmor_file_permission+0x1a/0x20
[28853.253814]  ? security_file_permission+0x65/0x190
[28853.255871]  __vfs_write+0x4f/0x90
[28853.257885]  vfs_write+0x14b/0x2d0
[28853.259744]  ksys_write+0xd9/0x180
[28853.261612]  ? __ia32_sys_read+0x50/0x50
[28853.263388]  ? mark_held_locks+0x29/0xb0
[28853.265119]  ? do_syscall_64+0x19/0x2c0
[28853.266842]  ? entry_SYSCALL_64_after_hwframe+0x49/0xbe
```



```
rdwr_tes-2438    4....  1060.741276: funcgraph_entry:              |  vfs_write() {
rdwr_tes-2438    4....  1060.741276: funcgraph_entry:              |    rw_verify_area() {
rdwr_tes-2438    4....  1060.741277: funcgraph_entry:              |      security_file_permission() {
rdwr_tes-2438    4....  1060.741277: funcgraph_entry:              |        apparmor_file_permission() {
rdwr_tes-2438    4....  1060.741277: funcgraph_entry:              |          common_file_perm() {
rdwr_tes-2438    4....  1060.741277: funcgraph_entry:     0.244 us |            aa_file_perm();
rdwr_tes-2438    4....  1060.741277: funcgraph_exit:      0.492 us |          }
rdwr_tes-2438    4....  1060.741277: funcgraph_exit:      0.715 us |        }
rdwr_tes-2438    4....  1060.741278: funcgraph_exit:      1.010 us |      }
rdwr_tes-2438    4....  1060.741278: funcgraph_exit:      1.273 us |    }
rdwr_tes-2438    4....  1060.741278: funcgraph_entry:              |    __vfs_write() {
rdwr_tes-2438    4....  1060.741278: funcgraph_entry:              |      write_miscdrv_rdwr() {
rdwr_tes-2438    4....  1060.741278: funcgraph_entry:              |        _dev_info() {
rdwr_tes-2438    4....  1060.741278: funcgraph_entry:              |          __dev_printk() {
```



```
rdwr_tes-2438    4....  1060.746698: funcgraph_entry:              |    schedule_timeout() {
rdwr_tes-2438    4....  1060.746698: funcgraph_entry:              |      lock_timer_base() {
rdwr_tes-2438    4....  1060.746698: funcgraph_entry:     0.110 us |        _raw_spin_lock_irqsave();
rdwr_tes-2438    4d...  1060.746698: funcgraph_exit:      0.318 us |      }
rdwr_tes-2438    4d...  1060.746698: funcgraph_entry:     0.104 us |      detach_if_pending();
rdwr_tes-2438    4d...  1060.746699: funcgraph_entry:     0.105 us |      get_nohz_timer_target();
rdwr_tes-2438    4d...  1060.746699: funcgraph_entry:              |      __internal_add_timer() {
rdwr_tes-2438    4d...  1060.746699: funcgraph_entry:     0.110 us |        calc_wheel_index();
rdwr_tes-2438    4d...  1060.746699: funcgraph_entry:     0.161 us |        enqueue_timer();
rdwr_tes-2438    4d...  1060.746699: funcgraph_exit:      0.588 us |      }
rdwr_tes-2438    4d...  1060.746699: funcgraph_entry:     0.106 us |      trigger_dyntick_cpu.isra.0();
rdwr_tes-2438    4d...  1060.746700: funcgraph_entry:     0.117 us |      __lock_text_start();
rdwr_tes-2438    4....  1060.746700: funcgraph_entry:              |      schedule() {
rdwr_tes-2438    4d...  1060.746700: funcgraph_entry:              |        rcu_note_context_switch() {
```



Legend

t0 : driver's read method called

sL : spin_lock(&slock);

t1 : read method enters critical section

t2 : read method leaves critical section

sU : spin_unlock(&slock);

t3 : read method finishes

→ *read method accessing shared writeable data*

t4 : interrupt (hardirq) handler entered

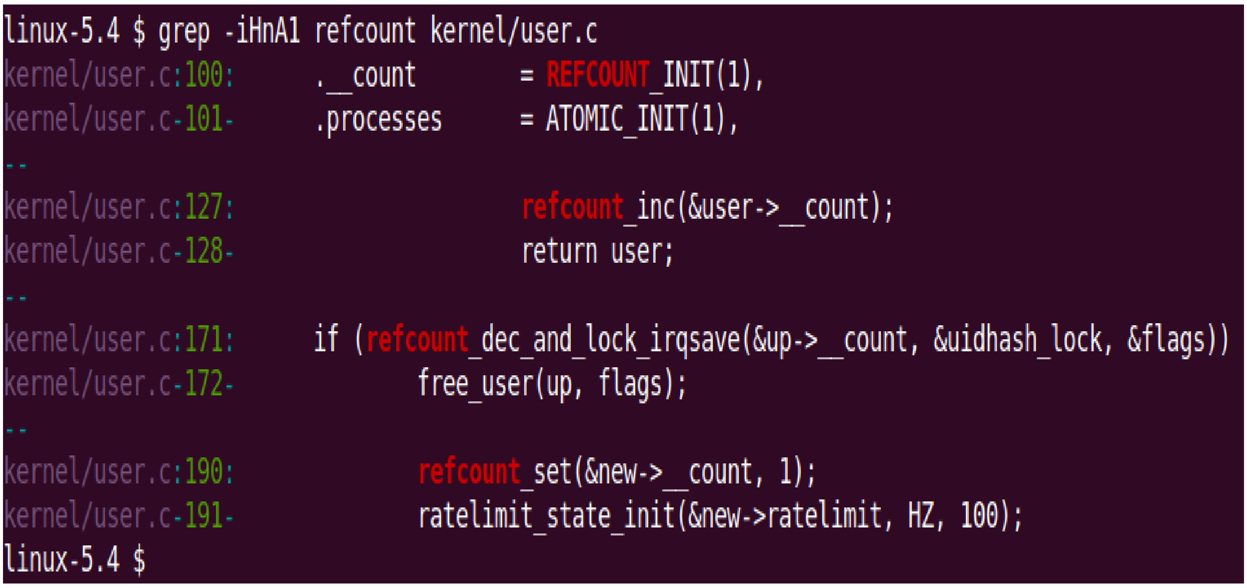
t5 : hardirq enters critical section

t6 : hardirq leaves critical section

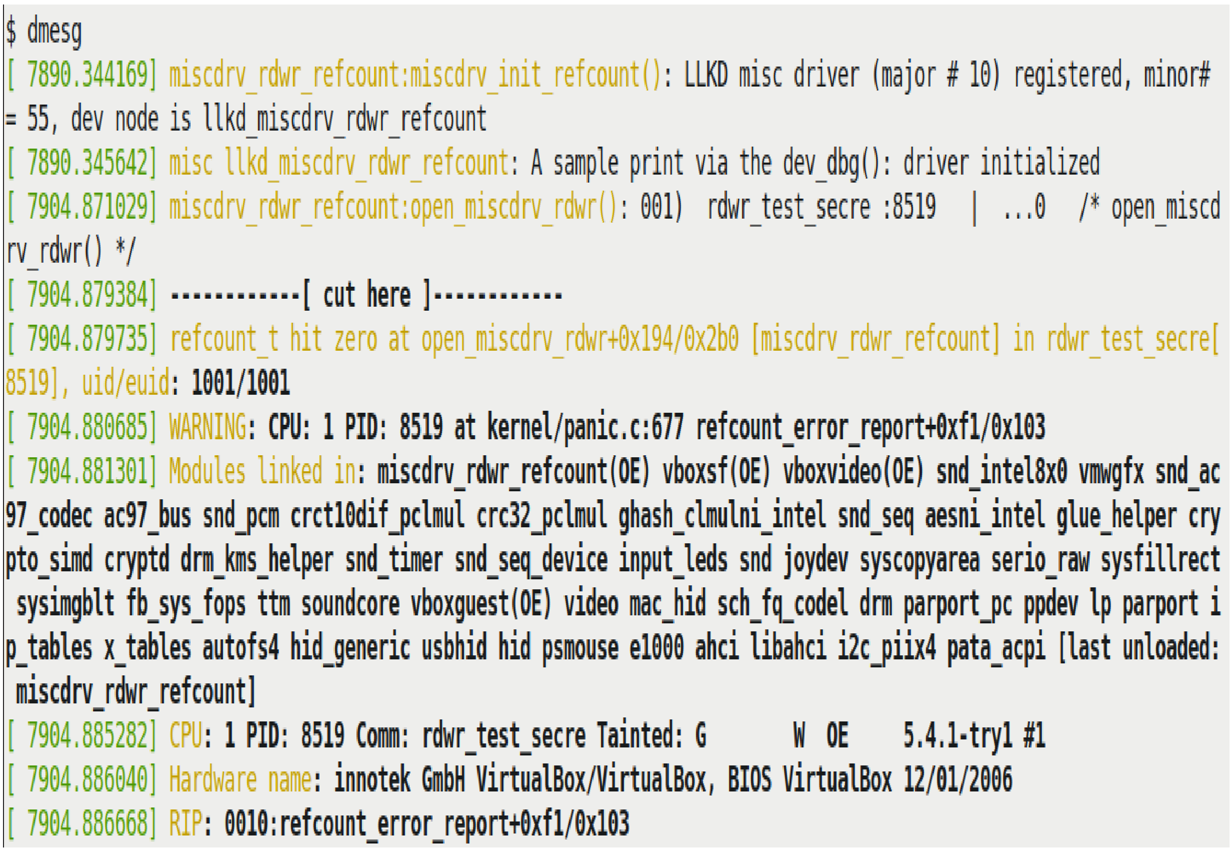
t5 : interrupt (hardirq) handler finishes

→ *hardirq handler accessing shared writeable data*

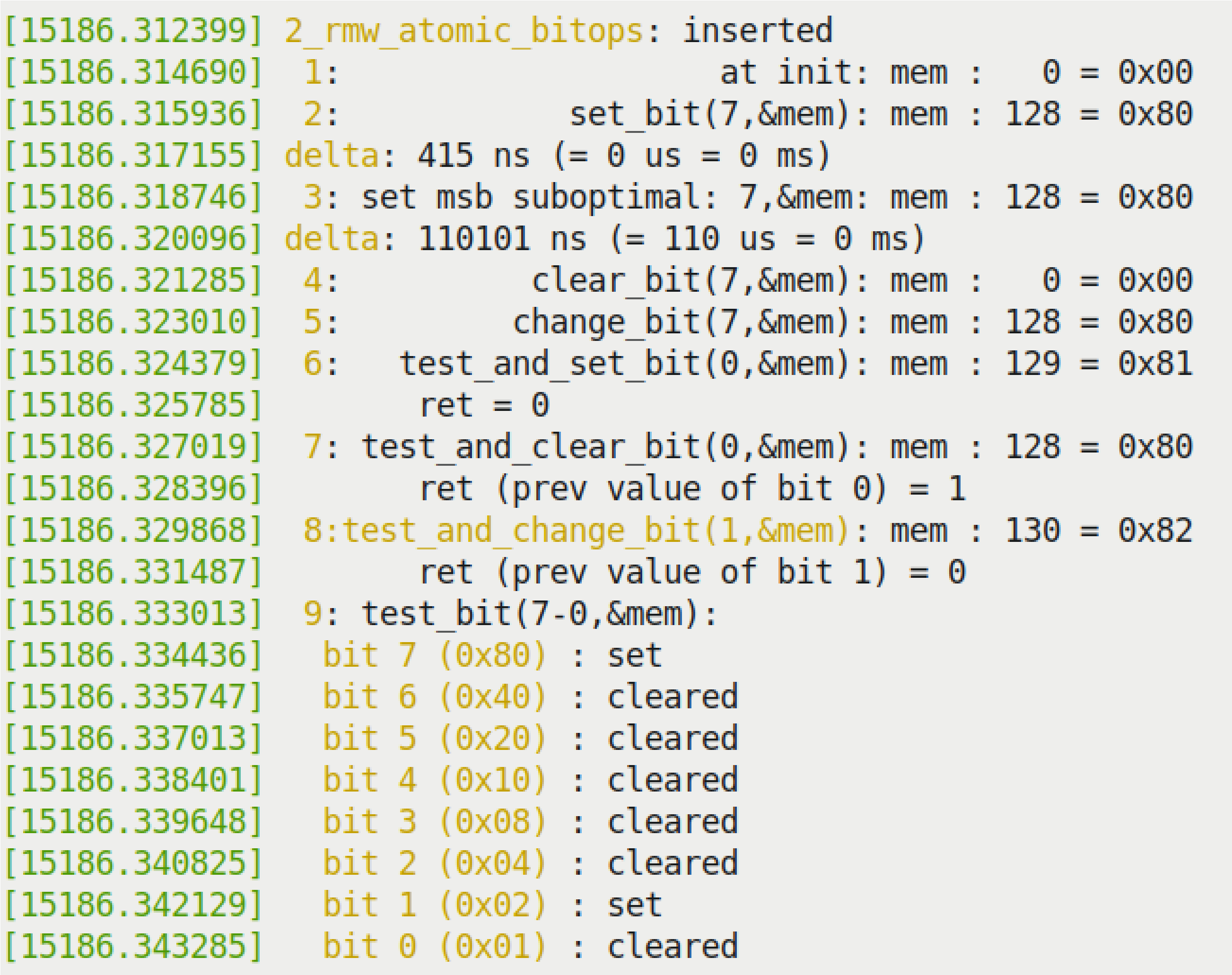
Chapter 7: Kernel Synchronization - Part 2



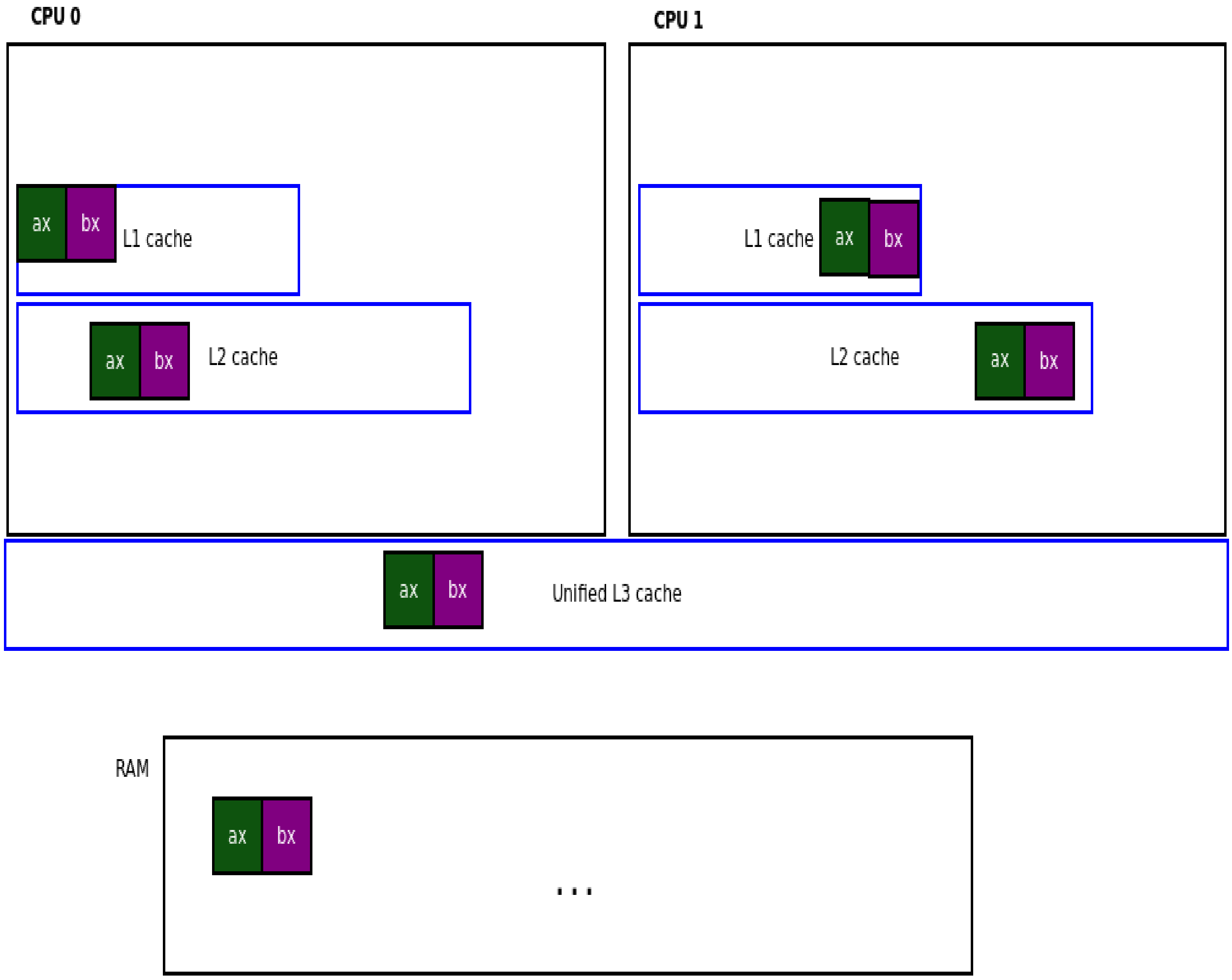
```
linux-5.4 $ grep -iHnA1 refcount kernel/user.c
kernel/user.c:100:      .__count        = REFCOUNT_INIT(1),
kernel/user.c-101-      .processes      = ATOMIC_INIT(1),
--
kernel/user.c:127:                      refcount_inc(&user->__count);
kernel/user.c-128-                      return user;
--
kernel/user.c:171:      if (refcount_dec_and_lock_irqsave(&up->__count, &uidhash_lock, &flags))
kernel/user.c-172-              free_user(up, flags);
--
kernel/user.c:190:              refcount_set(&new->__count, 1);
kernel/user.c-191-              ratelimit_state_init(&new->ratelimit, HZ, 100);
linux-5.4 $
```



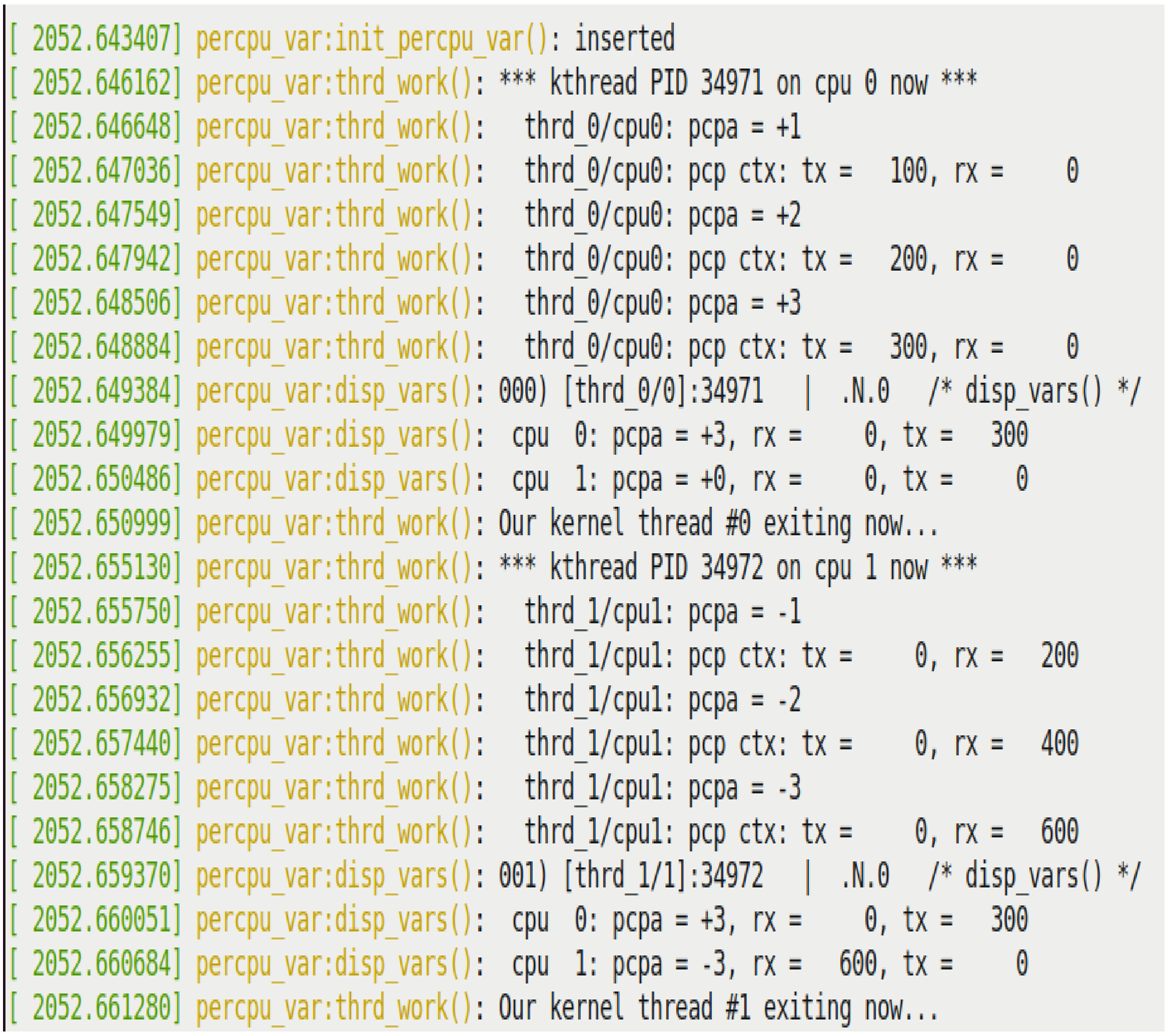
```
$ dmesg
[ 7890.344169] miscdrv_rdwr_refcount:miscdrv_init_refcount(): LLKD misc driver (major # 10) registered, minor#
= 55, dev node is llkd_miscdrv_rdwr_refcount
[ 7890.345642] misc llkd_miscdrv_rdwr_refcount: A sample print via the dev_dbg(): driver initialized
[ 7904.871029] miscdrv_rdwr_refcount:open_miscdrv_rdwr(): 001)  rdwr_test_secre :8519   |  ...0   /* open_miscd
rv_rdwr() */
[ 7904.879384] ------------[ cut here ]------------
[ 7904.879735] refcount_t hit zero at open_miscdrv_rdwr+0x194/0x2b0 [miscdrv_rdwr_refcount] in rdwr_test_secre[
8519], uid/euid: 1001/1001
[ 7904.880685] WARNING: CPU: 1 PID: 8519 at kernel/panic.c:677 refcount_error_report+0xf1/0x103
[ 7904.881301] Modules linked in: miscdrv_rdwr_refcount(OE) vboxsf(OE) vboxvideo(OE) snd_intel8x0 vmwgfx snd_ac
97_codec ac97_bus snd_pcm crct10dif_pclmul crc32_pclmul ghash_clmulni_intel snd_seq aesni_intel glue_helper cry
pto_simd cryptd drm_kms_helper snd_timer snd_seq_device input_leds snd joydev syscopyarea serio_raw sysfillrect
 sysimgblt fb_sys_fops ttm soundcore vboxguest(OE) video mac_hid sch_fq_codel drm parport_pc ppdev lp parport i
p_tables x_tables autofs4 hid_generic usbhid hid psmouse e1000 ahci libahci i2c_piix4 pata_acpi [last unloaded:
 miscdrv_rdwr_refcount]
[ 7904.885282] CPU: 1 PID: 8519 Comm: rdwr_test_secre Tainted: G        W  OE     5.4.1-try1 #1
[ 7904.886040] Hardware name: innotek GmbH VirtualBox/VirtualBox, BIOS VirtualBox 12/01/2006
[ 7904.886668] RIP: 0010:refcount_error_report+0xf1/0x103
```



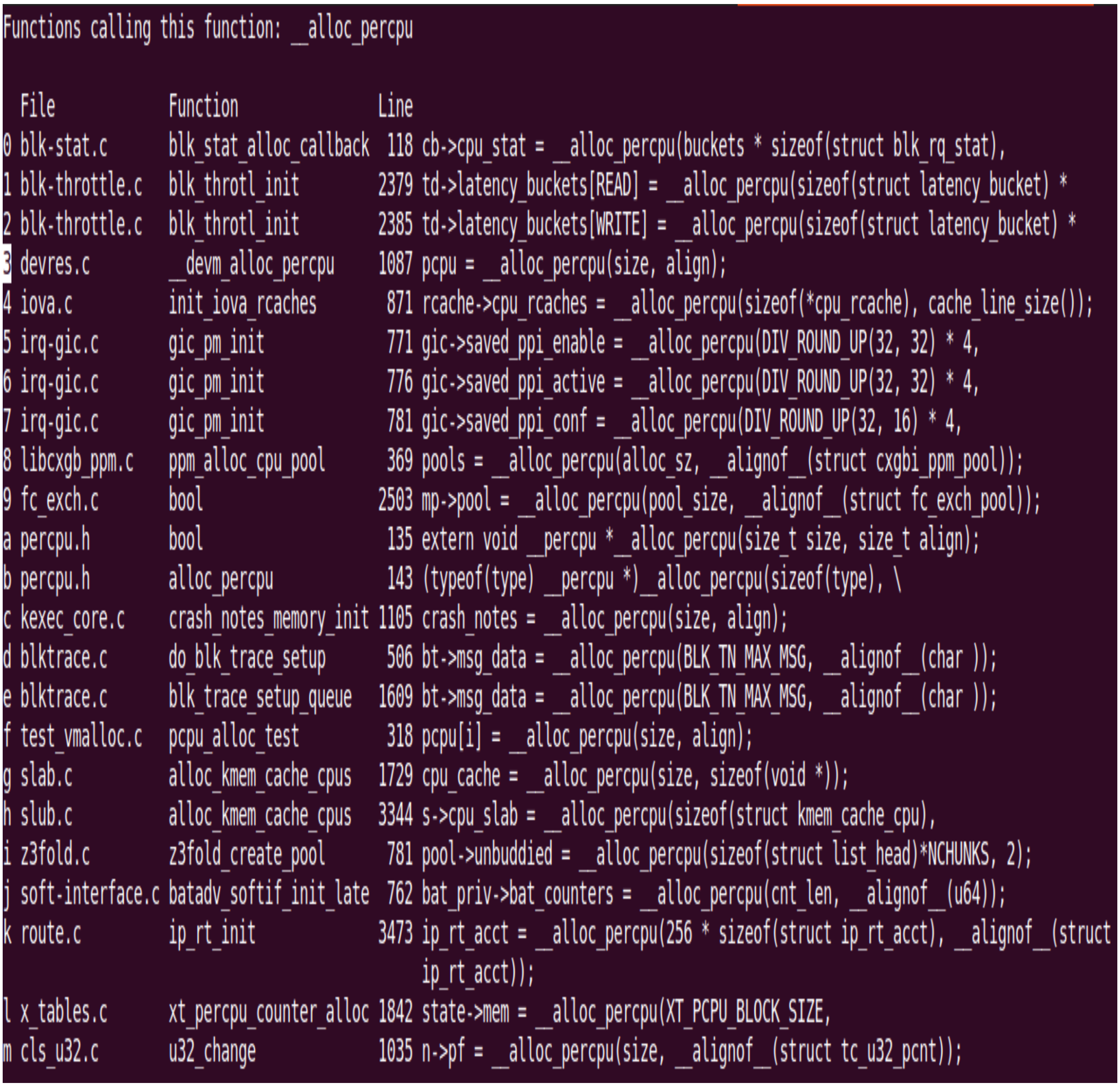
```
[15186.312399] 2_rmw_atomic_bitops: inserted
[15186.314690]  1:                    at init: mem :   0 = 0x00
[15186.315936]  2:            set_bit(7,&mem): mem : 128 = 0x80
[15186.317155] delta: 415 ns (= 0 us = 0 ms)
[15186.318746]  3: set msb suboptimal: 7,&mem: mem : 128 = 0x80
[15186.320096] delta: 110101 ns (= 110 us = 0 ms)
[15186.321285]  4:          clear_bit(7,&mem): mem :   0 = 0x00
[15186.323010]  5:         change_bit(7,&mem): mem : 128 = 0x80
[15186.324379]  6:   test_and_set_bit(0,&mem): mem : 129 = 0x81
[15186.325785]         ret = 0
[15186.327019]  7: test_and_clear_bit(0,&mem): mem : 128 = 0x80
[15186.328396]         ret (prev value of bit 0) = 1
[15186.329868]  8:test_and_change_bit(1,&mem): mem : 130 = 0x82
[15186.331487]         ret (prev value of bit 1) = 0
[15186.333013]  9: test_bit(7-0,&mem):
[15186.334436]   bit 7 (0x80) : set
[15186.335747]   bit 6 (0x40) : cleared
[15186.337013]   bit 5 (0x20) : cleared
[15186.338401]   bit 4 (0x10) : cleared
[15186.339648]   bit 3 (0x08) : cleared
[15186.340825]   bit 2 (0x04) : cleared
[15186.342129]   bit 1 (0x02) : set
[15186.343285]   bit 0 (0x01) : cleared
```



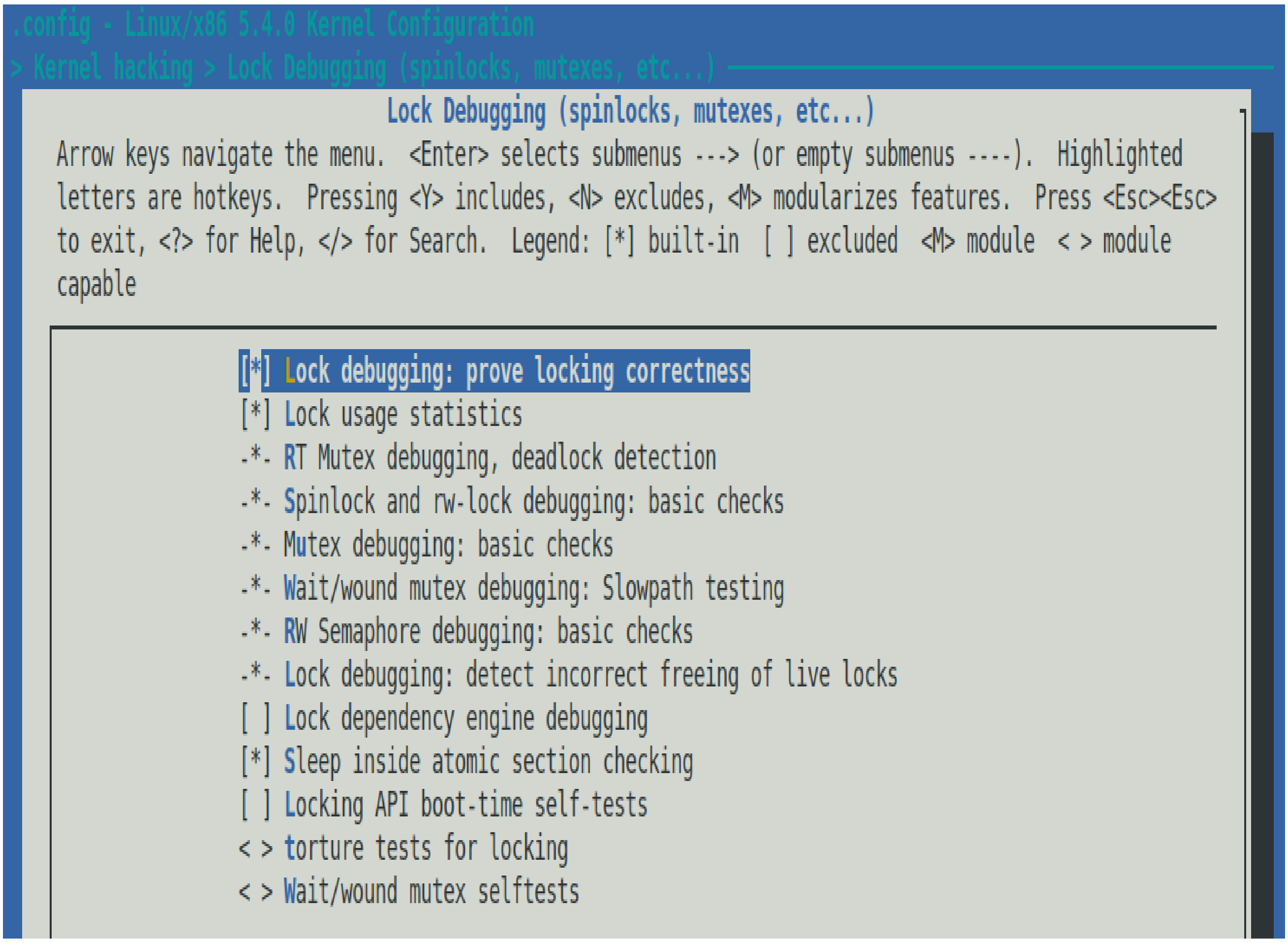
pcpa=0	pcpa=0	pcpa=0	pcpa=0
CPU 0	CPU 1	CPU 2	CPU 3



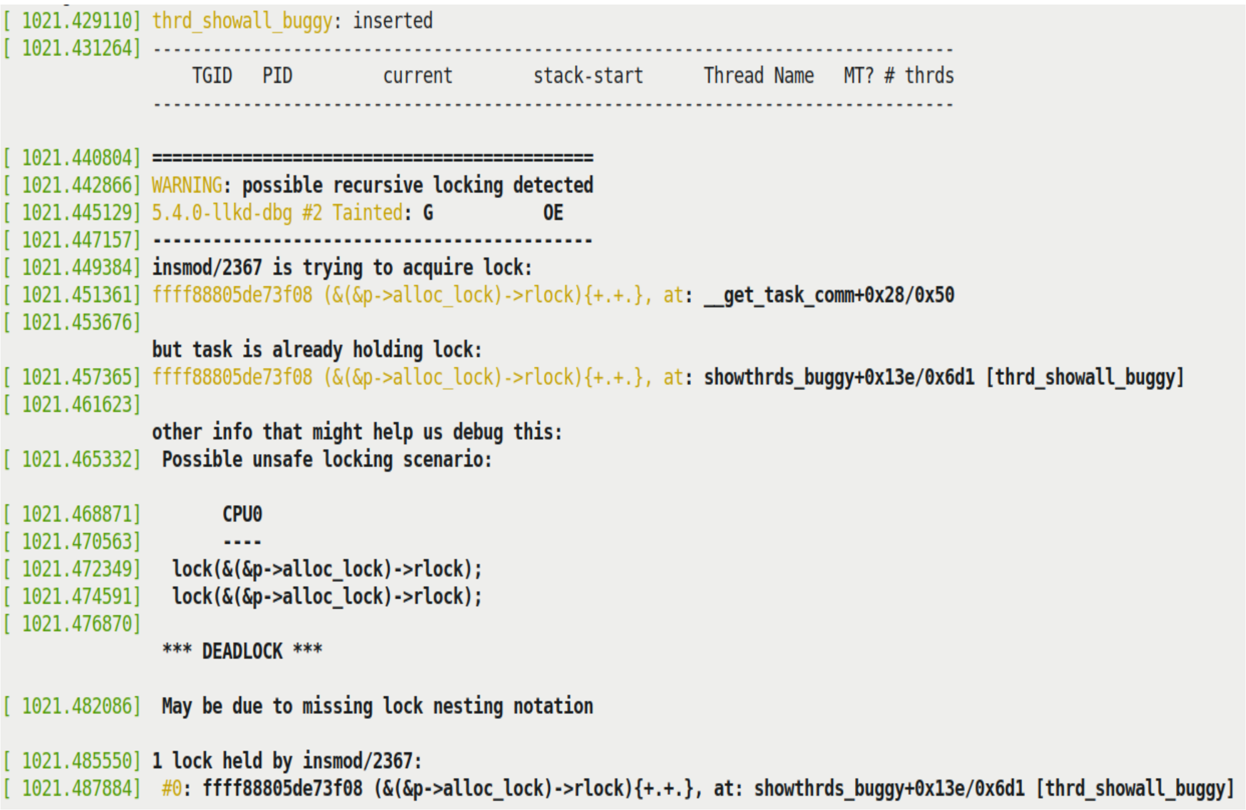
```
[ 2052.643407] percpu_var:init_percpu_var(): inserted
[ 2052.646162] percpu_var:thrd_work(): *** kthread PID 34971 on cpu 0 now ***
[ 2052.646648] percpu_var:thrd_work():   thrd_0/cpu0: pcpa = +1
[ 2052.647036] percpu_var:thrd_work():   thrd_0/cpu0: pcp ctx: tx =   100, rx =     0
[ 2052.647549] percpu_var:thrd_work():   thrd_0/cpu0: pcpa = +2
[ 2052.647942] percpu_var:thrd_work():   thrd_0/cpu0: pcp ctx: tx =   200, rx =     0
[ 2052.648506] percpu_var:thrd_work():   thrd_0/cpu0: pcpa = +3
[ 2052.648884] percpu_var:thrd_work():   thrd_0/cpu0: pcp ctx: tx =   300, rx =     0
[ 2052.649384] percpu_var:disp_vars(): 000) [thrd_0/0]:34971   |  .N.0   /* disp_vars() */
[ 2052.649979] percpu_var:disp_vars():  cpu  0: pcpa = +3, rx =     0, tx =   300
[ 2052.650486] percpu_var:disp_vars():  cpu  1: pcpa = +0, rx =     0, tx =     0
[ 2052.650999] percpu_var:thrd_work(): Our kernel thread #0 exiting now...
[ 2052.655130] percpu_var:thrd_work(): *** kthread PID 34972 on cpu 1 now ***
[ 2052.655750] percpu_var:thrd_work():   thrd_1/cpu1: pcpa = -1
[ 2052.656255] percpu_var:thrd_work():   thrd_1/cpu1: pcp ctx: tx =     0, rx =   200
[ 2052.656932] percpu_var:thrd_work():   thrd_1/cpu1: pcpa = -2
[ 2052.657440] percpu_var:thrd_work():   thrd_1/cpu1: pcp ctx: tx =     0, rx =   400
[ 2052.658275] percpu_var:thrd_work():   thrd_1/cpu1: pcpa = -3
[ 2052.658746] percpu_var:thrd_work():   thrd_1/cpu1: pcp ctx: tx =     0, rx =   600
[ 2052.659370] percpu_var:disp_vars(): 001) [thrd_1/1]:34972   |  .N.0   /* disp_vars() */
[ 2052.660051] percpu_var:disp_vars():  cpu  0: pcpa = +3, rx =     0, tx =   300
[ 2052.660684] percpu_var:disp_vars():  cpu  1: pcpa = -3, rx =   600, tx =     0
[ 2052.661280] percpu_var:thrd_work(): Our kernel thread #1 exiting now...
```



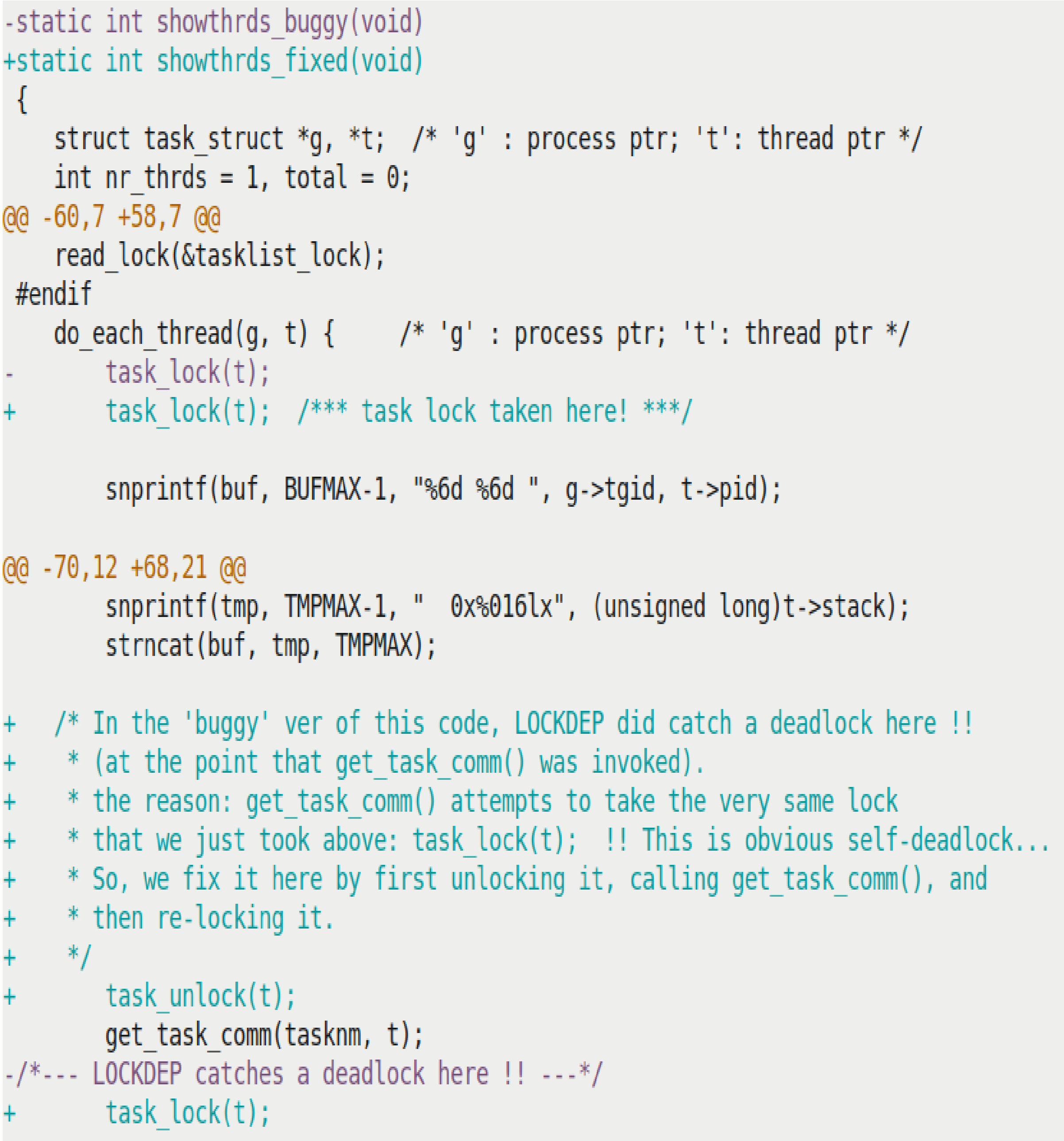
```
Functions calling this function: __alloc_percpu


  File             Function                Line
0 blk-stat.c       blk_stat_alloc_callback  118 cb->cpu_stat = __alloc_percpu(buckets * sizeof(struct blk_rq_stat),
1 blk-throttle.c   blk_throtl_init         2379 td->latency_buckets[READ] = __alloc_percpu(sizeof(struct latency_bucket) *
2 blk-throttle.c   blk_throtl_init         2385 td->latency_buckets[WRITE] = __alloc_percpu(sizeof(struct latency_bucket) *
3 devres.c         __devm_alloc_percpu     1087 pcpu = __alloc_percpu(size, align);
4 iova.c           init_iova_rcaches        871 rcache->cpu_rcaches = __alloc_percpu(sizeof(*cpu_rcache), cache_line_size());
5 irq-gic.c        gic_pm_init              771 gic->saved_ppi_enable = __alloc_percpu(DIV_ROUND_UP(32, 32) * 4,
6 irq-gic.c        gic_pm_init              776 gic->saved_ppi_active = __alloc_percpu(DIV_ROUND_UP(32, 32) * 4,
7 irq-gic.c        gic_pm_init              781 gic->saved_ppi_conf = __alloc_percpu(DIV_ROUND_UP(32, 16) * 4,
8 libcxgb_ppm.c    ppm_alloc_cpu_pool       369 pools = __alloc_percpu(alloc_sz, __alignof__(struct cxgbi_ppm_pool));
9 fc_exch.c        bool                    2503 mp->pool = __alloc_percpu(pool_size, __alignof__(struct fc_exch_pool));
a percpu.h         bool                     135 extern void __percpu *__alloc_percpu(size_t size, size_t align);
b percpu.h         alloc_percpu             143 (typeof(type) __percpu *)__alloc_percpu(sizeof(type), \
c kexec_core.c     crash_notes_memory_init 1105 crash_notes = __alloc_percpu(size, align);
d blktrace.c       do_blk_trace_setup       506 bt->msg_data = __alloc_percpu(BLK_TN_MAX_MSG, __alignof__(char ));
e blktrace.c       blk_trace_setup_queue   1609 bt->msg_data = __alloc_percpu(BLK_TN_MAX_MSG, __alignof__(char ));
f test_vmalloc.c   pcpu_alloc_test          318 pcpu[i] = __alloc_percpu(size, align);
g slab.c           alloc_kmem_cache_cpus   1729 cpu_cache = __alloc_percpu(size, sizeof(void *));
h slub.c           alloc_kmem_cache_cpus   3344 s->cpu_slab = __alloc_percpu(sizeof(struct kmem_cache_cpu),
i z3fold.c         z3fold_create_pool       781 pool->unbuddied = __alloc_percpu(sizeof(struct list_head)*NCHUNKS, 2);
j soft-interface.c batadv_softif_init_late  762 bat_priv->bat_counters = __alloc_percpu(cnt_len, __alignof__(u64));
k route.c          ip_rt_init              3473 ip_rt_acct = __alloc_percpu(256 * sizeof(struct ip_rt_acct), __alignof__(struct
                                                ip_rt_acct));
l x_tables.c       xt_percpu_counter_alloc 1842 state->mem = __alloc_percpu(XT_PCPU_BLOCK_SIZE,
m cls_u32.c        u32_change              1035 n->pf = __alloc_percpu(size, __alignof__(struct tc_u32_pcnt));
```



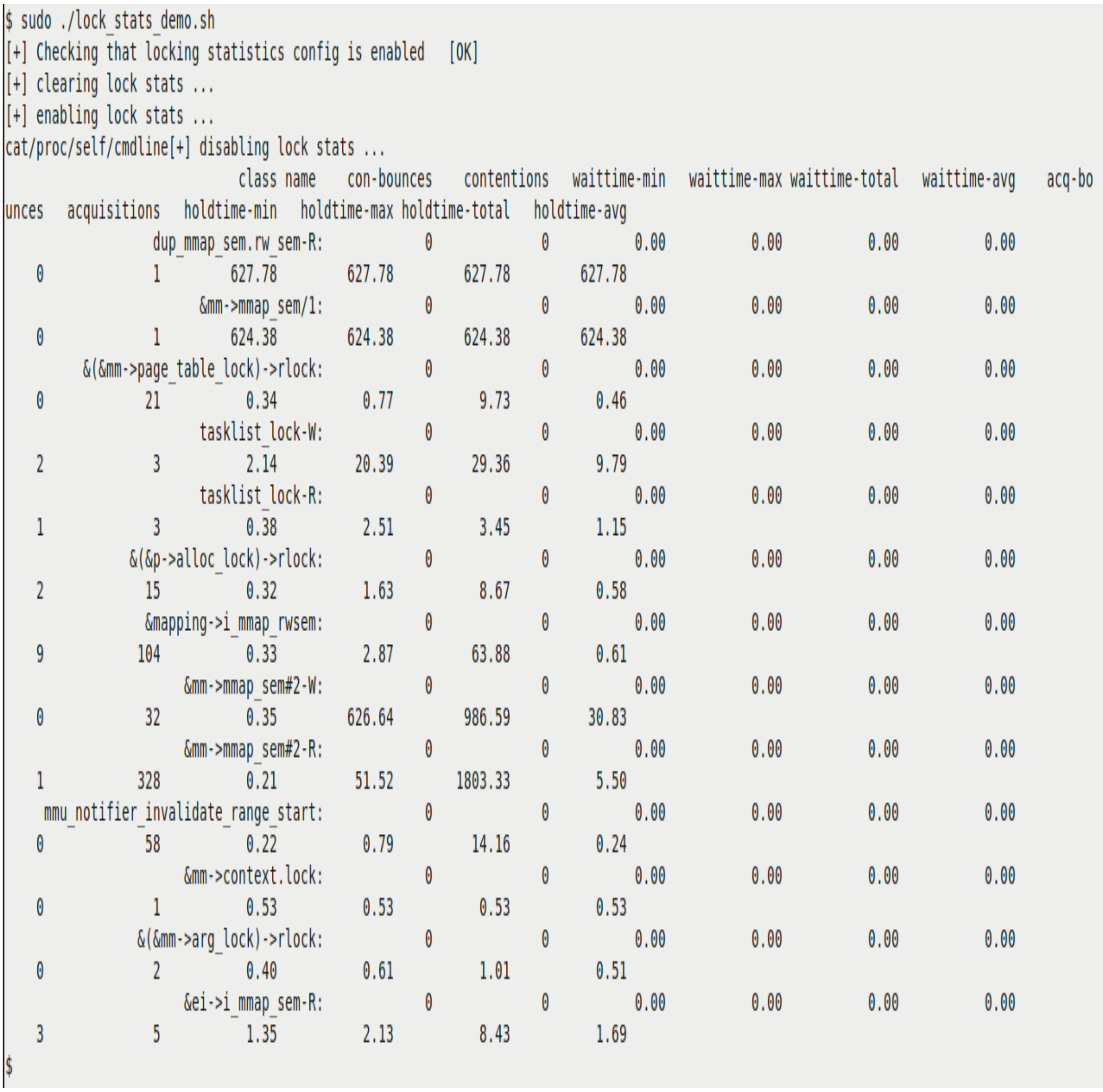
```
.config - Linux/x86 5.4.0 Kernel Configuration
> Kernel hacking > Lock Debugging (spinlocks, mutexes, etc...) ──────────────
                    Lock Debugging (spinlocks, mutexes, etc...)
  Arrow keys navigate the menu.  <Enter> selects submenus ---> (or empty submenus ----).  Highlighted
  letters are hotkeys.  Pressing <Y> includes, <N> excludes, <M> modularizes features.  Press <Esc><Esc>
  to exit, <?> for Help, </> for Search.  Legend: [*] built-in  [ ] excluded  <M> module  < > module
  capable

            [*] Lock debugging: prove locking correctness
            [*] Lock usage statistics
            -*- RT Mutex debugging, deadlock detection
            -*- Spinlock and rw-lock debugging: basic checks
            -*- Mutex debugging: basic checks
            -*- Wait/wound mutex debugging: Slowpath testing
            -*- RW Semaphore debugging: basic checks
            -*- Lock debugging: detect incorrect freeing of live locks
            [ ] Lock dependency engine debugging
            [*] Sleep inside atomic section checking
            [ ] Locking API boot-time self-tests
            < > torture tests for locking
            < > Wait/wound mutex selftests
```



```
[ 1021.429110] thrd_showall_buggy: inserted
[ 1021.431264] --------------------------------------------------------------------------------
                   TGID   PID         current        stack-start      Thread Name    MT? # thrds
               --------------------------------------------------------------------------------

[ 1021.440804] ============================================
[ 1021.442866] WARNING: possible recursive locking detected
[ 1021.445129] 5.4.0-llkd-dbg #2 Tainted: G           OE
[ 1021.447157] --------------------------------------------
[ 1021.449384] insmod/2367 is trying to acquire lock:
[ 1021.451361] ffff88805de73f08 (&(&p->alloc_lock)->rlock){+.+.}, at: __get_task_comm+0x28/0x50
[ 1021.453676]
               but task is already holding lock:
[ 1021.457365] ffff88805de73f08 (&(&p->alloc_lock)->rlock){+.+.}, at: showthrds_buggy+0x13e/0x6d1 [thrd_showall_buggy]
[ 1021.461623]
               other info that might help us debug this:
[ 1021.465332]  Possible unsafe locking scenario:

[ 1021.468871]        CPU0
[ 1021.470563]        ----
[ 1021.472349]   lock(&(&p->alloc_lock)->rlock);
[ 1021.474591]   lock(&(&p->alloc_lock)->rlock);
[ 1021.476870]
                *** DEADLOCK ***

[ 1021.482086]  May be due to missing lock nesting notation

[ 1021.485550] 1 lock held by insmod/2367:
[ 1021.487884]  #0: ffff88805de73f08 (&(&p->alloc_lock)->rlock){+.+.}, at: showthrds_buggy+0x13e/0x6d1 [thrd_showall_buggy]
```



```
-static int showthrds_buggy(void)
+static int showthrds_fixed(void)
 {
    struct task_struct *g, *t;  /* 'g' : process ptr; 't': thread ptr */
    int nr_thrds = 1, total = 0;
@@ -60,7 +58,7 @@
    read_lock(&tasklist_lock);
 #endif
    do_each_thread(g, t) {     /* 'g' : process ptr; 't': thread ptr */
-       task_lock(t);
+       task_lock(t);  /*** task lock taken here! ***/

        snprintf(buf, BUFMAX-1, "%6d %6d ", g->tgid, t->pid);

@@ -70,12 +68,21 @@
        snprintf(tmp, TMPMAX-1, "  0x%016lx", (unsigned long)t->stack);
        strncat(buf, tmp, TMPMAX);

+   /* In the 'buggy' ver of this code, LOCKDEP did catch a deadlock here !!
+    * (at the point that get_task_comm() was invoked).
+    * the reason: get_task_comm() attempts to take the very same lock
+    * that we just took above: task_lock(t);  !! This is obvious self-deadlock...
+    * So, we fix it here by first unlocking it, calling get_task_comm(), and
+    * then re-locking it.
+    */
+       task_unlock(t);
        get_task_comm(tasknm, t);
-/*--- LOCKDEP catches a deadlock here !! ---*/
+       task_lock(t);
```



```
$ sudo ./lock_stats_demo.sh
[+] Checking that locking statistics config is enabled   [OK]
[+] clearing lock stats ...
[+] enabling lock stats ...
cat/proc/self/cmdline[+] disabling lock stats ...
                              class name    con-bounces    contentions   waittime-min   waittime-max waittime-total   waittime-avg    acq-bo
unces   acquisitions   holdtime-min   holdtime-max holdtime-total   holdtime-avg
                   dup_mmap_sem.rw_sem-R:             0             0          0.00          0.00          0.00          0.00
    0              1         627.78         627.78         627.78         627.78
                         &mm->mmap_sem/1:             0             0          0.00          0.00          0.00          0.00
    0              1         624.38         624.38         624.38         624.38
          &(&mm->page_table_lock)->rlock:             0             0          0.00          0.00          0.00          0.00
    0             21           0.34           0.77           9.73           0.46
                         tasklist_lock-W:             0             0          0.00          0.00          0.00          0.00
    2              3           2.14          20.39          29.36           9.79
                         tasklist_lock-R:             0             0          0.00          0.00          0.00          0.00
    1              3           0.38           2.51           3.45           1.15
                &(&p->alloc_lock)->rlock:             0             0          0.00          0.00          0.00          0.00
    2             15           0.32           1.63           8.67           0.58
                  &mapping->i_mmap_rwsem:             0             0          0.00          0.00          0.00          0.00
    9            104           0.33           2.87          63.88           0.61
                       &mm->mmap_sem#2-W:             0             0          0.00          0.00          0.00          0.00
    0             32           0.35         626.64         986.59          30.83
                       &mm->mmap_sem#2-R:             0             0          0.00          0.00          0.00          0.00
    1            328           0.21          51.52        1803.33           5.50
     mmu_notifier_invalidate_range_start:             0             0          0.00          0.00          0.00          0.00
    0             58           0.22           0.79          14.16           0.24
                       &mm->context.lock:             0             0          0.00          0.00          0.00          0.00
    0              1           0.53           0.53           0.53           0.53
                 &(&mm->arg_lock)->rlock:             0             0          0.00          0.00          0.00          0.00
    0              2           0.40           0.61           1.01           0.51
                       &ei->i_mmap_sem-R:             0             0          0.00          0.00          0.00          0.00
    3              5           1.35           2.13           8.43           1.69
$
```