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 Syscord capturing system calls completely 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. We 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, we 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, RISC-V). we 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, we make the following contributions:

- We present a system call recording tool named SYSCORD on ARM platforms, which works with unmodified binary on ARM platform without hardware modification.

- We achieve high performance that allows the always-on trace for the production environment, which provides SYSCORD the ability to reconstruct the entire records.

- We 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. 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 run 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.

2.1 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*.

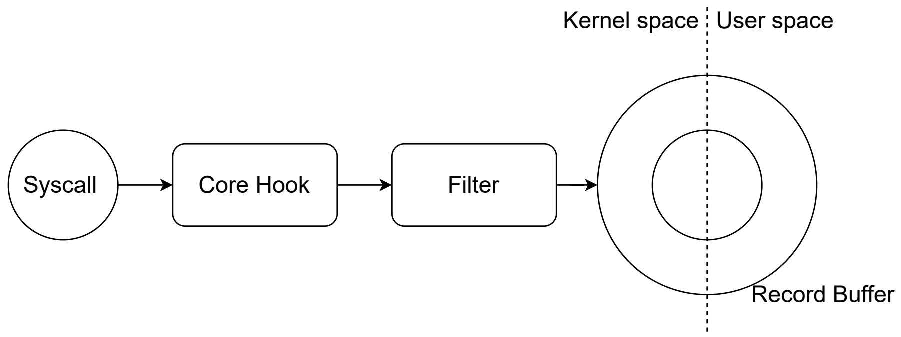


Figure 1: The Overview of SYSCORD

```
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* collect information for each system call, and then transfer 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 manage the buffer in kernel space and file write to user space.

2.2 Case Study

Figure 2 demonstrates a typical concurrency bug related to the 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 by subsequent control flow. Hence, it is necessary to record the content of syscall `read` to figure out the bug.

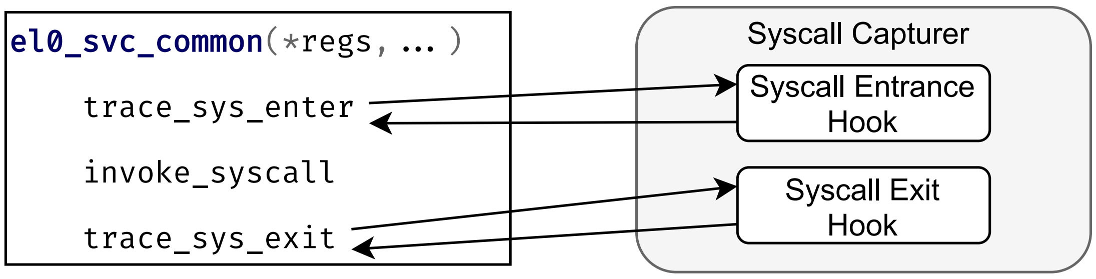


Figure 3: The Two hooks of *core hook*

Table 1: Part of Registers Used in Syscalls[11].

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
RISC-V	a7	a0	a1	a0	a1	a2

2.3 Core Hook

Core hook utilizes *Linux Kernel Tracepoints*[12], which provides lightweight hook points in critical positions over the kernel (includes the enter and exit of syscall), to achieve the capture. Figure 3 illustrates the procedure of syscalls (defined at function `el0_svc_common` for ARM64) and the recording for syscalls.

When entering a syscall, the kernel goes into *syscall entrance hook* which stores values of involved registers. 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 failure analysis. Therefore, we need to record the corresponding parameter at the *syscall entrance hook*, and further record the content addressed by `big_buf` at the *syscall exit hook*.

Before exiting a syscall, the kernel falls into *syscall exit hook*. This hook records different information for various types of syscalls. Firstly, if the capturer notices that a syscall only changes specific registers (e.g., `getpid` only changes return value), it could selectively save information to reduce the size of record file. Secondly, we need to record additional information for certain syscalls; for example, the syscall `read(int fd, void *buf, size_t count)` tries to fill `count` bytes to the memory region addressed by `buf`. In this case, we have to notice and record the write to this memory region.

For a better understanding of the impact of syscalls, we classify them into four types according to their effect on memory and registers.

- **R***eading* **S***tatus*: The *RS-Type* syscalls aim to read information related to system status. The results of these syscalls may be transferred by the return value (e.g., `getpid`), a pointer (e.g., `getitimer`), or shared memory (e.g., `getrandom`). For all syscalls in this category, we can directly record the memory or register they changed.

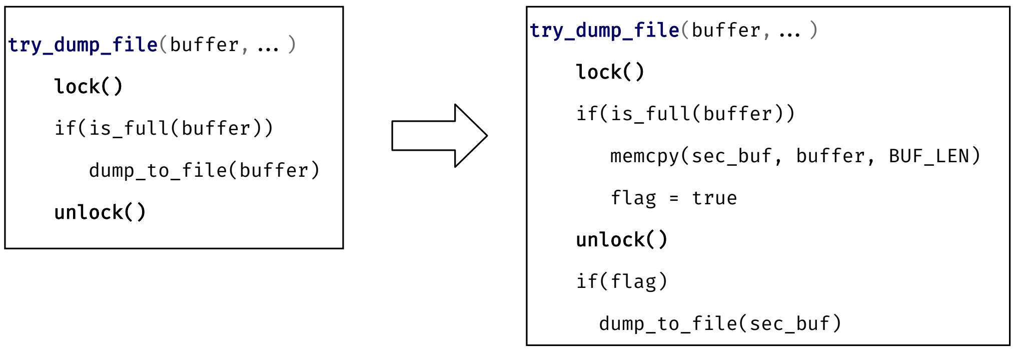
- **W***riting* **S***tatus*: The *WS-Type* syscalls changes the status of the system. As the *WS-Type* syscalls do not directly change the memory and registers of the program, we ignore them unless they fail and return an error code. For example, the syscall `epoll_create` in `epoll` API is ignored, but the influence of syscall `epoll_wait` is recorded since we classify it as an *RS-Type* syscall.
- **R***eading* **C***ontent*: The *RC-Type* syscalls read content from an external input, and the handling of *RC-Type* syscalls is similar to that of *RS-Type* syscalls. However, since the content is usually much larger than the status read in the *RS-Type* syscalls, we choose to truncate the content and record only the first 256 bytes due to performance requirements. Recording the entire content may become impractical in extreme cases (e.g., data center with tremendous uploads) as it results in a huge size of record file. Besides, we find that the truncated content is sufficient for bug diagnosis in most cases. A further evaluation is discussed in §4.4.4.
- **W***riting* **C***ontent*: The *WC-Type* syscalls write content to an external source. We consider that they would not affect the execution status of the target program. For example, the syscall `write` is ignored, and the written content is recorded if the *RC-Type* syscall `read` is used to read from the source again.

2.4 Filter

The *filter* is responsible for reducing the record size by applying filter conditions, and then only transfer relevant system calls to *record buffer*. The *filter* can filter the process who issued the syscall by its attributions (e.g., process id, process name, and parent process id), as it leverages `current` to collect relevant information about the issuer. The `current` points to the process descriptor `task_struct` which contains all the information about the executing process. During a procedure of system call, `current` pointer indicates the issuer of the syscall[13].

2.5 Record Buffer

The *record buffer* intends to transit between kernel space and user space. As Figure 4 demonstrates, considering that dumping to a file is relatively slow, the left design of Figure 4 causes `dump_to_file` to block system calls throughout the entire OS. In



```
try_dump_file(buffer, ... )
    lock()
    if(is_full(buffer))
       dump_to_file(buffer)
    unlock()
```

```
try_dump_file(buffer, ... )
    lock()
    if(is_full(buffer))
       memcpy(sec_buf, buffer, BUF_LEN)
       flag = true
    unlock()
    if(flag)
      dump_to_file(sec_buf)
```

Figure 4: The procedure for dumping file

contrast, the design on the right only locks the `memcpy`; after the memory copy, other simultaneous system calls can directly write to `buffer`.

There is a risk that the buffer fills up again before `dump_to_file(sec_buf)`, which makes part of the records lost. However, even in some extreme cases, this risk does not arise. The evaluation results are provided in §4.2.

3. Implementation

In this section, we describe our prototype of SYSCORD. We implement SYSCORD with more than `1.6k` lines C code on an Armv8 juno board based on Linux. In our prototype of SYSCORD, we analyze the semantics of the 60 commonly used syscalls. For each syscall, SYSCORD generates a description including return address, process id and its effects. We also use two 16 MB buffers to temporarily store the syscall records and then dump this buffer to files when it is full.

In the *Core Hook* part, As Figure 3 demonstrates, the entrance hook and the exit hook inspect the parameters and the impact of a syscall, respectively. At the entrance hook, SYSCORD stores the syscall number and the first two parameters (i.e., saved in `x0` and `x1`) into a hashmap. The key for the hashmap is the `thread_id` because a thread can only issue a single syscall until exit. At the exit hook, SYSCORD identifies the type of current syscall and applies specific capturing rules.

To avoid the uncertain influence of using Linux APIs (e.g., different versions with different implementations), I re-write some interfaces relevant to string operations (e.g., `strncpy`, `sprintf`). Besides, we further apply a simple encoding rule (i.e., encode the numbers to bytes) to reduce the recording size.

The *Filter* gets the filter conditions passed by the user, SYSCORD supports the filter of process id, process name, parent process id, and parent process name. The syscall satisfying all passed conditions then delivers to *Record Buffer.*

The *Record Buffer* maintain two global buffers to store records temporarily. Considering the parallelism of syscalls (i.e., different processor can tickle syscalls at the same time), writing to buffer or writing to file needs locks to prevent data races. I choose to use spinlock before each writes operation to buffer. For file writing, I use a pair of spinlock to protect a global flag and do not set any lock to protect writing to the file itself. Although simultaneous writing to file is unacceptable, SYSCORD usually can not fill a buffer before the last writing is finished. The evaluation about the writing file provides in §4.2

SYSCORD only incurs a low overhead in both time and space. Detailed measurement is presented in §4.4.

4. Evaluation

In this section, we first use a case study (§4.1) given in Figure 2 to show the complete workflow of SYSCORD. Next, we focus on evaluate SYSCORD with the four practical requirements discussed in §1.

We deploy SYSCORD on an Armv8 Juno r2 board equipped with 6 cores (2 Cortex-A72 cores and 4 Cortex-A53 cores) and 8GB RAM based on Linaro deliverables Linux 5.4.50. We equip the Juno board with an SSD and allocate 256 MB circular buffers for ETM tracing. We use this as the default setting for our experiment but also allowed developers to adjust it as required.

4.1 Case Study

To evaluate the functionality of SYSCORD, we test it with the program shown in Figure 2. As Table 2 illustrates, it is clear that Thread 2 checks the length of `big_buf` (Line 10, Block 2-A) before the first `read` (Line 4, Block 1-B) in Thread 1. Since we capture the data of `read`, we know that the `big_buf` contains a string with length 25. Therefore, the following `strcpy` (Line 12, Block 2-C) incurs a buffer overflow, and the variable `len` (Line 7) is unexpectedly changed to `875770417`, which finally leads to the failure of the `assert` (Line 14, Block 2-G).

It is notable that the program does not crash immediately after the buffer overflow. In contrast, it executes normally for the second cycle (Block D to F). In addition, the normal execution of the second time overwrites the values (`buf` and `len`) involved in the buffer overflow, which may confuse developers without the assistance of SYSCORD.

Table 2: Tailored Control Flow And Data Flow for Figure 2

Block	Thread 1	Thread 2	Data Values
A		bl 400910 <strlen@plt>	total=0, len=0, buf=?
B	bl 400960 <read@plt>		read, fd=3, size=64, res=25, data=”1234567890123456789012345”
C		bl 400970 <strcpy@plt>	total=0, len=875770417, buf= “123456789012345”
D	bl 400960 <read@plt>		read, fd=4, size=64, res=6, data=”123456”
E		bl 400910 <strlen@plt>	
F		bl 400970 <strcpy@plt>	total=875770423, len=6, buf=”123456”
G		bl 400940 <___assert_fail@plt>	

Table 3: Buffer usage when file dumps

Program	# dumping file	buffer usage		
		Min.	Max.	Avg.
nginx (5,000,000 requests)	96	25	360	183
large file read (40 GB)	79	39	7,550	4,042

Table 4: Syscalls issued from bugs recorded by SYSCORD. N/A=bugid is not available, E=reconstructed bug, R=real-world bug, OV=order violaion, SAV=single variable atomicity violation, MAV=multi variables atomicity violation, DL=deadlock, SEQ=sequential bug (non-concurrency bug), LOC=line of code

Program-BugID-GroupType	**bug type**	**LOC**	**Symptom**
shared_counter-N/A-E	SAV	45	assertion failure
log_proc_sweep-N/A-E	SAV	93	segmentation fault
bank_account-N/A-E	SAV	95	race condition fault
jdk1.4_StringBuffer-N/A-E	SAV	180	assertion failure
circular_list-N/A-E	MAV	155	race condition fault
mysql-169-E	MAV	120	assertion failure
mutex_lock-N/A-E	DL	51	deadlock
SQLite-1672-R	DL	80K	deadlock
memcached-127-R	SAV	18K	race condition fault
Python-35185-R	SAV	1256K	race condition fault
Python-31530-R	MAV	1256K	segmentation fault
aget-N/A-R	MAV	2.5K	assertion failure
pbzip2-N/A-R	OV	2K	use-after-free
curl-965-R	SEQ	160K	unhandled input pattern
cppcheck-2782-R	SEQ	120K	unhandled input pattern
cppcheck-3238-R	SEQ	138K	NULL pointer dereference

4.2 Completeness

Due to the implementation of the secondary buffer, the record may become incomplete because of not being transferred in time. We evaluate SYSCORD with two extreme scenarios: nginx with high concurrency and file reading with massive IO operations. As Table 3 shows, even in the worst case that 4,042 bytes, much less than 16 MB, is written to buffer before the secondary buffer dumping to file, SYSCORD ensures the completeness of the record.

4.3 Effectiveness

We show how effectively SYSCORD is for diagnosing the root cause of bugs. As listed in Table 4, we use 16 commonly C/C++ buggy programs[14-19] to evaluate SYSCORD. We divide these bugs into two groups, i.e., Group E and Group R. Group E contains 7 bugs reconstructed from applications[16, 17], and Group R includes 9 bugs in real-world applications[14, 15, 18, 19]. There are 13 concurrency bugs, of which 6 are single

Table 5: Syscord output of Cppcheck-2782

PID	Syscall	Parameters	Additional information
22571	getcwd	-	path=/home/root/cppcheck
22571	newfstatat	res=0	-
22571	fstat	res=0	fd=1
22571	write	res=29	-
22571	openat	res=3	dir=-100, path=./fail.cpp
22571	read	fd=3	data="int main() { return0; } #asm !while (val) mov bx #endasm"
22571	close	fd=3	res=0

variable atomicity violation (SAV), 4 are multi variables atomicity violation (MAV), 2 are deadlock (DL), and 1 is order violation (OV). There are also 3 non-concurrency bugs. These bugs are collected from a diverse set of real-world systems (e.g., Python, Memcached, SQLite, and Aget) and wide symptoms (e.g., NULL pointer dereference, use-after-free, and race).

We execute these programs separately in our system until the bug occurs and then use Syscord to record. And then anaylyze the root cause of bugs. Specifically, we manually analyze the bug by record and compare the related patches of these bugs. The result indicates that the failure reports generated by Syscord are exactly related to the root cause.

Out of those 16 bugs, we select one representative example to further demonstrate the effectiveness of Syscord.

Cppcheck-2782. In this case, we use Syscord to record a non-concurrency bug. We run the application with common C++ source code as its input until it crashes. Table 5 shows the syscalls recorded by Syscord. The C++ source code analyzed by Cppcheck which triggered the bug is loaded with the syscall `read`. We then find the source code is special for containing embedded assembly code. According to the public bug report, the Cppcheck-2782 is known as a bug caused by unhandled input pattern, for its disability in handling embedded assembly code. Syscord can help developer correctly locate the bug and easily analyze the root cause.

4.4 Efficiency

We show how efficiently Syscord can be used for bug analysis by first running Unixbench 5.1.2[20] to measure the performance impact on kernel such as syscall. We then run ApacheBench[21] with Nginx 1.20[22], representing a popular server program to simulate a high load scenario. We finally evaluate the runtime performance overhead of Syscord by running four real-world programs.

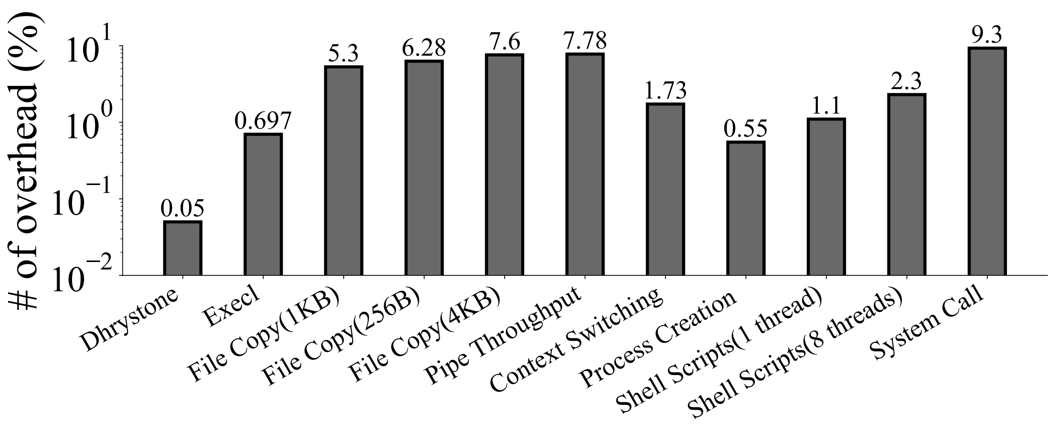


Figure 5: Performance overhead of running UnixBench with ETM tracing and non-deterministic event TheName.

4.4.1 Unixbench

We run UnixBench on Linux and show the performance results in Figure 5. For the tracing enabling, the performance overhead is 3.88% on average, and the highest performance overhead is for System Call at 9.3%. Specifically, there are three types of benchmarks: File Copy, Pipe Throughput, and System Call that have higher overhead. We believe that it is because these benchmarks have more frequent syscall invoking and I/O operations, which incur larger overhead than others.

4.4.2 Nginx

We use `nginx`[22] as a web server program to test the performance of SYSCORD in a high concurrency environment. We use `ab` (Apache HTTP server benchmarking tool)[21] to simulate user access behavior. Setting `nginx` to its default configuration, we operate performance testing with concurrency of 5,000 and request number of 500,000.

The average time cost for baseline (i.e., without SYSCORD) is 88.94s, and SYSCORD is 90.09s with 1.30% overhead. This shows that SYSCORD performs well even in high-pressure environments.

4.4.3 Performance Overhead on Real-world Programs

We use four real-world programs to test SYSCORD for performance overhead of normal executions, including `Pbzip2`, `Aget`, `SQLite`, and `Memcached`. We run different fine-grained tests on each of the programs to simulate three different load scenarios. We run `Pbzip2` to compress 10MB, 500MB, and 2GB files, respectively. We use `Aget` to download 50MB, 500MB, and 2GB files in the same network to avoid network speed interference. `SQLite` is evaluated by a `sqlite-bench`[23] to write $100,000$, $500,000$, and $2,000,000$ values in sequential key order in sync mode, respectively. A benchmark tool `Twemperf`[24] was used to test `Memcached`, which creates $20,000$, $300,000$, and

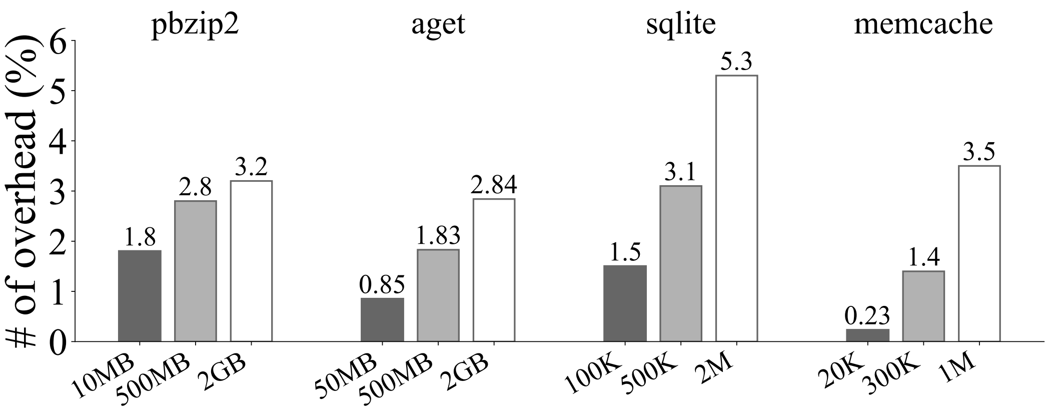


Figure 6: Performance overhead of 4 real-world prgrams execution enables ETM tracing and non-deterministic event TheName with library hook.

1, 000, 000 connections to a Memcached server running on localhost.

We show performance overhead results in Figure 6. Overall, the average performance overhead of all tests is 2.3%, and the highest overhead is 5.3% for `SQLite` when writing 2, 000, 000 values. The performance overhead has a slight improvement when test stress increases in all four programs. We believe it is caused by SYSCORD because there are many I/O operations and a large number of syscalls.

We compare the runtime performance overhead of real-world programs with the state-of-the-art systems[14, 15]. The results show that our overhead is slightly higher, mainly caused by the SYSCORD. Nevertheless, SYSCORD still incurs a low runtime performance overhead (2.3% on average). We conclude this overhead is acceptable and still suitable for practical deployment.

4.4.4 Trade-offs Between Performance and Accuracy

We design two versions of syscall capturing to record *RC-Type* syscalls: saving the whole content or truncating to the beginning 256 bytes . We test the time and space consumption of both versions by a program that continuously reads a 2 GB text file to simulate a highly concurrent environment on the server. The experimental results shown in Table 6 demonstrate that saving the entire content imposes a significant overhead (12.5%). In addition, an estimate for 24 hours of continuous recording generates nearly 1 TB files, which indicates that saving the entire content is also impractical for a larger throughput server. Therefore, we choose to truncate the records of *RC-Type* syscalls to the beginning 256 bytes.

4.5 Universality

In this section, we shows SYSCORD is a universal tool for almost all Linux devices. SYSCORD is basically a kernel module that does not need any dependency. All functions of SYSCORD are implemented in Linux kenrel, whcih indicates that SYSCORD is

Table 6: Space Consumption for Saving All Content or Truncation

Type	**Real Time**	**File Size**	**Estimated 24-hour File Size**
Baseline	2 min 50.3 s	-	-
All Content	3 min 11.552 s (+12.5%)	2.0 GB	902.1 GB
Truncation	2 min 57.675 s (+4.33%)	120 MB	52 GB

architecture-independent.

I also have verified on different platforms beyond ARM juno board. The specifications of these platforms are as follows.

- **ARM**: Raspberry Pi 3B, Debian GNU/Linux 10 with Linux 5.10.11-v8+.
- **x86**: Intel Core i5-10500, Ubuntu 18.04.5 LTS with 4.15.0-142-generic.
- **RISC-V**: Qemu 5.2.0 RISC-V virt, Linux 5.11.0.

Although SYSCORD currently does not support all features (e.g., recording for return address) on devices of other architectures, the core part of SYSCORD, i.e., collecting information and saving to file, works fine. Besides, there are some recent work to make extensions based on SYSCORD on RISC-V[25], which also indicates the universality of SYSCORD.

5. Discussion

This section discusses the limitations of SYSCORD and how I intend to address them in future work.

Currently, SYSCORD only considers 60 of 276 syscalls. Although I have already accounted for diverse types of system calls, handling more system calls is necessary to make SYSCORD becoming a practical system. Adding handlers for all system calls is the main task in the future.

Another limitation of SYSCORD is that it currently only supports system calls. Nevertheless, a similar approach with SYSCORD can handle other types of non-deterministic events. The interrupts and traps can be dealt with by similar kernel hooks, and the signals would require additional capturer in user-space. Moreover, the noninitialized variables are saved in memory, and therefore should appear in the coredump. Therefore, handling interrupts and traps is also one of my future work.

Currently, SYSCORD works perfectly on ARM devices. In my evaluation, it also runs normally on the x86 platform except for recording the return address. The only difference among these architectures is the approach to obtaining the registers differs. I will add supports for other platforms in the future.

Although SYSCORD performs well in common scenarios with only 4.88% overhead at most, I can still apply many optimizations. For example, different record formats for different system calls are still a problem, and my current solution is to name each item. Therefore, records likes `pid=14311, accept4, res=3` (with the attribute name `pid` and `res`) additionally consumes a considerable amount of space. In the future, I need to classify the format of the record to reduce the space occupation of records.

In addition, *filter* of SYSCORD can only support limited conditions. To make SYSCORD more practical, some basic logic operators (e.g., `AND`, `OR` and `NOT`) should also be supported.

Finally, SYSCORD needs to have the ability to support long-term running. The size of the record file should not increase unlimitedly. Therefore, processing of historical recording data (dumping, compression, overwriting, etc.) will also be a future work.

6. Related Work

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

Record and replay systems are a type of systems that record sufficient information to reconstruct the program execution and then replay this execution. Record and replay systems are undoubtedly necessary to handle system calls or not replay faithfully. Pinplay[26] is a record and replay system based on Pin[27], a binary instrumentation system. While it does record system calls, the instrument-based approach introduces significant overhead (up to 140x). REPT[9] chooses to not record non-deterministic events such as system calls. In contrast, REPT adopts inference algorithms to reconstruct the effects of system calls. However, inference-based approaches can not achieve a 100% reconstruction. For example, REPT can not deal with bugs like Figure 2.

RR[3] is one of the-state-of-the-practice record and replay systems. RR perform the recording to non-deterministic events via `ptrace`[28], a syscall that allows a process to inspect and control the execution of another one. By using `ptrace`, a supervisory process can easily observe and intercept syscalls issued by the target program. Nevertheless, `ptrace` brings a considerable overhead (up to 7.85$\times$) as the supervisor also runs in user space, which means that the supervisor needs to capture data with several context switches (between target process with supervisor).

Some **Record and replay systems** propose to perform the recording only in the application layer. //TODO(cyrus,r2,liblog)

Some **Linux troubleshooting tools** can also capture syscalls, as they monitor the target application and captures data from non-deterministic events, such as Sysdig[29], DTrace[30, 31], and Strace[32]. Nevertheless, Strace[32] also utilize `ptrace` to capture syscalls, which is similar to RR and inevitably introduces a considerable overhead.

DTrace is powerful and efficient, but it is too sophisticated to use and require technical knowledge to optimize[31]. Sysdig also leverages *Linux tracepoints* as SYSCORD to achieve a relevant reasonable overhead (about 15%); however, Sysdig sacrifices the completeness of its records, i.e., it does not guarantee that all system calls will always be recorded[33].

7. Conclusion

I propose a syscall record tool SYSCORD on ARM to satisfy practical requirements. SYSCORD takes advantage of the mechanisms in Linux to completely and efficiently capture and record system calls. We implement and deploy SYSCORD into various Linux systems on ARM. The experiments show that SYSCORD can successfully record the syscalls with a low overhead. SYSCORD achieves its design goals and is suitable as the basic part for record and replay systems.

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. 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. 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] MATHIEU D. Using the Linux Kernel Tracepoints —The Linux Kernel documentation[EB/OL]. 2021. https://www.kernel.org/doc/html/latest/trace/tracepoints.html.

[13] CORBET J, RUBINI A, KROAH-HARTMAN G, et al. Linux device drivers[M]. 3rd ed. [S.l.]: O'Reilly, 2005.

[14] CUI W, GE X, KASIKCI B, et al. REPT: Reverse debugging of failures in deployed software[C]//13th USENIX Symposium on Operating Systems Design and Implementation (OSDI'18). [S.l. : s.n.], 2018.

[15] KASIKCI B, CUI W, GE X, et al. Lazy Diagnosis of In-Production Concurrency Bugs[C]//Proceedings of the 26th Symposium on Operating Systems Principles. Shanghai China: ACM, 2017: 582-598.

[16] YU J, NARAYANASAMY S. A case for an interleaving constrained shared-memory multi-processor[J]. ACM SIGARCH Computer Architecture News, 2009.

[17] YU J, NARAYANASAMY S, PEREIRA C, et al. Maple: A coverage-driven testing tool for multithreaded programs[C]//Proceedings of the ACM international conference on Object oriented programming systems languages and applications. [S.l. : s.n.], 2012.

[18] KASIKCI B, SCHUBERT B, PEREIRA C, et al. Failure sketching: A technique for automated root cause diagnosis of in-production failures[C]//Proceedings of the 25th Symposium on Operating Systems Principles. [S.l. : s.n.], 2015.

[19] LIANG J, LI G, ZHANG C, et al. RIPT–An Efficient Multi-Core Record-Replay System[C]//Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security. [S.l. : s.n.], 2020.

[20] Kdlucas. Byte-unixbench[EB/OL]. 2018. https://github.com/kdlucas/byte-unixbench.

[21] The Apache Software Foundation. Apache HTTP server benchmarking tool[EB/OL]. 2018. https://httpd.apache.org/docs/2.2/programs/ab.html.

[22] NGINX. Nginx-1.20.0 stable version[EB/OL]. 2021. http://nginx.org/download/nginx-1.20.0.tar.gz.

[23] Retrage. SQLite3 Benchmark[EB/OL]. 2018. https://github.com/ukontainer/sqlite-bench.

[24] Thinkingfish. Twemperf[EB/OL]. 2014. https://github.com/twitter-archive/twemperf.

[25] Gdjs2. Gdjs2/Mysisdig[EB/OL]. 2021 [2021-05-12]. https://github.com/gdjs2/Mysisdig.

[26] PATIL H, PEREIRA C, STALLCUP M, et al. PinPlay: A Framework for Deterministic Replay and Reproducible Analysis of Parallel Programs[C]//CGO '10: Proceedings of the 8th Annual IEEE/ACM International Symposium on Code Generation and Optimization. Toronto, Ontario, Canada: Association for Computing Machinery, 2010: 2-11. DOI: 10.1145/1772954.1772958.

[27] REDDI V J, SETTLE A, CONNORS D A, et al. PIN: a binary instrumentation tool for computer architecture research and education[C/OL]//Proceedings of the 2004 workshop on Computer architecture education held in conjunction with the 31st International Symposium on Computer Architecture - WCAE '04. Munich, Germany: ACM Press, 2004: 22-es [2021-05-13]. http://dl.acm.org/citation.cfm?doid=1275571.1275600. DOI: 10.1145/1275571.1275600.

[28] THEODORE T, HEINRICH S, MICHAEL K. Ptrace(2) - Linux manual page[EB/OL]. 2021. https://man7.org/linux/man-pages/man2/getrandom.2.html.

[29] Sysdig, Inc. Draios/sysdig[EB/OL]. 2021. https://github.com/draios/sysdig.

[30] GREGG B. DTrace Tools[EB/OL]. 2019. http://www.brendangregg.com/dtrace.html.

[31] GREGG B, MAURO J. DTrace: Dynamic Tracing in Oracle Solaris, Mac OS X and FreeBSD[M]. 1st edition. [S.l. : s.n.], 2011.

[32] The strace developers. Strace/strace[EB/OL]. 2021. https://github.com/strace/strace.

[33] DEGIOANNI L. Sysdig vs DTrace vs Strace: A technical discussion.[EB/OL]. 2014 [2021-05-13]. https://sysdig.com/blog/sysdig-vs-dtrace-vs-strace-a-technical-discussion/.

Appendix

Source Code

SYSCORD is an open source proejct on Github: its source code can be found at https://github.com/Tert-butyllithium/syscord

Acknowledgement

SYSCORD is an important part of another project, which is submitted to ACM CCS 2021. Thanks to my collaborators, including Yiming Zhang, Wenxuan Shi, Yuxin Hu, and Xueying Zhang. I have learned a lot from the project. Besides, Yiming as the first user of SYSCORD has also raised many issues for SYSCORD, which makes SYSCORD better.

I would like to express my gratefulness to my advisor, Dr. Fengwei Zhang, for his continuous support and encouragement during my college life in SUSTech. Dr. Zhang's encouragement has rekindled my passion for research and eventually motivated me to continue on my research path. Dr. Zhang provided a lot of time to discuss with me when I was facing difficulties in research and life.

I am very appreciative to Dr. Zhenyu Ning for his guidance in our project. Dr. Ning's insightful and critical suggestions are an essential factor in making this project successful. Dr. Ning's enthusiasm, modest and gentle, has also influenced me in my life and research.

Moreover, I would like to thank all the group members in COMPASS Lab. I have joined COMPASS Lab a year ago and it has been a pleasant and impressive experience with every genius in the lab.

Also, I deeply appreciate to my girl friend, who has been absent for 21 years of my life. Therefore I can only focus on my work.