CLC _____ Number _____

UDC _____ Available for reference ☐Yes ☐No

SUSTech Southern University of Science and Technology

Undergraduate Thesis

Thesis Title: Real-time capturing of system calls on ARM

Student Name: Haonan Li

Student ID: 11712510

Department: Department of Computer Science and Engineering

Program: Computer Science and Technology

Thesis Advisor: Fengwei Zhang

Date: May 10, 2021

COMMITMENT OF HONESTY

1. I solemnly promise that the paper presented comes from my independent research work under my supervisor's supervision. All statistics and images are real and reliable.

2. Except for the annotated reference, the paper contents no other published work or achievement by person or group. All people making important contributions to the study of the paper have been indicated clearly in the paper.

3. I promise that I did not plagiarize other people's research achievement or forge related data in the process of designing topic and research content.

4. If there is violation of any intellectual property right, I will take legal responsibility myself.

Signature:

Date:

Real-time capturing of system calls on ARM

Haonan Li

(Department of Computer Science and Engineering Advisor: Fengwei Zhang)

[**ABSTRACT**]: Bug diagnosis is difficult. The first step of bug diagnosis is to reproduce the bug. In areas such as application development, developers usually can only rely on the report logs uploaded by the user to try to reproduce bugs. Unfortunately, it is still challenging to reproduce bugs that occurred in the production environment at the development environment. The primary obstacle of reproduction is non-deterministic events at runtime, such as system calls. Hence, the same execution may lead to different results.

In this thesis, I present Syscord, a practical tool for recording system calls on Linux. Syscord utilizes Linux tracepoints to hook system calls. Syscord collects relevant information for each system call related to their effects, which further helps developers reproduce and fix bugs. I implement a prototype of Syscord and evaluate it with real-world applications. The result demonstrates the Syscord capturing system calls entirely and efficiently.

[**Keywords**]: Linux, Syscall, Record

Contents

1. Intoduction

The program often fails. To sufficiently understand and prevent failures, developers requires firstly reproduce these bugs, which ensures the same output and bugs. However, directly re-exection is not suitable for non-deterministic failures, as they may not appear in a re-execution procedure. Non-deterministic failures are the consequence of non-deterministic instructions.

Instructions for running a program can be divided into two categories. One is deterministic, which means the behavior of deterministic instruction is determined in each execution. The other type is non-deterministic, meaning that execution in different situations will have different results. Although most of the CPU execution is deterministic (e.g., `ADD`), non-deterministic instructions (e.g., get user input) are also pervasive. Typical sources of nondeterminism include system calls, interrupts, signals, and data races for concurrency programs[1]. All these non-deterministic events can be futher classified into two types: inconstancy of the data flow - for example, certain system calls such as `getrandom` and `getpid`, and inconstancy of the control flow - for example, concurrency bug due to memory access in inconsistent order[2].

Record-and-replay is a type of approaches that addresses this challenge. Most Record-and-replay systems work by first recording non-deterministic events during the original run of a program and then substituting these records during subsequent re-execution. Record-and-replay system could ultimately guarante that each replay will be identical with the initial version. The fact that a number of replay systems have been built and put into use in recent years illustrates the value of record-and-replay systems in practice[3-6].

There is a rich amount of research on record-and-replay systems, and we can find their various treatments of non-deterministic records. Early record-and-replay systems tend to use virtualization techniques so as to observe and record the entire program non-deterministically on the hypervisior, but the virtual machine is very heavy[7, 8]. Some systems use dynamic binary instrumentation to get the results after running each instruction, but this is very inefficient[6]. There are some other systems that choose not to record at runtime in order to address the expensive cost of recording; instead, they infer these non-deterministic events based on the control flow and other information collected[5, 9]. However, inference often does not reproduce program execution as faithfully as records, and the time required for inference, which in the worst case is a search of the entire space, is a problem[4]. There are also systems that use custom hardware, which inevitably affects its usefulness in practice[10]. Recently there have been some practical systems that have adopted tools provided by Linux for tracing, thus achieving better efficiency. Nevertheless, it still introduces a considerable overhead (50%) and is therefore only available when the developer exactly needs to record and replay[3].

This thesis focuses on the data record part of record-and-replay systems, precisely, the recording of non-deterministic events caused by system calls. I argue that a *practical* record system should (1) run online, meaning that the recording has little performance impact on the execution of the target program, (2) log all data without any omission, (3) work on commercial off-the-shelf hardware, (4) not require any modification to the target program, and also (5) not require any modification to the kernel.

In this thesis, I propose SYSCORD, a practical solution for syscall capturing. It works with unmodified Linux programs on commercial off-the-shelf (OTS) hardware. My original design was on the ARM platform, but the system can be applied to other platforms as well (e.g. x86, riscv). I demonstrat its usefulness on both x86 and ARM platforms. SYSCORD consists of three component: *core hook*, *filter* and *record buffer*.

The *core hook* is a probe of system call. *core hook* inspects each system call, and collects the effects on memory and registers by considering the semantics of system calls. The *filter* stores relevant information of the process what issues the system call, and compares this information with the characteristics specified by the developer. The *record buffer* temporarily store the recording of system calls and dumps it to file.

We implement a prototype of SYSCORD and evaluate it with the aforementioned requirements in mind. The evaluation results show that SYSCORD completely records system calls. We also leverage SYSCORD to record 16 failed programs (7 code segments reconstructed from application and 9 real-world applications including Python, Memcached, and SQLite). The recording indicates that SYSCORD effectively records system calls with a performance overhead of up to 3.88% on average. Meanwhile, SYSCORD directly works on the unmodified binary of the target program and does not rely on any hardware modification.

In summary, I make the following contributions:

- I present a system call recording tool named SYSCORD on Arm platforms, which works with unmodified binary on Arm platform without hardware modification.
- I achieve high performance that allows the always-on trace for the production environment, which provides SYSCORD the ability to reconstruct the entire records.
- I implement a prototype of SYSCORD and evaluate it with real-world applications. The evaluation result demonstrates that SYSCORD successfully records various types of applications with up to 3.88% runtime performance overhead on average.

2. Background: Linux Trace

From trace maker to tracepoint

3. Design

The overarching target of SYSCORD is to capture all syscalls issued by target application with low overhead. Specifically, it has the following characteristics:

- **online:** Our scenario is that our entire record system can work simultaneously on the user side and eventually become a part of the log report. Therefore it must introduce only a extremely low overhead to guarantee that the user can run the desired program without perception.
- **complete:** We also need to ensure the integrity of record results, i.e. that every syscall is correctly captured with all the data needed for reproduction. Not only do we need to verify the correctness of our records on each syscall, but we also have to guarantee that the data in the buffer is fetched in a timely manner without any overflow.
- **off-the-shore:** An practical system should never make assumptions about the hardware. SYSCORD is desgined completely based on the off-the-shelf hardware.
- **without modification to application:** We do not make any changes to the source code or binary of the target application. The entire code of our system runs in kernel space, except for the transfer of logged results that need to be transferred to user space. This means that SYSCORD cares nothing about how the target application in user space is executed, but only about the event that this target application jumps to the kernel state.
- **without modification to kernel:** Modifying Kernel source may make all procedures much more easy. However, this approach will inevitably reduce the compatibility of SYSCORD, especially for devices whose kernels have been modified by manufacturers. Futhermore, to modify kernel, we would also need to synchronize upstream changes. Consequently, SYSCORD is baesd on kernel moudle and works as a loadbale driver.

SYSCORD realizes syscall capturing in three steps. (1) SYSCORD inspect and record each syscall. (2) Next, SYSCORD filter these records by some arrtibutions of its caller process. (3) Finally, SYSCORD transfer these filtered records.

3.1 Desgin Overview

In this section, I present the desgin of SYSCORD by focusing on how it addresses the above two key challenges. SYSCORD contians three parts: *core hook*, *filter*, and *record buffer*.

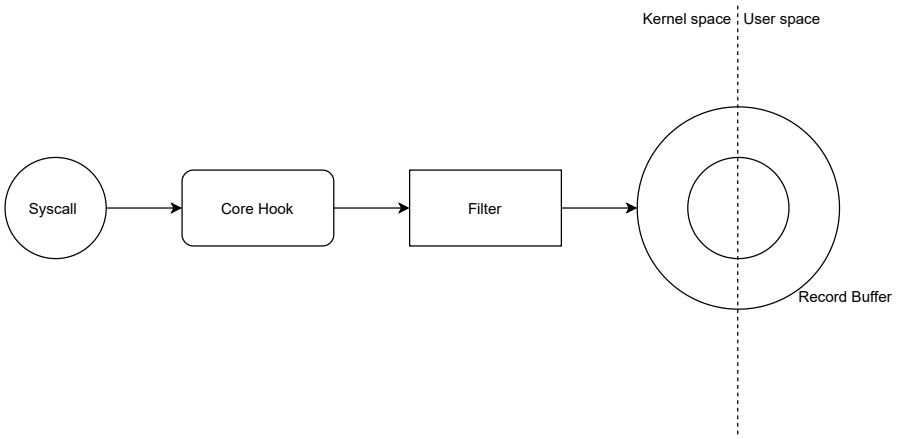

Figure 1: The Overview of SYSCORD

```
1    // Thread 1::
2    char big_buf[64];
3    while(1)
4      read(fd, big_buf, 64);
```

```
5    // Thread 2::
6    int total = 0;
7    int len = 0;
8    char buf[15];
9    for (short i=0; i < 2; ++i)
10     len = strlen(big_buf);
11     if (len < 15)
12       strcpy(buf, big_buf);
13       total += len;
14   assert(total<30)
```

Figure 2: Buffer Overflow Caused by Data Race

As Figure 1 shows, in the kernel space, *core hook* defines callback funtions in syscall, and will firslty inspect each syscall and then transfer relevant information to *filter* part. Subsequently, at the *filter* part, it will find process information from the kernel, and filter syscall records with specific features (e.g., process id or name) and finally pass to *record buffer*. The record buffer has two components: buffer management in kernel space and fetch daemon in user space. The daemon in user space will check the buffer periodically and dump these data form buffer to file.

3.2 Case Study

Figure 2 demonstrates a typical concurrency bug related to the non-deterministic event (syscall `read`). Assume that the loop (Line 10 to Line 13) in Thread 2 is executed twice. In the first iteration, the `read` of `big_buf` (Line 4) in Thread 1 is performed after the length check (Line 10 and Line 11) in Thread 2. The following `strcpy` (Line 12) in Thread 2 may lead to a buffer overflow and overwrite variables `len` and `total`. In the second iteration, no data race is involved, but the unpredictable `total` overwritten in the first iteration might be larger than 30 after the summation (Line 13) in Thread 2. This finally fails the `assert` (Line 14) in Thread 2 and crashes the program. In this example, the memory and registers indicating the root cause of the bug are overridden

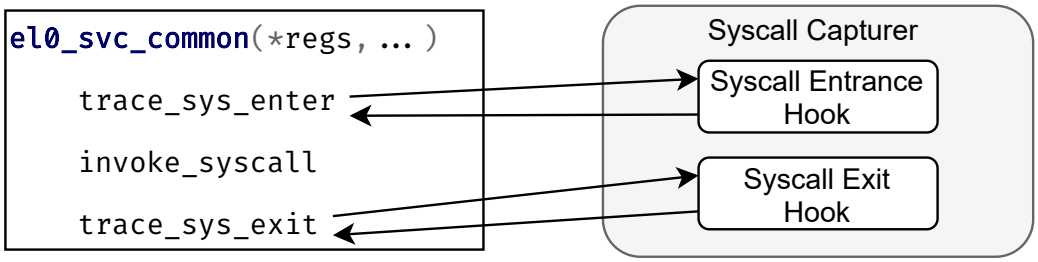

Figure 3: The Two hooks of *core hook*

Table 1: Part of Registers Used in Syscalls[12].

arch	syscall number	return value	return value 2	arg0	arg1	arg2
arm64	w8	x0	x1	x0	x1	x2
x86-64	rax	rax	rdx	rdi	rsi	rdx
riscv	a7	a0	a1	a0	a1	a2

by subsequent control flow. Hence, it is necessary to record the content of syscall `read` to figure out the bug.

3.3 Core Hook

Figure 3 shows the general workflow of *core hook*. It consists of two stages of hooks, at the beginning and end of kernel handling of syscall. For the stage 1, SYSCORD will follow the start of syscall handlers and save the value of first parameter, if this syscall may change the memory addressed by first argument.

The second stage takes on more responsibility, including the recording of other pointer type parameters (except for the first one), and return values. Besides, this stage should also get the relevant information collected by stage 1.

There is still a problem that, due to the concurrency of the system, there may be multiple system calls being processed at the same time. Especially considering that the system calls processed in a relatively long period. So, there will be multiple system calls going to different stages of *core hook*.

This problem is solved by the observation that syscall will block the thread in user space. Therefore, we can find a one-to-one mapping from thread number to a system call event at any moment. I maintain a table to save these correspondences in second stage, and also get its first parameters from stage 1.

3.3.1 Loss of Syscall Parameters

As Table 1 shows, the register `x0` and `x1` are used as both parameters and return values of syscalls on Arm64. Thus, these registers are overwritten by return values during syscall procedure. Consequently, we may not obtain the syscall parameters at the *syscall exit hook* directly. However, these parameters are critical in some cases.

For example, as Figure 2 demonstrates, the content of `big_buf`, addressed by the first parameter of `read`, is necessary for bug analysis.

3.4 Filter

The `filter` part is a relatively simple component that requires information about the caller of the syscall from the Linux kernel. Then it will perform filter by pre-passed conditions. Last, it passes this filtered information on to the next part.

3.5 Record Buffer

The *record buffer* is intended to act as a transit between kernel space and user space. Therefore it has two parts located in kernel space and user space respectively. One of the simplest designs is to maintain a daemon in user space constantly querying and retrieving the data stored in the buffer, and then dumping the data to a file. However, we note that this introduces a huge amount of overhead, mainly due to frequent file io. Placing a larger buffer in user space would also solve this problem, but SYSCORD do not want to introduce a large impact on the entire system.

Therefore, *record buffer* is designed to keep fetching the buffer occupancy, and then dumps the whole buffer only after finding that the buffer occupancy has reached a threshold.

4. Implementation

4.1 Core Hook

The part of *core hook* aims to hook system call, i.e., inject custom code at the begin and the end of a system call. By leveraging *core hook*, It is easy to get the return value of the system call and the changes to the parameters. Linux has provided many different approaches to achieve it, such as *ptrace*, *auditd*, *Kprobe* and *tracepoint*.

I choose *tracepoint* to implement our core hook, since it

4.2 Filter

We get the process descriptor of syscall issuer (the process using the syscall) via `current`, and compare it with conditions passed in.

4.3 Record buffer

The buffer is a circular queue. And there is also a number indicating the usage of the buffer.

The most sophisticated component is to share the buffer between kernel space to user space.

5. Related Work

There is a large amount of related work dedicated to cpaturing syscalls. In this section, I discuss some representative examples and describe how SYSCORD differs.

- **Pinplay** is a …
- **REPT** …
- **rr** …
- **DTrace** …
- **sysdig** …

Bibliography

[1] RONSSE M, DE BOSSCHERE K. RecPlay: a fully integrated practical record/replay system[J]. ACM Transactions on Computer Systems, 1999.

[2] MICHAEL H, DENYS V. Getrandom(2) —Linux manual page[EB/OL]. 2021 [2021-03-27]. https://man7.org/linux/man-pages/man2/ptrace.2.html.

[3] O' CALLAHAN R, JONES C, FROYD N, et al. Engineering Record and Replay for Deployability[C]//2017 USENIX Annual Technical Conference (USENIX ATC 17). Santa Clara, CA: USENIX Association, 2017: 377-389.

[4] CHEN Y, ZHANG S, GUO Q, et al. Deterministic Replay: A Survey[J]. ACM Comput. Surv., 2015, 48(2). DOI: 10.1145/2790077.

[5] ALTEKAR G, STOICA I. ODR: output-deterministic replay for multicore debugging[C]//Proceedings of the ACM SIGOPS 22nd symposium on Operating systems principles - SOSP '09. Big Sky, Montana, USA: ACM Press, 2009: 193 [2021-03-21]. DOI: 10.1145/1629575.1629594.

[6] BHANSALI S, CHEN W K, de JONG S, et al. Framework for instruction-level tracing and analysis of program executions[C]//VEE '06: Proceedings of the 2nd international conference on Virtual execution environments. New York, NY, USA: Association for Computing Machinery, 2006: 154-163.

[7] DUNLAP G W, KING S T, CINAR S, et al. ReVirt: enabling intrusion analysis through virtual-machine logging and replay[J]. ACM SIGOPS Operating Systems Review, 2003, 36(SI): 211-224.

[8] DUNLAP G W, LUCCHETTI D G, FETTERMAN M A, et al. Execution Replay of Multiprocessor Virtual Machines[C]//VEE '08: Proceedings of the Fourth ACM SIGPLAN/SIGOPS International Conference on Virtual Execution Environments. Seattle, WA, USA: Association for Computing Machinery, 2008: 121-130. DOI: 10.1145/1346256.1346273.

[9] CUI W, GE X, KASIKCI B, et al. REPT: Reverse Debugging of Failures in Deployed Software[C]//OSDI'18: Proceedings of the 13th USENIX Conference on Operating Systems Design and Implementation. Carlsbad, CA, USA: USENIX Association, 2018: 17-32.

[10] MONTESINOS P, HICKS M, KING S T, et al. Capo: A Software-Hardware Interface for Practical Deterministic Multiprocessor Replay[J]. SIGARCH Comput. Archit. News, 2009, 37(1): 73-84. DOI: 10.1145/2528521.1508254.

[11] University of California. Syscall(2) - Linux manual page[EB/OL]. 2021. https://man7.org/linux/man-pages/man2/syscall.2.html.

[12] University of California. Syscall(2) - Linux manual page[EB/OL]. 2021 [2021-03-28]. https://man7.org/linux/man-pages/man2/syscall.2.html.