



Deep I/O performance analysis of CVMFS using modern Linux tools

CERN openlab summer student lightning talk session

Shahnur Isgandarli, EP-SFT summer openlab student

Supervisors: Jakob Blomer, Gerardo Ganis

13/08/2019

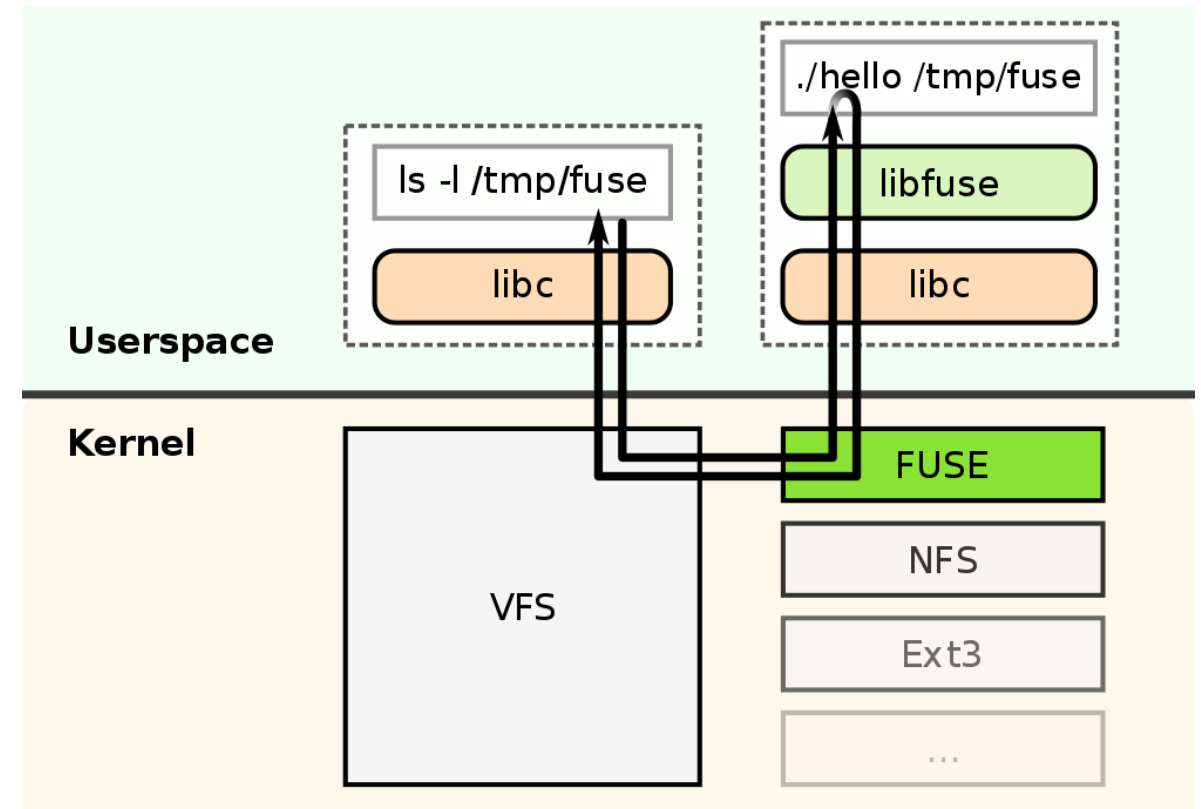
CernVM-File System

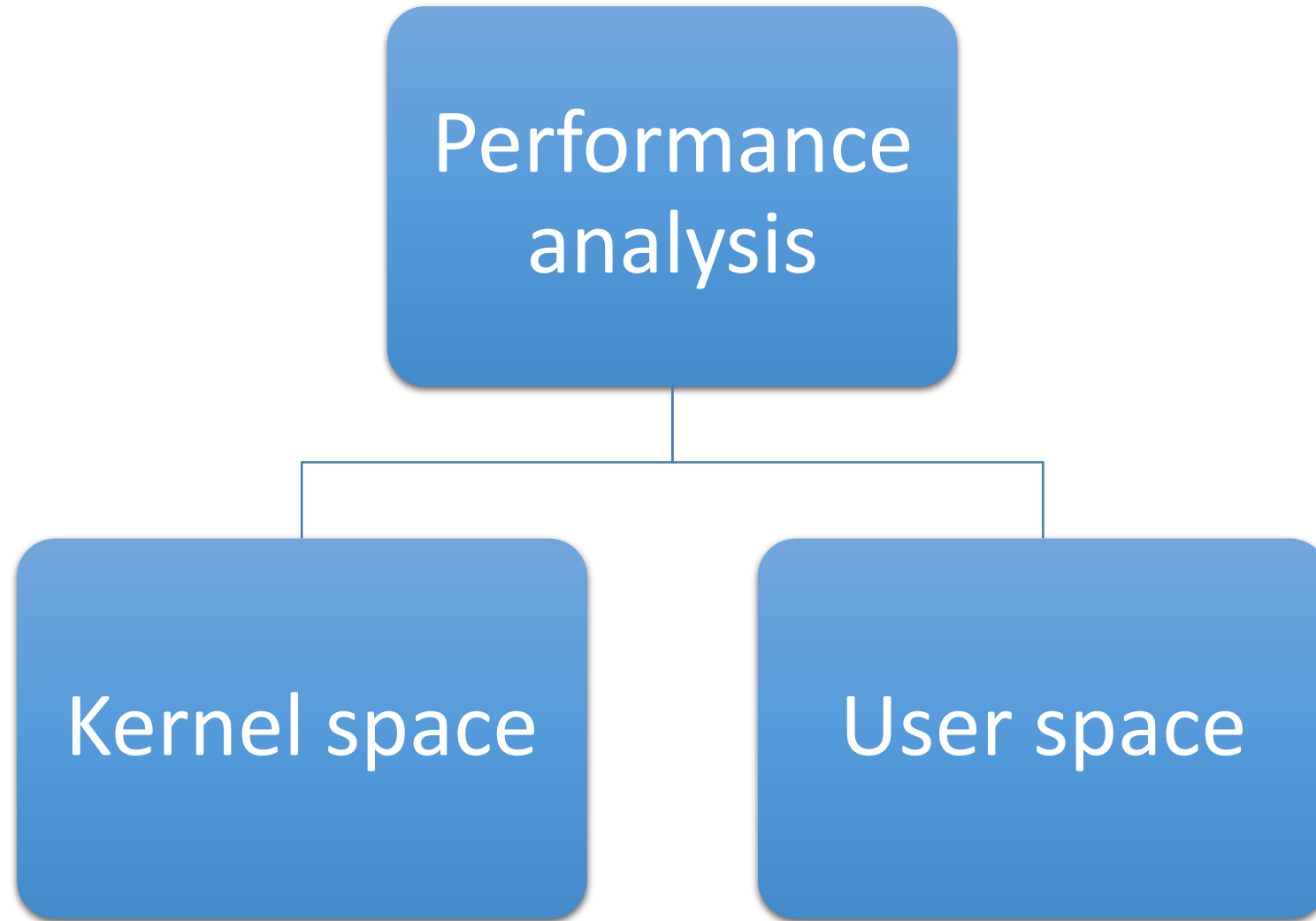
- A scalable, reliable and low-maintenance software distribution service.
- Developed to assist HEP collaborations to deploy software on the worldwide-distributed computing infrastructure used to run data processing applications



FUSE - Filesystem in Userspace

1. The request to list the files in the directory gets redirected by kernel through VFS to FUSE.
2. FUSE executes the registered handler program (./hello) and passes the request to it.
3. The handler program returns a response back to FUSE which is then redirected to the userspace program that originally made the request.





Kernel space

```
x64 irq 1,12
[ 1.085227] serio: i8042 KBD port at 0x60,0x64 irq 1
[ 1.086072] serio: i8042 AUX port at 0x60,0x64 irq 12
[ 1.087317] mousedev: PS/2 mouse device common for all mice
[ 1.089087] input: AT Translated Set 2 keyboard as /devices/platform/i8042/ser
rio0/input/input0
[ 1.091075] rtc_cmos rtc_cmos: registered as rtc0
[ 1.091929] rtc_cmos rtc_cmos: alarms up to one day, 114 bytes nvram
[ 1.094206] device-mapper: uevent: version 1.0.3
[ 1.096061] device-mapper: ioctl: 4.39.0-iocli (2018-04-03) initialised: dm-d
evel@redhat.com
[ 1.099224] Initializing XFRM netlink socket
[ 1.100206] NET: Registered protocol family 17
[ 1.101236] Key type dns_resolver registered
[ 1.103166] AUX version of gcm_enc/dec engaged.
[ 1.104095] AES CTR mode by8 optimization enabled
[ 1.132208] sched_clock: Marking stable (1132194021, 0)->(1696167804, -563973
783)
[ 1.134349] registered taskstats version 1
[ 1.135110] Loading compiled-in X.509 certificates
[ 1.143821] Key type encrypted registered
[ 1.887377] tsc: Refined TSC clocksource calibration: 2493.720 MHz
[ 1.888969] clocksource: tsc: mask: 0xffffffffffffffff max_cycles: 0x23f20d36
9de, max_idle_ns: 440795277732 ns
```



<https://en.wikipedia.org/wiki/Linux>
https://en.wikipedia.org/wiki/Linux_kernel

BCC – BPF Compiler Collection

- BCC - Tools for BPF¹-based Linux IO analysis, networking, monitoring, and more.
- Kprobe-based tracing
- Those tools were handy for analysis of the kernel space.

1 - Berkeley Packet Filters

<https://github.com/iovisor/bcc>

Kernel space analysis

```

^X[shahnur@fedora fuse_lookup_name_lat]$ sudo python fuse_lookup_name_lat.py 500
Tracing... Hit Ctrl-C to end.
^C
  usecs      : count   distribution
  0 -> 1     : 0
  2 -> 3     : 0
  4 -> 7     : 0
  8 -> 15    : 0
 16 -> 31    : 55
 32 -> 63    : 467      *****
 64 -> 127   : 2774     *****
128 -> 255   : 2472     *****
256 -> 511   : 8
512 -> 1023  : 5
1024 -> 2047 : 5
2048 -> 4095 : 2
[shahnur@fedora fuse_lookup_name_lat]$

```

Latency distribution of “lookup_name” call

ADDR	FUNC	COUNT
ffffffffc0aa1881	fuse_statfs	1
ffffffffc0a983f1	fuse_dentry_release	1
ffffffffc0aa1781	fuse_show_options	60
ffffffffc0a98731	fuse_do_getattr	279
ffffffffc0a989f1	fuse_perm_getattr	281
ffffffffc0aa10a1	fuse_inode_eq	29836
ffffffffc0aa1f31	fuse_inode_init_once	89578
ffffffffc0aa19f1	fuse_alloc_inode	89585
ffffffffc0aa25c1	fuse_iget	89585
ffffffffc0a9b1b1	fuse_init_dir	89585
ffffffffc0aa10c1	fuse_inode_set	89585
ffffffffc0a99f31	fuse_lookup_name	89586
ffffffffc0a9a131	fuse_lookup	89586

FUSE callback counters

User space



Source code instrumentation is needed!

William Pina: https://commons.wikimedia.org/wiki/File:Linux_Kernel_panic.png

Implementation

```
static void cvmfs_lookup(fuse_req_t req, fuse_ino_t parent, const char *name)
    HighPrecisionTimer guard_timer(file_system_>hist_fs_lookup());

    /* implementation */
}
static void cvmfs_opendir(fuse_req_t req, fuse_ino_t ino,
    struct fuse_file_info *fi)
{
    HighPrecisionTimer guard_timer(file_system_>hist_fs_opendir());

    /* implementation */
}
```

Measuring cvmfs_opendir and cvmfs_lookup

```
class HighPrecisionTimer : SingleCopy {
public:
    explicit HighPrecisionTimer(Log2Histogram *recorder)
        : timestamp_start_(platform_monotonic_time_ns())
        , recorder_(recorder)
    { }

    ~HighPrecisionTimer() {
        recorder_>Add(platform_monotonic_time_ns() - timestamp_start_);
    }

private:
    uint64_t timestamp_start_;
    Log2Histogram *recorder_;
};
```

HighPrecisionTimer data structure

Analysis (lookup call)

Lookup

	nsec	count	distribution
0 ->	1 :	0	
2 ->	3 :	0	
4 ->	7 :	0	
8 ->	15 :	0	
16 ->	31 :	0	
32 ->	63 :	0	
64 ->	127 :	0	
128 ->	255 :	0	
256 ->	511 :	0	
512 ->	1023 :	0	
1024 ->	2047 :	0	
2048 ->	4095 :	0	
4096 ->	8191 :	0	
8192 ->	16383 :	2242	
16384 ->	32767 :	12530	***
32768 ->	65535 :	106964	*****
65536 ->	131071 :	15820	****
131072 ->	262143 :	532	
262144 ->	524287 :	124	
524288 ->	1048575 :	5	
1048576 ->	2097151 :	4	
2097152 ->	4194303 :	10	
4194304 ->	8388607 :	3	
8388608 ->	16777215 :	1	
16777216 ->	33554431 :	0	
33554432 ->	67108863 :	0	
67108864 ->	134217727 :	0	
134217728 ->	268435455 :	0	
268435456 ->	536870911 :	0	
536870912 ->	1073741823 :	0	
overflow :		0	
total :		138235	

Without kernel-level caching

Lookup

	nsec	count	distribution
0 ->	1 :	0	
2 ->	3 :	0	
4 ->	7 :	0	
8 ->	15 :	0	
16 ->	31 :	0	
32 ->	63 :	0	
64 ->	127 :	0	
128 ->	255 :	0	
256 ->	511 :	0	
512 ->	1023 :	0	
1024 ->	2047 :	0	
2048 ->	4095 :	0	
4096 ->	8191 :	0	
8192 ->	16383 :	1	
16384 ->	32767 :	34	
32768 ->	65535 :	1226	*****
65536 ->	131071 :	5148	*****
131072 ->	262143 :	1253	*****
262144 ->	524287 :	0	
524288 ->	1048575 :	0	
1048576 ->	2097151 :	0	
2097152 ->	4194303 :	0	
4194304 ->	8388607 :	1	
8388608 ->	16777215 :	0	
16777216 ->	33554431 :	0	
33554432 ->	67108863 :	0	
67108864 ->	134217727 :	0	
134217728 ->	268435455 :	0	
268435456 ->	536870911 :	0	
536870912 ->	1073741823 :	0	
overflow :		0	
total :		7663	

With kernel-level caching

Summary of the project

Now we have:

- A powerful set of tools to look in user and kernel spaces of FUSE calls
- A toolset that enables fine-grained performance engineering of CVMFS client
- log2 histogram data structure for latency measurements is already merged into the CVMFS devel branch

Thank you for your attention!

Contact information

Shahnur Isgandarli

Computer Science student



CHALMERS
UNIVERSITY OF TECHNOLOGY



shahnur@student.chalmers.se

<https://www.linkedin.com/in/sisgandarli/>