CS380L: Advanced Operating Systems Lab #1

Zeyuan Hu [1], iamzeyuanhu@utexas.edu

EID:zh4378 Spring 2019

1 Environment

We use a Linux server for all the experiments. The server has 4 Intel(R) Xeon(R) CPU E3-1220 v5 @ 3.00GHz processors and 16GB of memory, and runs Ubuntu 16.04.2 LTS (kernel version 4.11.0). The CPU has 32KB of L1 data cache per core (8-way set associative) (found through `getconf -a | grep CACHE`). In addition, it has two-level TLBs. The first level (data TLB) has 64 entries (4-way set associative), and the second level has 1536 entries for both instructions and data (6-way set associative) (found through `cpuid | grep -i tlb`).

2 Memory map

`/proc/[pid]/maps` file contains process `[pid]`'s mapped memory regions and their access permissions [1]. We use the following code to read content of `/proc/self/maps` file [2]:

```
sprintf(filepath, "/proc/%u/maps", (unsigned)getpid());
FILE *f = fopen(filepath, "r");

printf("%-32s %-8s %-10s %-8s %-10s %s\n", "address", "perms", "offset", "dev", "
    inode", "pathname");
while (fgets(line, sizeof(line), f) != NULL) {
     sscanf(line, "%s%s%s%s%s%s", address, perms, offset, dev, inode, pathname);
     printf("%-32s %-8s %-10s %-8s %-10s %s\n", address, perms, offset, dev, inode,
          pathname);
}
fclose(f);
```

In the file, each line corresponds to a mapped memory region. There are six columns of each line, which represent six properties of the mapped memory region: `address`, `perms`, `offset`, `dev`, `inode`, and `pathname`. The result of running `memory_map.c` is below:

The `address` field gives range of virtual memory address of the mapped memory region. Access permission of each memory region is indicated by `perms` field. There are four bits in the field: `rwx`

[1] 30 hours spent on this lab.

[2] see `memory_map.c` for complete code

```
address                           perms offset   dev   inode    pathname
00400000-00401000                 r-xp  00000000 fd:01 12374202 /home/zeyuanhu/380L-Spring19/lab1/src/a.out
00600000-00601000                 r--p  00000000 fd:01 12374202 /home/zeyuanhu/380L-Spring19/lab1/src/a.out
00601000-00602000                 rw-p  00001000 fd:01 12374202 /home/zeyuanhu/380L-Spring19/lab1/src/a.out
022c1000-022e2000                 rw-p  00000000 00:00 0        [heap]
7fea45315000-7fea454d5000         r-xp  00000000 fd:01 24903836 /lib/x86_64-linux-gnu/libc-2.23.so
7fea454d5000-7fea456d5000         ---p  001c0000 fd:01 24903836 /lib/x86_64-linux-gnu/libc-2.23.so
7fea456d5000-7fea456d9000         r--p  001c0000 fd:01 24903836 /lib/x86_64-linux-gnu/libc-2.23.so
7fea456d9000-7fea456db000         rw-p  001c4000 fd:01 24903836 /lib/x86_64-linux-gnu/libc-2.23.so
7fea456db000-7fea456df000         rw-p  00000000 00:00 0        /lib/x86_64-linux-gnu/libc-2.23.so
7fea456df000-7fea45705000         r-xp  00000000 fd:01 24903834 /lib/x86_64-linux-gnu/ld-2.23.so
7fea458e0000-7fea458e3000         rw-p  00000000 00:00 0        /lib/x86_64-linux-gnu/ld-2.23.so
7fea45904000-7fea45905000         r--p  00025000 fd:01 24903834 /lib/x86_64-linux-gnu/ld-2.23.so
7fea45905000-7fea45906000         rw-p  00026000 fd:01 24903834 /lib/x86_64-linux-gnu/ld-2.23.so
7fea45906000-7fea45907000         rw-p  00000000 00:00 0        /lib/x86_64-linux-gnu/ld-2.23.so
7ffe67d7e000-7ffe67d9f000         rw-p  00000000 00:00 0        [stack]
7ffe67da3000-7ffe67da5000         r--p  00000000 00:00 0        [vvar]
7ffe67da5000-7ffe67da7000         r-xp  00000000 00:00 0        [vdso]
ffffffffff600000-ffffffffff601000 r-xp  00000000 00:00 0        [vsyscall]
```

Figure 1: Output of memory_map.c

represents read, write, and executable respectively; the last bit (`p` or `s`) represents whether the region is private or shared. `offset` field represents the offset in the mapped file. `dev` field indicates the device (represented with format of `major:minor`) that the mapped file resides . There are two kinds of value in this column for our case: `fd:01` and `00:00`. The former one is the device id (in hex) of `/` (checked with `mountpoint -d /`) and the latter one represents no device associated with the file. `inode` field represents the inode number of the file on the device. `0` means no file is associated with the mapped memory region. `pathname` field gives the absolute path to the file associated with the mapped memory region. It can be some special values like `[heap]`, `[stack]`, `[vdso]`, etc.

To locate the start of the text section of the executable, we invoke `objdump -h` on the binary and get `0000000000400600`. Output of `/proc/self/maps` shows that the start address of `libc` is `7fea45315000`. The reason for these two addresses are different is `libc` is dynamic loaded library, which is loaded during the runtime of executable, which is not compiled and linked as part of executable. The code segment contains the executable instruction, not the dynamic loaded library.

One interesting thing happens between runs of the executable: the content of `/proc/self/maps` is different. Addresses of all mapped memory regions are different except for the regions mapped to the executable and `[vsyscall]`. The root cause behind this phenomenon is Address Space Layout Randomization (ASLR) [2] for programs in user space. This feature is enabled by default and can be seen via the content of `/proc/sys/kernel/randomize_va_space` file. In our case, the value is 2, which means the positions of stack itself, virtual dynamic shared object (VDSO) page, shared memory regions, and data segments are randomized [3].

3 Obtaining and building the kernel

We first obtain the Linux Kernel source via `wget https://cdn.kernel.org/pub/linux/kernel/v4.x/linux-4.20.4.tar.xz`. Then, we extract the files using `tar -xJf linux-4.20.4.tar.xz`. We make a new directory `kbuild` as the build directory for kernel and `cd kbuild`, we generate `.config` file using `yes "" | make -C ../linux-4.20.4/ O=$(pwd)x86_64_defconfig`. Note that generating the `.config` file like this automatically set `CONFIG_SATA_AHCI=y`. We run `make -j4` [3]to build the kernel.

4 Installing and Copying Kernel Modules

We install the newly-built kernel by first making a new directory called `kinstall` as a sibling of `kbuild`. `kinstall` will contain the built kernel modules. Inside `kbuild`, we run `make INSTALL_MOD_PATH =../kinstall modules_install`.

We can see `lib` directory inside `kinstall`, which has to be copied to the root file system of the VM. We notice there are two symbolic links `build` and `source` inside `kinstall/lib/modules/4.20.4`, which links to the built kernel image and the source of the kernel. They are useless and may cause problems when copying files to the VM. Thus we just delete them. Next, we copy the entire 4.20.4 directory to `/lib/modules` in the guest system by doing `scp -P 4444 -r 4.20.4/ ubuntu@localhost:/home/ubuntu` and inside the guest sytem, do `sudo mv 4.20.4/ /lib/modules /`.

5 Booting KVM with your new Kernel

We can now start VM with our own Linux kernel. The shell command we run now:

```
qemu-system-x86_64 \
-enable-kvm -curses \
-m 512 -smp 4 -redir tcp:4444::22 \
-hda my-disk.img -hdb my-seed.img \
-kernel ~/3801-lab0/kbuild/arch/x86_64/boot/bzImage \
-append "root=/dev/sda1" \
-cpu host
```

Note that we append two new options `-kernel` and `-append` to `QEMU`. `-kernel` option tells the location of the kernel to use, and `-append` option suggests the parameters to start the kernel. The

[3]`-j4` means 4 threads are used, which can speed up the build process

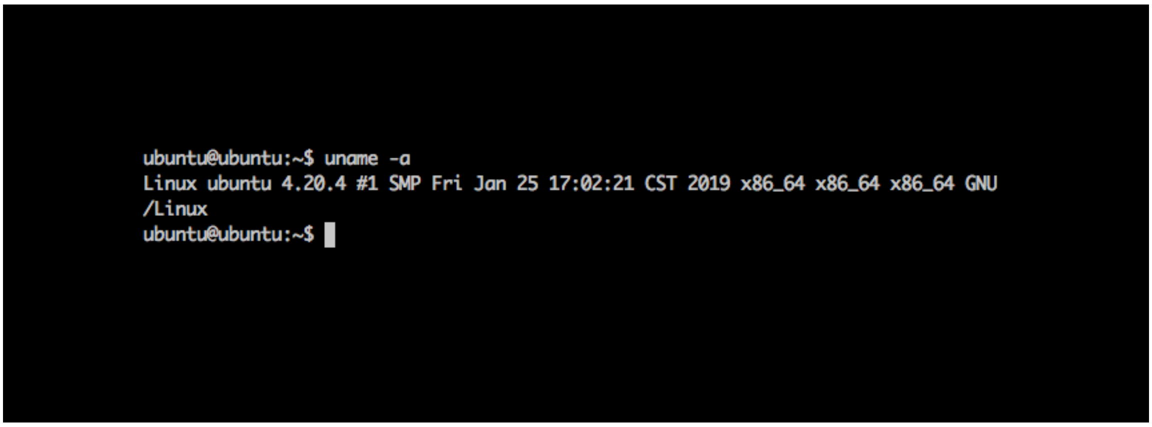


Figure 2: VM with our newly-built kernel

`root` parameter suggests the disk partition used as root file system. After login, use `uname -a` to check the kernel version string, which is shown in Figure 2.

6 Booting, kernel modules, and discovering devices

The wall clock time (tracked using a stopwatch) for our boot takes 34.08 seconds while the time reported by the Kernel takes 28.81 seconds. This difference may be due to the human delay on stopping the stopwatch and also due to a disagreement between human and OS on how to define boot finish status. Here, we stop our stopwatch when we see the login prompt but the last line of `dmesg` [4] shows:

```
[ 28.811823] new mount options do not match the existing superblock, will be
    ignored
```

To eliminate the potential human error, we use real-time clock in Linux system to time the difference between the wall clock time and the time reported by Kernel.

```
$ dmesg -T | grep "RTC time"
[Fri Jan 25 23:54:33 2019] RTC time: 23:54:32, date: 01/25/19
```

`RTC` stands for "real-time clocks" [5]. We find that the time reported by Kernel is 1 second slower than the real-time clock at that moment. "RTC vs system clock" section in `man rtc` explains possible root cause for this 1 second difference: when the system is in a low power state, only RTC work not the system clock. The system clock is mantained by kernel implemented as counting of timer interrupts and the system clock will set to the wall clock time once the system boots and out of low power state. Thus, one possible explanation of the 1 second difference is due to the slower

[4] `dmesg` is used to inspect the kernel ring buffer, which contains the system log during kernel boot.
[5] definition of RTC can be found via `man rtc`

frequency of timer interrupts and another possible explanation is because the system clock has not aligned well with the wall clock time yet.

We also inspect the discovery of PCI devices at boot time from the boot log. We use the command `lspci` and there are 6 PCI devices in the VM:

```
$ lspci
00:00.0 Host bridge: Intel Corporation 440FX - 82441FX PMC [Natoma] (rev 02)
00:01.0 ISA bridge: Intel Corporation 82371SB PIIX3 ISA [Natoma/Triton II]
00:01.1 IDE interface: Intel Corporation 82371SB PIIX3 IDE [Natoma/Triton II]
00:01.3 Bridge: Intel Corporation 82371AB/EB/MB PIIX4 ACPI (rev 03)
00:02.0 VGA compatible controller: Device 1234:1111 (rev 02)
00:03.0 Ethernet controller: Intel Corporation 82540EM Gigabit Ethernet Controller
     (rev 03)
```

We can search the boot log with the pattern of `0000:ID` (e.g., `0000:00:00.0`)from `lspci` to learn how the kernel discovers and identifies these devices during the boot process and the log message helps us to decide what kind of the device is.

```
$ dmesg | grep "0000:00:00.0"
[ 0.244811] pci 0000:00:00.0: [8086:1237] type 00 class 0x060000
[ 0.579080] pci 0000:00:00.0: Limiting direct PCI/PCI transfers

$ dmesg | grep "0000:00:01.0"
[ 0.245549] pci 0000:00:01.0: [8086:7000] type 00 class 0x060100
[ 0.578484] pci 0000:00:01.0: PIIX3: Enabling Passive Release
[ 0.586375] pci 0000:00:01.0: Activating ISA DMA hang workarounds

$ dmesg | grep "0000:00:01.1"
[ 0.246566] pci 0000:00:01.1: [8086:7010] type 00 class 0x010180
[ 0.250524] pci 0000:00:01.1: reg 0x20: [io 0xc040-0xc04f]
[ 0.252018] pci 0000:00:01.1: legacy IDE quirk: reg 0x10: [io 0x01f0-0x01f7]
<-- snip -->

$ dmesg | grep "0000:00:01.3"
[ 0.256256] pci 0000:00:01.3: [8086:7113] type 00 class 0x068000
[ 0.257044] pci 0000:00:01.3: quirk: [io 0x0600-0x063f] claimed by PIIX4 ACPI
[ 0.257208] pci 0000:00:01.3: quirk: [io 0x0700-0x070f] claimed by PIIX4 SMB
```

```
$ dmesg | grep "0000:00:02.0"
[ 0.258317] pci 0000:00:02.0: [1234:1111] type 00 class 0x030000
[ 0.259810] pci 0000:00:02.0: reg 0x10: [mem 0xfd000000-0xfdffffff pref]
[ 0.262214] pci 0000:00:02.0: reg 0x18: [mem 0xfebb0000-0xfebb0fff]
<-- snip -->

$ dmesg | grep "0000:00:03.0"
[ 0.267327] pci 0000:00:03.0: [8086:100e] type 00 class 0x020000
[ 0.268194] pci 0000:00:03.0: reg 0x10: [mem 0xfeb80000-0xfeb9ffff]
[ 0.268973] pci 0000:00:03.0: reg 0x14: [io 0xc000-0xc03f]
<-- snip -->
```

7 Tracing the kernel

7.1 Make a debug build

To trace the kernel, we need to make a debug build of the kernel by modifying several debug options. Make a new directory `debug_bld2` for holding the debug build. In the created directory, run

```
make -C ../linux-4.20.4 O=$(pwd) x86_64_defconfig
make -C ../linux-4.20.4 O=$(pwd) kvmconfig
make -C ../linux-4.20.4 O=$(pwd) menuconfig
```

The last command will bring up a configuration menu and we change the options as follow [4]:

- Kernel hacking
  - Compile-time checks and compiler options
    - Compile the kernel with debug info (check this)
      - Generate dwarf4 debuginfo (check this)
      - Provide GDB scripts for kernel debugging (check this)
  - KGDB: kernel debugger (check this)
- General setup
  - Configure standard kernel features (expert users) (check this)
- Processor type and features
  - Build a relocatable kernel (uncheck this)

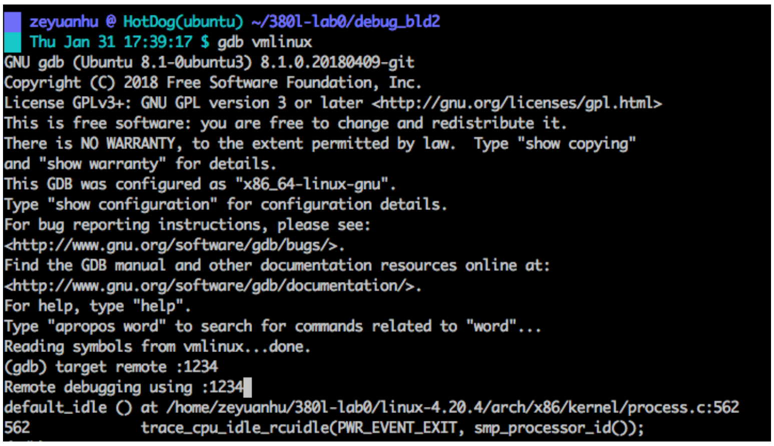


Figure 3: Fire up GDB and be ready to debug kernel

We also want to explict set `CONFIG_DEBUG_INFO_REDUCED=n` explicitly in `.config` of `debug_bld2`. Then we compile the kernel `make -j16` and start the VM as

```
sudo qemu-system-x86_64 -enable-kvm -nographic -m 512 -smp 4 -redir tcp:4444::22 -
    s -hda my-disk.img -hdb my-seed.img -kernel ~/380l-lab0/debug_bld2/arch/x86_64
    /boot/bzImage -append "root=/dev/sda1" -cpu hos
```

Note that we add an option `-s`, which tells QEMU to start a GDB server on port 1234 for debugging [?] [6]. we can start GDB in `debug_bld2` directory via `gdb vmlinux`, and type `target remote :1234` to connect gdb to the kgdb server in the guest system. Figure 3 shows a screenshot of the GDB that is ready to debug the kernel.

7.2 Tracing the kernel

Next, we create a program `testprog.c` on the guest system like the following [7]:

```
#include<unistd.h>
#include<fcntl.h>
int main()
{
    int fd = open("/dev/urandom", O_RDONLY);
    char data[4096];
    read(fd, &data, 4096);
```

[6] We also use `-nographic` instead of `-curses` because we find out that typing `./testprog` can be quite sluggish on the guest system (due to the constant checking of the breakpoint) and using `-nographic` instead of `-curses` to boot up the VM and login the VM via SSH helps to alleviate this effect.

[7] We modify the program by appending extra line `while (1){}`. Doing so make sure that the breakpoint will be hit evetually when the program is being executed (since the program is non-terminal). Since the program is fairly short and the execution is very quick. If we do not add this line, sometimes the program will finish execution without the breakpoint getting hit and that hurts reproducibility

```
    close(fd);
    fd = open("/dev/null", O_WRONLY);
    write(fd, &data, 4096);
    close(fd);
    while (1) {}
}
```

Compile it with gcc: `gcc -o testprog -g testprog.c`. Now, we want to trace into the kernel when the process contains `testprog` is running [8]. To do so, we set a conditional breakpoint in `spin_lock` in kernel code that will only stop execution if the above process is running. `spin_lock` is an inline Macro and the actual symbol name is `__raw_spin_lock`, which is defined in `include/linux/spinlock_api_smp.h`. To ensure the breakpoint only be triggered during the execution of `testprog`, we have to add a condition to the breakpoint. We use the helper script provided by kernel to figure out the PID of `testprog`. We achieve so via `$lx_current()`, which reads `task_struct` of current task in GDB and `task_struct` contains all the information we need to identify the current proces. Specifically, `$lx_current().pid` gives the PID of the current running process and `$lx_current().comm` gives the command line content, which we will use it to identify the process.

The command we run is the following, where 2297 is the pid of the program. [9]

```
b __raw_spin_lock if $lx_current().pid == 2297
```

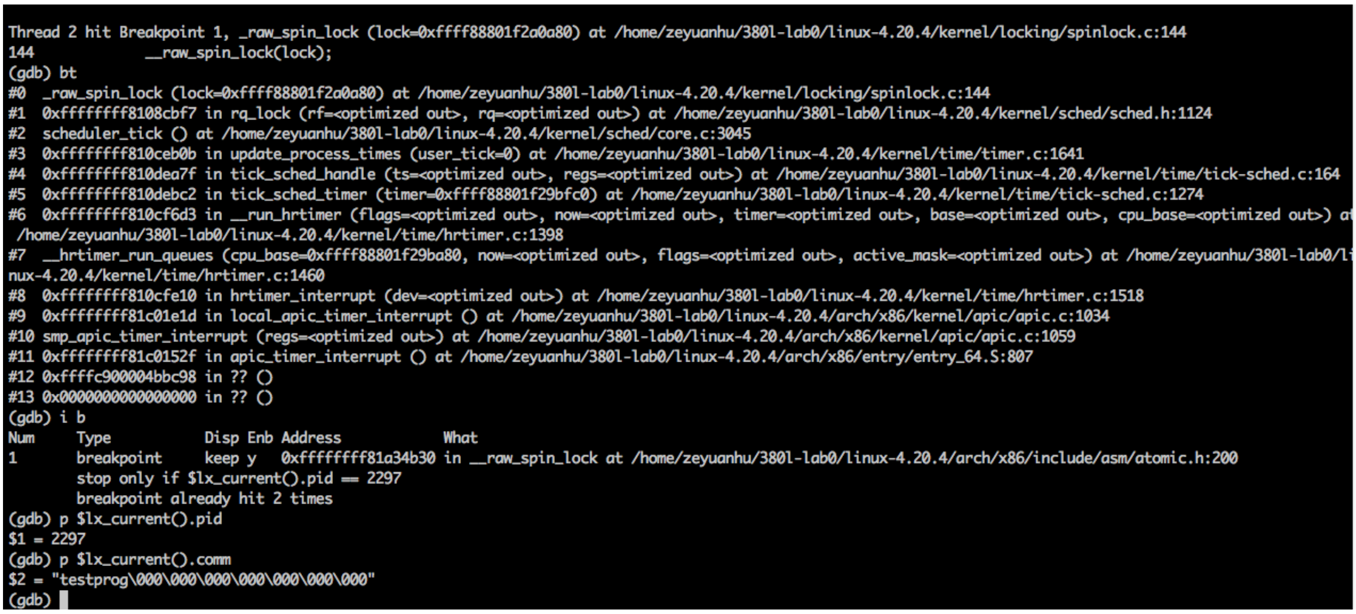
Figure 4 shows the result of `testprog` hits the breakpoint. From the figure we can see that `$lx current().pid` gives 2297 and `$lx current().comm` gives `"testprog\000\000\000\000\000\000\000"`, which confirm that we are in `testprog` process when we hit `spin_lock` breakpoint. Then, we use `bt` to examine the call stack.

The kernel backtrace of the first instance we find looks like below:

```
#0 _raw_spin_lock (lock=0xffff88801f2a0a80) at /home/zeyuanhu/380l-lab0/linux
    -4.20.4/kernel/locking/spinlock.c:144
#1 0xffffffff8108cbf7 in rq_lock (rf=<optimized out>, rq=<optimized out>) at /home
    /zeyuanhu/380l-lab0/linux-4.20.4/kernel/sched/sched.h:1124
#2 scheduler_tick () at /home/zeyuanhu/380l-lab0/linux-4.20.4/kernel/sched/core.c
    :3045
#3 0xffffffff810ceb0b in update_process_times (user_tick=0) at /home/zeyuanhu/380l
    -lab0/linux-4.20.4/kernel/time/timer.c:1641
```

[8] We first run `target remote :1234` and then we setup the breakpoint. Afterward, we issue `continue` in the GDB so that we can run `testprog` on the guest system.

[9] Alternatively, we can `b __raw_spin_lock if $_streq($lx_current().comm, "testprog")`. This command directly compares the process name with the target program, which has more overhead than comparing process id. This is quite noticeable for a frequently-used function like spin lock in this case and for a machine with less powerful CPU. The detailed steps of using the comparing-process-id approach can be found in Appendix.



```
Thread 2 hit Breakpoint 1, _raw_spin_lock (lock=0xffff88801f2a0a80) at /home/zeyuanhu/380l-lab0/linux-4.20.4/kernel/locking/spinlock.c:144
144             __raw_spin_lock(lock);
(gdb) bt
#0  _raw_spin_lock (lock=0xffff88801f2a0a80) at /home/zeyuanhu/380l-lab0/linux-4.20.4/kernel/locking/spinlock.c:144
#1  0xffffffff8108cbf7 in rq_lock (rf=<optimized out>, rq=<optimized out>) at /home/zeyuanhu/380l-lab0/linux-4.20.4/kernel/sched/sched.h:1124
#2  scheduler_tick () at /home/zeyuanhu/380l-lab0/linux-4.20.4/kernel/sched/core.c:3045
#3  0xffffffff810ceb0b in update_process_times (user_tick=0) at /home/zeyuanhu/380l-lab0/linux-4.20.4/kernel/time/timer.c:1641
#4  0xffffffff810dea7f in tick_sched_handle (ts=<optimized out>, regs=<optimized out>) at /home/zeyuanhu/380l-lab0/linux-4.20.4/kernel/time/tick-sched.c:164
#5  0xffffffff810debc2 in tick_sched_timer (timer=0xffff88801f29bfc0) at /home/zeyuanhu/380l-lab0/linux-4.20.4/kernel/time/tick-sched.c:1274
#6  0xffffffff810cf6d3 in __run_hrtimer (flags=<optimized out>, now=<optimized out>, timer=<optimized out>, base=<optimized out>, cpu_base=<optimized out>) at
 /home/zeyuanhu/380l-lab0/linux-4.20.4/kernel/time/hrtimer.c:1398
#7  __hrtimer_run_queues (cpu_base=0xffff88801f29ba80, now=<optimized out>, flags=<optimized out>, active_mask=<optimized out>) at /home/zeyuanhu/380l-lab0/li
nux-4.20.4/kernel/time/hrtimer.c:1460
#8  0xffffffff810cfe10 in hrtimer_interrupt (dev=<optimized out>) at /home/zeyuanhu/380l-lab0/linux-4.20.4/kernel/time/hrtimer.c:1518
#9  0xffffffff81c01e1d in local_apic_timer_interrupt () at /home/zeyuanhu/380l-lab0/linux-4.20.4/arch/x86/kernel/apic/apic.c:1034
#10 smp_apic_timer_interrupt (regs=<optimized out>) at /home/zeyuanhu/380l-lab0/linux-4.20.4/arch/x86/kernel/apic/apic.c:1059
#11 0xffffffff81c0152f in apic_timer_interrupt () at /home/zeyuanhu/380l-lab0/linux-4.20.4/arch/x86/entry/entry_64.S:807
#12 0xffffc900004bbc98 in ?? ()
#13 0x0000000000000000 in ?? ()
(gdb) i b
Num     Type           Disp Enb Address            What
1       breakpoint     keep y   0xffffffff81a34b30 in __raw_spin_lock at /home/zeyuanhu/380l-lab0/linux-4.20.4/arch/x86/include/asm/atomic.h:200
        stop only if $lx_current().pid == 2297
        breakpoint already hit 2 times
(gdb) p $lx_current().pid
$1 = 2297
(gdb) p $lx_current().comm
$2 = "testprog\000\000\000\000\000\000\000"
(gdb)
```

Figure 4: `testprog` hits breakpoint

```
#4 0xffffffff810dea7f in tick_sched_handle (ts=<optimized out>, regs=<optimized
    out>) at /home/zeyuanhu/380l-lab0/linux-4.20.4/kernel/time/tick-sched.c:164
#5 0xffffffff810debc2 in tick_sched_timer (timer=0xffff88801f29bfc0) at /home/
    zeyuanhu/380l-lab0/linux-4.20.4/kernel/time/tick-sched.c:1274
#6 0xffffffff810cf6d3 in __run_hrtimer (flags=<optimized out>, now=<optimized out
    >, timer=<optimized out>, base=<optimized out>, cpu_base=<optimized out>) at /
    home/zeyuanhu/380l-lab0/linux-4.20.4/kernel/time/hrtimer.c:1398
#7 __hrtimer_run_queues (cpu_base=0xffff88801f29ba80, now=<optimized out>, flags=<
    optimized out>, active_mask=<optimized out>) at /home/zeyuanhu/380l-lab0/linux
    -4.20.4/kernel/time/hrtimer.c:1460
#8 0xffffffff810cfe10 in hrtimer_interrupt (dev=<optimized out>) at /home/zeyuanhu
    /380l-lab0/linux-4.20.4/kernel/time/hrtimer.c:1518
#9 0xffffffff81c01e1d in local_apic_timer_interrupt () at /home/zeyuanhu/380l-lab0
    /linux-4.20.4/arch/x86/kernel/apic/apic.c:1034
#10 smp_apic_timer_interrupt (regs=<optimized out>) at /home/zeyuanhu/380l-lab0/
    linux-4.20.4/arch/x86/kernel/apic/apic.c:1059
#11 0xffffffff81c0152f in apic_timer_interrupt () at /home/zeyuanhu/380l-lab0/
    linux-4.20.4/arch/x86/entry/entry_64.S:807
<-- snip -->
```

Here, the kernel acquires a lock on the run queue and charge one tick to the current process (`update_process_time`). Then, the kernel runs handler for the timer interrupt. If we take a look at function `hrtimer_interrupt` in `kernel/time/hrtimer.c`, we know the `hrtimer_bases`, a per-CPU variable [?], acquired a lock [10].

The second instance that we take a look at is the following:

```
#0 _raw_spin_lock (lock=0xffff88801f2a0a80) at /home/zeyuanhu/380l-lab0/linux
    -4.20.4/kernel/locking/spinlock.c:144
#1 0xffffffff8108bb81 in rq_lock (rf=<optimized out>, rq=<optimized out>) at /home
    /zeyuanhu/380l-lab0/linux-4.20.4/kernel/sched/sched.h:1124
#2 ttwu_queue (wake_flags=<optimized out>, cpu=<optimized out>, p=<optimized out>)
     at /home/zeyuanhu/380l-lab0/linux-4.20.4/kernel/sched/core.c:1845
#3 try_to_wake_up (p=0xffff88801eddd780, state=<optimized out>, wake_flags=0) at /
    home/zeyuanhu/380l-lab0/linux-4.20.4/kernel/sched/core.c:2057
#4 0xffffffff8108bcbc in wake_up_process (p=<optimized out>) at /home/zeyuanhu/380
    l-lab0/linux-4.20.4/kernel/sched/core.c:2129
```

Here, kernel tries to awaken a sleeping process through calling `try_to_wake_up()`. `try_to_wake_up()` function wakes a sleeping or stopped process by setting its state to `TASK_RUNNING` and inserting it into the runqueue of the local CPU [5].

The third instance is the following:

```
#0 _raw_spin_lock (lock=0xffff88001cc1ec6c) at /home/zeyuanhu/linux-4.20.4/kernel/
    locking/spinlock.c:144
<-- snip -->
#3 0xffffffff81167316 in pud_alloc (address=<optimized out>, p4d=<optimized out>,
    mm=<optimized out>) at /home/zeyuanhu/linux-4.20.4/include/linux/mm.h:1733
#4 __handle_mm_fault (vma=<optimized out>, address=6295640, flags=<optimized out>)
     at /home/zeyuanhu/linux-4.20.4/mm/memory.c:4008
#5 0xffffffff811678ad in handle_mm_fault (vma=<optimized out>, address=<optimized
    out>, flags=<optimized out>) at /home/zeyuanhu/linux-4.20.4/mm/memory.c:4104
#6 0xffffffff8104bede in __do_page_fault (regs=0xffffc90000317ce8, error_code=2,
    address=6295640) at /home/zeyuanhu/linux-4.20.4/arch/x86/mm/fault.c:1426
#7 0xffffffff81a0168b in async_page_fault () at /home/zeyuanhu/linux-4.20.4/arch/
    x86/entry/entry_64.S:1118
```

At this place, kernel tries to handle the page fault. In `/arch/x86/entry/entry_64.S`, we can see `async_page_fault()` is invoked when we're in the KVM guest environment (i.e., QEMU this case). `do_page_fault()` function, which is the Page Fault interrupt service routine for the x86 architecture, compares the linear address that caused the Page Fault against the memory regions of the current

[10] In GDB, the helper script also provides a function `$lx_per_cpu` to obtain per-CPU variables (actually `$lx_current()` is a shorthand to `$lx_per_cpu("current task")`)

process and determines the proper way to handle the exception. In this case, `handle_mm_fault()` is invoked to allocate a new page frame [5].

8 Differences between `/dev/random` and `/dev/urandom`

Both `/dev/random` and `/dev/urandom` are interfaces to the kernel's random number generator [?] and both of them are fed by the same cryptographically secure pseudorandom number generator [6]. However, they are different on how they handle their repective entropy pool when the pool is empty. `/dev/random` will block the reads if its entropy pool is empty and the reads will be blocked until additional environmental noise is gathered. However, `/dev/urandom` will not block waiting for more entropy and as a result, the returned values may have theoretical vulunerability. There is an argument on when to use which and some suggests that use `/dev/urandom` is strictly better as the thoeretical vulunerability may not lead to computational vulunerability [6] and thus should be used all the time. But, `man` page seems to suggest that it is a case-by-case situation [?].

References

[1] "proc(5) - linux man page." `http://man7.org/linux/man-pages/man5/proc.5.html`.

[2] "Address space layout randomization (aslr)." `https://en.wikipedia.org/wiki/Address_space_layout_randomization`, 2018.

[3] "Linux and aslr: kernel/randomize_va_space." `https://linux-audit.com/linux-aslr-and-kernelrandomize_va_space-setting`, 2016.

[4] "Cs380l: Advanced operating systems lab 0." `https://www.cs.utexas.edu/~rossbach/380L/lab/lab0.html#debug-config`, 2018.

[5] D. P. Bovet and M. Cassetti, *Understanding the Linux Kernel.* Sebastopol, CA, USA: O'Reilly & Associates, Inc., 2000.

[6] "Myths about /dev/urandom." `https://www.2uo.de/myths-about-urandom/`.

Appendices

A Detailed steps to let program hit breakpoint

Let $T1$ denotes the tab with `gdb vmlinux`, let $T1$ and $T2$ denote the tabs that we ssh into the guest system. Then we proceed as the following:

1. `gdb vmlinux` ($T1$)
2. `target remote :1234` ($T1$)
3. `c` ($T1$)
4. `gdb testprog` ($T2$)
5. `b main` ($T2$)
6. `r` ($T2$)
7. `ps aux | grep test` ($T3$) to obtain pid of *testprog*
8. `b __raw_spin_lock if $lx_current().pid == 2297` (`2297` is the pid we find out in the earlier step) ($T1$)
9. `c` ($T1$)
10. `c` ($T2$)

Now, at some point, $T1$ will show that the breakpoint is hit.