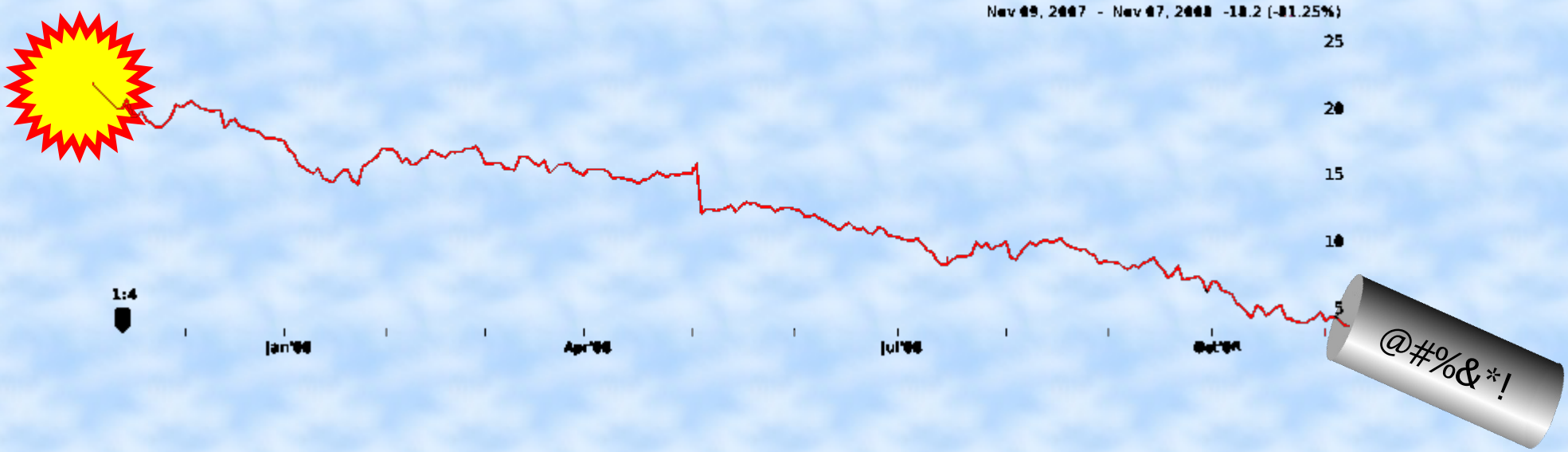


FUSE – Filesystems in Userspace

London OpenSolaris Usergroup, Nov'08

Frank Hofmann, fr.ch.hf@gmail.com

Userspace filesystems – why ?

Advantages of writing fs code in userspace

- Stable / documented filesystem interface
  well ... unless you're maintaining a kernel piece
- Portable filesystems become possible
  - Same code for Linux/*BSD/OpenSolaris/MacOSX
- All userland APIs available to your fs code:
  - Write a filesystem in Perl, Java, ...
  - Debug filesystem using userspace tools
  - Cache management: just let **`mmap()`** do it ...

Userspace filesystems - why ?

Advantages of writing fs code in userspace

- System security:
  - Filesystem as unprivileged user process/daemon
- System stability:
  - Crashing filesystems don't crash the kernel
  - Hanging filesystem code can simply be killed
  - Greedy filesystems can be resource-controlled
- Sidestep the licensing flamewars:
  - Run a CDDL filesystem with a GPL kernel
  - Run a GPL filesystem with a CDDL/BSD kernel

Userspace filesystems – how ?

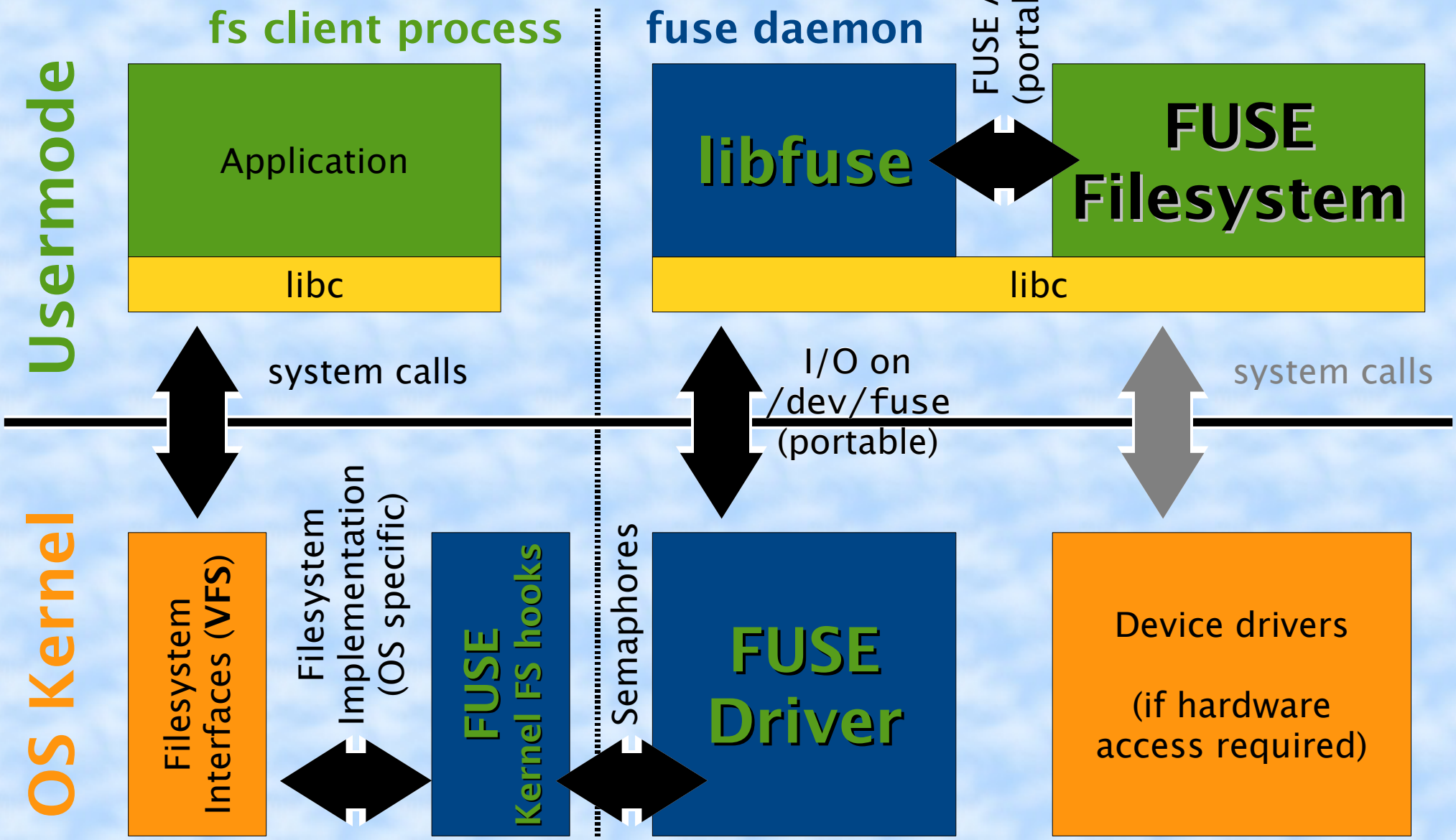
Some examples

- Intercept system calls via LD_PRELOAD
  - AVFS: http://www.inf.bme.hu/~mszeredi/avfs/
  - PlasticFS: http://plasticfs.sourceforge.net/
- Userland NFS server, 'virtual fs' backend:
  - Belenix' FSWmiscfs:
    http://www.genunix.org/distributions/belenix_site/binfiles/README.FSWfsmisc.txt
- Kernel / userland redirectors
  - LUFS (defunct), Coda, FUSE, ...

What is FUSE ?

- **F**ilesystem in **use**rspace
- OS-independent framework (abstraction layer) to implement a filesystem
  - Kernel component : Native OS fs interfaces
  - Userland component : FUSE ABI / library
  - Filesystem modules : portable !
- Available for Linux, *BSD, Solaris, MacOSX

FUSE Architecture



OpenSolaris FUSE, presented by Frank Hofmann

FUSE Filesystems

- Anything under the Sun / in the shade ...
  - NTFS-3G
  - ZFS-Linux
  - Lustre/GlobalFS
- Many special / virtual filesystems
  - Access archives (tar, cpio, zip, ...) as filesystems
  - GmailFS presents Google mail in a filesystem
- Most comprehensive list:

  http://fuse.sourceforge.net/wiki/index.php/FileSystems

FUSE filesystem example

```
//  FUSE: Filesystem in Userspace
//  Copyright (C) 2001-2005  Miklos Szeredi <miklos@szeredi.hu>
//  This program can be distributed under the terms of the GNU GPL.
//  See the file COPYING.
#define  FUSE_USE_VERSION 26
#include <fuse.h>
#include <stdio.h>
#include <string.h>
#include <errno.h>
#include <fcntl.h>

static const char *hello_str = "Hello World!\n", *hello_path = "/hello";

static int hello_open(const char *path, struct fuse_file_info *fi)
{
    if(strcmp(path, hello_path) != 0)
        return -ENOENT;
    if((fi->flags & 3) != O_RDONLY)
        return -EACCES;
    return 0;
}
```

FUSE filesystem example

```
static int hello_readdir(const char *path, void *buf, fuse_fill_dir_t filler,
                         off_t offset, struct fuse_file_info *fi)
{
    if(strcmp(path, "/") != 0)
        return -ENOENT;
    filler(buf, ".", NULL, 0);
    filler(buf, "..", NULL, 0);
    filler(buf, hello_path + 1, NULL, 0);
    return 0;
}

static int hello_getattr(const char *path, struct stat *stbuf)
{
    memset(stbuf, 0, sizeof(struct stat));
    if(strcmp(path, "/") == 0) {
        stbuf->st_mode = S_IFDIR | 0755;
        stbuf->st_nlink = 2;
    } else if(strcmp(path, hello_path) == 0) {
        stbuf->st_mode = S_IFREG | 0444;
        stbuf->st_nlink = 1;
        stbuf->st_size = strlen(hello_str);
    }
    return stbuf->st_nlink ? 0 : -ENOENT;
}
```

FUSE filesystem example

```
static int hello_read(const char *path, char *buf, size_t size, off_t offset,
                      struct fuse_file_info *fi)
{
    size_t len = strlen(hello_str);
    if(strcmp(path, hello_path) != 0)
        return -ENOENT;
    if (offset > len)
        return 0;
    size = MIN(size, len - offset);
    memcpy(buf, hello_str + offset, size);
    return size;
}

static struct fuse_operations hello_oper = {
    .getattr    = hello_getattr,    .readdir  = hello_readdir,
    .open       = hello_open,       .read     = hello_read,
};

int main(int argc, char *argv[])
{
    return fuse_main(argc, argv, &hello_oper);
}
```

"FUSE ABI" – similar to vfs / vnode ops

FUSE filesystem example

- Using the “hello” filesystem:

```
$ mkdir /tmp/fuse
$ ./hello /tmp/fuse
$ ls -la /tmp/fuse
total 9
drwxr-xr-x   2 root     root           0 Jan  1  1970 .
drwxrwxrwt  10 root     sys         1063 Nov 12 10:59 ..
-r--r--r--   1 root     root          13 Jan  1  1970 hello
$ cat /tmp/fuse/hello
Hello World!
$ df >/dev/null
Arithmetic Exception (core dumped)
$ fusermount -u /tmp/fuse
$ rmdir /tmp/fuse
```

Turns itself into a daemon process !

Internals: Daemon Idle State

```
# echo '::pgrep hello|::walk thread|::findstack; !pstack `pgrep hello`' | mdb -k
   stack pointer for thread ffffff01ce2a0140: ffffff000875fb10
   [ ffffff000875fb10 _resume_from_idle+0xf1() ]
     ffffff000875fb50 swtch+0x17f()
     ffffff000875fbc0 sema_p_sig+0x2a5()
     ffffff000875fc20 fuse_dev_read+0x92()
     ffffff000875fc50 cdev_read+0x3c()
     ffffff000875fcd0 spec_read+0x275()
     ffffff000875fd40 fop_read+0x69()
     ffffff000875fe90 read+0x28b()
     ffffff000875fec0 read32+0x1e()
     ffffff000875ff10 _sys_sysenter_post_swapgs+0x14b()
   1746:   ./hello /tmp/fuse/
    feebe605 read     (3, 807ec28, 21000)
    fef71f59 fuse_kern_chan_receive (8047034, 807ec28, 21000) + 79
    fef76fec fuse_chan_recv (8047034, 807ec28, 21000) + 4c
    fef722a2 fuse_session_loop (80626f8, 0) + 92
    fef710dc fuse_loop (80630d0, 0) + 1c
    fef77b53 fuse_main_common (2, 8047114, 80613a0, 98, 0, 0) + 53
    fef77bc1 fuse_main_real (2, 8047114, 80613a0, 98, 0, fee408bc) + 21
    080510c5 main     (2, 8047114, 8047120) + 36
    08050cb8 _start   (2, 8047258, 8047260, 0, 804726b, 80472cc) + 80
```

Internals: Watch it in action !

- Dtrace the communication:

```
$ dtrace -n '
fbt::fuse_dev_read:entry {self->t=1}
fbt::fuse_dev_read:return {self->t=0}
fbt::sema_p*:return /self->t/ {stack(); ustack(); }

fbt::fuse_queue_request_nowait:entry { stack(); ustack() }
fbt::frd_on_request_complete_wakeup:entry { stack(); ustack() }
'  -c "ls -l /tmp/fuse/hello"
```

Internals: Watch it in action !

- Client (ls): post request to FUSE filesystem

```
 CPU     ID                    FUNCTION:NAME
  1  60417  fuse_queue_request_nowait:entry
              fuse`fuse_queue_request_wait+0x3c
              fuse`fuse_access_i+0x1ae
              fuse`fuse_access+0x2f
              fuse`fuse_lookup+0x148
              genunix`fop_lookup+0xf2
              genunix`lookuppnvp+0x351
              genunix`lookuppnat+0x125
              genunix`lookupnameat+0x82
              genunix`cstatat_getvp+0x160
              genunix`cstatat64_32+0x7d
              genunix`lstat64_32+0x31
              unix`_sys_sysenter_post_swapgs+0x14b

              libc.so.1`lstat64+0x15
              ls`main+0x7da
              ls`_start+0x7a
```

Internals: Watch it in action !

- Daemon (hello): wakeup !

```
1  15125                    sema_p_sig:return
              fuse`fuse_dev_read+0x92
              genunix`cdev_read+0x3c
              specfs`spec_read+0x275
              genunix`fop_read+0x69
              genunix`read+0x28b
              genunix`read32+0x1e
              unix`_sys_sysenter_post_swapgs+0x14b

              libc.so.1`__read+0x15
              libfuse.so.2.7.1`fuse_kern_chan_receive+0x79
              libfuse.so.2.7.1`fuse_chan_recv+0x4c
              libfuse.so.2.7.1`fuse_session_loop+0x92
              libfuse.so.2.7.1`fuse_loop+0x1c
              libfuse.so.2.7.1`fuse_main_common+0x53
              libfuse.so.2.7.1`fuse_main_real+0x21
              hello`main+0x36
              hello`_start+0x80
```

Internals: Watch it in action !

- Daemon (hello): post answer

```
  1  60253 frd_on_request_complete_wakeup:entry
                fuse`fuse_dev_write+0x3b8
[ ... ]
                genunix`writev+0x38d
                genunix`writev32+0x1b
                unix`_sys_sysenter_post_swapgs+0x14b

                libc.so.1`__writev+0x15
                libfuse.so.2.7.1`fuse_kern_chan_send+0x30
                libfuse.so.2.7.1`fuse_chan_send+0x1a
[ ... ]
                libfuse.so.2.7.1`fuse_lib_access+0xe7
                libfuse.so.2.7.1`do_access+0x34
[ ... ]
                libfuse.so.2.7.1`fuse_loop+0x1c
                libfuse.so.2.7.1`fuse_main_common+0x53
                libfuse.so.2.7.1`fuse_main_real+0x21
                hello`main+0x36
                hello`_start+0x80
```

Internals: Watch it in action !

- Demo time !

References

- OpenSolaris FUSE project:

  http://opensolaris.org/os/project/fuse/
- FUSE website:

  http://fuse.sourceforge.net/
- Developing Filesystems with FUSE:

  http://www.ibm.com/developerworks/linux/library/l-fuse/
- NTFS-3G for OpenSolaris, install instructions:

  http://forums.opensolaris.org/message.jspa?messageID=1239

References

- Want ZFS for Linux ? Go ZFS Fuse !

  http://www.wizy.org/wiki/ZFS_on_FUSE

  https://developer.berlios.de/projects/zfs-fuse/

- Thanks for FUSE/OpenSolaris go to:
  - Mark Phalan & others