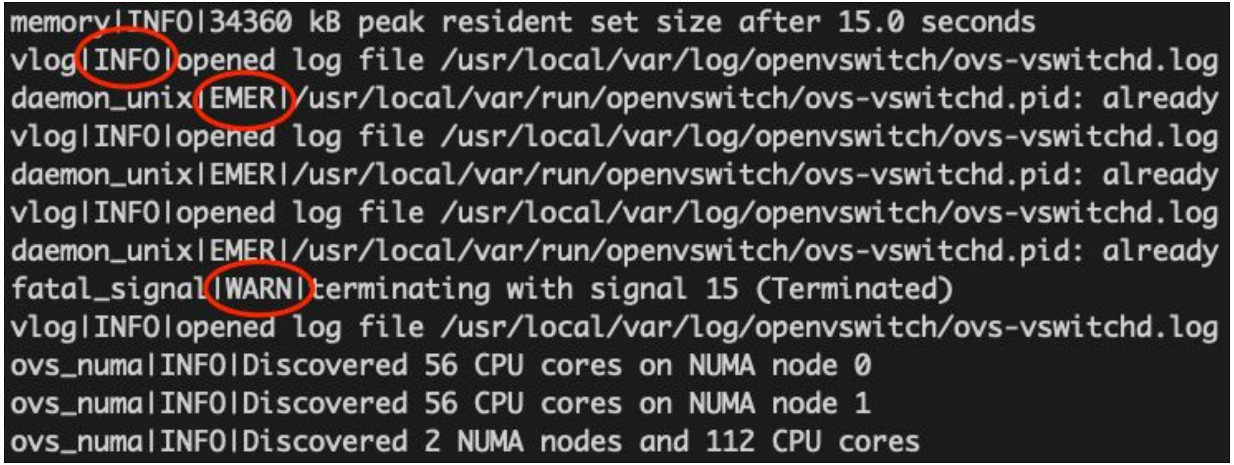
System Debugging & Profiling

Tutorial_07

Scripts & Code: https://github.com/130B848/ipads-tutorial07.git

Print-based Debugging & Logging

- Multiple log levels
    - INFO, DEBUG, WARN, ERROR, etc.
    - An example from OVS:

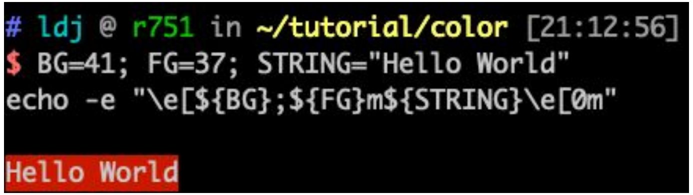


```
memory|INFO|34360 kB peak resident set size after 15.0 seconds
vlog|INFO|opened log file /usr/local/var/log/openvswitch/ovs-vswitchd.log
daemon_unix|EMER|/usr/local/var/run/openvswitch/ovs-vswitchd.pid: already
vlog|INFO|opened log file /usr/local/var/log/openvswitch/ovs-vswitchd.log
daemon_unix|EMER|/usr/local/var/run/openvswitch/ovs-vswitchd.pid: already
vlog|INFO|opened log file /usr/local/var/log/openvswitch/ovs-vswitchd.log
daemon_unix|EMER|/usr/local/var/run/openvswitch/ovs-vswitchd.pid: already
fatal_signal|WARN|terminating with signal 15 (Terminated)
vlog|INFO|opened log file /usr/local/var/log/openvswitch/ovs-vswitchd.log
ovs_numa|INFO|Discovered 56 CPU cores on NUMA node 0
ovs_numa|INFO|Discovered 56 CPU cores on NUMA node 1
ovs_numa|INFO|Discovered 2 NUMA nodes and 112 CPU cores
```

Coloring: Make Things More Readable

- Red background & white foreground

```
BG=41; FG=37; STRING="Hello World"
echo -e "\e[${BG};${FG}m${STRING}\e[0m"
```



```
# ldj @ r751 in ~/tutorial/color [21:12:56]
$ BG=41; FG=37; STRING="Hello World"
echo -e "\e[${BG};${FG}m${STRING}\e[0m"

Hello World
```

- Try this!

```
for R in $(seq 0 20 255); do
    for G in $(seq 0 20 255); do
        for B in $(seq 0 20 255); do
            printf "\e[38;2;${R};${G};${B}m█\e[0m";
        done
    done
done
```

Bash's color and formatting: https://misc.flogisoft.com/bash/tip_colors_and_formatting

Understand Kernel Behaviors

Stop Trying to Reinvent the Wheel

Kernel Messages

Debug FS

Strace

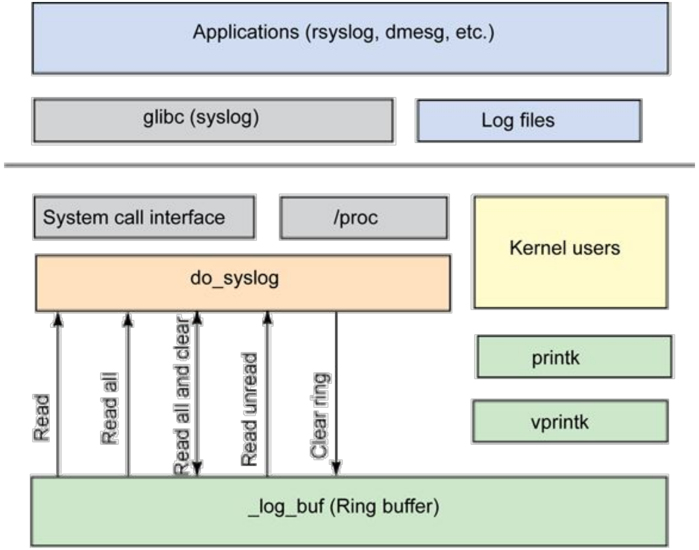
Ftrace

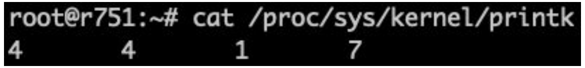
Linux *perf*

Flamegraph

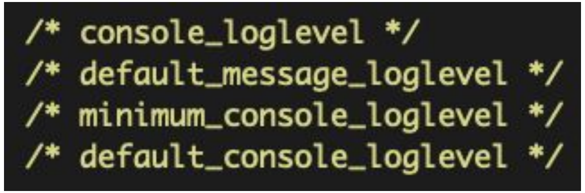
Kernel Messages

- **`printk/pr_warn/pr_err…`**
  - Ring buffer size: **`CONFIG_LOG_BUF_SHIFT=18`**
- Show kernel messages
  - `syslogd/klogd` → **`/var/log/messages`**
  - **`dmesg`** → `stdout`
- Set log level
  - **`/proc/sys/kernel/printk`**

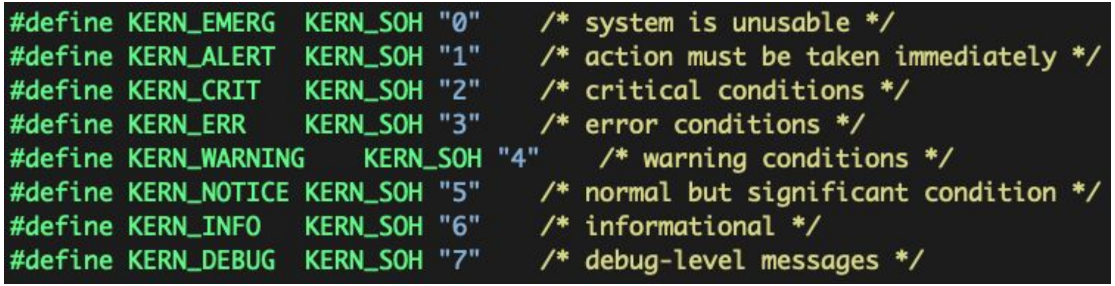




```
root@r751:~# cat /proc/sys/kernel/printk
4       4       1       7
```



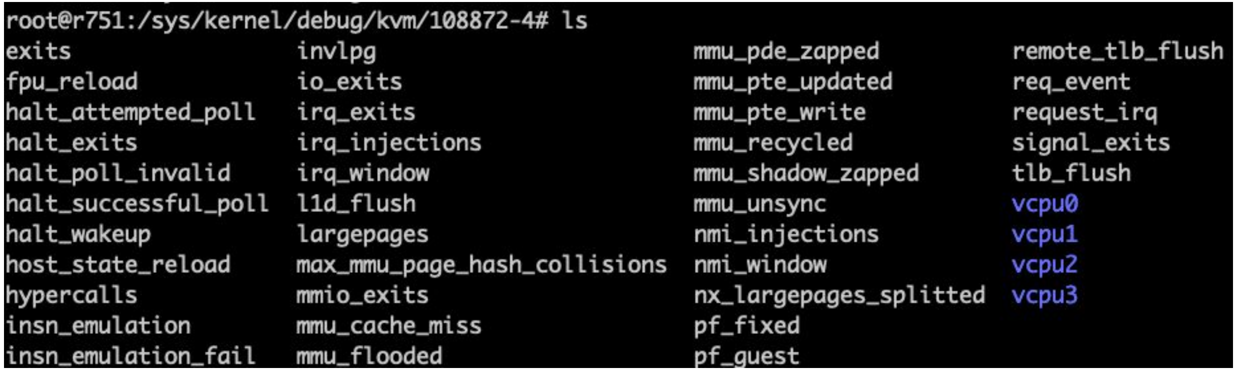
```
 /* console_loglevel */
 /* default_message_loglevel */
 /* minimum_console_loglevel */
 /* default_console_loglevel */
```



```
#define KERN_EMERG   KERN_SOH "0"    /* system is unusable */
#define KERN_ALERT   KERN_SOH "1"    /* action must be taken immediately */
#define KERN_CRIT    KERN_SOH "2"    /* critical conditions */
#define KERN_ERR     KERN_SOH "3"    /* error conditions */
#define KERN_WARNING     KERN_SOH "4"    /* warning conditions */
#define KERN_NOTICE KERN_SOH "5"    /* normal but significant condition */
#define KERN_INFO    KERN_SOH "6"    /* informational */
#define KERN_DEBUG   KERN_SOH "7"    /* debug-level messages */
```

Debug FS

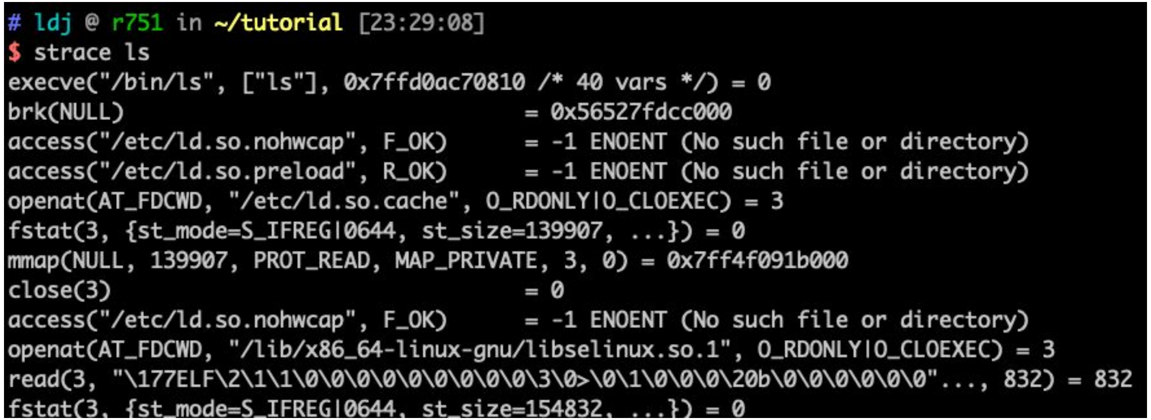
- Kconfig
  - `CONFIG_DEBUG_FS=y`
- An example from KVM
  - Count # of stage-2 page fault
  - Read: **`cat /path/to/debugfs/pf_fixed`**
  - Clear: **`echo 0 > /path/to/debugfs/pf_fixed`**



```
root@r751:/sys/kernel/debug/kvm/108872-4# ls
exits                 invlpg                         mmu_pde_zapped          remote_tlb_flush
fpu_reload            io_exits                       mmu_pte_updated         req_event
halt_attempted_poll   irq_exits                      mmu_pte_write           request_irq
halt_exits            irq_injections                 mmu_recycled            signal_exits
halt_poll_invalid     irq_window                     mmu_shadow_zapped       tlb_flush
halt_successful_poll  l1d_flush                      mmu_unsync              vcpu0
halt_wakeup           largepages                     nmi_injections          vcpu1
host_state_reload     max_mmu_page_hash_collisions   nmi_window              vcpu2
hypercalls            mmio_exits                     nx_largepages_splitted  vcpu3
insn_emulation        mmu_cache_miss                 pf_fixed
insn_emulation_fail   mmu_flooded                    pf_guest
```

Strace

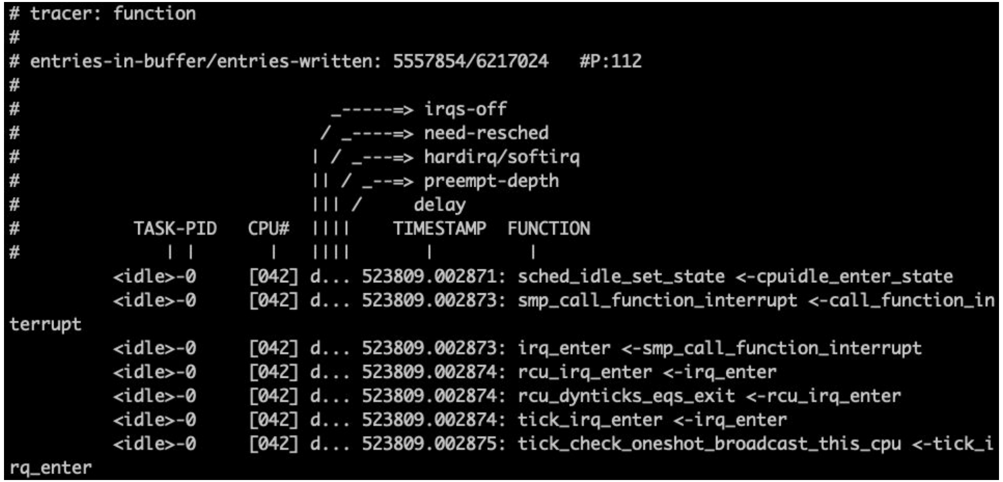
- Trace system calls
  - `strace ls`
- Attach to a process
  - `strace -p`
- Log to file
  - `strace -o`



```
# ldj @ r751 in ~/tutorial [23:29:08]
$ strace ls
execve("/bin/ls", ["ls"], 0x7ffd0ac70810 /* 40 vars */) = 0
brk(NULL)                               = 0x56527fdcc000
access("/etc/ld.so.nohwcap", F_OK)      = -1 ENOENT (No such file or directory)
access("/etc/ld.so.preload", R_OK)      = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/etc/ld.so.cache", O_RDONLY|O_CLOEXEC) = 3
fstat(3, {st_mode=S_IFREG|0644, st_size=139907, ...}) = 0
mmap(NULL, 139907, PROT_READ, MAP_PRIVATE, 3, 0) = 0x7ff4f091b000
close(3)                                = 0
access("/etc/ld.so.nohwcap", F_OK)      = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/lib/x86_64-linux-gnu/libselinux.so.1", O_RDONLY|O_CLOEXEC) = 3
read(3, "\177ELF\2\1\1\0\0\0\0\0\0\0\0\0\3\0>\0\1\0\0\0\20b\0\0\0\0\0\0"..., 832) = 832
fstat(3, {st_mode=S_IFREG|0644, st_size=154832, ...}) = 0
```

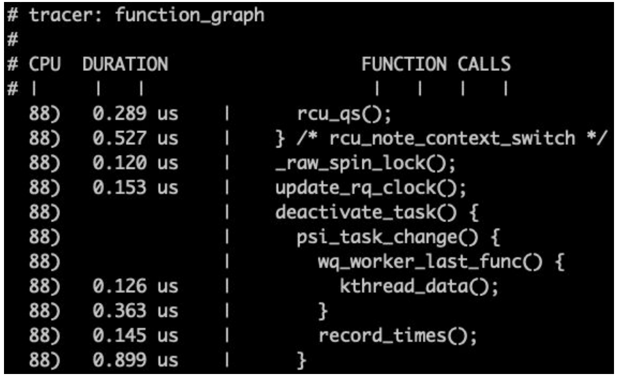
Ftrace

- Function
  - Function ← parent function



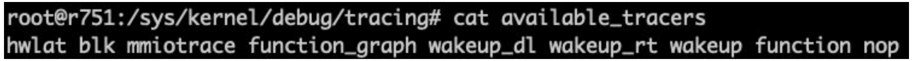
```
# tracer: function
#
# entries-in-buffer/entries-written: 5557854/6217024   #P:112
#
#                              _-----=> irqs-off
#                             / _----=> need-resched
#                            | / _---=> hardirq/softirq
#                            || / _--=> preempt-depth
#                            ||| /     delay
#           TASK-PID   CPU#  ||||    TIMESTAMP  FUNCTION
#              | |       |   ||||       |         |
          <idle>-0     [042] d... 523809.002871: sched_idle_set_state <-cpuidle_enter_state
          <idle>-0     [042] d... 523809.002873: smp_call_function_interrupt <-call_function_in
terrupt
          <idle>-0     [042] d... 523809.002873: irq_enter <-smp_call_function_interrupt
          <idle>-0     [042] d... 523809.002874: rcu_irq_enter <-irq_enter
          <idle>-0     [042] d... 523809.002874: rcu_dynticks_eqs_exit <-rcu_irq_enter
          <idle>-0     [042] d... 523809.002874: tick_irq_enter <-irq_enter
          <idle>-0     [042] d... 523809.002875: tick_check_oneshot_broadcast_this_cpu <-tick_i
rq_enter
```

- Function graph
  - + → > 10us, ! → > 100us



```
# tracer: function_graph
#
# CPU  DURATION                  FUNCTION CALLS
# |     |   |                     |   |   |   |
 88)   0.289 us    |        rcu_qs();
 88)   0.527 us    |      } /* rcu_note_context_switch */
 88)   0.120 us    |      _raw_spin_lock();
 88)   0.153 us    |      update_rq_clock();
 88)               |      deactivate_task() {
 88)               |        psi_task_change() {
 88)               |          wq_worker_last_func() {
 88)   0.126 us    |            kthread_data();
 88)   0.363 us    |          }
 88)   0.145 us    |          record_times();
 88)   0.899 us    |        }
```

- Filter
  - `/sys/kernel/debug/tracing/set_ftrace_filter`

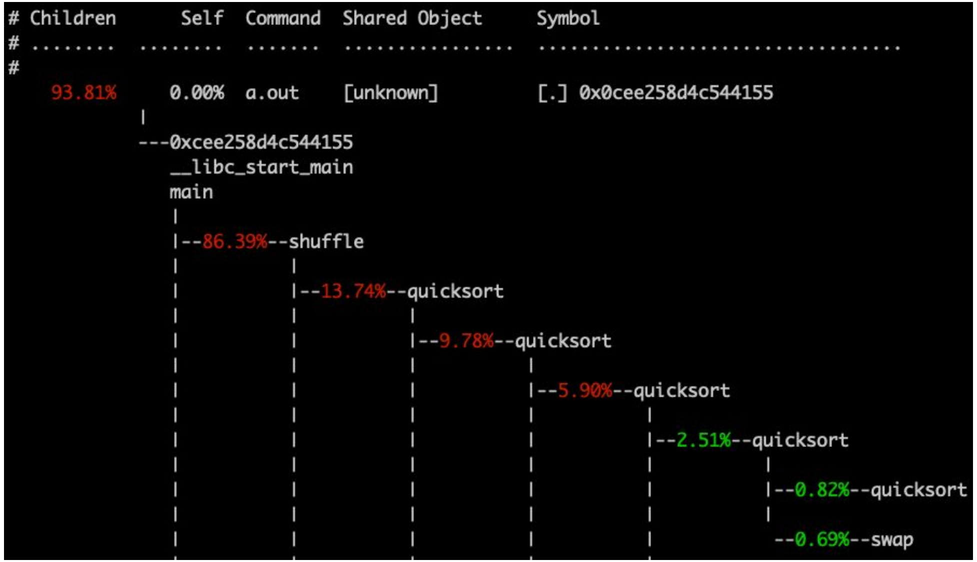


```
root@r751:/sys/kernel/debug/tracing# cat available_tracers
hwlat blk mmiotrace function_graph wakeup_dl wakeup_rt wakeup function nop
```

Linux *perf*

- Event-based
  - Gather PMU events: **`perf stat`**

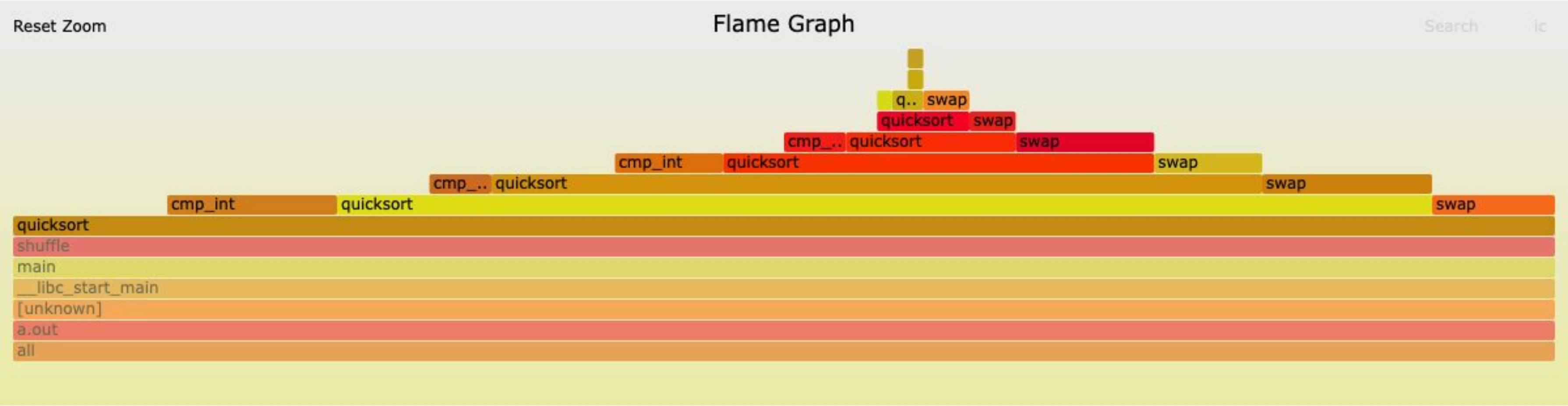
- Sample-based
  - Record profile: **`perf record`**
  - Display the profile: **`perf report`**



```
# Children      Self  Command  Shared Object     Symbol
# ........  ........  .......  ................  ................................
#
    93.81%     0.00%  a.out    [unknown]         [.] 0x0cee258d4c544155
            |
            ---0xcee258d4c544155
               __libc_start_main
               main
               |
               |--86.39%--shuffle
               |          |
               |          |--13.74%--quicksort
               |          |          |
               |          |          |--9.78%--quicksort
               |          |          |          |
               |          |          |          |--5.90%--quicksort
               |          |          |          |          |
               |          |          |          |          |--2.51%--quicksort
               |          |          |          |          |          |
               |          |          |          |          |          |--0.82%--quicksort
               |          |          |          |          |          |
               |          |          |          |          |           --0.69%--swap
               |          |          |          |          |
```

Flamegraph

- Visualization
  - Backtrace & execution time
  - Generate trace output: **`perf script`**
  - Draw flame graph: **`stackcollapse-perf.pl, flamegraph.pl`**



“The most effective debugging tool is still careful thought, coupled with judiciously placed print statements”

—— Brian Kernighan, Unix for Beginners.

Hack the Kernel

Nothing Ventured,
Nothing Gained

GDB & QEMU

Kernel Module

VFS: ioctl, mmap, ...

Shared Memory

Serial Port

The GNU Debugger (GDB) & QEMU

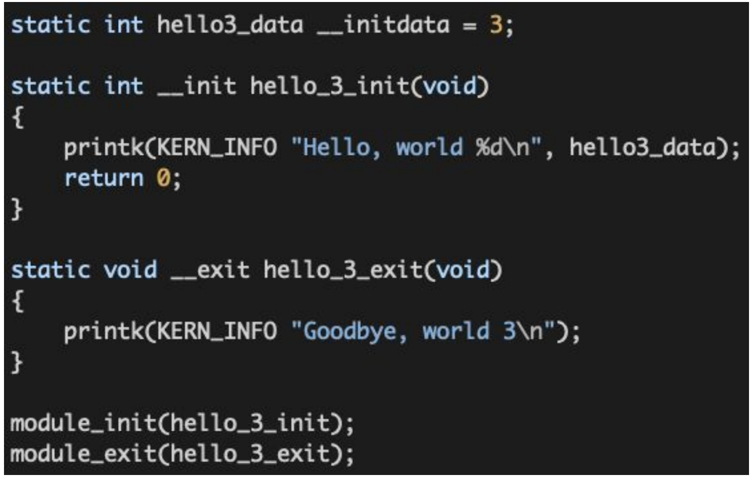
- Basic: list, break, run, print, step, etc.
  - You MUST be familiar with them after ICS & ChCore labs
- Kconfig
  - **`CONFIG_DEBUG_INFO=y`**
- Startup scripts (.gdbinit)
  - E.g., **`target remote :1234, b start_kernel`**

Breakpoint is Not a Panacea

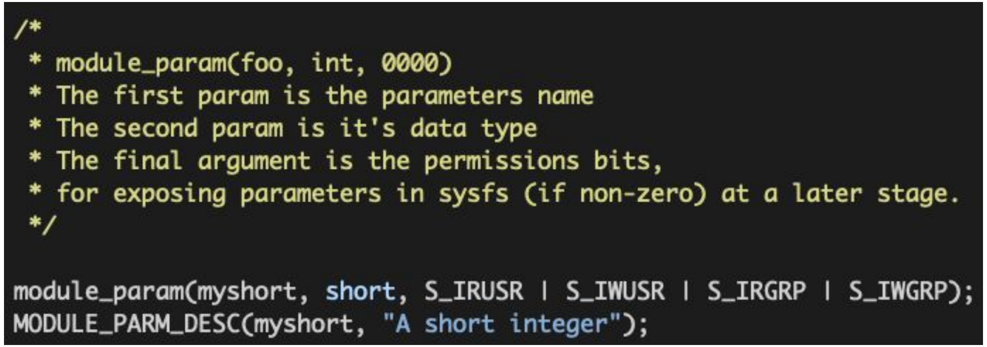
- Symbol not found
  - Compiler optimizations (e.g., stage-2 page fault code in ARM64)
- Step instruction does not work
  - Hardware breakpoint should work, but...
- Printk can change the execution flow
- Dead loop

Kernel Module

- Set a flag to indicate start/end of debugging
- Passing arguments
  - `insmod args.ko mystring="bebop" myintArray=233,666`



```c
static int hello3_data __initdata = 3;

static int __init hello_3_init(void)
{
    printk(KERN_INFO "Hello, world %d\n", hello3_data);
    return 0;
}

static void __exit hello_3_exit(void)
{
    printk(KERN_INFO "Goodbye, world 3\n");
}

module_init(hello_3_init);
module_exit(hello_3_exit);
```

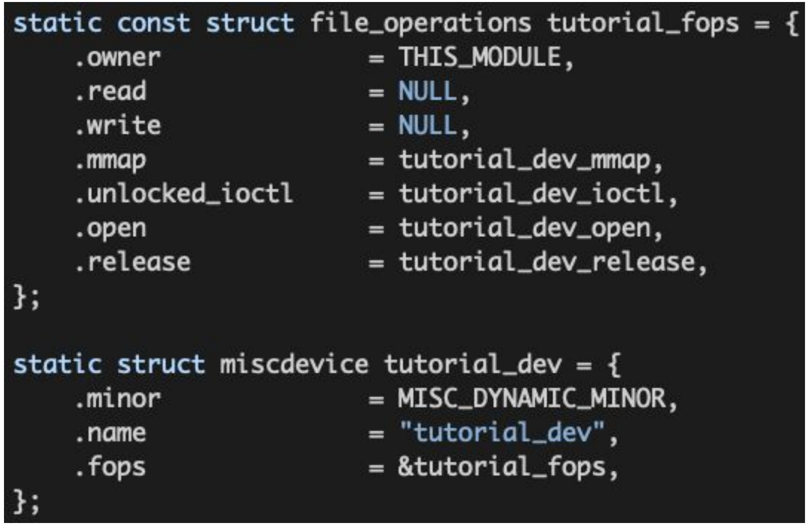


```c
/*
 * module_param(foo, int, 0000)
 * The first param is the parameters name
 * The second param is it's data type
 * The final argument is the permissions bits,
 * for exposing parameters in sysfs (if non-zero) at a later stage.
 */

module_param(myshort, short, S_IRUSR | S_IWUSR | S_IRGRP | S_IWGRP);
MODULE_PARM_DESC(myshort, "A short integer");
```

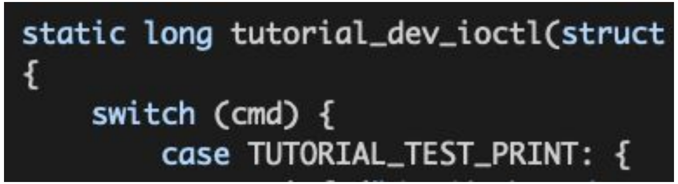
Ref: *https://tldp.org/LDP/lkmpg/2.4/html/x281.htm*

VFS

- Invoke in the user application
  - open, ioctl, mmap, etc.



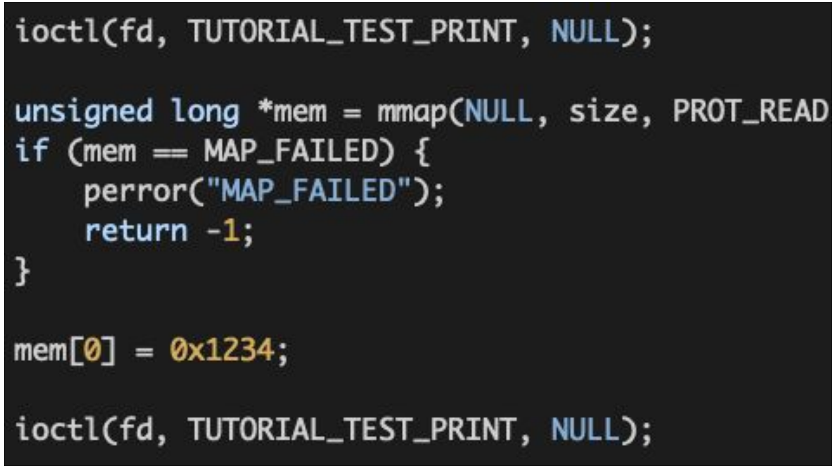
```
static const struct file_operations tutorial_fops = {
    .owner                  = THIS_MODULE,
    .read                   = NULL,
    .write                  = NULL,
    .mmap                   = tutorial_dev_mmap,
    .unlocked_ioctl         = tutorial_dev_ioctl,
    .open                   = tutorial_dev_open,
    .release                = tutorial_dev_release,
};

static struct miscdevice tutorial_dev = {
    .minor                  = MISC_DYNAMIC_MINOR,
    .name                   = "tutorial_dev",
    .fops                   = &tutorial_fops,
};
```



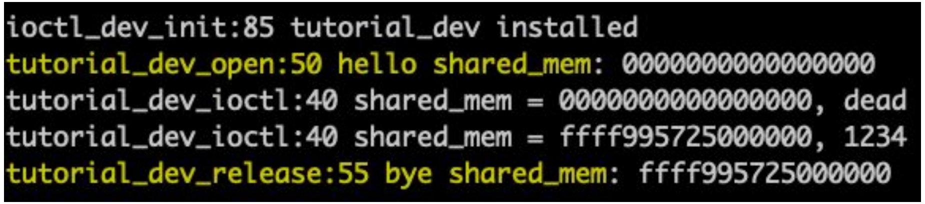
```
static long tutorial_dev_ioctl(struct
{
    switch (cmd) {
        case TUTORIAL_TEST_PRINT: {
```

Shared Memory

- Cross-ring breakdown
- Set up a shared memory between userspace/kernel/VM
  - Userspace: **`mmap(..., size, …)`**
  - Kernel: **`remap_pfn_range(..., pfn, size, …)`**



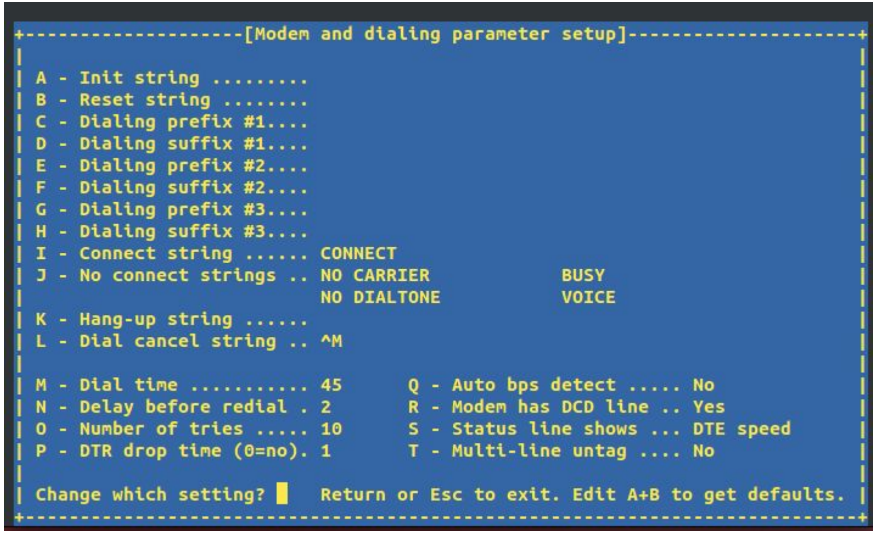
```
ioctl(fd, TUTORIAL_TEST_PRINT, NULL);

unsigned long *mem = mmap(NULL, size, PROT_READ
if (mem == MAP_FAILED) {
    perror("MAP_FAILED");
    return -1;
}

mem[0] = 0x1234;

ioctl(fd, TUTORIAL_TEST_PRINT, NULL);
```



```
ioctl_dev_init:85 tutorial_dev installed
tutorial_dev_open:50 hello shared_mem: 0000000000000000
tutorial_dev_ioctl:40 shared_mem = 0000000000000000, dead
tutorial_dev_ioctl:40 shared_mem = ffff995725000000, 1234
tutorial_dev_release:55 bye shared_mem: ffff995725000000
```

Serial Port

- System crash ➔ no log is saved
- Serial port
  - Host GRUB: **`console=ttyS0,115200,8n1 kgdboc=ttyS0,115200`**
  - Serial machine: Minicom, Screen, ...



```
+--------------------[Modem and dialing parameter setup]---------------------+
|                                                                            |
| A - Init string .........                                                  |
| B - Reset string ........                                                  |
| C - Dialing prefix #1....                                                  |
| D - Dialing suffix #1....                                                  |
| E - Dialing prefix #2....                                                  |
| F - Dialing suffix #2....                                                  |
| G - Dialing prefix #3....                                                  |
| H - Dialing suffix #3....                                                  |
| I - Connect string ...... CONNECT                                          |
| J - No connect strings .. NO CARRIER            BUSY                       |
|                           NO DIALTONE           VOICE                      |
| K - Hang-up string ......                                                  |
| L - Dial cancel string .. ^M                                               |
|                                                                            |
| M - Dial time ........... 45       Q - Auto bps detect ..... No           |
| N - Delay before redial . 2        R - Modem has DCD line .. Yes          |
| O - Number of tries ..... 10       S - Status line shows ... DTE speed    |
| P - DTR drop time (0=no). 1        T - Multi-line untag .... No           |
|                                                                            |
| Change which setting?     Return or Esc to exit. Edit A+B to get defaults. |
+----------------------------------------------------------------------------+
```

Understand the code before you dive into it

Control variable: Phenomenon ➔ Assumption ➔ Experiment ➔ …

Reproduce bugs: Unit tests + CI + Logs

Thanks!