Abstract

Code-reuse attacks continue to pose a significant threat to systems security, from resource constrained environments to data-centers. Current mitigation techniques suffer from significant performance and energy overheads especially in the embedded domain. A viable alternative which, thus far, has been a software only mitigation, is Execution Path Randomization (EPR). The basic idea in EPR is to frequently switch between two distinct program variants forcing the attacker to gamble on which code to reuse. In this paper, we propose a novel technique, the Phantom Addressed Instructions that are Randomly Selected (PAIRS) that enables an efficient implementation of EPR with minor hardware adjustments and no source code modifications. For certain modes of operation, PAIRS does not require program re-compilation, thus is applicable to legacy binaries. Compared to state-of-the-art defenses, PAIRS has negligible performance overheads making it viable for embedded IoT devices up to server-grade systems.

1. Introduction

Modern software stacks, from servers to embedded systems, consist of (oftentimes critical) components, such as hypervisors, OS kernels, language runtime environments, etc., which are typically written in memory unsafe languages, like C/C++. The same is also true for the majority of legacy code. The use of memory unsafe languages, however, in settings that constantly operate on inputs of unknown trustworthiness, and provenance, facilitates attacks that rely on memory corruption vulnerabilities [1].

Though it is possible to retrofit memory safety (at least) in C/C++ applications [2, 3], these solutions can lead to compatibility and performance issues [4]. Hardware extensions, like Intel MPX [5, 6] and SPARC ADI [7], try to alleviate the incurred runtime performance overheads, but they also come with their own limitations [8]. For instance, they still require source code to be recompiled in order to make use of the special hardware.

Instead of addressing the root cause viz., memory (un)safety, recent studies have attempted to enforce Control-Flow Integrity (CFI) [9], to potentially vulnerable applications, in order to mitigate the end-effect of certain attacks. However, both coarse- and fine-grained CFI suffer from limitations that allow enough leeway to attackers for bypassing the enforced control-flow policies [10, 11, 12, 13, 14, 15].

Security researchers have recently found a reasonable compromise between memory safety (root cause removal) and CFI (exploit mitigation). Cryptographic control-flow integrity (CCFI) [16], applies memory safety partially to code regions only with the aim of preventing unauthorized modifications to structures that are critical to maintaining control-flow of a program. For instance, for every code pointer, CCFI stores a cryptographically-secure authentication code in the pointer’s unused most significant bits. Checking the authentication code of a pointer before indirect branches prevents control-flow hijacking, as shown in Figure 1(a). Recently, ARM introduced the Pointer Authentication Code (PAC) feature in Armv8.3A as a hardware primitive to enable CCFI with lower overheads on 64-bit systems [17].

Unfortunately, CCFI incurs higher overheads on resource-constrained 32-bit systems as now every pointer load/store operation will need to be accompanied by an additional load/store to retrieve and save the authentication code, respectively. As 32-bit systems are widely used in Internet-of-Things and Cyber Physical Systems, there is a need for low overhead solutions.

An alternative concept to CFI approaches that can be equally applied to 32 and 64-bit systems is Execution Path Randomization (EPR) [18]. Unlike CFI which is a deterministic defense that provides a precise failure model, EPR is a probabilistic defense that makes it very likely for the program to crash in the event of an attack. As shown in Figure 1(b), two different copies of the program that provide the same functionality using different instructions are maintained in memory; then at some control-flow granularity, typically function calls, a security manager switches between the copies. Since the attacker cannot guess which copy will be executed, it reduces the chance of a successful code-reuse attack (CRA). Isomeron, a software-only implementation of EPR was first introduced by Davi et al. [18]. Although it is effective against CRAs, Isomeron on average incurs a 200% runtime slowdown due to code duplication and dynamic binary instrumentation (DBI) usage [19].
In this paper, we introduce a novel concept, called *Phantom Addressed Instructions that are Randomly Selected* (PAIRS), which can support Execution Path Randomization on 32 and 64-bit systems without incurring large overheads. Instead of maintaining two copies of the program, we create a *phantom copy*—a copy that does not exist in physical memory—of the program and switch between it and the original, as shown in Figure 1(c). The locations of program instructions in the phantom copy are shifted by some number of bytes compared to the locations of the same instructions in the original program. On every fetch, the hardware randomly decides if it should fetch an instruction from the phantom or the original. Since the attacker is not privy to the random guesses, with high probability they cannot carry out a CRA. This approach makes the runtime overhead close to baseline execution. PAIRS is realized entirely in the microarchitecture: no binary modification or code recompilation is required in contrast to state-of-the-art CFI techniques.

PAIRS effectively reduces the runtime performance overhead of Isomeron from 200% to 0%, over a vulnerable native execution. We further show that PAIRS can be augmented with a lightweight pointer encryption scheme to protect all control transfers (not just returns) even for *resource-constrained devices* [20, 21] (e.g., in IoT settings) with less than 6% performance overhead. Experimental results show PAIRS’s effectiveness against CRAs.

The remainder of the paper is organized as follows. We provide background on code-reuse attacks in Section 2. Our threat model is illustrated in Section 3. We introduce PAIRS in Section 4 and discuss its security extensions in Section 5. The microarchitectural details are presented in Section 6. We evaluate the performance overheads of PAIRS in Section 7 and its security guarantees in Section 8. We discuss the practicality of PAIRS deployment in Section 9. Section 10 reviews related work. Finally, we conclude our work in Section 11.

2. Background

In this section, we provide background information regarding code-reuse attacks, an execution path randomization system (i.e., Isomeron), and live re-randomization systems.

2.1. Code-reuse Attacks

Attacks that chain together gadgets whose last instruction is a *return* are known as return oriented programming (ROP) attacks [22, 23]. To mount a ROP attack, the attacker has to first analyze the code to identify the respective gadgets, which are sequences of instructions in the victim program (including any linked-in libraries) that end with a return. Second, the attacker uses a memory corruption vulnerability to inject a sequence of return addresses corresponding to a sequence of gadgets. When the function returns, it returns to the location of the first gadget. As that gadget terminates with a return, the return address is that of the next gadget, and so on. AsROP executes legitimate instructions belonging to the program, it is not prevented by W^X [24]. Note that variants of ROP that use indirect *jmp* or *call* instructions, instead of *ret*, to chain the execution of small instruction sequences together also exist, dubbed jump-oriented programming (JOP) [25] and call-oriented programming (COP) [10], respectively.

The standard mitigation technique against ROP/JOP/COP, and pretty much every CRA variant, is address space layout randomization (ASLR), which is currently a well-adopted de-
fense, enabled on (pretty much) every contemporary OS [26]. Essentially, ASLR forces the attacker to first disclose the code layout (e.g., via a code pointer) to determine the addresses of gadgets. Snow et al. [27] observed that typical programs have multiple memory disclosure vulnerabilities. They developed a just-in-time ROP (JIT-ROP) compiler that explores the program’s memory, disassembling any code it finds (in memory), as well as, searching for API and system calls. Then, they construct a compatible code-reuse payload on the fly. Note that, in principle, JIT-ROP is not restricted to dynamically stitching together only ROP payloads; it can also compile JOP, COP, or any other code-reuse payload.

2.2. Execution Path Randomization

Execution Path Randomization (EPR) is considered an effective mitigation to conventional ROP/JOP/COP and JIT-ROP attacks. EPR breaks the chain of gadgets the attacker expects to execute, as it is not predictable which copy of the program will be invoked. As a result, the adversary is therefore unable to reliably construct a working payload, even with full knowledge of the memory layout.

A recent example of such a system is Isomeron [18]. The high-level idea of Isomeron is as follows: (1) a program is duplicated within a single virtual address space (e.g., function foo is duplicated into fooA and fooB); (2) the code is randomized (i.e., fooAdiv and fooBdiv) using any fine-grained ASLR technique [28, 29, 30], generating two variants; (3) function calls are instrumented, with a diversifier, which introduces execution path randomization, by choosing, arbitrarily at runtime, between the variants of every function.

There are two fundamental sources of overhead with Isomeron, and an artefactual overhead. First, Isomeron increases the memory pressure on the system, as it requires two duplicate copies of the code to be loaded into the program’s address space. As a result, this adds more TLB and I-Cache misses. Second, the random diversifier wreaks havoc on the effectiveness of the branch prediction unit in hardware resulting in low prediction accuracy, amounting to wasted execution in out-of-order processors. Finally, the authors of the original paper proposed using Dynamic Binary Instrumentation (DBI) to create variants and hook on to control-transfer instructions. These overheads alone can be close to 5% up to 50% [19] depending on the complexity of the instrumented program. The overheads of software Isomeron are 200% on average. Crucially, the above are only exacerbated in resource-constrained settings, such as IoT devices, as the majority of these platforms use much smaller address spaces and simpler CPUs.

2.3. Live Re-randomization

More recent solutions such as runtime (live) code re-randomization solutions [31, 32, 33] seek to re-randomize, and invalidate, leaked information, before the attacker has a chance to use it (i.e., craft a (JIT-ROP) CRA payload). Existing solutions, such as Shuffler [31] and CodeArmor [32], periodically re-randomize the code layout within milli- and micro-seconds, respectively. Morpheus [33] on the other hand, encrypts code and pointers and relocates pointers in memory using a secret displacement that is changed on the order of milli-seconds. Although such techniques greatly reduce the attack surface by minimizing the time window available to the attacker (for successfully mounting a CRA), they typically introduce considerable energy and runtime overheads. PAIRS overcomes these obstacles as it uses low-complexity hardware modifications that can be realized from the most basic microcontrollers (MCUs) to server-grade systems.

3. Threat Model

Adversarial Capabilities. We consider an adversary model that is consistent with previous work on code-reuse attacks and mitigations [16, 18, 33, 34]. We assume that the adversary is aware of the applied defenses and has access to the source code, or binary image, of the target program. Furthermore, the target program suffers from a memory corruption vulnerability that allows the adversary to read from, and write to, arbitrary memory addresses. The attacker’s objective is to (ab)use the memory corruption vulnerability, mount a CRA, and achieve privilege escalation.

Hardening Assumptions. We assume that the underlying operating system (OS) enables both ASLR and W^X protection—i.e., no code injection is allowed (non-executable data), and all code sections are non-writable (immutable code). Thus, attacks that modify program code at runtime, such as rowhammer [35], are out of scope. We also do not consider non-control data attacks [1], such as Data-Oriented Programming [36]. This class of attacks only (ab)use memory load and store operations, without inducing any unintended control flows in the program. This limitation also applies to prior CFI work [9].

4. Phantom Addressed Instructions that are Randomly Selected (PAIRS)

Figure 2 shows the high level construction of PAIRS. The goal is to assign different virtual addresses, Phantoms, to the same basic block—or more generally, any contiguous group of instructions—randomly at runtime.

Traditionally, one address is allocated to a static instruction in the binary. Thus, an address creates a unique name/identifier for an instruction. In our system, we allocate multiple addresses or names for a static instruction. In doing so, we decouple the runtime address/name of an instruction and a program instruction.

Assigning multiple addresses/names needs to be done with care to avoid collisions. We facilitate this using a new abstraction called the phantom address space: a range of addresses in the virtual address (VA) space of a program that are used to create aliases for program instructions. Associated with this phantom address space is a mapping function \( f \) that represents a many to one transformation, which operates on addresses, \( pa, \)
of a physical address (PA) space and maps them to addresses, \( \text{va} \), of a virtual address (VA) space. This mapping creates phan-
toms of a given physical address space. In other words, a basic
block, \( \text{BBL}_i \), is mapped by \( f \) to many phantoms, \( \text{Phantom}_i \), in VA. The function \( f \) does not have to be kept a secret, as
the security is purely derived from the random selection between
two known addresses.

4.1. PAIRS Construction

There are three main operations to realize PAIRS: Populate, Randomize, and Resolve.

Populate. PAIRS creates multiple phantoms of basic blocks
and populates them in the VA space. The left-hand side of Fig-
ure 2 shows a program with two Phantoms. The two Phan-
toms are separated by a domain offset, \( \Delta \), in the VA space.
To add discrepancy between the Original basic block and its Phantom
copy, we introduce a minor security shift, \( \delta \), so that they are
not perfectly overlapped after removing \( \Delta \). By doing so, each
basic block gets two different \( \text{VAS} \) (Original and Phantom).
The rest of the address space setup remains unchanged.

Randomize. At runtime, we modify the hardware to random-
ize program execution between the Phantoms. For example,
some basic blocks will be executed from Original while other
basic blocks will be executed from Phantom. Correctness is
unaltered because both Phantoms provide the same functional-
ity by construction.

Resolve. Accessing different \( \text{VAS} \) at runtime can incur addi-
tional performance overheads as each \( \text{VA} \) needs to be translated
to a PA before usage. To mitigate this problem, PAIRS uses the
inverse mapping function, \( f^{-1} \), to resolve the different Phan-
toms to their archetype basic block. By doing so, the processor
back-end continues to operate as if there is only one copy of
the program in the VA space.

Many approaches can be used to populate the Phantoms.
One way to realize PAIRS is to use the most significant bit
(MSB) to separate the phantom replicas of the program in VA
space (\( \Delta \) of 0x8000_0000 on 32-bit systems). Then, a map-
ing function, \( f \), will set/unset the MSB at control-flow transi-
tions to randomize the execution at runtime. More complicated
mapping functions can be used as long as \( f \) is a linear trans-
formation that satisfies the bijective property.

One downside of PAIRS is that it reduces the size of \( \text{VA} \)
to one half. However, for classes of 32-bit systems, namely
MCUs, the codes are typically of small sizes (and so is the
SRAM and FLASH). For 64-bit systems, the problem does not
exist because the maximum size of \( \text{VA} \) space is \( 2^{64} \) locations
whereas the \( \text{VA} \) space is limited to \( 2^{48} \) locations. So, any
higher order bit, from 48 to 63, can be used for phantom
region creation.

We note that PAIRS can support any granularity, starting
from a single instruction to a whole program. In the rest of the
paper, we use basic blocks as our elements of interest. We do
not evaluate finer granularities here due to the lack of a strong
security need. We define the basic block (BBL) as a single
entry, single exit region of code. Thus, any instruction that
changes the PC register (referred to by control-flow instruc-
tions, such as \text{jmp, call, ret}) terminates a BBL and starts
a new one.\(^1\)

4.2. PAIRS Correctness

In order to discuss the correctness of PAIRS, we consider the
structured programming theorem [37], in which, any program
can be composed of one out of four control structures: \text{Se-
quence, Selection, Iteration, and Recursion}. We show that
PAIRS does not affect the four structures. First, the \text{Sequence}
structure represents a series of ordered statements or subrou-
tines executed in sequence. PAIRS guarantees that property by
executing the statements (instructions) of a BBL in the same
domain of execution (either Original or Phantom).

For handling the \text{Selection} structure, let us assume that a
program is represented by a binary tree with a branching factor
of 2. Hence, we can define two types of such trees. Type I
is the tree where nodes are represented by the BBLs of \text{com-
mitted} instructions. Type II is the tree where each node has
the address of the first instruction in the \text{executed} BBL. The
edges are given by the direction taken by the last instruction of
each BBL. The root of the tree is the first instruction fetched
from the \text{_start()} section of a binary (or the address of this
instruction in the address based tree). The leaf node is the
last BBL of the program (or the address of this BBL in trees
of Type II). In the case of PAIRS, every taken branch on the
tree of Type I is the same as every taken branch on the tree
of Type II, i.e., program functional decisions are not affected.
However, the contents of the tree nodes in Type II trees would
be different for each program execution as each BBL will be
fetched from either Original or Phantom domain and addresses
of these domains differ by \( \Delta \). In other words, after the branch
is resolved to be taken or not taken, PAIRS operates on the
outcome and randomly chooses the domain of execution for
the next basic block.

The above argument also applies for the \text{Iteration} control-
structure, in which the same basic block is executed multiple
times. PAIRS does not change the functionality of the basic

\(^1\)Some compilers, such as LLVM, deviate from this definition and treat
\text{call} instructions as part of the BBL.
will cause performance overheads with almost no additional
As described in Section 2.1, the major class of CRAs is return
main offset, $\Delta$
that they are not perfectly overlapped after removing the do-
δ
security (beyond changing the gadget addresses in the two
the archetype basic block from the different Phantoms. This
aggressive randomization approach will complicate the inverse
flow of instructions has not changed (Type I trees are unique).
This enables a substantial security benefit as we show next.

4.3. PAIRS for CRA Protection

As described in Section 2.1, the major class of CRAs is return oriented programming (ROP). PAIRS mitigates ROP by ensuring that the addresses of the ROP gadgets in the gadget chain change after the chain is built. This will result in undefined behavior of the payload (likely leading to a program crash).

Let us consider the example of Figure 2. PAIRS simultaneously populates two pairs (copies) of the program code in one VA space. One copy is the original program code, Original, while the other is the phantom program code, Phantom. To successfully thwart the ROP gadget chain, the location of the ROP gadgets in both copies should be different [18].

Ideally, any in-place code layout randomization approach can be used to generate Phantom code. However, using an aggressive randomization approach will complicate the inverse mapping function, $f^{-1}$, which is responsible for recovering the archetype basic block from the different Phantoms. This will cause performance overheads with almost no additional security (beyond changing the gadget addresses in the two copies). PAIRS adopts a more efficient code layout randomization technique by introducing a security shift, $\delta$, between the individual basic block Original and its Phantom copy, so that they are not perfectly overlapped after removing the domain offset, $\Delta$. As a result, the instruction offsets of any two twin BBLs pairs are shifted by $\delta$, while keeping $f^{-1}$ simple (as will be shown in Section 6) and maintaining code locality.

While the program is executing, PAIRS randomly decides which copy of the program should be executed next. Figure 3(a) shows the normal execution of a program, where Inst 10 changes the control-flow of the program to a different BBL (starting with Inst 71). After the called BBL is executed, the control-flow is transmitted to the original landing point (Inst 11 via a ret instruction). Figure 3(b) shows a successful CRA via ROP, in which the attacker uses a memory safety vulnerability to overwrite the return address stored on the stack and divert the control flow to Inst 24 upon executing the ret instruction. Figure 3(c) shows the diversified execution of a program with PAIRS. For simplicity, we use a security shift, $\delta$, sized to one instruction. Each control flow instruction can arbitrary choose to change the execution domain or not. Here, the Randomize action decides to execute Inst 71 from the Phantom domain. As the attacker cannot predict this runtime decision in advance, ze provides the wrong gadget address on the stack (now shifted by $\delta$). Thus, ze will end-up executing a WRONG instruction, as shown in Figure 3(d).

This WRONG instruction may belong to a different BBL or divert the execution to a new undesired BBL. In general, if the attacker makes the wrong guess, they will execute one less (or one more) instruction compared to the desired gadget. If $\delta$ is smaller than the instruction size, the attacker will skip a portion of the instruction resulting in a possibly totally incorrect instruction decoding.

5. PAIRS Security Extensions

In this section, we discuss three PAIRS enhancements that boost its security guarantees against state-of-the-art CRAs.

5.1. Secret Domain Stack

In order for PAIRS to completely mitigate ROP, the current domain of execution should not be leaked to the attacker via any architectural state. Otherwise, the attacker will be able to
adjust the addresses of the desired gadgets on-the-fly to bypass the protection. Normal programs push return addresses to the architectural stack to help returning from non-leaf function calls. The attacker may leak the domain of execution, Original or Phantom, by monitoring the stack contents at runtime using arbitrary read or write vulnerabilities.

To avoid attackers leaking information from the architectural stack, we split the return addresses between the architectural stack and a new micro-architectural structure called the Secret Domain Stack (SDS), which by construction is immutable to external writes. SDS achieves this goal by splitting the return address into two parts; the MSB, which represents the domain bit, \( \delta \), and the lower order bits of the address. With each function call instruction, the lower order bits of the return address are pushed to the architectural (software) stack, whereas the MSB is pushed onto the SDS. A return instruction pops the most recent \( b \) bit from the top of SDS and concatenates it with the return address stored on the architectural stack in memory. While under attack, the state of the two stacks will thus become inconsistent. Deployment issues with the SDS such as sizing, overflows, multithreading, etc. are described in Section 9.

### 5.2. TRAP Instructions

To further limit the attack surface of PAIRS, we add TRAP instructions. These instructions are inserted at the beginning of every basic block. While PAIRS is enabled, the security shift, \( \delta \), will cause the TRAP instruction that exists in the beginning of a BBL in Original domain to appear at the end of the same BBL in Phantom domain, as shown in Figure 4. This way we provide the ability to catch attackers that guess the incorrect diversification on the TRAP instruction while targeting BBL boundaries.

Let us consider the example of Figure 3(d). Here, the attacker tries to divert the control flow to \texttt{Inst 24} upon executing the \texttt{ret} instruction. Our Randomize action decides to execute \texttt{Inst 71} from the Phantom domain. As a result, the attacker will step on the instruction that follows \texttt{Inst 24} in Phantom, which will be a TRAP instruction in the new model. Executing a TRAP instruction results in program termination, thwarting the attacker. Programs cannot execute TRAP instructions in normal conditions (under no attack) as there exist no control-flow transfer to them.\(^2\)

### 5.3. Lightweight Pointer Encryption

Besides ROP, CRA variants also extensively rely on pointer corruption (e.g., JOP/COOP [25, 34]) to subvert a program’s intended control flow. There also exist many software-based mitigations for JOP/COOP-like attacks [38, 39, 40, 41]. In this paper, we use a hardware-based technique for hardening PAIRS against them. Since the attacker needs to overwrite legitimate pointers used by indirect branches to launch the attack, we encrypt the contents of the pointer upon creation and only decrypt it upon usage (at a call site). Consequently, attackers cannot correctly overwrite it.

To achieve the above goal our Lightweight Pointer Encryption (PtrEnc) scheme adds two new instructions: \texttt{ENCP} and \texttt{DECP}. The two instructions can either be emitted by the compiler (if re-compiling the program is possible) or inserted by a binary rewriter.

**Encrypt Pointer** \texttt{ENCP RegX}. The mnemonic \texttt{ENCP} indicates an encryption instruction. \( \texttt{RegX} \) is the register containing the pointer (e.g., \texttt{vptr}) value. The register that holds the encryption key is hardware-based and never appears in the program binary.

**Decrypt Pointer** \texttt{DECP RegX}. The mnemonic \texttt{DECP} indicates a decryption instruction. \( \texttt{RegX} \) is the register containing the pointer (e.g., \texttt{vptr}) value (it can only be specified at runtime). The register that holds the decryption key is hardware-based and does not appear in the program binary. As a result, the attacker cannot directly leak the key’s value.

In contrast to full CCFI solutions [16, 17], which use pointer authentication to protect all code pointers including return addresses, our approach only guards pointer usages (loads and stores). Return addresses are handled by PAIRS randomization, reducing the overall performance overheads, as will be shown in Section 7.

### 6. Microarchitecture

Figure 5 summarizes the modifications to support PAIRS. The changes are limited to structures that operate on program counter (PC) values.

\(^2\)We modify the hardware to handle the case of a fall-through BBL to prevent legitimately stepping on a TRAP instruction.
6.1. Diversifier

The Diversifier ($D$) is responsible for adjusting the PC before executing any new BBL so that the execution flow is unexpected to an attacker. Specifically, $D$ takes the predicted target for a branch ($PC_{new}$) with control-flow signal, $s$, as input. $s$ is set to one if the Branch Predictor Unit (BPU) has a predicted target for this instruction and set to zero otherwise. $D$ generates the nextPC as the output. If $s$ equals one, $D$ flips a coin to generate a random value $b$. This can be implemented using a single metastable flip-flop [42]. Based on $b$, $D$ adjusts the nextPC according to Table 1.

| MSB(PC) | Diversifier ($b$) | nextPC         |
|---------|-------------------|----------------|
| 0 (Original) | 0 (Original) | $PC_{new}$     |
| 0 (Original) | 1 (Phantom)   | $PC_{new} + (\Delta - \delta)$ |
| 1 (Phantom)  | 1 (Phantom)   | $PC_{new}$     |
| 1 (Phantom)  | 0 (Original)  | $PC_{new} - (\Delta - \delta)$ |

Table 1: Address translation scenarios for PAIRS.

Table 1 shows the four cases for PAIRS address translation. MSB(PC) defines current domain of execution, while Diversifier ($b$) directs the execution either to Original or Phantom. nextPC does not change if $D$ decides to remain in the same execution domain. On the other hand, $(\Delta - \delta)$ is used to adjust nextPC when $D$ randomly switches the domain in order to maintain the correctness of benign programs. As the attacker lacks the domain information, ze will cause a wrong transition, as shown in Figure 3(d).

Performance Optimization #1. Placing $D$ as described above adds one cycle latency to the nextPC calculations in the fetch stage. To resolve this issue, we move $D$ to the commit stage.

At the commit stage, the target of the branch instruction is known and the branch is committed. The resolved target address is sent back to the fetch stage to update (train) the BPU buffers. At this point, $D$ will adjust the target address by using $b$, as explained above and update the BPU buffers with nextPC. This ensures that the next execution for this control-flow instruction will be random and unexpected to an attacker. To bootstrap the first control-flow instruction that is executed, we consider the two possible cases: correct prediction or incorrect one. If the first occurrence of the control-flow instruction is correctly predicted to be PC + 4 (failing through), then $D$ will keep using the current domain of execution (unknown for the attacker) for the next BBL. If the first occurrence of the control-flow instruction is incorrectly predicted, it would be detected later on in the commit stage and the pipeline will be flushed. In this case, $D$ will adjust the resolved target address by using $b$ and update the BPU buffers with nextPC. Placement of $D$ at the commit stage allows us to hide latency overheads needed for target addresses adjustments so it does not affect performance.

6.2. Branch Prediction Unit (BPU)

The branch prediction unit stores a record of previous target addresses in the branch target buffer (BTB), and the recent return addresses in the return address stack (RAS). For the current PC value, the BPU checks if the corresponding entry exists in the BTB via indexing. If so, the found target address becomes the nextPC. Otherwise, nextPC is incremented to PC + 4 (or PC + Instruction size). If the predicted target address turns out to be incorrect later in the instruction pipeline, the processor re-fetches the instruction with the correct target address (available usually at the execute stage of the branch instruction) and nullifies the instructions fetched with the predicted target address.

Performance Optimization #2. In a PAIRS system, we have different addresses for the same control-flow instruction. In this case, we will have multiple entries in the prediction tables for the same effective instruction; this reduces the capacity by 50%. To handle this issue, we modify the hashing function that indexes BPU tables to ignore the MSB of the incoming PC and to always adjust PC to the Original domain. This way we guarantee that all PCPhantoms map to the same table entry. After performing the indexing step, we get the desired values from the prediction tables. As explained in Section 6.1, the nextPC values stored in the BTB are already chosen at random from the last successful commit of this control-flow instruction (or any of its phantoms). The only exception is the RAS. Unlike the other BPU structures, the RAS uses a LIFO approach for indexing, thus no PAIRS modifications are needed in this case. However, PAIRS ensures that the stored return addresses in the RAS are randomized before storage similar to the BTB case. The branch direction prediction results (Taken vs. Not Taken) in the branch direction buffer (BDB) remain the same.

6.3. Translation Look-aside Buffer (TLB)

Performance Optimization #3. Similar to the BPU buffers issue, the fact that we have multiple variants of every BBL with multiple different VAs may lead to multiple different VA-to-PA entries in the TLB for the same Original translation, reducing its capacity by 50%. To avoid potential performance degradation, we mask the MSB of the incoming VA and adjust it to the Original domain before indexing or updating the ITLB. Thus the translations related to all BBLPhantoms map to a single entry in the ITLB. We do not modify the PAS so that the stored PA part of the translation remains unaffected.

6.4. Instruction Cache

Performance Optimization #4. The main reason for the inefficiency of Execution Path Randomization [18] is the large number of L1-I$ misses. Creating two variants of the code sections for each program means that the L1-I$ capacity would be effectively reduced to one half. As shown in Figure 6, we mask the MSB in the tag comparison so that BBLOriginal or BBLPhantom map to the same I-cache line. Second, we add
We evaluate the optimal size of SDS in Section 7.

This structure does not introduce additional latency as it is accessed in parallel to the normal architectural stack access.

Performance Optimization #5. If the target architecture allows forwarding the PC register through the pipeline for regular instructions, we make sure that the PC register is always mapped to the Original domain before operating on it. This mapping may introduce additional latency for the execute stage as it should be done before/after it. To hide such latency, one solution is to add a new control signal to the execution units in order to select whether to operate on Phantom or Original version of the PC register. Both versions should be forwarded to the desired execution units. Although such a solution can completely hide the adjustment latency, it slightly increases the area of the execution unit.

6.6. Secret Domain Stack

Performance Optimization #6. Unlike prior work, which stores a complete version of the return addresses (e.g., 32-bit on AARCH32) in what is called a shadow stack [43], we only store a single bit per each return address. This saves silicon area within the processor and facilities managing the SDS. This structure does not introduce additional latency as it is accessed in parallel to the normal architectural stack access. We evaluate the optimal size of SDS in Section 7.

7. Performance Evaluation

In this section, we compare PAIRS against the state-of-the-art software EPR, Isomeron, and a hardware CCFI implementation. As we focus on resource constrained devices, we use ARM to demonstrate PAIRS as it dominates the embedded and mobile markets with its 32-bit ARMv5–8 instruction set architecture (ISA). However, the concept of PAIRS can be applied to any other ISA (e.g., RISC-V).
both instructions [49]. This latency is to emulate the effect of the actual encryption/decryption operations in addition to the latency of the additional load/store pointer operation.3

**PtrEncFull.** In this approach, we instrument code pointer load/store operations in addition to function entry/exit points to protect return addresses for non-leaf functions. Conceptually, this solution is similar to ARM Pointer Authentication Code [50]. However, due to the absence of CCFI support in Gem5 (and for 32-bit ARM architectures in general), we only perform behavioral simulation for comparison purposes, without keeping track of the actual pointer metadata.

**Software-EPR.** For the sake of completeness and fair comparison, we also implement a static version of software Execution Path Randomization, Isomeron [18], (at the BBL granularity) as opposed to the original work, which uses a DBI framework with function granularity. We create the two copies by introducing a shift of TRAP instruction size in one of them. At a high-level our implementation works as follows: (1) clone functions using an LLVM IR pass (2) LLVM backend pass to insert TRAPS for cloned functions (3) instruct the LLVM backend to globalize BBL labels (4) emit a diversifier BBL for every BBL (5) and finally rewrite branch instruction targets to point to the diversifier. In comparison to the original work, which not only has to contend with the overhead of the DBI framework in addition to the diversifier logic; the static version only pays the penalty of two jumps, one to the diversifier and a second to the main code.

Of the 16 C/C++ benchmarks, 14 compile with all different toolchain modifications, parest has compatibility issues with musl due to exception handling usages, while povray failed to run on Gem5. For Software-EPR, gcc, xalancbmk, and x264 present compilation and/or linking issues.

### 7.2. Experimental Results

We run all benchmarks to completion with the test input set on our augmented Gem5. We verified the correctness of the outputs against the reference output. Figure 7 shows the performance overheads of the different design approaches (all normalized to Baseline). As expected, PAIRS has identical performance to Baseline. The overhead of PAIRS-TRAP is minimal, approximately 4%. Adding support for PtrEnc increases the performance overheads of PAIRS-PtrEncLite to 6%. The perlbench benchmark suffers from a relatively high overhead due to its extensive use of function pointers and indirect branches. On the other hand, fully protecting the binaries with a deterministic defense such as PtrEncFull encounters a 91% overhead on average (geometric mean of 62%). Our static implementation of software EPR introduces an arithmetic average overhead of 31% (geometric mean of 26%).

In contrast to the DBI software Isomeron, the overheads for our implementation are primarily attributed to the indirection every BBL branch must make to the diversifier.

As illustrated in Section 6, the required PAIRS modifications do not add additional cycle latency to the processor pipeline. However, we performed an additional set of experiments with a more conservative assumption of having one additional cycle latency for all instructions in fetch stage, or one more cycle for accessing L1 instruction cache, or both. We show results compared to an unmodified baseline in Figure 8. We notice an average performance overhead of 1% for stalling the fetch stage. However, stalling the instruction cache for one cycle (hit latency is originally two cycles) is more harmful to the performance. Thus, the IS optimizations are mandatory, as described in Section 6.

Finally, the call depths listed in Table 2 show that SPEC programs do not exceed a depth of 244 (leela), indicating that a 512-entry hardware Secret Domain Stack is sufficient to handle the common execution cases.

### 8. Security Analysis and Evaluation

In this section, we present our security evaluation results and discuss the security guarantees of PAIRS against CRAs.

---

3Our Software-EPR implementation does not instrument external libraries (only the main application code) due to compilation issues. This leads to overheads that are less than intuitively expected.
Table 2: Maximum call depth for SPEC C/C++ benchmarks.

| Bench Name | Call Depth | Bench Name | Call Depth |
|------------|------------|------------|------------|
| perlbench  | 24         | x264       | 15         |
| gcc        | 28         | deepsjeng  | 48         |
| mcf        | 28         | leela       | 244        |
| omnetpp    | 196        | xz          | 16         |
| xalan/bmk  | 77         | namd        | 12         |

Table 3: ROP gadget-chain reduction with PAIRS deployment for SPEC C/C++ benchmarks. PAIRS and PAIRS represent the number of valid ROP chains before and after PAIRS.

| Bench Name | PAIRS Chains | PAIRS Chains | Bench Name | PAIRS Chains | PAIRS Chains |
|------------|--------------|--------------|------------|--------------|--------------|
| perlbench  | 17           | 0            | ibn        | 23           | 0            |
| gcc        | 23           | 11           | blender    | 23           | 0            |
| mcf        | 11           | 0            | leela      | 23           | 0            |
| omnetpp    | 23           | 0            | xz         | 23           | 0            |
| xalan/bmk  | 15           | 0            | namd       | 23           | 0            |

8.1. Quantitative Security Analysis

ROP-Gadget Chain Evaluation. To evaluate PAIRS against real-world ROP attacks we use Ropper [51], a tool that can find gadgets and build ROP chains for a given binary. A common ROP attack is to target the execute function with /bin/sh as an input to launch a shell. As the chain-creation functionality in Ropper is only available for x86 [51], we analyze SPEC2017 x86 binaries for this particular exploit and report the number of available gadget chains (PAIRS).

To emulate the effect of PAIRS, we modified the Ropper code to extend each gadget length by one byte, decode the gadget, and check if the new gadget is semantically equivalent to the old one or not. This emulates the effect of an attacker targeting a particular address, but instead executing the one before due to the PAIRS security shift, δ. As shown in Table 3, PAIRS corrupts all the gadget-chains found by Ropper. Extending the Ropper chain-creation functionality to the ARM ISA is part of our future work. Intuitively, the results would be even worse for the attacker in ARM as the state-space is more constrained due to instruction alignment requirements.

8.2. Qualitative Security Evaluation

Just-In-Time Return-Oriented Programming. Although JIT-ROP [27] permits the attacker to construct a compatible code-reuse payload on the fly, ze cannot modify the gadget chain after the control flow has been hijacked. As a result, the attacker needs to guess the domain of execution of the entire JIT-ROP gadget-chain in advance. So, PAIRS mitigates JIT-ROP similarly to how it mitigates (static) ROP/JOP/COP: i.e., by removing the attacker’s ability to put together (either in advance or on the fly) a valid code-reuse payload.

Whole-function Reuse. Unlike ROP attacks, which (re)use short instruction sequences, entire functions are invoked, in this case, to manipulate the control-flow of the program. This type of attack includes counterfeit object-oriented programming (COOP) attacks, in which whole C++ functions are invoked through code pointers in read-only memory, such as vtables [34]. PAIRS relies on PtrEnc to prevent the attacker from manipulating pointers (vptr) that point to vtables—a necessary step for mounting a COOP attack.

Another example for whole function reuse attacks is ret2libc, in which the attacker tries to execute entire libc functions [54, 55]. Recent work aims at reducing the set of available functions (and gadgets) for ret2libc attackers by either (1) whitelisting legitimate invocations to security-critical APIs (in shared libraries) and blocking every other invocation at runtime (e.g., Shredder [56]), or (2) safely removing code (including security-critical functions) that is never used by the main binary [57, 58, 59, 60]. Specifically, Nibbler safely removes code bloat from shared libraries, without introducing any extra runtime overhead.

Our analysis of real-world exploits shows that executing a ret2libc attack incurs multiple steps in order for the attacker to (1) prepare the function arguments based on the calling convention, (2) jump to the desired function entry, (3) silence any side-effects that occur due to executing the whole function, and (4) reliably continue (or gracefully terminate) the victim program without noticeable crashes. (1) and (3) generally requires code-reuse (ROP) gadgets, as demonstrated by the following publicly-available exploits: (a) ROP + ret2libc-based exploit against mcrypt, (b) ROP + ret2libc-based exploit against Nginx, (c) ROP + ret2libc + shellcode.
based exploit for Apache + PHP⁸ and (d) ROP + ret2libc-based exploit against Netperl⁹. Thus, if the ROP part of the exploit requires n gadgets, the probability for successfully exploiting the program is \( P_{\text{success}} \leq 2^{-n} \). That is because the attacker will have to guess the domain of execution, Original or Phantom, of every gadget.

**Side-Channel Attacks.** PAIRS takes multiple steps to be resilient to side channel attacks. Firstly, PAIRS purposefully avoids timing variances introduced due to hardware modifications, in order to limit timing-based side channel attacks. Additionally, the attacker cannot leak the random decision, \( b \), as it is unreadable from both user and kernel mode—it exists within the processor only. Similarly, the execution domain, Original or Phantom, cannot be leaked to the attacker through the architectural stack, as PAIRS keeps it within the hardware in the SDS.

### 9. PAIRS Deployment

In this section, for completeness, we outline design changes required to deploy a PAIRS general-purpose system.

**Sizing.** Although our Secret Domain Stack only stores one bit in hardware, it still has a limited size that can not be dynamically increased as the architectural stack. This means programs with deeply nested function calls may result in a SDS overflow. To handle this issue, we add two new hardware exception signals: `hardware-stack-overflow` and `hardware-stack-underflow`. The former is raised when the SDS overflows. In this case, the OS (or another trusted entity), copies the contents of the SDS to the kernel memory. This kernel memory location will be a stack of stacks and every time a stack is full it will be appended to the previous full stack. The second exception will be raised when the SDS is empty to page in the last saved full-stack from the kernel memory.

**Stack Unwinding.** Since addresses are split across the architectural (software) stack and the Secret Domain Stack it is vital to keep them in sync for correct operation. Earlier, we described how normal LIFO `call/rets` are handled. In some cases, however, the stack can be reset arbitrarily by `setjmp/longjmp` or C++ exception handling. To ensure the stack cannot be disclosed/manipulated maliciously during non-LIFO operations, we change the runtime to encrypt the `jmp_buffer` before storing it to memory. Additionally, we also store the current index of the SDS. When a `longjmp` is executed, we decrypt the contents of the `jmp_buffer` and use the decrypted SDS index to re-synchronize it with the architectural stack. The same approach can be applied to the C++ exception handling mechanism by instrumenting the appropriate APIs.

**Debugging Support.** Debugging tools typically include a dump of the stack contents to help the developers understand the root cause of user programs crashes. In our secure PAIRS environment, the return addresses that are stored in the software stack are split to prevent the attacker from leaking information regarding the current domain of execution. To provide a consistent view during debugging, we can use a trusted system call that can be invoked only through the debugger with the capability of reconstructing the correct stack contents by reading the contents of SDS and concatenating them with the return addresses stored in the architectural stack before generating the report.

**Process Management.** During process creation, a copy of the `text` and `stack` sections are created for the child process from the parent process. All non-shared memory pages are copied from the parent process to the child process using a copy-on-write mechanism. During this process, the contents of the SDS are copied as well to the child process by the OS, however in this case there is no out-of-synchronization problem. The forked process can continue to work normally in the PAIRS environment. Similarly, for context switching, the SDS of the current process is stored in the Process Control Block. In terms of cost, the typical size of the SDS is 512-bits (64 bytes). Moving this number of bytes between the SDS and memory during context switch requires just a couple of `load` and `store` instructions, which consumes a few cycles. This overhead is negligible with respect to the overhead of the rest of the context switch (which happens infrequently; every tens of milliseconds).

**Multithreading.** Although multi-threaded programs can run normally in the PAIRS environment, they can pose some serious issues to the SDS, which is shared among all threads. Let us consider the following scenario, thread \( T_1 \) executes a `call` instruction, and the return address \( R_1 \) is pushed into the architectural stack, where as its MSB, \( b_1 \) is pushed into the SDS. Then, another thread, \( T_2 \), executes a different `call` instruction, and the return address \( R_2 \) is pushed into the architectural stack, where as its MSB, \( b_2 \) is pushed into SDS. In this case, if thread \( T_1 \) executes the return instruction before \( T_2 \), \( R_1 \) will be popped from the architectural stack and concatenated with \( b_2 \) from SDS. This may lead to an execution corruption even if it happens legitimately. One possible solution is to expand the SDS with a thread identifier on resource-constrained platforms that permit multithreading. The full development of this is left for future work.

### 10. Related Work

As a response to CRAs, especially the ROP variant, many hardware mitigations have been proposed over the last decade. In this section, we discuss the pros and cons of some representative mitigations compared to PAIRS (summarized in Table 4).

**CRA Mitigations for Resource Constrained Devices.** Nyman et al. [6] introduced CaRE, an interrupt-aware control-flow integrity (CFI) scheme for low-end microcontrollers that leverages TrustZone-M security extensions. CaRE instruments binaries in a manner which removes all function calls and

---

⁸https://www.exploit-db.com/exploits/40142  
⁹https://www.exploit-db.com/exploits/46997
indirect branches and replaces them with dispatch instructions that trap control flow to a branch monitor. Although the branch monitor eliminates control-flow attacks that utilize such branches, the performance overheads introduced range between 13% and 513%, which is huge compared to PAIRS. In the case of indirect calls, Care matches the branch target against a record of valid subroutine entry points. Unlike PAIRS-PtEncLite, this coarse-grained approach does not protect against whole function reuse attacks.

**Hardware-based CRA Mitigations.** Intel architectures offer a hardware-based CFI technology named Control-flow Enforcement Technology (CET) that is to be available in future x86 processors [62, 64]. CET adds a new ENDBRANCH instruction, which is placed at the entry of each BBL that can be invoked via an indirect branch. When an indirect forward branch occurs, the following instruction is expected to be an ENDBRANCH, otherwise an attack is assumed. CET provides only coarse-grained protection where any of the possible indirect targets are allowed at every indirect control-flow transfer. Thus, an attacker can still reuse the whole BBL and store the address of the ENDBRANCH of the desired BBL in the stack as before. Additionally, CET protects call-return instructions using a full shadow stack, that resides in virtual memory, while PAIRS uses a single-bit hardware SDS.

Another CRA mitigation is introduced by Lee et al. [63]. They added a mis-prediction validation unit (MVU) to the regular BPU for validating every control-flow transfer instance in terms of (indirect branch location, target address) sets. The BPU validates the recently encountered sets (as they are stored in BTBs), whereas the MVU only validates the mis-predicted indirect branches. The MVU checks a reference CFG in order to distinguish whether a branch mis-prediction is caused by a legitimate control-flow transfer or a compromised one. Similar to PAIRS, this technique introduces negligible performance overheads. However, preparing a precise CFG and storing it a priori represents a major challenge as the authors only use traces from previous runs to generate it.

**Live Re-randomization.** In addition to the live re-randomization techniques mentioned in Section 2.3, Gallagher et al. proposed Morpheus, an architecture that (1) dynamically displaces code and data pointers in the address space and (2) randomizes the representation of code, code pointers, and data pointers using encryption [33]. Compared to PAIRS, Morpheus protects code pointers without the need for PtrEnc. However, this protection comes with the cost of program re-compilation to insert memory tags (to differentiate code/data pointer accesses from data only ones). Additionally, these tags increases memory overhead along the entire pipeline, with a 6% increase in the physical memory for a 32-bit system. On the other hand, PAIRS modifications are transparent to the pipeline and have no additional physical memory overheads (only the fixed cost of the SDS). Furthermore, PAIRS provides more fine grained randomization compared to Morpheus’s randomization interval (every 50ms). This period is sufficient to execute thousands of BBLs on a device running at 1GHz, where as in PAIRS, every BBL can be randomly executed in a different domain (i.e., nano-seconds granularity). Finally, Morpheus flushes the pipeline upon randomization, which may have severe consequences on the performance of an out-of-order processor, while PAIRS does not have such a side-effect.

### 11. Conclusion

For the foreseeable future, code-reuse attacks will continue to plague systems security. The increased proliferation of resource constrained systems that cannot deal with the performance overheads of server-grade defenses calls for more efficient mitigations. In this paper, we proposed a novel hardware design, PAIRS, which can be used to mitigate CRAs with no performance overheads or binary modifications. We showed that PAIRS only requires minimal extensions to the processor front-end. Furthermore, we discussed three PAIRS extensions that boost its security guarantees against state-of-the-art CRAs. Experimental results showed that PAIRS incurs negligible performance degradation compared to software-based EPR and hardware-based CCFI. Thus, PAIRS provides a cheaply deployable hardware technique that strengthens control flow protection uniformly across embedded and server ecosystems.

### References

[1] Laszlo Szekeres, Mathias Payer, Tao Wei, and Dawn Song. SoK: Eternal war in memory. In *Proceedings of the 2013 IEEE Symposium on Security and Privacy*, S&P ’13, pages 48–62, San Francisco, CA, USA, 2013.
[39] Chao Zhang, Scott A. Carr, Tongxin Li, Yu Ding, Chengyu Song, Mathias Payer, and Dawn Song. Vtrust: Regaining trust on virtual calls. In Proceedings of the 2016 Network and Distributed System Security Symposium, NDSS ‘16, San Diego, CA, USA, February 2016.

[40] Dimitar Bounov, Rami Gokhan Kici, and Sorin Lerner. Protecting C++ dynamic dispatch through VTable interleaving. In Proceedings of the 2016 Network and Distributed System Security Symposium, NDSS ‘16, San Diego, CA, USA, February 2016.

[41] Nathan Burow, Derrick McKee, Scott A. Carr, and Mathias Payer. CFIXX: Object type integrity for C++ virtual dispatch. In Proceedings of the 2018 Network and Distributed System Security Symposium, NDSS ’18, San Diego, CA, USA, February 2018.

[42] Lee-Sup Kim and Robert W. Dutton. Metastability of CMOS latch/flip-flop. IEEE Journal of Solid-State Circuits, 25(4):942–951, Aug 1990.

[43] Nathan Burow, Xingping Zhang, and Mathias Payer. SoK: Shining light on shadow stacks. In Proceedings of the 2019 IEEE Symposium on Security and Privacy, S&P ’19, May 2019.

[44] Nathan Binkert, Derek Bradford Beckmann, Gabriel Black, Steven K. Reinhardt, Ali Saidi, Arkaprava Basu, Joel Hestness, Derek R. Hower, Tushar Krishna, Somuayeh Sardashti, Rathijit Sen, Korey Sewell, Muhammad Shoaib, Nilay Vaish, Mark D. Hill, and David A. Wood. The Gem5 simulator. SIGARCH Computer Architecture News, 2011.

[45] James Bueck, Klaus-Dieter Lange, and Hao kim v. Kistowski. SPEC CPU2017: Next-generation compute benchmark. In Proceedings of the 2018 ACM/IEEE International Conference on Performance Engineering, ICPE ’18, pages 41–42, Berlin, Germany, April 2018.

[46] musl libc. http://www.musl-libc.org/. [Online; accessed 5-November-2019].

[47] Taegyu Kim, Chung Hwan Kim, Hongjun Choi, Yonghi Kwon, Brending Saltaromaggi, Xiangyu Zhang, and Dongyan Xu. RevARM: A platform-agnostic ARM binary rewriter for security applications. In Proceedings of the 33rd Annual Computer Security Applications Conference, ACSAC ’17, pages 412–424, Orlando, FL, USA, 2017.

[48] Dongsoo Ha, Wenhui Jin, and Heekuck Oh. REPICA: Rewriting position independent code of ARM. IEEE Access, 6:50488–50500, 2018.

[49] Roberto Avanzi. The QARMA block cipher family. almost MDS matrices over rings with zero divisors, nearly symmetric even-mansour constructions with non-involutory central rounds, and search heuristics for low-latency S-boxes. IACR Transactions on Symmetric Cryptology, 2017(1):4–44, Mar. 2017.

[50] Hans Liljestrand, Thomas Nyman, Kui Wang, Carlos Chinea Perez, Jan-Erik Ekberg, and N. Asokan. PAC it up: Towards pointer integrity using ARM pointer authentication. In Proceedings of the 28th USENIX Security Symposium, (USENIX Security 19), pages 177–194, Santa Clara, CA, USA, August 2019.

[51] Sascha Schirra. Ropper. https://github.com/sashs/Ropper. [Online; accessed 5-November-2019].

[52] John Wilander, Nick Nikiforakis, Yves Younan, Marian Kamkar, and Wouter Joosen. RIPE: Runtime intrusion prevention evaluator. In Proceedings of the 27th Annual Computer Security Applications Conference, ACSAC ’11, pages 41–50, Orlando, Florida, USA, 2011.

[53] Yueqiang Cheng, Zongwei Zhou, Miao Yu, Xuhua Ding, and Robert H. Deng. Ropecker: A generic and practical approach for defending against ROP attacks. In Proceedings of the 2014 Network and Distributed System Security Symposium, NDSS ’16, San Diego, CA, USA, February 2014.

[54] Solar Designer. Getting around non-executable stack (and fix). http://seclists.org/bugtraq/1997/Aug/63, August 1997.

[55] Nergal. The advanced return-into-lib(c) exploits: Pax case study. https://phrack.org/issues/58/4.html, 2001. [Online; accessed 5-November-2019].

[56] Shachee Mishra and Michalis Polychronakis. Shredder: Breaking exploits through API specialization. In Proceedings of the 34th Annual Computer Security Applications Conference, ACSAC ’18, pages 1–16, San Juan, PR, USA, Dec 2018.

[57] Kihong Heo, Woosuk Lee, Pardis Pushakhanloo, and Mayur Naik. Effective program deboloating via reinforcement learning. In Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, CCS ’18, pages 380–394, Toronto, Canada, 2018. ACM.

[58] Hashim Sharif, Muhammad Abubakar, Ashish Gehani, and Fareed Zafar. TRIMMER: Application specialization for code deboloating. In Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering, ASE 2018, pages 329–339, Montpellier, France, 2018. ACM.

[59] Anh Quach, Aravind Prakash, and Lok Yan. Deboloating software through piece-wise compilation and loading. In Proceedings of the 27th USENIX Conference on Security Symposium, SEC 18, pages 869–886, Baltimore, MD, USA, 2018. USENIX Association.

[60] Ioannis Agadakos, Di Jin, David Williams-King, Vasileios P. Kemerlis, and Georgios Portokalidis. Nibbler: Deboloating binary shared libraries. In Proceedings of the 35th Annual Computer Security Applications Conference, ACSAC ’19, pages 1–16, San Juan, PR, USA, Dec 2018.

[61] Thomas Nyman, Jan-Erik Ekberg, Lucas Davi, and N. Asokan. CaRE: Hardware-supported call and return enforcement for commercial microcontrollers. In Marc Dacier, Michael Bailey, Michalis Polychronakis, and Manos Antonakakis, editors, Research in Attacks, Intrusions, and Defenses, pages 259–284, Cham, 2017. Springer International Publishing.

[62] Intel. Intel control-flow enforcement technology preview. https://software.intel.com/sites/default/files/managed/4d/2a/control-flow-enforcement-technology-preview.pdf, 2017. [Online; accessed 5-November-2019].

[63] Yongsu Lee and Gyungho Lee. Detecting code reuse attacks with branch prediction. Computer, 51(4):40–47, April 2018.

[64] Vedvyas Shanbhogue, Deepak Gupta, and Ravi Sahita. Security analysis of processor instruction set architecture for enforcing control-flow integrity. In Proceedings of the 8th International Workshop on Hardware and Architectural Support for Security and Privacy, HASP ’19, Phoenix, AZ, USA, 2019.