Record-Replay Architecture as a General Security Framework

Yasser Shalabi, Mengjia Yan, Nima Honarmand,[†] Ruby B. Lee,[‡] and Josep Torrellas

University of Illinois at Urbana-Champaign
http://iacoma@cs.uiuc.edu

[†]Stony Brook University
nhonarmand@cs.stonybrook.edu

[‡]Princeton University
rblee@princeton.edu

ABSTRACT

Hardware security features need to strike a careful balance between design intrusiveness and completeness of methods. In addition, they need to be flexible, as security threats continuously evolve. To help address these requirements, this paper proposes a novel framework where Record and Deterministic Replay (RnR) is used to *complement* hardware security features. We call the framework *RnR-Safe*. RnR-Safe reduces the cost of security hardware by allowing it to be less precise at detecting attacks, potentially reporting false positives. This is because it relies on on-the-fly replay that transparently verifies whether the alarm is a real attack or a false positive. RnR-Safe uses two replayers: an always-on, fast *Checkpoint* replayer that periodically creates checkpoints, and a detailed-analysis *Alarm* replayer that is triggered when there is a threat alarm.

As an example application, we use RnR-Safe to thwart Return Oriented Programming (ROP) attacks, including on the Linux kernel. Our design augments the Return Address Stack (RAS) with relatively inexpensive hardware. We evaluate RnR-Safe using a variety of workloads on virtual machines running Linux. We find that RnR-Safe is very effective. Thanks to the judicious RAS hardware extensions and hypervisor changes, the checkpointing replayer has an execution speed comparable to the recorded execution. Also, the alarm replayer needs to handle very few false positives.

Keywords: Record and Deterministic Replay; Hardware Security; Return Oriented Programming.

1. INTRODUCTION

As security attacks become more frequent and varied, there is increasing interest in augmenting processor and system hardware with security features. As a result, processor manufacturers have developed new hardware architectures, such as Intel's MPX [1], AMD's Secure Processor [2], and ARM TrustZone technology [3].

A general difficulty in this area is that security threats are continuously evolving, circumventing existing security defenses. What used to be an effective defense yesterday is less effective today. Hence, defense systems have to be flexible. For example, to defend against code injection attacks, W$\oplus$X features [2, 4] have been widely deployed in processors. They prevent the execution of data by enforcing the invariant that memory pages are either executable or writable, but never both. Unfortunately, new attacks have appeared that do not need code injection. In particular, an attack based on code reuse called Return Oriented Programming (ROP) [5] is now the preferred technique. It builds attack code by chaining together multiple snippets of code from the victim program, hence bypassing W$\oplus$X defenses.

An intriguing primitive that can help defend against security threats is Record and Deterministic Replay (RnR) (e.g., [6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22]). With RnR, a workload's initial execution creates a log, which can be deterministically replayed on another machine. RnR has been used for security purposes, most often off-line, to provide insight into how and when an attack took place (e.g., [8, 11]).

In this paper, we explore an approach to hardware security design where RnR is used to *complement* hardware security features—to offload intrusion checks and/or to eliminate check imprecision. We call the framework *RnR-Safe*. In RnR-Safe, we *reduce* the cost of security hardware, by allowing the hardware to be less precise at detecting attacks, potentially reporting false positives. This is because we rely on an *on-the-fly* replayer that transparently verifies whether the alarm is a real attack or a false positive. The result is a very general security framework that can be combined with a variety of relatively inexpensive security hardware.

RnR-Safe relies on two types of on-the-fly replayers running on a different machine: an always-on fast replayer that periodically creates checkpoints of the monitored execution (*Checkpointing* replayer), and an analyzing replayer—triggered by an alarm—that starts from a checkpoint and analyzes the execution to determine whether the alarm was due to a real attack or a false positive (*Alarm* replayer). The alarm replayer can execute multiple times, with different levels of analysis, until the attack is fully understood.

As an example application, this paper then applies this approach to thwart ROP attacks—including *on the kernel*, a challenging target to defend. A ROP attack causes a Return Address Stack (RAS) misprediction. However, a RAS misprediction is an imprecise ROP detector, as it may also occur for benign software. Hence, rather than augmenting the RAS hardware to guarantee perfect detection, RnR-Safe makes simple modifications to the RAS hardware to eliminate the vast majority of the false positives. The few remaining false positives are identified by the alarm replayer, thus minimizing hardware changes. This follows the RnR philosophy.

To evaluate RnR-Safe, we execute a set of varied workloads on Virtual Machines (VMs) running Linux. We find that RnR-Safe is an effective hardware-software co-design point. Thanks to the judicious RAS hardware extensions and hypervisor changes, the checkpointing replayer has a speed that is comparable to that of the recorder, and can be replaying continuously. In addition, the alarm replayer has to handle only very few false positives.

The contribution of this paper is threefold. First, we propose a novel RnR-based hardware-software approach to enhance security called RnR-Safe. Second, we tailor RnR-Safe to detect ROPs—with a focus in the kernel—and discuss the

challenges encountered during its implementation. Finally, we evaluate RnR-Safe in this use.

Assumed System and Threat Models. ROP attacks can occur within the kernel or user contexts, and RnR-Safe can secure both. The target most difficult to secure is the kernel, as it is privileged and multi-context. The protected system (kernel and applications) runs inside a VM whose execution is continuously recorded. The recorded execution is concurrently being replayed on a different machine, where ROP attacks may be uncovered.

We assume the attacker can launch a ROP attack against the kernel via any combination of address disclosure and memory corruption vulnerabilities to hijack and corrupt the kernel stack. We assume that the host machine OS and hypervisor (in the recording and replaying machines) are benign, and that they can safeguard against compromised guest VMs through traditional memory page permissions.

2. BACKGROUND

2.1 Record and Replay

Record and deterministic Replay (RnR) of workloads is a popular architectural technique (e.g., [6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22]). As a workload runs, RnR records all the non-deterministic events that can affect the execution and stores them in a log. Then, in a potentially different platform, the workload is re-run. At this time, the system injects the recorded events at the correct times, enforcing a deterministic execution (*Replay*). Typically, the non-deterministic events are the inputs to the workload and, in parallel programs, the interleaving of memory accesses.

RnR can be done at different abstraction layers. In this work, we use VM-level RnR [7, 8, 15, 14]. Moreover, we consider uniprocessor hardware. As a result, the sources of non-determinism are interrupts raised and data copied by virtual devices into the guest machine. We also assume the widely used model of hypervisor-mediated I/O, as used in Xen [23] or Qemu [24]. These assumptions are not necessarily limitations, as RnR approaches compatible with multiprocessing [9, 21] and virtualized I/O [25] exist.

There are several papers that investigate the use of RnR in a security-related scenario (e.g., [8, 11, 13, 14, 19, 26]). ReVirt [8] shows an example of using VM-level RnR for post-facto offline analysis of time-of-check to time-of-use race conditions in the Linux kernel. IntroVirt[11] explores using VM-level RnR to determine if systems were previously exploited once zero-day attacks are discovered. Speck[13] explores using a combination of OS-level speculation and program-level RnR to remove security checks from the critical path of a program. ParanoidAndroid [19] and Secloud [26] explore the possibility of maintaining replicas of mobile devices in the cloud, and perform program-level RnR in the cloud. Finally, Aftersight [14] suggests using VM-level RnR to perform online dynamic analysis of a system's execution. However, it does not address several important hardware-software design issues of such a model, including one key contribution of our work: separation between the fast checkpointing replayer and the exhaustive alarm replayer. We discuss the details in Section 9.

2.2 Example Threat: Return Oriented Programming (ROP)

RnR-Safe is a general framework that can be used to thwart a variety of attacks. As an illustration, in this paper, we consider ROP [5] attacks, focusing on those on the kernel. Appendix A describes ROP attacks.

At a high level, ROP attacks can be detected with a structure called *Shadow Stack* (e.g., [27, 28, 29, 30]). When a call instruction is executed, the address of the instruction following the call is pushed to the shadow stack. At return instructions, the top of the shadow stack is popped. ROP attacks can be detected when the return address used by the processor mismatches the one popped from the shadow stack.

However, shadow stacks present several implementation challenges. First, the validity of the shadow stack hinges on its integrity. Hence, the shadow stack must be secured against the very software it protects. This includes protecting it against the *kernel* itself, which might be compromised and malicious. Also, codes can be highly nested (e.g., recursive), multi-context (e.g., the kernel), or imperfectly nested (e.g., error and exception handling). Since false positives are unacceptable, proper handling of all these corner cases is necessary. Most solutions add instructions to manipulate the shadow stack in these cases. Such instructions, even if privileged, provide a security vulnerability.

There have been a variety of approaches that focus on protection against ROP attacks (e.g., [31, 32, 28, 29, 30, 27, 33, 34, 35, 36, 37, 38, 39, 40]). Unfortunately, while some of them are provable prevention techniques, they appear insufficient from a practical point of view. Indeed, it is a matter of fact that systems today still remain vulnerable to such attacks. We discuss the reasons in the next section.

2.3 ROP is Still Unsolved Today

There are three reasons why existing ROP protection techniques have limitations from a practical point of view:
Hardware Intrusiveness. Some approaches [27, 28, 34, 37] require intrusive hardware changes. For example, both SmashGuard [28] and SRAS [27] add a hardware stack used to verify the return targets. The hardware needs to be very carefully designed, as it intimately interacts with the speculation mechanism of modern processors [28]. In addition, it uses instructions in case of stack overflow and context switch [27, 28]. Such instructions, even if privileged, provide a security vulnerability. The PUMP [37] processor supports general metadata propagation. This can be used to implement various safety checks, including Control Flow Integrity (CFI). However, each stage of the pipeline is modified, to support tag storage and/or rule execution. REV [34] hashes the instruction sequences within a basic block to verify a program's control flow. It requires an additional 32KB first level cache dedicated to cache signatures, to avoid prohibitive slow-downs.
Software Impact. The completeness of instrumentation-based CFI-enforcing solutions such as [31, 32] is attractive. However, *securely* maintaining a shadow stack at call/ret boundaries via binary instrumentation adds overheads of over 100% [32]. If the source code is available, compiler based solutions can reduce the overhead substantially [41]. Other approaches [36, 42] propose recompiling the kernel and application to target a secure virtual instruction architecture. This

architecture is emulated by a compiler-based virtual machine (similar to the Java Virtual Machine). Aside from performance costs, this approach also requires source code, which limits its applicability.

Completeness. Other proposals explore alternative approaches by either detecting ROP attack characteristics [35, 38, 39] or by making the ROP attacks difficult to mount [43, 44, 45, 46]. By monitoring control flow characteristics that are indicators of ROP execution [35, 38, 47], some ROP payloads can be detected. Unfortunately, ROP payloads may be able to blend their signature to match that of benign code to evade detection [48, 49]. Similar criticisms can be levied against other proposals which provide probabilistic defences through randomization [46] and encryption [50, 51, 52]. These defences will randomize code locations or encrypt on-stack return addresses, thus complicating critical steps in attack code. In particular, Address Space Layout Randomization (ASLR) [43, 44, 45, 46] is a widely deployed defence adopting this approach. While useful in the short-term, these approaches remain insufficient in the long-term, as attackers have learned to circumvent them [53, 54, 55]. We discuss ASLR in more detail in Section 9.

Recently, Intel has introduced Control-Flow Enforcement Technology (CET) [30], which has shadow stacks that verify return targets. The shadow stack pages can only be accessed through special privileged instructions. However, to our understanding, since these shadow stack pages are in memory, and because they can be accessed by certain privileged instructions, it is conceivable that an attacker can subvert the system and update the stacks. The same concern arises for Griffin [56]—a CFI verification technique based on analysis of branch traces. The traces are provided by Intel's Processor Trace, and are stored in the system memory directly by the hardware.

In this paper, we use RnR-Safe to support ROP protection in a different manner, with a great deal of flexibility, and a different set of tradeoffs than prior methods. In particular, we have no hardware shadow stack. Effectively, the *replayer* in a *secure* machine implements a shadow stack *in software*.

2.4 Return Address Stack

Modern processors use a hardware Return Address Stack (RAS) to predict the target of return instructions. When a procedure call executes, the hardware pushes the address of the instruction that follows it into the top of the RAS. When a return instruction is decoded, the hardware pops the entry at the top of the RAS and uses its value as the predicted target of the return. The RAS is not accessible by software. The IBM POWER7 [57] and POWER8 [58] processors have a RAS with 32 and 64 entries, respectively.

ROP attacks cause RAS mispredictions because the attacker forces a return to an unexpected instruction. However, a RAS misprediction cannot by itself be used as an indicator of ROP attacks because the RAS sometimes mispredicts in the course of benign program execution.

3. AN RNR SECURITY FRAMEWORK

Figure 1 shows the organization of RnR-Safe, our envisioned security framework. On the left side, a workload runs on a *Recorded* VM. Its hypervisor records all the non deterministic events of the execution in a software log. Recording adds only modest overhead—less than 15% on average, according to Pokam et al. [21]. Note that we record at the VM level to also protect the operating system.

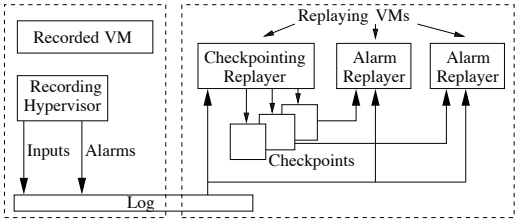


Figure 1: RnR-Safe organization.

The designer has augmented the hardware in the recorded VM (e.g., processor and memory system) with support to detect certain classes of attacks, with potentially some false positives. When this hardware or the recording hypervisor suspect an attack, the hypervisor inserts an alarm marker in the log. At this point—and depending on the risk tolerance of the workload—the recorded VM may be stopped until the alarm is analyzed, or allowed to continue.

On the right side, a *Checkpointing Replayer* VM re-executes the workload natively. It uses the log to inject all the non-deterministic events. As a result, the execution deterministically follows the original one. If an alarm is found in the log, an *Alarm Replayer* VM characterizes the alarm, detecting either an attack or a false positive.

3.1 RnR-Safe Modes of Execution

In RnR-Safe, monitored recording consists of normal execution. Transparently to the recorded VM, all the non deterministic inputs are being recorded by the hypervisor in a log, and the hardware and the hypervisor monitor for safety violations. If a violation is found or suspected, an alarm entry is inserted in the log. A key detail is that, in order to claim complete protection, the detector must catch all potential threats. In other words, false negatives are not acceptable.

In RnR-Safe, the replay execution is performed with two types of replayers. One is the *Checkpointing Replayer*. Such replayer runs all the time, at roughly recording speeds. It uses the log to determistically replay the workload while creating state checkpoints at regular intervals. When an alarm marker is found in the log, the checkpointing replayer launches the execution of an *Alarm Replayer* out of a recent (typically the latest) checkpoint. Old checkpoints and log entries are regularly discarded to save storage.

The second type of replayer is the *Alarm Replayer*. An alarm replayer replays log entries from a given checkpoint until an alarm marker, while performing an extensive, attack-specific analysis of the replayed execution. Its goal is to resolve an alarm, either to show that it is a false positive or to characterize the attack. It can be much slower than the checkpointing replayer. Typically, alarms are rare events.

3.2 What RnR-Safe Offers

RnR-Safe provides three security benefits.

Robustness at Relatively Modest Hardware. Perfect detection accuracy often necessitates very intrusive hardware. RnR-Safe minimizes intrusiveness by separating alarm detection from attack verification using RnR. False positive

alarms and rare corner cases are handled by software-based replay. Thus, RnR restores robustness to a system built out of imprecise security hardware. The one requirement of the alarm mechanism is to avoid any false negatives.

Flexibility. RnR-Safe is flexible. As attackers devise new attacks, defenders can augment the recorded VM and its hypervisor with hardware and software for new alarm generation, and the replaying VMs with software for new analysis techniques. The addition of new replayers is particularly compelling, as the analysis is performed in software. Multiple types of attacks can be tracked at the same time.

Execution Auditing. If desired, RnR-Safe also allows detailed analysis of executions offline. An execution context can be replayed to audit the code and data state. This is a general mechanism for identifying security violations by auditing sensitive flows in the system.

3.3 How to Use RnR-Safe

The RnR-Safe framework can be tailored to detect different attacks. For each attack, we first need a first-line of detection that targets that threat in the machine. The key advantage of RnR-Safe is that such a detection technique, implemented in hardware or software, can be imprecise—i.e., it can suffer false positives. This property often makes the design less intrusive or complicated.

While it is unclear which types of attacks would be best suited for RnR-Safe to target, Table 1 outlines three example attacks: ROP, jump-oriented programming (JOP) [59] and denial of service (DOS) [60]. For each, the table outlines the alarm trigger, possible first detection technique, and the role of replay. The first entry is ROP, which will be discussed in detail in the rest of this paper. The alarm trigger is a RAS misprediction. As we will see, the first detection technique is based on managing a multithreaded RAS, and using a whitelist of acceptable RAS mispredictions. The role of replay is to model a kernel-compatible shadow stack.

Attack	Alarm Trigger	Possible First Detection Technique	Role of Replay
ROP (this paper)	RAS misprediction	Manage a multithreaded RAS, use a whitelist	Execute a kernel-compatible shadow stack algorithm
Jump Oriented Programming (JOP)	Stray indirect branch or call	Table of begin and end addresses of the most common functions	Verify if the target is one of the less common functions
Denial of Service (DOS)	Kernel scheduler inactivity	Counter of number of context switches	Identify reason for low switching frequency

Table 1: Examples of potential RnR-Safe uses.

JOP is a related class of attacks, mounted by redirecting indirect branches or call instructions to execute the victim's code. Preventing such attack requires preserving the CFI of call and branch instructions. A first detection technique against JOP can be a table of begin and end addresses of the most common functions. An indirect branch or call target is compared to the table and is legal if the target is the first instruction of a function. Indirect branch targets within the current function are also fine. Otherwise, an alarm is triggered, and the replay verifies the same conditions for the less common functions.

A DOS attack on the OS can be detected with a counter that increments every time the kernel performs a context switch. If the counter has not increased much for a while, an alarm is raised, and the replay analyzes the code that has dominated the system's execution time. Due to space limitations, this paper focuses only on the ROP attack next.

4. EXAMPLE: TARGETING KERNEL ROPS

4.1 Main Idea

The architecture primitive that we use to help detect ROPs is the RAS. The RAS stores the addresses of the predicted targets of return instructions, and is not accessible by software; not even the kernel. A ROP attack, by causing returns to unexpected addresses, induces RAS mispredictions.

To use RAS mispredictions to thwart ROP attacks in RnR-Safe requires that there be no false negatives. Fortunately, execution of ROP payloads is guaranteed to cause RAS mipredictions, making false negatives impossible. Furthermore, for the RAS to be useful in RnR-Safe, false positive alarms should be infrequent. In practice, there are a few major sources of false positives in the basic RAS operation. We explain these sources with Linux kernel examples, when relevant.

First, there is the effect of multithreading. In a multithreaded environment, on a context switch from Thread *i* to Thread *j*, the RAS may still retain some entries belonging to Thread *i*. When executing code in Thread *j*, these entries might be incorrectly popped and used for prediction. If so, not only will Thread *j* encounter mispredictions, but also Thread *i*'s entries will no longer be available for their use when *i* executes again. Hence, Thread *i* will also mispredict.

A second effect is non-procedural returns in the kernel. Sometimes—e.g., during a context switch—the kernel inserts an address into the software stack, which will later be used by a return instruction as target. Since there was no prior call from a matching address, the RAS will not contain the corresponding entry and will mispredict.

RAS underflows are a third source of imprecision. If the code executes many nested procedure calls, the RAS may evict some of the earlier return addresses. Later, when the execution returns from the inner calls and tries to pop entries corresponding to the outer calls, the RAS will be empty (*underflow*) and will mispredict.

Finally, imperfect nesting of procedure calls is another reason for RAS mispredictions—a situation where a procedure is called but never returned from. Within the kernel, these are events that typically only take place as part of bug recovery processes in the kernel. When the kernel execution encounters a recoverable bug, it initiates a recovery process, as part of which it terminates the current thread of execution, leaving all the RAS entries of the current thread orphaned. For user-mode code, these events occur relatively more frequently—specifically, in the setjmp/longjmp calls, typically used in exception handling.

These effects show that the RAS is a detector of ROPs with many false positives. For RnR-Safe to use it as the initial defense, we need to robustify the RAS detection capability with simple support to minimize the false positive rate. Hence, RnR-Safe adds support for the first two sources of false positives: multithreading and non-procedural returns in the kernel. However, to completely eliminate false positives

would require major hardware and software changes. Hence, RnR-Safe raises alarms for the last two sources of false positives (underflow and imperfect nesting), and relies on the replayer to flag the events as false positives, and discard them. We discuss each of these cases in the next few sections.

4.2 Basic Design

As shown in Figure 1, the workload (applications plus kernel) runs in a *Recorded* VM. As the workload runs, the hypervisor creates an input log that is sent to and is consumed by the *Checkpointing Replayer* VM. If the workload executes a return instruction and the hardware finds a mismatch between the predicted target in the RAS and the actual return target, a VM exit is triggered. Then, the hypervisor inserts a ROP alarm entry in the input log. Depending on its configuration, the hypervisor may or may not stop the recorded VM until the alarm is fully processed in the replaying VM.

As the checkpointing replayer consumes the log, if it finds an alarm entry in the log, it triggers the execution of an *Alarm Replayer*, starting from the most recent checkpoint. The alarm replayer determines whether the alarm is a false alarm or a real ROP.

This basic RnR-Safe design does not miss an attack, but suffers from many false alarms. Next, we extend this basic design to reduce its false positives.

4.3 Supporting a Multithreaded Environment

In a multithreaded environment, as a thread is de-scheduled, it may leave return addresses in the RAS. Such addresses may be popped by subsequent threads, which will suffer mispredictions. Further, when the original thread is re-scheduled, it may pop RAS entries belonging to other threads, which will also cause mispredictions. The result is false ROP alarms.

To address this problem, RnR-Safe augments the microcoded virtualization hardware so that it additionally performs the following operations on a context switch. First, it saves the current RAS into a safe memory area outside of the reach of the kernel. Then, it restores the RAS state as needed for the upcoming running thread. To know the correct memory area to move data to and from, the hardware uses a new hardware pointer that the hypervisor sets.

The structures are shown in Figure 2. The software structure in memory is an array of backed-up RASes (*BackRAS array*). Each entry in the array belongs to a thread, and has a RAS and a counter with the number of entries in the RAS. The counter is needed to know the number of entries that need to be reloaded later on. The processor hardware includes a pointer (*BackRASptr*) that points to the backed-up RAS of the currently-running thread. The pointer is set by the hypervisor.

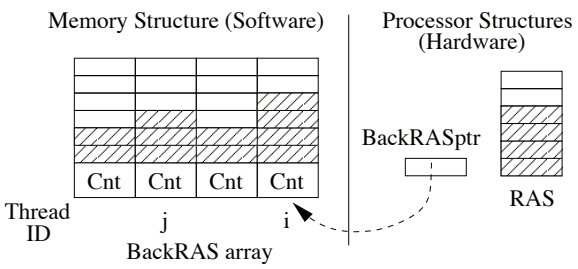


Figure 2: Structures used to support multiple threads.

Figure 3 shows the logic used. On a context switch, as part of the transition to the hypervisor, the hardware saves the RAS to the BackRAS entry pointed to by BackRASptr. In addition, it stores the count of saved entries. Our measurements show that a transition to the hypervisor takes about 1,000 cycles. We estimate that backing-up the RAS will add about 200 cycles. Later, when the hypervisor runs, it changes BackRASptr to point to the BackRAS entry for the new thread. Finally, as part of the transition back to the guest, the microcoded hardware loads the correct BackRAS entry into the RAS, taking another 200 cycles.

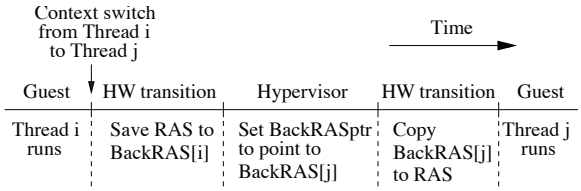


Figure 3: Algorithm and timeline to handle multiple threads.

To program the BackRASptr, the hypervisor needs to be informed of context switches in the guest kernel and identify the new thread to be scheduled. Section 5.2.1 explains how this is done without modifying the guest kernel.

With this support, when a thread is scheduled, it finds its correct state in the RAS, thus eliminating many false alarms.

4.4 Supporting Non-Procedural Returns

Sometimes, the kernel inserts an address into the software stack, and then executes a return that uses that address as target. Since there was no corresponding procedure call, the RAS did not push an entry, and will mispredict. Consequently, in these cases, the RAS should not be popped, as doing so would corrupt the RAS state.

In the Linux version we use, this use of returns occurs in one place, namely when a context switch is complete. At that point, right before launching the next thread, the kernel executes such a return in order to start executing code on behalf of the new thread. This code is written in assembly and directs the control flow to three well-defined locations in the kernel code. These locations complete the task switching, depending on whether it involves forking a thread, executing a kernel thread, or rescheduling a task.

To address this problem, RnR-Safe extends the processor hardware with a table of "whitelisted" addresses. There is a single-entry return whitelist (*RetWhitelist*) with the PC of the instruction with the non-procedural return, and a target whitelist (*TarWhitelist*) with the PC of the three instructions that can be the target of this return. During return address prediction, if a return PC and its target PC match entries in the tables, then the RAS is not popped, and no alarm is raised. These lists are only writable by the hypervisor.

The logic used and its timeline are as follows. When an instruction is decoded and identified as a return, the hardware checks if its PC is in the RetWhitelist. If so, the RAS is not popped and a *Whitelisted* flag is set. Later, when the target address is generated, if the Whitelisted flag is set, the hardware checks if its PC is in the TarWhitelist. If it is not, a VM exit is triggered.

The whitelisted addresses can be found by analyzing the binary image of the guest kernel. The hypervisor can popu-

late RetWhiteList and TarWhiteList using the identified addresses when entering the VM as explained in Section 5.1.

4.5 RAS Underflows and Imperfect Nesting

For false positives that occur very infrequently, RnR-Safe's approach is to avoid complicated hardware in the recorded VM and, instead, raise alarms and rely on the replayer VM to handle them. RnR-Safe uses this approach for the remaining two cases: RAS underflows and imperfect nesting.

The first case occurs when the kernel or an application executes enough nested procedure calls to cause the RAS to evict an earlier return address. Later, when the hardware tries to pop that entry from the RAS in a return instruction, it will find the RAS empty. This is a RAS underflow, and is flagged as a RAS misprediction.

RnR-Safe handles this case as follows. When a RAS entry is about to be evicted in the recorded VM, the hardware triggers a VM exit and dumps the about-to-be-evicted RAS entry to a special memory that is out of the kernel's reach. Then, the hypervisor places this data as an *Evict record* in the log. Later, when RAS underflow causes a RAS misprediction, the hardware will raise an alarm and the hypervisor will place an alarm record in the log.

The second case is imperfect procedure nesting, such as that caused by setjump and longjump. Imperfect procedure nests occur rarely, but require very complicated hardware to handle transparently [28]. Consequently, when an imperfect procedure nest causes a RAS misprediction, RnR-Safe simply places an alarm record the log.

The replayer has enough information to handle both cases. The replayer will be able to match evict records with alarm records due to RAS underflow. Further, the replayer will be able to identify setjumps and longjumps easily and fix its software RAS. Overall, while these false positive alarms will slow down the workload execution when they happen, they are rare enough to have a negligible total performance impact. The proven effectiveness of the RAS to predict return targets suggests that the RAS will soon re-synchronize with the program's true return addresses.

4.6 Replaying Platform

The input log is passed on-the-fly to another platform, where a VM running the checkpointing replayer deterministically replays the execution.

4.6.1 Checkpointing Replayer

To understand the operation of the Checkpointing Replayer (CR), we first describe the contents of a checkpoint. Figure 4 shows three checkpoints. Each checkpoint has three components. The first one is all the VM state. It includes all the memory pages of the VM, a page with the processor state at the time of checkpoint (PC, stack pointer, and the rest of the registers), and the virtual disk image contents. The latter is the state that the VM has written to the virtual disk. We need to checkpoint it because, if the execution later reads this data, the data will not appear in the input log. Note, however, that the state checkpoints are *incremental*. Since we take regular checkpoints, a given checkpoint keeps copies of only the pages and blocks that have been modified since the previous checkpoint; for each unmodified page or block, it keeps a pointer to it in the latest checkpoint that modified it.

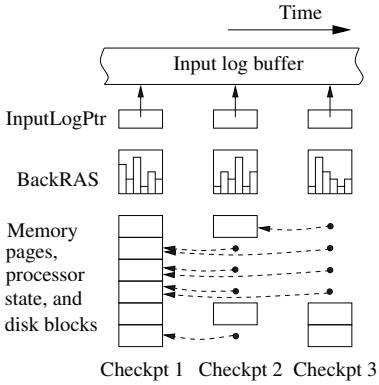


Figure 4: Checkpoints created by the checkpointing replayer.

The second component of a checkpoint is a pointer to the input log buffer (*InputLogPtr*). The pointer points to the next input log record to be processed after the checkpoint. Finally, the third component is the BackRAS at the time of the checkpoint. We will see in Section 4.6.2 why it is needed.

With this background, we can describe the CR operation. The CR re-executes the workload deterministically. When the time since the last checkpoint becomes higher than a certain threshold and a VM exit occurs, a checkpoint is taken. In a checkpoint, first, the hardware automatically saves the RAS into the BackRAS. Then, the CR dumps the processor state (PC, stack pointer, and all registers) into a memory page. Then, the CR creates the checkpoint by saving: (1) the page with the processor state, and all the memory pages and disk blocks modified since the prior checkpoint (together with pointers to the unchanged ones), (2) the current BackRAS, and (3) the current InputLogPtr. Then, the CR marks all pages copy-on-write, and resumes execution. During execution, when a page is written for the first time since the last checkpoint, a copy is made and used from then on.

The CR regularly recycles checkpoints. However, it can only recycle a memory page or disk block if it is not pointed to by a later checkpoint.

The replay takes place in a safe platform, where the hardware's ability to trigger ROP alarms is disabled. This is because replay does not create alarms. However, the hardware still dumps the RAS at context switching points. This ensures that, at the point of a checkpoint, the CR can reconstruct the up-to-date state of the complete BackRAS to stash in the checkpoint.

4.6.2 Alarm Replayer

When the CR encounters an alarm, it launches an alarm replayer (AR) VM starting from the checkpoint immediately preceding the alarm. The AR will determine whether the alarm is caused by a ROP or is a false alarm. The AR starts by initializing the VM state using the checkpoint. It marks all the pages and blocks in the checkpoint as copy-on-write to avoid modifying the initial state. Then, it reads the checkpoint's BackRAS into a software data structure that it uses to simulate the RAS. Next, it loads the saved processor state from memory into the processor registers. Finally, it starts execution, reading from the log starting from the checkpointed InputLogPtr.

The AR executes the original workload natively, in a deterministic manner, consuming the input log until it reaches the

alarm marker. The AR traps at every call and return instruction, inducing VM exits and transferring control to the hypervisor. There, an unbounded RAS is modeled in software, with our extensions for multithreading and non-procedural returns. The hardware on which the AR runs neither dumps the RAS state nor triggers ROP alarms. Both capabilities are disabled because they are not needed.

Once the AR encounters the alarm in the log, it checks whether the RAS mismatch can only be explained as an ROP attack. If so, the AR provides the state of the system at the point of the ROP attack. An expert programmer can study the state to glean information about the attack. Moreover, the AR can be re-run multiple times, with increasing levels of instrumentation, or starting at different checkpoints, to fully characterize the attack.

The case of evict records in the input log and the resulting underflow alarm records in the input log is handled in a special manner. While it can be handled by ARs, it is simpler if the CR handles this special case itself. Specifically, as the CR processes the input log and finds evict records, it saves them in a software structure. When the CR finds an underflow alarm record in the input log, it compares it against the latest evict record from the same thread. If they match, the alarm was a false one, and the CR removes the corresponding evict record from its structure. Otherwise, the CR launches an AR to handle this alarm.

Section 6 shows an attack. Note that our design allows running multiple ARs concurrently, to analyze the same or different ROP alarms in parallel.

5. IMPLEMENTATION ISSUES

This section summarizes the hardware and hypervisor support required for the architecture described. Following Intel's VT terminology, we use VMCS (VM Control Structure) to refer to the in-memory control structure through which the hypervisor communicates with and configures the virtualization hardware. We use VMEnter to mean transferring execution from the hypervisor to the VM, and VMExit to mean the opposite transfer.

5.1 Hardware Support

The hardware support required by RnR-Safe includes a replay platform, small extensions to the RAS hardware, the BackRASptr register, and the two whitelist tables. The requirement for RnR can be considered the most substantial addition. However, RnR is well understood and accepted as a useful primitive for debugging and program analysis. Also, the RnR infrastructure can be reused for a large variety of debugging and security analyses. The RAS extensions consist of triggering an exception when the RAS is about to evict an entry, and microcode to dump into memory and restore the BackRAS array, and to dump into memory the evicted RAS entries. There is also microcode to maintain the BackRASptr and whitelist tables.

We extend the VMCS with three new fields for the BackRASptr and the two whitelist tables. Microcode reads these fields to program these three processor hardware structures. Then, microcode uses the value of BackRASptr to dump the RAS contents into the active BackRAS entry in certain VMExits, and to read the active BackRAS entry into the RAS in certain VMEnters.

5.2 Hypervisor Support

5.2.1 Programming BackRASPtr on a Context Switch

The hypervisor needs to interpose on all context switches in the guest kernel during both recording and replay. In Linux, there is a single instruction where the stack pointer is changed from pointing to the current thread's stack to pointing to the next thread's stack. By setting a trap on this instruction, the hypervisor forces a VMExit when the guest executes this instruction. As part of the VMExit's microcode, the hardware dumps the RAS into the memory location pointed to by BackRASPtr.

Once the VMExit is complete and the control is transfered to the hypervisor, the latter can introspect the state of the guest kernel to identify the next thread to be scheduled. In Linux, a thread's descriptor (called *task_struct*) can be easily found if the thread's stack pointer is known. Since we set the trap on the instruction that changes the processor's stack pointer, we can find the next thread's stack pointer by examining the register content of the VM—which is available in the VMCS after a VMExit. Using this stack pointer, we find the corresponding *task_struct* descriptor in the VM's memory, and from that descriptor, read the next thread's ID.

The hypervisor stores the BackRAS in a memory area inaccessible to the guest machine. It stores it as a hash table mapping a thread's ID ("key") to its BackRAS entry ("value"). Using this organization, once the next thread's ID is found, the hypervisor checks the map to determine its BackRAS entry. Then, the hypervisor sets the BackRASPtr field of the VMCS to point to the BackRAS entry.

5.2.2 Recycling BackRAS Entries

In Linux, threads can be killed and their IDs may be reused. To keep the BackRAS consistent, we need to remove from the BackRAS a thread's entry when the thread is killed. Similarly to the case of context switching, the hypervisor sets a trap on the function that implements this functionality in the guest kernel to force a VMExit when it is executed. At that point, the thread ID can be found by introspection and then used to delete the corresponding BackRAS entry. A similar approach is followed when a thread is created and its BackRAS entry needs to be allocated.

6. MOUNTING A KERNEL ROP ATTACK

We built and mounted the ROP attack of Figure 10. In the recorded VM, as the workload calls the *Vulnerable* procedure of Figure 10(c), the hardware pushes into the RAS the return address at the call site (call it *Return Address*). This is the same address that is stored above the buffer in the software stack of Figure 10(e). After the malicious string copy, the software stack becomes Figure 10(f). As the program executes the return of the Vulnerable procedure, the hardware uses the RAS to predict that execution will transfer to *Return Address*. In reality, the target of the return is resolved to be the address of gadget G1, as shown in Figure 10(f). This mismatch causes the recorded VM to raise an alarm.

The recorded VM hypervisor then inserts an alarm marker in the log and may decide to stall the VM. When the check-

pointing replayer sees the alarm marker in the log, it starts an alarm replayer from the most recent checkpoint. As the alarm replayer executes, it models the RAS in software. At the point of the alarm, it observes the mismatch between the return's predicted target (in the RAS) and the actual target (in the software stack), hence declaring an ROP attack.

At this point, the hypervisor can analyze the system. It can use VM introspection to analyze the VM state, which has not been polluted by the execution of any gadget. It can also launch additional alarm replayers further back in time to perform a deeper analysis of the system.

One question that replay analysis can answer is: how was the attack possible to begin with? The hypervisor uses the return instruction that caused the alarm to determine that the attack occurred in the Vulnerable procedure. It uses the address at the top of the simulated RAS to determine the call site. An analyis of the Vulnerable procedure can conclude the presence of buffer overflow.

Another question is who attacked the machine? The hypervisor can determine the thread ID of the current thread, extract which users are logged in, and determine which network connections are established.

Yet another question is what did the attacker do? An analysis of the software stack reveals the gadgets used by the attacker. In this case, they did not execute. If they did, the hypervisor can use VM introspection to analyze what files were touched, what sockets were utilized, and what processes were forked [61]. This information is easy to get now because the workload is not running.

7. EXPERIMENTAL SETUP

7.1 Goal of the Evaluation

In this paper, the goal of our evaluation is to assess the overhead of recording, and of replay using the checkpointing and alarm replayers. We also want to know the rate of log generation, the bandwidth consumed to save/restore the RAS, and the frequency of alarms. Additional information includes the time window between attack and detection, the log generated during this window, and the number of checkpoints that the system needs to retain.

In our work, we mount only the kernel ROP attack that we describe in Section 6. Such attack is representative of ROP attacks, as they all use similar gadget-based patterns. Collecting and analyzing multiple real-world kernel ROP attacks is left as future work.

7.2 Evaluation Environments

To evaluate RnR-Safe, we use two evaluation environments. The first one evaluates the performance of our recording and replaying modes. For this, we use *Insight* [62], a VM RnR tool based on a modified Linux KVM hypervisor and QEMU devices. Since the KVM hypervisor can leverage Intel VTx extensions to virtualize the processor in hardware, the performance numbers from this setup are representative of real-world machines.

The second environment evaluates the correctness of our techniques and the functional characteristics of our proposed hardware. For this, we use QEMU in emulation mode. In this mode, QEMU also emulates the processor using dynamic translation of the systems software. This mode makes it easy to simulate our hardware and evaluate its function.

Table 2 shows the system configuration we use for our performance evaluation, and Table 3 shows our benchmarks. The benchmarks are: *fileio* and *mysql* from SysBench [63], *apache*, *make* (a compilation of the Linux kernel), and *radiosity* from SPLASH-2 [64].

Host machine	
CPU: Xeon E3-64bit,4-cores,3.1GHz	Memory: 8 Gbytes
OS: Ubuntu, Linux kernel 2.6.38-rc8	
Guest machine	
CPU: uniprocessor	Memory: 1 Gbyte
OS: Debian, Linux kernel 3.19.0	Disk: 32 Gbytes

Table 2: System configuration for performance evaluation.

Benchmark	Parameters
apache	-n100000 -c20
fileio	–file-total-size=6G –file-test-mode=rndrw –file-extra-flags=direct –max-requests=10000
make	linux-4.0 config with all-no
mysql	–test=oltp –oltp-test-mode=simple –max-requests=500000 –table-size=4000000
radiosity	-p1 -bf 0.005 -batch -largeroom

Table 3: Benchmarks executed.

7.3 Handling Non-Deterministic Events

Synchronous Non-Deterministic Events. Instructions such as *rdtsc* (read time stamp counter) or *rdrand* (read random number generator) return non-deterministic results. Accesses to memory regions like Memory Mapped IO (MMIO) are also non-deterministic. The VMCS controls when the processor will perform a VMExit. During recording, we configure the controls to synchronously trap these non-deterministic accesses, allowing the hypervisor to log their results. With similar configuration of the controls on the replaying system, these events are deterministically reproduced during replay.

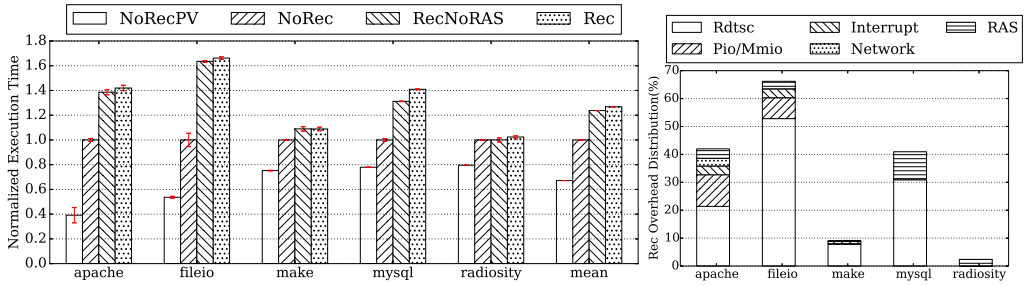
Network inputs are a special case and are also synchronous in our system. The arrival of network packets to the physical NIC is inherently asynchronous but the data is delivered to the VM at the boundaries of synchronous VMExits. Thus, this simplifies the recording and replaying of network events.

Asynchronous Non-Deterministic Events. Asynchronous events are more challenging to replay. They occur from external interrupts. These interrupts originate from other processors or from physical devices like disks. The VMCS structure can also be configured to cause a VMExit on these events. These VMExits, however, are asynchronous and will not repeat on the same instruction during replay. Therefore, for faithful replay, replay has to manually recreate them.

Trapping the VM at the same processor context is not straightforward. *Insight* uses performance counters to cause a VMExit as close as possible to the required point in replay. From there, the processor is single-stepped until execution reaches the desired injection point. Each step will suffer the overhead of a VMExit ($\approx$1,000 cycles).

7.4 Evaluating Replay Overhead

To evaluate the overhead of checkpointing replay, we reuse the Linux copy-on-write implementation used during fork



(a) (b)

Figure 5: Execution time of recording setups (a) and breakdown of the *Rec* overhead over *NoRec* (b).

system calls. Virtual memory belonging to the VM is allocated within a user-space QEMU process running on the host machine. With minor modifications, a checkpoint can be created by forking the QEMU process.

The alarm replayer models the RAS at every call and return instruction. Unfortunately, current Intel VTx extensions do not support trapping call and return instructions. Hence, to measure the performance impact of alarm replay, we modified GCC to instrument binaries by inserting a debug exception before kernel context switches, and before call and return instructions. The debug exception is a single byte opcode (0xCC) used to trap instructions by raising debug exceptions. The VMCS is configured to cause VMExits on debug exceptions. This allows us to mimic the behavior of the alarm replayer, modulo a minor performance impact due to a 0.11% increase in the size of the Linux binary.

7.5 Evaluating the Proposed Hardware

In binary translation mode, QEMU virtualizes the processor using software only. This mode is significantly slower, but it allows for simulation of hardware. We use this mode to evaluate our proposed hardware modifications in RnR-Safe. We simulate a 48-entry RAS by default.

8. EVALUATION

8.1 Recording

Our recording scheme generates the log and also saves and restores the RAS at context switches. Recall that we require hypervisor mediated I/O, which prevents the use of paravirtualized (PV) network drivers. Figure 5(a) compares the execution time of our scheme (called *Rec*) to that of three other setups: no recording with PV drivers (*NoRecPV*), no recording and no PV drivers (*NoRec*), and recording without saving and restoring the RAS at context switches (*RecNoRAS*). Each benchmark is normalized to *NoRec*.

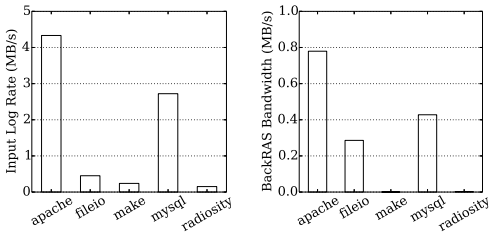
We see that disabling PV increases the execution time of these benchmarks by 25-150%. Apache and fileio are affected the most, while mysql is not impacted much, as it avoids disk accesses by caching recently-accessed tables in memory. Note, however, that RnR has been successfuly applied to PV drivers [25]; applying those techniques in our solution would eliminate this overhead from our system.

Recording (*Rec*) takes, on average, 27% longer than *NoRec*. Recording without saving and restoring the RAS (*RecNoRAS*) takes 24% longer than *NoRec*. These overheads are modest, and will decrease in a reasonably-optimized implementation of recording—e.g., Pokam et al. [21] measure that their implementation of recording adds only 13% overhead.

To understand the source of overheads, Figure 5(b) takes the slowdown of *Rec* over *NoRec* and breaks it down into their sources, namely recording timer reads (*rdtsc*), port and memory-mapped I/O accesses (*pio/mmio*), interrupts, network packet contents, and saving and restoring the RAS at context switches.

We see that the dominant overhead across all benchmarks is due to recording *rdtsc*. This event occurs very frequently, especially in fileio and mysql, where the application itself issues many timer reads to measure transaction speed. In addition, fileio issues disk command and control signals using pio. It also has DMA activity, which causes interrupt events to signal file access completion. Apache receives network packets and uses mmio accesses to the NIC to retrieve the packets. The more computation-intensive benchmarks (make and radiosity) have little overhead. Finally, saving and restoring the RAS induces only 4% overhead on average.

Figures 6(a) and (b) show the input log generation rate, and the bandwidth of RAS saving and restoring, respectively, for all our benchmarks. We do not compress the data. We see that the rates of log generation are low. Apache has the highest input log rate (4 MB/s) because of recording network packet contents. Also, the bandwidth to save and restore the RAS at context switches is very small. Overall, the impact of the architecture on the memory system is modest.

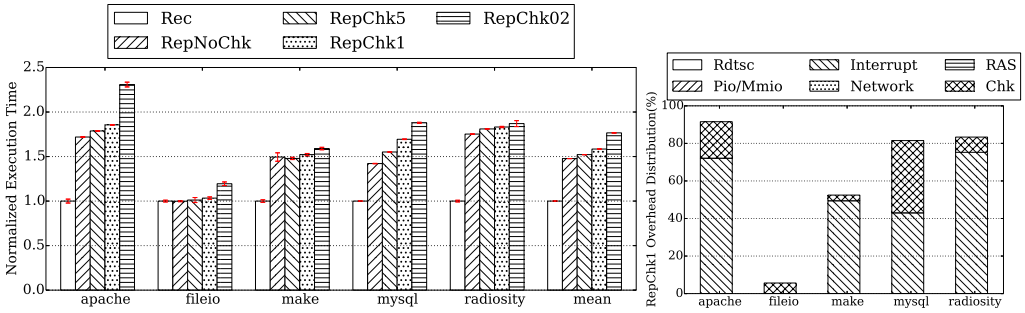


(a) (b)

Figure 6: Input log generation rate (a) and bandwidth to save and restore the RAS at context switches (b).

8.2 Minimizing False Alarms in the Kernel

The RnR-Safe hardware eliminates most of the false alarms in the kernel, allowing only a few false alarms to be reported to the replayers. Figure 8 shows the number of kernel false alarms, broken down into those suppressed with the whitelist and with the BackRAS, and those reported to the replayers (*FalseAlarm*). The figure shows the number per million instructions. Since the number of false alarms passed to the re-



(a) (b)

Figure 7: Execution time of checkpointing replay setups (a) and breakdown of the *RepChk1* overhead over *Rec* (b).

players is so small, the *FalseAlarm* category cannot be seen, and we put the number on top of the bars. All the benchmarks except Apache pass practically no kernel false alarm. Apache passes a few false alarms, which are RAS underflows. They are caused by the deep procedure nesting of the network driver code under extreme loads. Both the whitelist and the BackRAS are very effective at supressing false alarms.

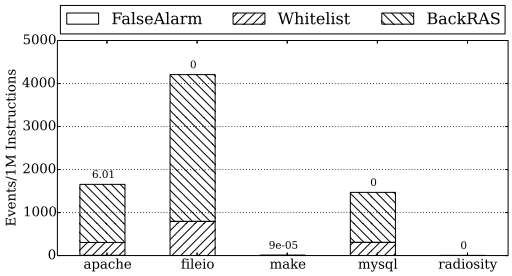


Figure 8: Kernel false alarms suppressed (*Whitelist* and *BackRAS*) and reported to the replayers (*FalseAlarm*).

8.3 Replaying

8.3.1 Checkpointing Replay

Figure 7(a) compares the execution time of various checkpointing replay setups to the recording setup (*Rec*). The replay setups use no checkpointing (*RepNoChk*) or checkpoint every 5, 1, or 0.2 seconds (*RepChk5*, *RepChk1*, and *RepChk02*, respectively). The bars are normalized to *Rec*.

From the data, we see that checkpointing every 1 second (*RepChk1*) increases the execution time over *Rec* by 59% on average. These results show that checkpointing replay runs at a speed that is roughly comparable to that of recording. As a result, checkpointing replay can be *on* all the time. While checkpointing replay is a bit slower, it can easily catch up with recording because even busy machines are rarely 100% utilized—they are often waiting for multiple reasons. During that time, recording slows down but replay can continue. If the replay gets significantly behind, we can use back pressure to temporarily slow down recorded execution.

The figure also shows that increasing or decreasing the checkpoint period changes the replay speed. However, even without checkpointing, replay already takes on average 48% longer that *Rec*.

To understand these effects, Figure 7(b) takes the slowdown of *RepChk1* over *Rec* and breaks it down into its sources. The sources are the events that we saw in Figure 5(b) for the recording, plus creating checkpoints (*Chk*).

The breakdown in the figure shows that creating checkpoints contributes noticeably to the total overhead. This is why the frequency of checkpoints matters. The actual overhead depends on the memory write characteristics of the workload; poor memory locality causes more page copies, increasing checkpointing overhead.

Interestingly, we see that interrupt overhead dominates. The reason is that interrupts are asynchronous events, while *rdtsc*, *pio/mmio*, and *network* are synchronous. Identifying the instruction that should get the asynchronous interrupt injected during replay is time consuming. As indicated in Section 7.3, it requires single-stepping VMExits over several instructions. This is the reason for the overhead of Figure 7(b). It also explains that replaying without checkpointing (*RepNoChk*) already has significant overhead over *Rec*.

8.3.2 Alarm Replay

We now consider the speed of an alarm replayer that checks for ROP attacks in the kernel. Figure 9 compares the execution time of such alarm replay (*RepAlarm*) to previously-shown environments: checkpointing replay (*RepChk1*) and recording (*Rec*). The bars are normalized to *Rec*. Alarm replay needs to trap on every kernel call and return instruction. Hence, the slowdown of this mode directly relates to how many kernel call and return instructions were executed. We see that alarm-replaying make and mysql takes 30-40x longer than recording them. For apache, it takes 50x. On the other hand, for radiosity, with its modest kernel activity, it takes 2.8x.

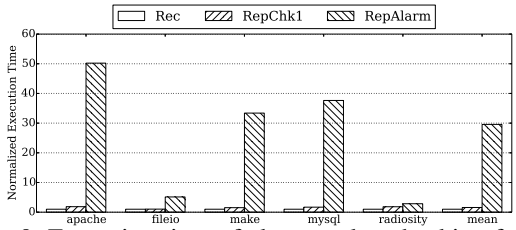


Figure 9: Execution time of alarm replay checking for ROP attacks in the kernel.

8.4 Time Window to Respond to an Attack

The amount of time it takes to detect a ROP is the difference between the time when the alarm replayer confirms a ROP, and the time when the recorded execution logged the alarm. Such time window and the size of the resulting log that was generated in between the two times depend on multiple factors. Two of the most important ones are the work-

load characteristics and the number of machines dedicated to replay. For our system, we measured that the time window is on average a few seconds, and the log size several MBs (Figure 6(a)).

The number of checkpoints that the system needs to retain depends on how far back we want execution to roll to fully understand the attack. Strictly speaking, to reproduce the state at the point of the attack, the alarm replayer only needs to start from the most recent checkpoint. In *RepChk1*, such checkpoint is, at worst, one second old. If that is all that is desired, RnR-Safe only needs to keep as many checkpoints as the duration of the time window mentioned above in seconds plus two—this is to ensure that the correct checkpoint is not prematurely overwritten.

However, if the user wants to analyze the last *N* seconds of execution before the attack was triggered to understand the context of the attack, RnR-Safe needs to keep an additional *N* checkpoints. Finally, checkpoints can be stored indefinitely, if the user wants the entire history recorded. The user can be motivated to do this as the recorded history can be used for forensics or to audit prior executions to detect intrusions.

9. RELATED WORK

Control Flow Integrity. Enforcing Control Flow Integrity (CFI) [31] is the sound technique to prevent code reuse attacks. It requires preventing branch destinations disallowed by the Control Flow Graph (CFG) and/or the shadow stack. Relaxed approaches [48] avoid the shadow stack and/or CFG by relaxing the definition of valid branch targets. Valid branch targets depend on either the type or location of the destination instruction. For example, Intel CET [30] re-purposes a multi-byte NOP instruction to mark valid destinations for indirect branches. Other approaches [33, 34] define validity via the proximity of the branch destination with respect to function boundaries. In general, such approaches fail to completely eliminate gadgets [48, 65], permitting ROP payload construction. Moreover, shadow stack integrity and longevity of CFI are additional points of concerns for CFI [66, 67].

Address Space Layout Randomization. ASLR hardens systems against ROP attacks by randomizing the locations of the stack, heap, and program instructions. Thus, attackers must first discover the location of the code and stack via address disclosure attacks [55, 54, 53, 68]. This additional requirement compounds the difficulty of mounting ROP attacks. Additionally, to further strengthen ASLR, there are proposals for hardening systems against address disclosure attacks [69, 70, 71]. In summary, ASLR is a practical, effective, and widely deployed hardening technique which indeed makes ROP attacks more difficult to mount. However, until the address disclosure attack surface is eliminated, ASLR cannot fully eliminate ROP attacks.

Record and Deterministic Replay (RnR) for Security. Bezoar [72] uses taint tracking hardware to identify network inputs originating from an attacker. Then, the VM is replayed while skipping the malicious network inputs.

The closest previous work to ours is Aftersight [14]. It suggests using VM-level RnR to perform online dynamic analysis of a system's execution. Although it lays out a general direction for VM-level RnR for online analysis, Aftersight does not address some important aspects of such a model.

Unlike RnR-Safe, Aftersight assumes that replay analysis in full is constantly running and is able to catch up with (or only modestly slow down) the recording; otherwise, it loses precision and might introduce false positives. This is not a reasonable assumption in case of heavy-weight analysis, as needed for ROP detection.

Instead, RnR-Safe's architecture presents the key practical aspects of online RnR security analysis. They are: (1) Co-designed hardware-software mechanisms (e.g., the RAS extensions are co-designed with the capabilities of the replayers) to achieve reasonable overhead while keeping hardware changes simple; (2) separate checkpointing and alarm replayers; and (3) *need-based* triggering of the alarm replayers, as opposed to constantly running the analysis.

Memory Safety. Guaranteeing memory safety can entirely eliminate code reuse attacks. A memory access is safe when it is conducted through a valid memory reference, and when it does not violate the bounds of the original allocation. Ensuring the first property requires tracking the lifetime of every pointer, detecting any dereference of a pointer that has been freed. Ensuring the second property requires checking every pointer dereference to ensure it is within the boundaries of the structure [73, 74]. Recently, Intel has begun to include in its processors hardware accelerated bounds checking via the Memory Protection eXtensions (MPX) [1]. While MPX is the most significant step towards providing memory safety on commodity systems, it does not protect against the first property mentioned before. Without this property, the use-after-free attack surface remains. Moreover, it has at least two usability concerns. One is that bounds updates can race with memory accesses. This may lead to time-of-check-to-time-of-use vulnerabilities in multi-threaded codes. Another is that a compromised kernel can bypass MPX by disabling or corrupting its checks. Hence, it is difficult to depend on MPX for kernel protection.

10. CONCLUSIONS

This paper proposed RnR-Safe, a framework where RnR is used to complement hardware security features, allowing such features to be imprecise and suffer false positives. This property often makes the features less intrusive or complicated. RnR-Safe uses two on-the-fly replayers: a checkpointing one and an alarm one. As an example, we applied RnR-Safe to thwart ROP attacks—with a focus on the kernel, the most difficult target to secure. RnR-Safe augments the RAS hardware to eliminate false positives due to multithreading and non-procedural returns. We evaluated RnR-Safe on a VM running Linux. We found that RnR-Safe is a very effective co-design. The checkpointing replayer has comparable execution speed as the recorder, and can be replaying all the time. Also, the alarm replayer has to handle only very few false positives.

Acknowledgments

This work was supported in part by NSF grants CNS 15-26493, CCF 16-29431, CCF 16-49432, and CCF 17-25734.

APPENDIX A: WHAT IS A ROP ATTACK?

The objective of attackers is to execute malware on a victim machine. In the past, attackers injected malware machine

code into memory allocated for data, and hijacked execution to fetch instructions from there. The W⊕X policy [2, 4, 75, 76, 77] was designed to counter this specific attack vector. By enforcing that memory pages are either executable or writable—but never both—malware injected into memory can no longer be executed. To bypass W⊕X, *Code Reuse* based attacks were proposed. For these attacks, existing correct code unwittingly provides malware instructions. ROP [5] is the dominating example of this approach.

Conceptually, an ROP attack executes multiple snippets of code from the victim program or software environment (e.g. libc) called *Gadgets*. Each gadget is terminated with a return—a branching instruction whose target is popped from the software stack. The attacker first loads into the software stack the addresses of the desired gadgets. Then, to trigger the attack, control flow is forced to the first gadget. As the first gadget terminates, its return instruction pops the next entry from the software stack, redirecting execution to the next gadget. Thus, by writing onto the stack the addresses of gadgets, the attacker can stitch together a desired sequence of gadgets required to achieve the desired malicious effects.

This type of attack is dangerous for several reasons. First, it has been shown that the right set of gadgets can construct a Turing-complete language [5], enabling an ROP compiler to translate malware from any other Turing-complete language (like C) to one expressed entirely in gadgets. Second, this attack bypasses the prevalent W⊕X defense techniques, because there is no data being written and then directly executed: the malware executes existing code. Finally, any simple bug in the code enabling attackers to corrupt the stack can trigger the execution of a sophisticated chain of gadgets.

Figure 10 shows an example of an ROP attack that exploits a buffer overflow to execute three gadgets. We use a buffer overflow bug for simplicity; any bug that allows stack modification can be used to launch an ROP attack.

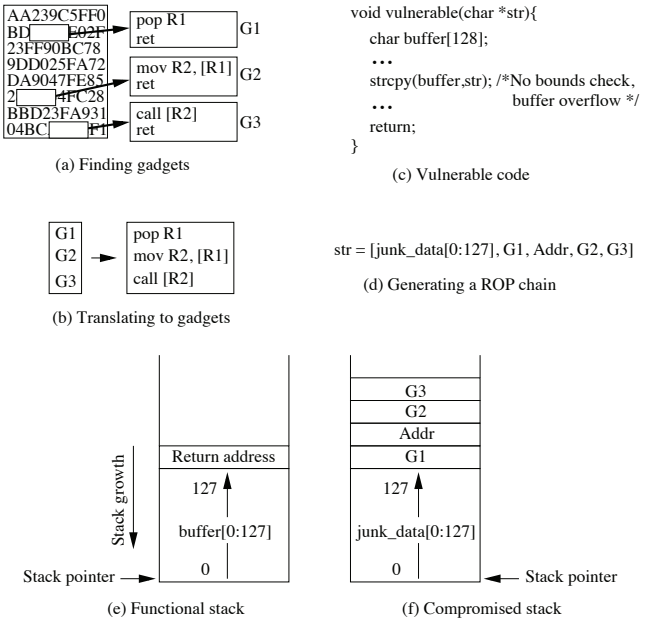


Figure 10: Example of Return Oriented Programming attack.

In Figure 10(a), the executable is scanned for instances of the return (*ret*) instruction. We decode a few bytes before three returns creating three gadgets (G1-G3). Executing the three gadgets in sequence is equivalent to executing the code in Figure 10(b). The code will result in a subroutine call to a function pointer loaded from a memory location stored on the stack. If this is executed during kernel execution, it can be a call to code giving the user root privileges.

Figure 10(c) shows code that is vulnerable to a buffer overflow attack. The code copies a string into a 128-byte buffer without verifying that it can fit in the buffer. Figure 10(d) shows how a payload can be constructed to exploit this code to execute ROP malware. Figure 10(e) shows the benign state of the stack, and Figure 10(f) its state after being corrupted by the malicious input string. Now, returning from the vulnerable function takes us to G1, which will pop Addr into R1 and then return. The return will lead to G2, which will load into R2 and return to G3. Then, G3 will perform the call.

11. REFERENCES

[1] Intel Corporation, *Intel 64 and IA-32 Architectures Software Developer's Manual, Volume 1*. 2017.

[2] American Micro Devices, "AMD64 Architecture Programmer's Manual Volume 2: System Programming," 2006.

[3] J. Winter, "Trusted Computing Building Blocks for Embedded Linux-based ARM Trustzone Platforms," in *Workshop on Scalable Trusted Computing*, 2008.

[4] Intel Corporation, *Intel 64 and IA-32 Architectures Software Developer's Manual*. No. 253669-033US, December 2015.

[5] H. Shacham, "The Geometry of Innocent Flesh on the Bone: Return-into-libc Without Function Calls (on the x86)," in *Conference on Computer and Communications Security*, 2007.

[6] T. J. LeBlanc and J. M. Mellor-Crummey, "Debugging Parallel Programs with Instant Replay," *IEEE Trans. Comp.*, April 1987.

[7] T. Bressoud and F. Schneider, "Hypervisor-Based Fault-Tolerance," *ACM Transactions on Computer Systems*, vol. 14, February 1996.

[8] G. W. Dunlap, S. T. King, S. Cinar, M. A. Basrai, and P. M. Chen, "ReVirt: Enabling Intrusion Analysis Through Virtual-machine Logging and Replay," *SIGOPS Oper. Syst. Rev.*, vol. 36, Dec. 2002.

[9] M. Xu, R. Bodik, and M. D. Hill, "A "Flight Data Recorder" for Enabling Full-System Multiprocessor Deterministic Replay," ISCA, June 2003.

[10] S. Narayanasamy, G. Pokam, and B. Calder, "BugNet: Continuously Recording Program Execution for Deterministic Replay Debugging," in *International Symposium on Computer Architecture (ISCA)*, 2005.

[11] A. Joshi, S. T. King, G. W. Dunlap, and P. M. Chen, "Detecting Past and Present Intrusions Through Vulnerability-specific Predicates," in *Symposium on Operating Systems Principles*, SOSP '05, 2005.

[12] P. Montesinos, L. Ceze, and J. Torrellas, "DeLorean: Recording and Deterministically Replaying Shared-Memory Multiprocessor Execution Efficiently," ISCA, June 2008.

[13] E. B. Nightingale, D. Peek, P. M. Chen, and J. Flinn, "Parallelizing Security Checks on Commodity Hardware," in *International Conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS XIII, 2008.

[14] J. Chow, T. Garfinkel, and P. M. Chen, "Decoupling Dynamic Program Analysis from Execution in Virtual Environments," USENIX ATC, June 2008.

[15] G. W. Dunlap, D. G. Lucchetti, M. A. Fetterman, and P. M. Chen, "Execution Replay of Multiprocessor Virtual Machines," VEE, March 2008.

[16] G. Altekar and I. Stoica, "ODR: Output-Deterministic Replay for Multicore Debugging," SOSP, October 2009.

[17] S. Park, Y. Zhou, W. Xiong, Z. Yin, R. Kaushik, K. H. Lee, and S. Lu, "PRES: Probabilistic Replay with Execution Sketching on Multiprocessors," SOSP, October 2009.

[18] H. Patil, C. Pereira, M. Stallcup, G. Lueck, and J. Cownie, "PinPlay:

A Framework for Deterministic Replay and Reproducible Analysis of Parallel Programs," CGO, April 2010.

[19] G. Portokalidis, P. Homburg, K. Anagnostakis, and H. Bos, "Paranoid Android: Versatile Protection for Smartphones," in *Computer Security Applications Conference*, ACSAC '10, 2010.

[20] K. Veeraraghavan, D. Lee, B. Wester, J. Ouyang, P. M. Chen, J. Flinn, and S. Narayanasamy, "DoublePlay: Parallelizing Sequential Logging and Replay," ASPLOS, March 2011.

[21] G. Pokam, K. Danne, C. Pereira, R. Kassa, T. Kranich, S. Hu, J. Gottschlich, N. Honarmand, N. Dautenhahn, S. T. King, and J. Torrellas, "QuickRec: Prototyping an Intel Architecture Extension for Record and Replay of Multithreaded Programs," ISCA, June 2013.

[22] N. Honarmand and J. Torrellas, "RelaxReplay: Record and Replay for Relaxed-Consistency Multiprocessors," ASPLOS, March 2014.

[23] P. Barham, B. Dragovic, K. Fraser, S. Hand, T. Harris, A. Ho, R. Neugebauer, I. Pratt, and A. Warfield, "Xen and the Art of Virtualization," *ACM SIGOPS Operating Systems Review*, vol. 37, no. 5, 2003.

[24] "QEMU Open Source Process Emulator." `http://qemu.org`.

[25] A. Burtsev, D. Johnson, M. Hibler, E. Eide, and J. Regehr, "Abstractions for Practical Virtual Machine Replay," in *International Conference on Virtual Execution Environments*, VEE '16, 2016.

[26] S. Zonouz, A. Houmansadr, R. Berthier, N. Borisov, and W. Sanders, "Secloud: A Cloud-based Comprehensive and Lightweight Security Solution for Smartphones," *Comput. Secur.*, vol. 37, Sept. 2013.

[27] R. B. Lee, D. K. Karig, J. P. McGregor, and Z. Shi, "Enlisting Hardware Architecture to Thwart Malicious Code Injection," in *International Conference on Security in Pervasive Computing*, March 2003.

[28] H. Ozdoganoglu, T. Vijaykumar, C. Brodley, B. Kuperman, and A. Jalote, "SmashGuard: A Hardware Solution to Prevent Security Attacks on the Function Return Address," *IEEE Transactions on Computers*, vol. 55, Oct 2006.

[29] T. H. Dang, P. Maniatis, and D. Wagner, "The Performance Cost of Shadow Stacks and Stack Canaries," in *Symposium on Information, Computer and Communications Security*, 2015.

[30] Intel Corporation, *Control-flow Enforcement Technology Preview*. No. 253669-033US, June 2017.

[31] M. Abadi, M. Budiu, U. Erlingsson, and J. Ligatti, "Control-flow Integrity," in *Conference on Computer and Communications Security*, CCS '05, 2005.

[32] L. Davi, A.-R. Sadeghi, and M. Winandy, "ROPdefender: A Detection Tool to Defend Against Return-oriented Programming Attacks," in *Symposium on Information, Computer and Communications Security*, ASIACCS '11, 2011.

[33] M. Kayaalp, M. Ozsoy, N. Abu-Ghazaleh, and D. Ponomarev, "Branch Regulation: Low-overhead Protection from Code Reuse Attacks," in *International Symposium on Computer Architecture (ISCA)*, 2012.

[34] E. Aktas, F. Afram, and K. Ghose, "Continuous, Low Overhead, Run-Time Validation of Program Executions," in *International Symposium on Microarchitecture*, MICRO-47, 2014.

[35] Y. Cheng, Z. Zhou, Y. Miao, X. Ding, and H. Deng, "ROPecker: A Generic and Practical Approach for Defending against ROP Attack," in *Network and Distributed System Security Symposium (NDSS'14)*, 2014.

[36] J. Criswell, N. Dautenhahn, and V. Adve, "KCoFI: Complete Control-Flow Integrity for Commodity Operating System Kernels," in *Symposium on Security and Privacy (SP)*, May 2014.

[37] U. Dhawan, C. Hritcu, R. Rubin, N. Vasilakis, S. Chiricescu, J. M. Smith, T. F. Knight, Jr., B. C. Pierce, and A. DeHon, "Architectural Support for Software-Defined Metadata Processing," in *International Conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS '15, 2015.

[38] V. Pappas, M. Polychronakis, and A. D. Keromytis, "Transparent ROP Exploit Mitigation Using Indirect Branch Tracing.," in *USENIX Security*, 2013.

[39] Y. Xia, Y. Liu, H. Chen, and B. Zang, "CFIMon: Detecting Violation of Control Flow Integrity Using Performance Counters," in *International Conference on Dependable Systems and Networks (DSN)*, 2012.

[40] J. Li, Z. Wang, X. Jiang, M. Grace, and S. Bahram, "Defeating Return-oriented Rootkits with "Return-Less" Kernels," in *European Conference on Computer Systems*, EuroSys '10, 2010.

[41] C. Cowan, C. Pu, D. Maier, J. Walpole, P. Bakke, S. Beattie, A. Grier, P. Wagle, Q. Zhang, and H. Hinton, "StackGuard: Automatic Adaptive Detection and Prevention of Buffer-Overflow Attacks," in *Usenix Security*, vol. 98, 1998.

[42] J. Criswell, A. Lenharth, D. Dhurjati, and V. Adve, "Secure Virtual Architecture: A Safe Execution Environment for Commodity Operating Systems," in *Symposium on Operating Systems Principles*, SOSP '07, 2007.

[43] J. Hiser, A. Nguyen-Tuong, M. Co, M. Hall, and J. W. Davidson, "ILR: Where'd My Gadgets Go?," in *Symposium on Security and Privacy*, May 2012.

[44] C. Giuffrida, A. Kuijsten, and A. S. Tanenbaum, "Enhanced Operating System Security Through Efficient and Fine-grained Address Space Randomization," in *USENIX Security Symposium (USENIX Security 12)*, 2012.

[45] R. Wartell, V. Mohan, K. W. Hamlen, and Z. Lin, "Binary Stirring: Self-randomizing Instruction Addresses of Legacy x86 Binary Code," in *Conference on Computer and Communications Security*, CCS '12, 2012.

[46] PaX Team, "PaX Address Space Layout Randomization (ASLR)," 2003. https://pax.grsecurity.net/docs/aslr.txt.

[47] M. Kayaalp, T. Schmitt, J. Nomani, D. Ponomarev, and N. Abu-Ghazaleh, "SCRAP: Architecture for Signature-based Protection from Code Reuse Attacks," in *International Symposium on High Performance Computer Architecture (HPCA2013)*, 2013.

[48] L. Davi, D. Lehmann, A.-R. Sadeghi, and F. Monrose, "Stitching the Gadgets: On the Ineffectiveness of Coarse-grained Control-flow Integrity Protection," in *USENIX Security Symposium*, 2014.

[49] F. Schuster, T. Tendyck, J. Pewny, A. Maaß, M. Steegmanns, M. Contag, and T. Holz, "Evaluating the Effectiveness of Current Anti-ROP Defenses," in *Research in Attacks, Intrusions and Defenses*, Springer, 2014.

[50] C. Cowan, S. Beattie, J. Johansen, and P. Wagle, "PointGuard: Protecting Pointers from Buffer Overflow Vulnerabilities," in *Conference on USENIX Security Symposium*, vol. 12, 2003.

[51] N. Tuck, B. Calder, and G. Varghese, "Hardware and Binary Modification Support for Code Pointer Protection From Buffer Overflow," in *International Symposium on Microarchitecture*, Dec 2004.

[52] B. Spengler, "Grsecurity," 2006. http://www.grsecurity.net.

[53] H. Shacham, M. Page, B. Pfaff, E.-J. Goh, N. Modadugu, and D. Boneh, "On the Effectiveness of Address-space Randomization," in *Conference on Computer and Communications Security*, CCS '04, 2004.

[54] R. Hund, C. Willems, and T. Holz, "Practical Timing Side Channel Attacks against Kernel Space ASLR," in *Symposium on Security and Privacy*, May 2013.

[55] K. Z. Snow, F. Monrose, L. Davi, A. Dmitrienko, C. Liebchen, and A. R. Sadeghi, "Just-In-Time Code Reuse: On the Effectiveness of Fine-Grained Address Space Layout Randomization," in *Symposium on Security and Privacy*, May 2013.

[56] X. Ge, W. Cui, and T. Jaeger, "GRIFFIN: Guarding Control Flows Using Intel Processor Trace," in *International Conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS '17, 2017.

[57] B. Sinharoy, R. Kalla, W. J. Starke, H. Q. Le, R. Cargnoni, J. A. Van Norstrand, B. J. Ronchetti, J. Stuecheli, J. Leenstra, G. L. Guthrie, D. Q. Nguyen, B. Blaner, C. F. Marino, E. Retter, and P. Williams, "IBM POWER7 Multicore Server Processor," *IBM Journal of Research and Development*, vol. 55, May 2011.

[58] B. Sinharoy, J. Van Norstrand, R. Eickemeyer, H. Le, J. Leenstra, D. Nguyen, B. Konigsburg, K. Ward, M. Brown, J. Moreira, D. Levitan, S. Tung, D. Hrusecky, J. Bishop, M. Gschwind, M. Boersma, M. Kroener, M. Kaltenbach, T. Karkhanis, and K. Fernsler, "IBM POWER8 Processor Core Microarchitecture," *IBM Journal of Research and Development*, vol. 59, Jan 2015.

[59] S. Checkoway, L. Davi, A. Dmitrienko, A.-R. Sadeghi, H. Shacham, and M. Winandy, "Return-oriented Programming Without Returns," in *Conference on Computer and Communications Security*, CCS '10, 2010.

[60] "CVE-2015-5364." Available from MITRE, CVE-ID CVE-2015-5364, Dec. 2015.

[61] S. T. King and P. M. Chen, "Backtracking Intrusions," SOSP, October 2003.

[62] R. Senthilkumaran and P. Kulkarni, "InSight: A Framework for Application Diagnosis using Virtual Machine Record and Replay," Tech. Rep. TR-CSE-2014-57, Department of Computer Science and Engineering, Indian Institute of Technology Bombay, January 2014.

[63] A. Kopytov, "SysBench Manual," 2004. https://www.scribd.com/document/267352718/Sysbench-Manual.

[64] S. C. Woo, M. Ohara, E. Torrie, J. P. Singh, and A. Gupta, "The SPLASH-2 Programs: Characterization and Methodological Considerations," in *International Symposium on Computer Architecture (ISCA)*, 1995.

[65] N. Carlini and D. Wagner, "ROP is Still Dangerous: Breaking Modern Defenses," in *USENIX Security Symposium (USENIX Security 14)*, 2014.

[66] M. Abadi, M. Budiu, U. Erlingsson, and J. Ligatti, "Control-flow Integrity Principles, Implementations, and Applications," *ACM Trans. Inf. Syst. Secur.*, vol. 13, Nov. 2009.

[67] N. Carlini, A. Barresi, M. Payer, D. Wagner, and T. R. Gross, "Control-flow Bending: On the Effectiveness of Control-flow Integrity," in *USENIX Conference on Security Symposium*, SEC'15, 2015.

[68] D. Evtyushkin, D. Ponomarev, and N. Abu-Ghazaleh, "Jump over ASLR: Attacking Branch Predictors to Bypass ASLR," in *International Symposium on Microarchitecture (MICRO)*, Oct 2016.

[69] M. Backes, T. Holz, B. Kollenda, P. Koppe, S. Nürnberger, and J. Pewny, "You Can Run but You Can't Read: Preventing Disclosure Exploits in Executable Code," in *SIGSAC Conference on Computer and Communications Security*, CCS '14, 2014.

[70] J. Werner, G. Baltas, R. Dallara, N. Otterness, K. Z. Snow, F. Monrose, and M. Polychronakis, "No-Execute-After-Read: Preventing Code Disclosure in Commodity Software," in *Asia Conference on Computer and Communications Security*, ASIA CCS '16, 2016.

[71] A. Tang, S. Sethumadhavan, and S. Stolfo, "Heisenbyte: Thwarting Memory Disclosure Attacks Using Destructive Code Reads," in *SIGSAC Conference on Computer and Communications Security*, CCS '15, 2015.

[72] D. A. S. de Oliveira, J. R. Crandall, G. Wassermann, S. Ye, S. F. Wu, Z. Su, and F. T. Chong, "Bezoar: Automated Virtual Machine-based Full-system Recovery from Control-flow Hijacking Attacks," in *Network Operations and Management Symposium*, April 2008.

[73] J. Devietti, C. Blundell, M. M. K. Martin, and S. Zdancewic, "Hardbound: Architectural Support for Spatial Safety of the C Programming Language," in *International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, 2008.

[74] S. Nagarakatte, J. Zhao, M. M. Martin, and S. Zdancewic, "SoftBound: Highly Compatible and Complete Spatial Memory Safety for C," in *Conference on Programming Language Design and Implementation (PLDI)*, 2009.

[75] S. Andersen and V. Abella, "Data Execution Prevention. Changes to Functionality in Microsoft Windows XP Service Pack 2, Part 3: Memory Protection Technologies," 2004.

[76] D. Seal, *ARM Architecture Reference Manual*. Pearson Education, 2001.

[77] PaX Team, "Non Executable Data Pages," 2004. https://pax.grsecurity.net/docs/noexec.txt.