

Debugging Linux Kernel by Ftrace

AceLan Kao

[Http://people.ubuntu.com/~acelan/coscup-2010/](http://people.ubuntu.com/~acelan/coscup-2010/)



Ftrace Introduction

- Lightweight, flexible function and tracepoint tracer, profiler
- Useful for data gathering, debugging, and performance tuning
- In Ubuntu 9.10 and later releases
- No need for kernel recompile or separate flavour!
- Documentation/trace/{ftrace.txt,ftrace-design.txt}

Ftrace's trick

- Use gprof hooks. Add mcount() call at entry of each kernel function.
- Require kernel to be compiled with -pg option
- During compilation the mcount() call-sites are recorded.
- Convert the mcount() call to a NOP at boot time

The Debugfs

- Debugfs officially be mounted at
 - `/sys/kernel/debug`
- Ftrace
 - `/sys/kernel/debug/tracing`

The Tracing Directory

```
acelan@acelan-nb /sys/kernel/debug % ls tracing
```

available_events	per_cpu/	trace_clock
available_filter_functions	printk_formats	trace_marker
available_tracers	README	trace_options
buffer_size_kb	saved_cmdlines	trace_pipe
current_tracer	set_event	trace_stat/
dyn_ftrace_total_info	set_ftrace_filter	tracing_cpumask
Events/	set_ftrace_notrace	tracing_enabled
Failures	set_ftrace_pid	tracing_on
function_profile_enabled	set_graph_function	
Options/	trace	

Available Tracers

- `available_tracers` (Lucid)
 - `blk` – for blk device
 - `function` – trace entry of all kernel functions
 - `function_graph` – trace on both entry and exit of all functions. And provides C style of calling graph
 - `mmiotrace` – In-kernel memory-mapped I/O tracing
 - `sched_switch` – context switches and wakeups between tasks
 - `nop` – trace nothing

The Function Tracer

```
acelan@acelan-nb /sys/kernel/debug/tracing % sudo su -c "echo function > current_tracer"
acelan@acelan-nb /sys/kernel/debug/tracing % cat trace | head -n 15
# tracer: function
#
#      TASK-PID    CPU#    TIMESTAMP    FUNCTION
#      ||      |           |
zsh-6345 [001] 18446730808.793822: page_add_file_rmap <-__do_fault
zsh-6345 [001] 18446730808.793822: native_set_pte_at <-__do_fault
zsh-6345 [001] 18446730808.793822: unlock_page <-__do_fault
zsh-6345 [001] 18446730808.793823: page_waitqueue <-unlock_page
zsh-6345 [001] 18446730808.793823: __wake_up_bit <-unlock_page
zsh-6345 [001] 18446730808.793823: up_read <-do_page_fault
zsh-6345 [001] 18446730808.793824: _spin_lock_irqsave <-__up_read
zsh-6345 [001] 18446730808.793824: _spin_unlock_irqrestore <-__up_read
zsh-6345 [001] 18446730808.793826: down_read_trylock <-do_page_fault
zsh-6345 [001] 18446730808.793826: _spin_lock_irqsave <-__down_read_trylock
zsh-6345 [001] 18446730808.793826: _spin_unlock_irqrestore <-__down_read_trylock
```

Ftrace Filter

```
acelan@acelan-nb /sys/kernel/debug/tracing % cat available_filter_functions | head -n 5
```

```
hypercall_page  
do_one_initcall  
run_init_process  
init_post  
name_to_dev_t
```

```
acelan@acelan-nb /sys/kernel/debug/tracing % cat available_filter_functions | wc -l  
28296
```

```
acelan@acelan-nb /sys/kernel/debug/tracing % sudo su -c "echo ext4* >  
set_ftrace_filter"
```

```
acelan@acelan-nb /sys/kernel/debug/tracing % cat set_ftrace_filter | wc -l  
346
```

```
acelan@acelan-nb /sys/kernel/debug/tracing % cat set_ftrace_filter | head -n 5
```

```
ext4_get_group_no_and_offset  
ext4_bg_has_super  
ext4_bg_num_gdb  
ext4_new_meta_blocks  
ext4_has_free_blocks
```

Ftrace Filter (cont.)

```
acelan@acelan-nb /sys/kernel/debug/tracing % sudo su -c "echo function > current_tracer"
acelan@acelan-nb /sys/kernel/debug/tracing % cat trace | head -n 15
# tracer: function
#
#      TASK-PID   CPU#    TIMESTAMP  FUNCTION
#      ||      |      |      |
nepomukservices-3920 [000] 18446732636.142310: ext4_check_dir_entry <-
htree_dirblock_to_tree
nepomukservices-3920 [000] 18446732636.142311: ext4_rec_len_from_disk <-
ext4_check_dir_entry
nepomukservices-3920 [000] 18446732636.142312: ext4fs_dirhash <-htree_dirblock_to_tree
nepomukservices-3920 [000] 18446732636.142312: ext4_htree_store_dirent <-
htree_dirblock_to_tree
nepomukservices-3920 [000] 18446732636.142313: ext4_check_dir_entry <-
htree_dirblock_to_tree
nepomukservices-3920 [000] 18446732636.142314: ext4_rec_len_from_disk <-
ext4_check_dir_entry
nepomukservices-3920 [000] 18446732636.142314: ext4fs_dirhash <-htree_dirblock_to_tree
nepomukservices-3920 [000] 18446732636.142315: ext4_htree_store_dirent <-
htree_dirblock_to_tree
nepomukservices-3920 [000] 18446732636.142316: ext4_check_dir_entry <-
htree_dirblock_to_tree
nepomukservices-3920 [000] 18446732636.142316: ext4_rec_len_from_disk <-
ext4_check_dir_entry
nepomukservices-3920 [000] 18446732636.142317: ext4fs_dirhash <-htree_dirblock_to_tree
```

Acceptable Globs

- `value*`
 - Select all functions that begin with "value"
- `*value*`
 - Select all functions that contain the text "value"
- `*value`
 - Select all functions that end with "value"
- `set_ftrace_notrace`

Filter Modules

```
acelan@acelan-nb /sys/kernel/debug/tracing % sudo su -c "echo :mod:nvidia >
set_ftrace_filter"
acelan@acelan-nb /sys/kernel/debug/tracing % cat set_ftrace_filter | head -n 10
__nv_setup_pat_entries
__nv_restore_pat_entries
nv_verify_page_mappings
nv_set_dma_address_size
nv_guest_pfn_list
nv_dma_to_mmap_token
nv_unlock_rm
nv_no_incoherent_mappings
nv_get_adapter_state
nv_get_smu_state
acelan@acelan-nb /sys/kernel/debug/tracing % cat set_ftrace_filter | wc -l
168
```

Filter Modules (cont.)

```
acelan@acelan-nb /sys/kernel/debug/tracing % sudo su -c "echo function > current_tracer"
acelan@acelan-nb /sys/kernel/debug/tracing % cat trace | head -n 15
# tracer: function
#
#      TASK-PID    CPU#    TIMESTAMP    FUNCTION
#      ||      |           |
Xorg-1300 [001] 18446733467.812990: os_memcpy_to_user <-_nv006635rm
Xorg-1300 [001] 18446733467.812991: os_free_mem <-_nv006612rm
Xorg-1300 [001] 18446733467.812991: os_release_sema <-_nv006510rm
Xorg-1300 [001] 18446733467.812992: nv_verify_pci_config <-rm_set_interrupts
Xorg-1300 [001] 18446733467.812997: os_acquire_spinlock <-_nv006653rm
Xorg-1300 [001] 18446733467.812997: os_release_spinlock <-_nv006509rm
Xorg-1300 [001] 18446733467.812999: nv_kern_unlocked_ioctl <-vfs_ioctl
Xorg-1300 [001] 18446733467.812999: nv_kern_ioctl <-nv_kern_unlocked_ioctl
Xorg-1300 [001] 18446733467.812999: nv_printf <-nv_kern_ioctl
Xorg-1300 [001] 18446733467.813000: nv_check_pci_config_space <-nv_kern_ioctl
Xorg-1300 [001] 18446733467.813000: os_get_current_time <-_nv006601rm
```

Function Graph Tracer

```
acelan@acelan-nb /sys/kernel/debug/tracing % sudo su -c "echo function_graph >
current_tracer"
```

```
acelan@acelan-nb /sys/kernel/debug/tracing % cat trace | head -n 15
```

```
# tracer: function_graph
```

```
#
```

```
# CPU DURATION FUNCTION CALLS
```

```
# | | | | |
1) 1.295 us | }
1) 3.100 us | }
1) | pick_next_task_fair() {
1) | set_next_entity() {
1) 0.432 us | update_stats_wait_end();
1) 0.443 us | __dequeue_entity();
1) 2.135 us | }
1) 0.425 us | hrtick_start_fair();
1) 3.885 us | }
1) + 10.265 us | }
1) + 14.835 us | }
```

Function Graph Tracer (cont.)

```
acelan@acelan-nb /sys/kernel/debug/tracing % sudo su -c "echo function_graph > current_tracer"
acelan@acelan-nb /sys/kernel/debug/tracing % sudo su -c "echo sys_read > set_graph_function"
acelan@acelan-nb /sys/kernel/debug/tracing % cat trace | head -n 20
# tracer: function_graph
#
# CPU DURATION FUNCTION CALLS
# | | | | |
1) 0.767 us | }
1) 0.798 us | _spin_lock_irq();
1) + 14.276 us | }
1) 0.929 us | inotify_inode_queue_event();
1) 0.773 us | __fsnotify_parent();
1) 0.791 us | inotify_dentry_parent_queue_event();
1) 0.827 us | fsnotify();
1) + 28.812 us | }
1) + 34.927 us | }
1) | sys_read() {
1) 0.713 us | fget_light();
1) | vfs_read() {
1) | rw_verify_area() {
1) | security_file_permission() {
1) 0.762 us | apparmor_file_permission();
1) 2.284 us | }
```

Tracing a Process

```
acelan@acelan-nb ~ % cat ~/bin/ftrace-me
#!/bin/sh
DEBUGFS=`grep debugfs /proc/mounts | awk '{ print $2; }'`
sudo su -c "\
    echo 0 > $DEBUGFS/tracing/tracing_on; \
    echo $$ > $DEBUGFS/tracing/set_ftrace_pid; \
    echo function_graph > $DEBUGFS/tracing/current_tracer; \
    echo 1 > $DEBUGFS/tracing/tracing_on"
exec $*
sudo su -c "\
    echo -1 > $DEBUGFS/tracing/set_ftrace_pid; \
    echo 0 > $DEBUGFS/tracing/tracing_on"
```

Who Call Me

```
acelan@acelan-nb /sys/kernel/debug/tracing % sudo su -c "echo kfree >
set_ftrace_filter"
acelan@acelan-nb /sys/kernel/debug/tracing % cat set_ftrace_filter
kfree
acelan@acelan-nb /sys/kernel/debug/tracing % sudo su -c "echo
function > current_tracer"
acelan@acelan-nb /sys/kernel/debug/tracing % sudo su -c "echo 1 >
options/func_stack_trace"
acelan@acelan-nb /sys/kernel/debug/tracing % cat trace | tail -5
=> __fput
=> fput
=> remove_vma
=> do_munmap
=> sys_munmap
acelan@acelan-nb /sys/kernel/debug/tracing % sudo su -c "echo 0 >
options/func_stack_trace"
acelan@acelan-nb /sys/kernel/debug/tracing % sudo su -c "echo >
set_ftrace_filter"
```

Reference

- Debugging the kernel using Ftrace – part 1
 - <http://lwn.net/Articles/365835/>
- Debugging the kernel using Ftrace – part 2
 - <http://lwn.net/Articles/366796/>
- Secrets of the Ftrace function tracer
 - <http://lwn.net/Articles/370423/>

Q/A

Ubuntu Hardware Summit

- Ubuntu Hardware Summit 2010
 - <http://www.ubuntu.com/uhs2010-zh>
- 日期及時間：
 - 9 月 24 日 (星期五) 早上 9:00AM
 -
- 地點：
 - 維多麗亞酒店 (美麗華購物中心旁)
 - 10466 台北市中山區敬業四路 168 號