Low-Level Software Security for Compiler Developers

Version 0-147-g19a7017

Contents

Chapter 1

Introduction

Compilers, assemblers and similar tools generate all the binary code that processors execute. It is no surprise then that these tools play a major role in security analysis and hardening of relevant binary code.

Often the only practical way to protect all binaries with a particular security hardening method is to have the compiler do it. And, with software security becoming more and more important in recent years, it is no surprise to see an ever increasing variety of security hardening features and mitigations against vulnerabilities implemented in compilers. Indeed, compared to a few decades ago, today's compiler developer is much more likely to implement security features than not.

Furthermore, with the ever-expanding range of techniques implemented, it's very hard to gain a basic understanding of all security features implemented in typical compilers.

This poses a practical problem: compiler developers must be able to work on security hardening features, yet it's hard to gain a good, basic understanding of such compiler features.

There are a lot of materials that explain individual vulnerabilities or attack vectors. There are also lots of presentations explaining specific exploits. But there seems to be a limited set of materials that give a structured overview of all vulnerabilities and exploits against which a code generator plays a role in protecting.

This book aims to provide such a structured, broad overview. It does not necessarily go into full details, instead aiming to give a thorough description of all relevant high-level aspects of attacks, vulnerabilities, mitigations, and hardening techniques. For further details, this book provides pointers to materials with more details on specific techniques.

The purpose of this book is to serve as a guide to every compiler developer that needs to learn about software security relevant to compilers. Even though the focus is on compiler developers, we expect that this book will also be useful to people working on other low-level software.

1.1 Why an open source book?

The idea for this book emerged out of a frustration of not finding a good overview on this topic. Kristof Beyls and Georgia Kouveli, both compiler engineers working on security features, wished a book like this would exist. After not finding such a book, they decided to try and write one themselves. They immediately realized that they do not have all necessary expertise themselves to complete such a daunting task. So they decided to try and create this book in an open source style, seeking contributions from many experts.

As you read this, the book remains unfinished. This book may well never be finished, as new vulnerabilities continue to be discovered regularly. Our hope is that developing the book as an open source project will allow for it to continue to evolve and improve. The open source development process of this book increases the likelihood that it remains relevant as new vulnerabilities and mitigations emerge.

Kristof and Georgia, the initial authors, are far from experts on all possible vulnerabilities. So what is the plan to get high quality content to cover all relevant topics? It is two-fold.

First, by studying specific topics, they hope to gain enough knowledge to write up a good summary for this book.

Second, they very much invite and welcome contributions. If you're interested in potentially contributing content, please go to the home location for the open source project at https://github.com/llsoftsec/llsoftsecbook.

As a reader, you can also contribute to making this book better. We highly encourage feedback, both positive and constructive criticisms. We prefer feedback to be received through https://github.com/llsoftsec/llsoftsecbook.

Add section describing the structure of the rest of the book.

Chapter 2

Memory vulnerability based attacks

2.1 A bit of background on memory vulnerabilities

Memory access errors describe memory accesses that, although permitted by a program, were not intended by the programmer. These types of errors are usually defined (Hicks 2014) by explicitly listing their types, which include:

- buffer overflow
- null pointer dereference
- use after free
- use of uninitialized memory
- illegal free

Memory vulnerabilities are an important class of vulnerabilities that arise due to these types of errors, and they most commonly occur due to programming mistakes when using languages such as C/C++ . These languages do not provide mechanisms to protect against memory access errors by default. An attacker can exploit such vulnerabilities to leak sensitive data or overwrite critical memory locations and gain control of the vulnerable program.

Memory vulnerabilities have a long history. The Morris worm in 1988 was the first widely publicized attack exploiting a buffer overflow. Later, in the mid-90s, a few famous write-ups describing buffer overflows appeared (Aleph One 1996). Stack buffer overflows were mitigated with stack canaries and non-executable stacks. The answer was more ingenious ways to bypass these mitigations: code reuse attacks, starting with attacks like return-into-libc (Solar Designer 1997). Code reuse attacks later evolved to Return-Oriented Programming (ROP) (Shacham 2007) and even more complex techniques.

To defend against code reuse attacks, the Address Space Layout Randomization (ASLR) and Control-Flow Integrity (CFI) measures were introduced. This interaction between offensive and defensive security research has been essential

to improving security, and continues to this day. Each newly deployed mitigation results in attempts, often successful, to bypass it, or in alternative, more complex exploitation techniques, and even tools to automate them.

Memory safe (Hicks 2014) languages are designed with prevention of such vulnerabilities in mind and use techniques such as bounds checking and automatic memory management. If these languages promise to eliminate memory vulnerabilities, why are we still discussing this topic?

On the one hand, C and C++ remain very popular languages, particular in the implementation of low-level software. On the other hand, programs written in memory safe languages can themselves be vulnerable to memory errors as a result of bugs in how they are implemented, e.g. a bug in their compiler. Can we fix the problem by also using memory safe languages for the compiler and runtime implementation? Even if that were as simple as it sounds, unfortunately there are types of programming errors that these languages cannot protect against. For example, a logical error in the implementation of a compiler or runtime for a memory safe language can lead to a memory access error not being detected. We will see examples of such logic errors in compiler optimizations in a later section.

Given the rich history of memory vulnerabilities and mitigations and the active developments in this area, compiler developers are likely to encounter some of these issues over the course of their careers. This chapter aims to serve as an introduction to this area. We start with a discussion of exploitation primitives, which can be useful when analyzing threat models . We then continue with a more detailed discussion of the various types of vulnerabilities, along with their mitigations, presented in a rough chronological order of their appearance, and, therefore, complexity.


Discuss threat models elsewhere in book and refer to that section here #161

2.2 Exploitation primitives

Newcomers to the area of software security may find themselves lost in many blog posts and other publications describing specific memory vulnerabilities and how to exploit them. Two very common, yet unfamiliar to a newcomer, terms that appear in such publications are *read primitive* and *write primitive*. In order to understand memory vulnerabilities and be able to design effective mitigations, it's important to understand what these terms mean, how these primitives could be obtained by an attacker, and how they can be used.

An *exploit primitive* is a mechanism that allows an attacker to perform a specific operation in the memory space of the victim program. This is done by providing specially crafted input to the victim program.

A *write primitive* gives the attacker some level of write access to the victim's memory space. The value written and the address written to may be controlled by the attacker to various degrees. The primitive, for example, may allow:

- writing a fixed value to an attacker-controlled address, or
- writing to an address consisting of a fixed base and an attacker-controlled offset limited to a specific range (e.g. a 32-bit offset) , or
- writing to an attacker-controlled base address with a fixed offset.


Consider describing in more detail why the range limitation matters#162

Primitives can be further classified according to more detailed properties. See slide 11 of (Miller, n.d.) for an example.

The most powerful version of a write primitive is an *arbitrary write* primitive, where both the address and the value are fully controlled by the attacker.

A *read primitive*, respectively, gives the attacker read access to the victim's memory space. The address of the memory location accessed will be controlled by the attacker to some degree, as for the write primitive. A particularly useful primitive is an *arbitrary read* primitive, in which the address is fully controlled by the attacker.

The effects of a write primitive are perhaps easier to understand, as it has obvious side-effects: a value is written to the victim program's memory. But how can an attacker observe the result of a read primitive?

This depends on whether the attack is interactive or non-interactive (Hu et al. 2016).

- In an *interactive attack*, the attacker gives malicious input to the victim program. The malicious input causes the victim program to perform the read the attacker instructed it to, and to output the results of that read. This output could be any kind of output, for example a network packet that the victim transmits. The attacker can observe the result of the read primitive by looking at this output, for example parsing this network packet. This process then repeats: the attacker sends more malicious input to the victim, observes the output and prepares the next input. You can see an example of this type of attack in (Beer 2020), which describes a zero-click radio proximity exploit.
- In a *non-interactive (one-shot) attack*, the attacker provides all malicious input to the victim program at once. The malicious input triggers multiple primitives one after the other, and the primitives are able to observe the effects of the preceding operations through the victim program's state. The input could be, for example, in the form of a JavaScript program (Groß 2020), or a PDF file pretending to be a GIF (Beer and Groß 2021).

The references in this section describe complicated modern exploits. Consider linking to simpler exploits, as well as some tutorial-level material. #163

How does an attacker obtain these kinds of primitives in the first place? The details vary, and in some cases it takes a combination of many techniques, some of which are out of scope for this book. But we will be describing a few of them in this chapter. For example a stack buffer overflow results in a (restricted) write primitive when the input size exceeds what the program expected.

As part of an attack, the attacker will want to execute each primitive more than once, since a single read or write operation will rarely be enough to achieve their end goal (more on this later). How can primitives be combined to perform multiple reads/writes?

In the case of an interactive attack, preparing and sending input to the victim

program and parsing the output of the victim program are usually done in an external program that drives the exploit. The attacker is free to use a programming language of their choice, as long as they can interact with the victim program in it. Let's assume, for example, an exploit program in C, communicating with the victim program over TCP. In this case, the primitives are abstracted into C functions, which prepare and send packets to the victim, and parse the victim's responses. Using the primitives is then as simple as calling these functions. These calls can be easily combined with arbitrary computations, all written in C, to form the exploit.

For this cycle of repeated input/output interactions to work, the state of the victim program must not be lost between the different iterations of providing input and observing output. In other words, the victim process must not be restarted.

It's interesting to note that while the read/write primitives consist of carefully constructed inputs to the victim program, the attacker can view these inputs as *instructions* to the victim program. The victim program effectively implements an interpreter unintentionally, and the attacker can send instructions to this interpreter. This is explored further in (Dullien 2020).

In the case of a non-interactive attack, all computation happens within the victim program. The duality of input data and code is even more obvious in this case, as the malicious input to the victim can be viewed as the exploit code. There are cases for which the input is obviously interpreted as code by the victim application as well, as in the case of a JavaScript program given as input to a JavaScript engine. In this case, the read/write primitives would be written as JavaScript functions, which when called have the unintended side-effect of accessing arbitrary memory that a JavaScript program is not supposed to have access to. The primitives can be chained together with arbitrary computations, also expressed in JavaScript.

There are, however, cases where the correspondence between data and code isn't as obvious. For example, in (Beer and Groß 2021), the malicious input consists of a PDF file, masquerading as a GIF. Due to an integer overflow bug in the PDF decoder, the malicious input leads to an unbounded buffer access, therefore to an arbitrary read/write primitive. In the case of JavaScript engine exploitation, the attacker would normally be able to use JavaScript operations and perform arbitrary computations, making exploitation more straightforward. In this case, there are no scripting capabilities officially supported. The attackers, however, take advantage of the compression format intricacies to implement a small computer architecture, in thousands of simple commands to the decoder. In this way, they effectively *introduce* scripting capabilities and are able to express their exploit as a program to this architecture.

So far, we have described read/write primitives. We have also discussed how an attacker might perform arbitrary computations:

- in an external program in the case of interactive attacks, or
- by using scripting capabilities (whether originally supported or introduced by the attacker) in non-interactive attacks.

Assuming an attacker has gained these capabilities, how can they use them to

achieve their goals?

The ultimate goal of an attacker may vary: it may be, among other things, getting access to a system, leaking sensitive information or bringing down a service. Frequently, a first step towards these wider goals is arbitrary code execution within the victim process. We have already mentioned that the attacker will typically have arbitrary computation capabilities at this point, but arbitrary code execution also involves things like calling arbitrary library functions and performing system calls.

Some examples of how the attacker may use the obtained primitives:

- Leak information, such as pointers to specific data structures or code, or the stack pointer.
- Overwrite the stack contents, e.g. to perform a ROP attack.
- Overwrite non-control data, e.g. authorization state. Sometimes this step is sufficient to achieve the attacker's goal, bypassing the need for arbitrary code execution.

Once arbitrary code execution is achieved, the attacker may need to exploit additional vulnerabilities in order to escape a process sandbox, escalate privilege, etc. Such vulnerability chaining is common, but for the purposes of this chapter we will focus on:

- Preventing memory vulnerabilities in the first place, thus stopping the attacker from obtaining powerful read/write primitives.
- Mitigating the effects of read/write primitives, e.g. with mechanisms to maintain Control-Flow Integrity (CFI).

2.3 Stack buffer overflows

A buffer overflow occurs when a read from or write to a data buffer exceeds its boundaries. This typically results in adjacent data structures being accessed, which has the potential of leaking or compromising the integrity of this adjacent data.

When the buffer is allocated on the stack, we refer to a stack buffer overflow. In this section we focus on stack buffer overflows since, in the absence of any mitigations, they are some of the simplest buffer overflows to exploit.

The stack frame of a function includes important control information, such as the saved return address and the saved frame pointer. Overwriting these values unintentionally will typically result in a crash, but the overflowing values can be carefully chosen by an attacker to gain control of the program's execution.

Here is a simple example of a program vulnerable to a stack buffer overflow[1]:

```
#include <stdio.h>
#include <string.h>

void copy_and_print(char* src) {
```

[1] This is an oversimplified example for illustrative purposes. However, as this is a wide class of vulnerabilities, many real-world examples can be found and studied.

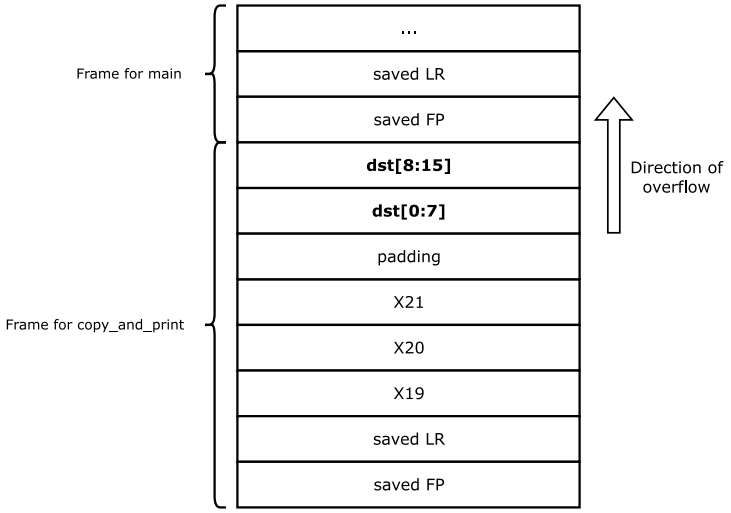
```
  char dst[16];

  for (int i = 0; i < strlen(src) + 1; ++i)
    dst[i] = src[i];
  printf("%s\n", dst);
}

int main(int argc, char* argv[]) {
  if (argc > 1) {
    copy_and_print(argv[1]);
  }
}
```

In the code above, since the length of the argument is not checked before copying it into `dst`, we have a potential for a buffer overflow.

When looking at code generated for AArch64 with GCC 11.2[2], the stack layout looks like this:



Stack frame layout for stack buffer overflow example

The exact details of the stack frame layout, including the ordering of variables and the exact control information stored, will depend on the specific compiler version you use and the architecture you compile for.

As can be seen the stack diagram, an overflowing write in function `copy_and_print` can overwrite the saved frame pointer (FP) and link register (LR) in `main`'s frame. When `copy_and_print` returns, execution continues in `main`. When `main` returns, however, execution continues from the address stored in the saved LR, which has been overwritten. Therefore, when an attacker can choose the value that overwrites the saved LR, it's possible to control where the program resumes execution after returning from `main`.

[2]The code is generated with the `-fno-stack-protector` option, to ensure GCC's stack guard feature is disabled. We also used the `-O1` optimization level.

Before non-executable stacks were mainstream, a common way to exploit these vulnerabilities would be to use the overflow to simultaneously write shellcode[3] to the stack and overwrite the return address so that it points to the shellcode. (Aleph One 1996) is a classic example of this technique.

The obvious solution to this issue is to use memory protection features of the processor in order to mark the stack (along with other data sections) as non-executable[4]. However, even when the stack is not executable, more advanced techniques can be used to exploit an overflow that overwrites the return address. These take advantage of code that already exists in the executable or in library code, and will be described in the next section.

Stack canaries are an alternative mitigation for stack buffer overflows. The general idea is to store a known value, called the stack canary, between the buffer and the control information (in the example, the saved FP and LR), and to check this value before leaving the function. Since an overflow that would overwrite the return address is going to overwrite the canary first, a corruption of the return address through a stack buffer overflow will be detected.

This technique has a few limitations: first of all, it specifically aims to protect against stack buffer overflows, and does nothing to protect against stronger primitives (e.g. arbitrary write primitives). Control-flow integrity techniques, which are described in the next section, aim to protect the integrity of stored code pointers against any modification.

Secondly, since a compiler needs to generate additional instructions for ensuring the canary's integrity, heuristics are usually employed to determine which functions are considered vulnerable. The additional instructions are then generated only for the functions that are considered vulnerable. Since heuristics aren't always perfect, this poses another potential limitation of the technique. To address this, compilers can introduce various levels of heuristics, ranging from applying the mitigations only to a small proportion of functions, to applying it universally. See, for example, the `-fstack-protector`, `-fstack-protector-strong` and `-fstack-protector-all` options offered by both GCC and Clang.

Another limitation is the possibility of leaks of the canary value. The canary value is often randomized at program start but remains the same during the program's execution. An attacker who manages to obtain the canary value at some point might, therefore, be able to reuse the leaked canary value and corrupt control information while avoiding detection. Choosing a canary value that includes a null byte (the C-style string terminator) might help in limiting the damage of overflows coming from string manipulation functions, even when the value is leaked.

Many buffer overflow vulnerabilities result from the use of unsafe library functions, such as `gets`, or from the unsafe use of library functions such as `strcpy`. There is extensive literature on writing secure C/C++ code, for example (Seacord 2013) and (Dowd, McDonald, and Schuh 2006). A different approach to limiting the effects of overflows is library function hardening, which aims to detect buffer

[3] A shellcode is a short instruction sequence that performs an action such as starting a shell on the victim machine.

[4] Note that the use of nested functions in GCC requires trampolines which reside on an executable stack. The use of nested functions, therefore, poses a security risk.

overflows and terminate the program gracefully. This involves the introduction of feature macros like `_FORTIFY_SOURCE` (Sharma 2014).

Finally, it's important to mention that not all buffer overflows aim to overwrite a saved return address. There are many cases where a buffer overflow can overwrite other data adjacent to the buffer, for example an adjacent variable that determines whether authorization was successful, or a function pointer that, when modified, can modify the program's control flow according to the attacker's wishes.

Some of these vulnerabilities can be mitigated with the measures described in this section, but often more general measures to ensure memory safety or Control-Flow Integrity are necessary. For example, in addition to the hardening of specific library functions, compilers can also implement automatic bounds checking for arrays where the array bound can be statically determined (`-fsanitize=bounds`), as well as various other "sanitizers". We will describe these measures in following sections.

2.4 Code reuse attacks

In the early days of memory vulnerability exploitation, attackers could simply place shellcode of their choice in executable memory and jump to it. As non-executable stack and heap became mainstream, attackers started to reuse code already present in an application's binary and linked libraries instead. A variety of different techniques to this effect came to light.

The simplest of these techniques is return-to-libc (Solar Designer 1997). Instead of returning to shellcode that the attacker has injected, the return address is modified to return into a library function, such as `system` or `exec`. This technique is simpler to use when arguments are also passed on the stack and can therefore be controlled with the same stack buffer overflow that is used to modify the address.

2.4.1 Return-oriented programming

Return-to-libc attacks restrict an attacker to whole library functions. While this can lead to powerful attacks, it has also been demonstrated that it is possible to achieve arbitrary computation by combining a number of short instruction sequences ending in indirect control transfer instructions, known as **gadgets**. The indirect control transfer instructions make it easy for an attacker to execute gadgets one after another, by controlling the memory or register that provides each control transfer instruction's target.

In return-oriented programming (ROP) (Shacham 2007), each gadget performs a simple operation, for example setting a register, then pops a return address from the stack and returns to it. The attacker constructs a fake call stack (often called a ROP chain) which ensures a number of gadgets are executed one after another, in order to perform a more complex operation.

This will hopefully become more clear with an example: a ROP chain for AArch64 Linux that starts a shell, by calling `execve` with `"/bin/sh"` as an argument.

The prototype of the `execve` library function, which wraps the exec system call, is:

```
int execve(const char *pathname, char *const argv[],
           char *const envp[]);
```

For AArch64, `pathname` will be passed in the `x0` register, `argv` will be passed in `x1`, and `envp` in `x2`. For starting a shell, it is sufficient to:

- Make `x0` contain a pointer to `"/bin/sh"`.
- Make `x1` contain a pointer to an array of pointers with two elements:
  - The first element is a pointer to `"/bin/sh"`.
  - The second element is zero (`NULL`).
- Make `x2` contain zero (`NULL`).

This can be achieved by chaining gadgets to set the registers `x0`, `x1`, `x2`, and then returning to `execve` in the C library.

Let's assume we have the following gadgets:

1. A gadget that loads `x0` and `x1` from the stack:

```
gadget_x0_x1:
  ldp x0, x1, [sp]
  ldp x20, x19, [sp, #64]
  ldp x29, x30, [sp, #32]
  ldr x21, [sp, #48]
  add sp, sp, #0x50
  ret
```

2. A gadget that sets `x2` to zero, but also clears `x0` as a side-effect:

```
gadget_x2:
  mov x2, xzr
  mov x0, x2
  ldp x20, x19, [sp, #32]
  ldp x29, x30, [sp]
  ldr x21, [sp, #16]
  add sp, sp, #0x30
  ret
```

Explain how these gadgets could result from C/C++ code. The current versions are slightly tweaked by hand to have more manageable offsets. #164

Both gadgets also clobber several uninteresting registers, but since `gadget_x2` also clears `x0`, it becomes clear that we should use a ROP chain that:

1. Returns to `gadget_x2`, which sets `x2` to zero.
2. Returns to `gadget_x0_x1`, which sets `x0` and `x1` to the desired values.
3. Returns to `execve`.

Figure 2.1 shows this control flow.

We can achieve this by constructing the fake call stack shown in figure 2.2, where "Original frame" marks the frame in which the address of `gadget_x2` has

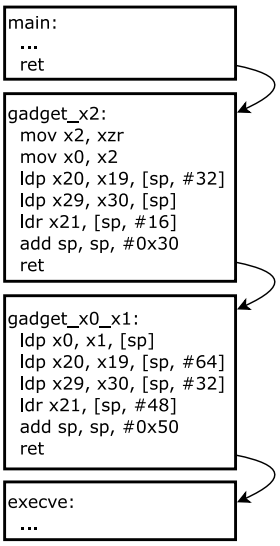


Figure 2.1: ROP example control flow

replaced a saved return address that will be loaded and returned to in the future. As an alternative, an attacker could place this fake call stack somewhere else, for example on the heap, and use a primitive that changes the stack pointer's value instead. This is known as stack pivoting.

Note that this fake call stack contains NULL bytes, even without considering the exact values of the various return addresses included. An overflow bug that is based on a C-style string operation would not allow an attacker to replace the stack contents with this fake call stack in one go, since C-style strings are null-terminated and copying the fake stack contents would stop once the first NULL byte is encountered. The ROP chain would therefore need to be adjusted so that it doesn't contain NULL bytes, for example by initially replacing the NULL bytes with a different byte and adding some more gadgets to the ROP chain that write zero to those stack locations.

A question that comes up when looking at the stack diagram is "how do we know the addresses of these gadgets"? We will talk a bit more about this in the next section.

ROP gadgets like the ones used here may be easy to identify by visual inspection of a disassembled binary, but it's common for attackers to use "gadget scanner" tools in order to discover large numbers of gadgets automatically. Such tools can also be useful to a compiler engineer working on a code reuse attack mitigation, as they can point out code sequences that should be protected and have been missed.

2.4.2 Jump-oriented programming

Jump-oriented programming (JOP) (Bletsch et al. 2011) is a variation on ROP, where gadgets can also end in indirect branch instructions instead of return instructions. The attacker chains a number of such gadgets through a dispatcher

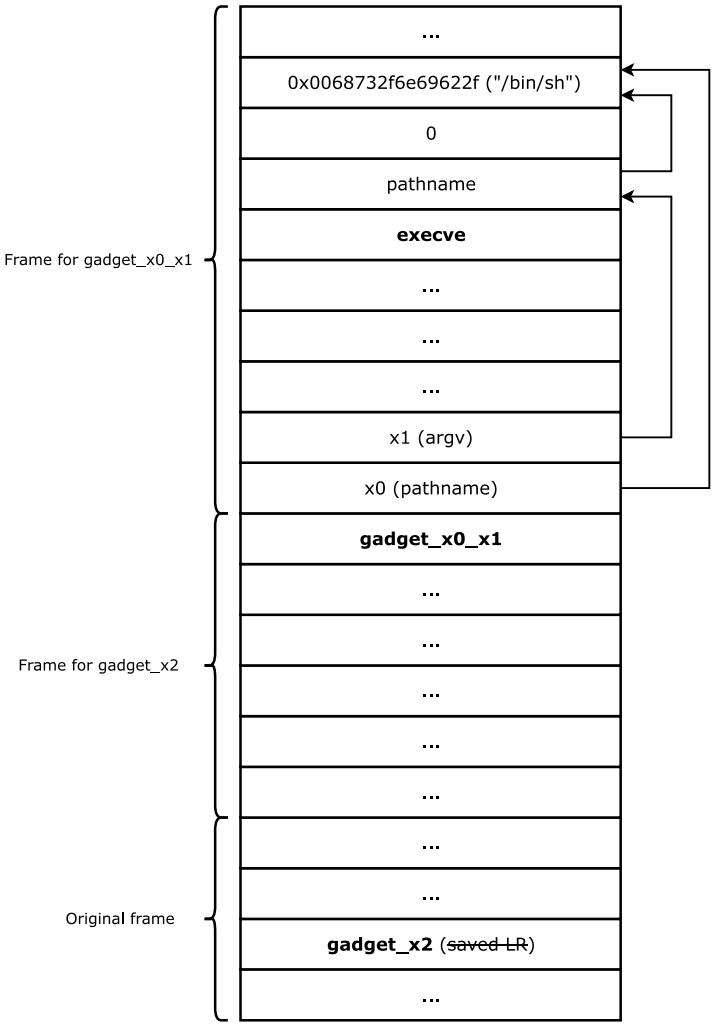


Figure 2.2: ROP example fake call stack

gadget, which loads pointers one after another from an array of pointers, and branches to each one in return. The gadgets used must be set up so that they branch or return back to the dispatcher after they're done. This is demonstrated in figure 2.3.

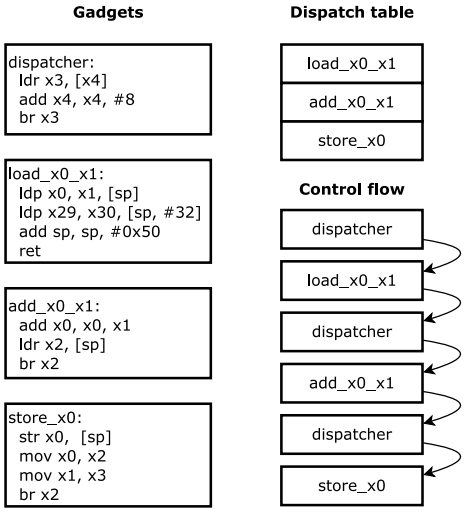


Figure 2.3: JOP example

The gadgets in the figure are made up, chosen to highlight that each gadget can end in a different type of indirect control flow transfer instruction. Consider replacing them with more realistic ones. #165

In figure 2.3, `x4` initially points to the "dispatch table", which has been modified by the attacker to contain the addresses of the three gadgets they want to execute. The dispatcher gadget loads each address in the dispatch table one by one and branches to them. The first gadget loads `x0` and `x1` from the stack, where the attacker has placed the inputs of their choice. It then loads its return address, also modified by the attacker so that it points back to the dispatcher gadget, and returns to it. The dispatcher branches to the next gadget, which adds `x0` and `x1` and leaves the result in `x0`, branching back to the dispatcher through another value loaded from the stack into `x2`. The final gadget stores the result of the addition, which remains in `x0`, to the stack, before branching to `x2`, which still points to the dispatcher gadget.

2.4.3 Counterfeit Object-oriented programming

Counterfeit Object-oriented programming (COOP) (Schuster et al. 2015) is a code reuse technique that takes advantage of C++ virtual function calls. A COOP attack takes advantage of existing virtual functions and vtables, and creates fake objects pointing to these existing vtables. The virtual functions used as gadgets in the attack are called vfgadgets. To chain vfgadgets together, the attacker uses a "main loop gadget", similar to JOP's dispatcher gadget, which is

itself a virtual function that loops over a container of pointers to C++ objects and invokes a virtual function on these objects. (Schuster et al. 2015) describes the attack in more detail. It is specifically mentioned here as an example of an attack that doesn't depend on directly replacing return addresses and code pointers, like ROP and JOP do. Such language-specific attacks are important to consider when considering mitigations against code reuse attacks, which will be the topic of the next section.

2.4.4 Sigreturn-oriented programming

One last example of a code reuse attack that is worth mentioning here is sigreturn-oriented programming (SROP) (Bosman and Bos 2014). It is a special case of ROP where the attacker creates a fake signal handler frame and calls `sigreturn`. `sigreturn` is a system call on many UNIX-type systems which is normally called upon return from a signal handler, and restores the state of the process based on the state that has been saved on the signal handler's stack by the kernel previously, on entry to the signal handler. The ability to fake a signal handler frame and call `sigreturn` gives an attacker a simple way to control the state of the program.

2.5 Mitigations against code reuse attacks

When discussing mitigations against code reuse attacks, it is important to keep in mind that there are two capabilities the attacker must have for such attacks to work:

- the ability to overwrite return addresses or function pointers
- knowledge of the target addresses to overwrite them with (e.g. libc function entry points).

When code reuse attacks were first described, programs used to contain absolute code pointers, and needed to be loaded at fixed addresses. The stack base was predictable, and libraries were loaded in predictable memory locations. This made code reuse attacks simple, as all of the addresses needed for a successful exploit were easy to discover.

2.5.1 ASLR

Address space layout randomization (ASLR) makes this more difficult by randomizing the positions of the memory areas containing the executable, the loaded libraries, the stack and the heap. ASLR requires code to be position-independent. Given enough entropy, the chance that an attacker would successfully guess one or more addresses in order to mount a successful attack will be greatly reduced.

Does this mean that code reuse attacks have been made redundant by ASLR? Unfortunately, this is not the case. There are various ways in which an attacker can discover the memory layout of the victim program. This is often referred to as an "info leak" (Serna 2012).

Since we can not exclude code reuse attacks solely by making addresses hard to guess, we need to also consider mitigations that prevent attackers from

overwriting return addresses and other code pointers. Some of the mitigations described earlier, like stack canaries and library function hardening, can help in specific situations, but for the more general case where an attacker has obtained arbitrary read and write primitives, we need something more.

2.5.2 CFI

Control-flow integrity (CFI) is a family of mitigations that aim to preserve the intended control flow of a program. This is done by restricting the possible targets of indirect branches and returns. A scheme that protects indirect jumps and calls is referred to as forward-edge CFI, whereas a scheme that protects returns is said to implement backward-edge CFI. Ideally, a CFI scheme would not allow any control flow transfers that don't occur in a correct program execution, however different schemes have varying granularities. They often rely on function type checks or use static analysis (points-to analysis) to identify potential control flow transfer targets. (Burow et al. 2017) compares a number of available CFI schemes based on the precision. For forward-edge CFI schemes, for example, schemes are classified based on whether or not they perform, among others, flow-sensitive analysis, context-sensitive analysis and class-hierarchy analysis.

2.5.2.1 Clang CFI

Clang's CFI includes a variety of forward-edge control-flow integrity checks. These include checking that the target of an indirect function call is an address-taken function of the correct type and checking that a C++ virtual call happens on an object of the correct dynamic type

For example, assume we have a class `A` with a virtual function `foo` and a class `B` deriving from `A`, and that these classes are not exported to other compilation modules:

```
class A {
public:
  virtual void foo() {}
};

class B : public A {
public:
  virtual void foo() {}
};

void call_foo(A* a) {
  a->foo();
}
```

When compiling with `-fsanitize=cfi -flto -fvisibility=hidden`,[5] the code for `call_foo` would look something like this:

```
00000000004006b4 <call_foo(A*)>:
  4006b4:       a9bf7bfd        stp     x29, x30, [sp, #-16]!
  4006b8:       910003fd        mov     x29, sp
```

[5] The LTO and visibility flags are required by Clang's CFI.

```
  4006bc:       f9400008        ldr     x8, [x0]
  4006c0:       90000009        adrp    x9, 400000 <_init-0x558>
  4006c4:       91216129        add     x9, x9, #0x858
  4006c8:       cb090109        sub     x9, x8, x9
  4006cc:       d1004129        sub     x9, x9, #0x10
  4006d0:       93c91529        ror     x9, x9, #5
  4006d4:       f100093f        cmp     x9, #0x2
  4006d8:       540000a2        b.cs    4006ec <call_foo(A*)+0x38>
  4006dc:       f9400108        ldr     x8, [x8]
  4006e0:       d63f0100        blr     x8
  4006e4:       a8c17bfd        ldp     x29, x30, [sp], #16
  4006e8:       d65f03c0        ret
  4006ec:       d4200020        brk     #0x1
```

This code looks complicated, but what it does is check that the virtual table pointer (vptr) of the argument points to the vtable of `A` or of `B`, which are stored consecutively and are the only allowed possibilities. The checks generated for different types of control-flow transfers are similar.

Another implementation of forward-edge CFI is Windows Control Flow Guard, which only allows indirect calls to functions that are marked as valid indirect control flow targets.

2.5.2.2 Clang Shadow Stack

Clang also implements a backward-edge CFI scheme known as Shadow Stack. In Clang's implementation, a separate stack is used for return addresses, which means that stack-based buffer overflows cannot be used to overwrite return addresses. The address of the shadow stack is randomized and kept in a dedicated register, with care taken so that it is never leaked, which means that an arbitrary write primitive cannot be used against the shadow stack unless its location is discovered through some other means.

As an example, when compiling with `-fsanitize=shadow-call-stack -ffixed-x18`,[6] the code generated for the `main` function from the earlier stack buffer overflow example will look something like:

```
main:
    cmp w0, #2
    b.lt    .LBB1_2
    str x30, [x18], #8
    stp x29, x30, [sp, #-16]!
    mov x29, sp
    ldr x0, [x1, #8]
    bl  copy_and_print
    ldp x29, x30, [sp], #16
    ldr x30, [x18, #-8]!
.LBB1_2:
    mov w0, wzr
    ret
```

[6] The `-ffixed-x18` flag results in treating the `x18` register as reserved, and is required by `-fsanitize=shadow-call-stack` on some platforms.

You can see that the shadow stack address is kept in `x18`. The return address is also saved on the "normal" stack for compatibility with unwinders, but it's not actually used for the function return.

2.5.2.3 Pointer Authentication

In addition to software implementations, there are a number of hardware-based CFI implementations. A hardware-based implementation has the potential to offer improved protection and performance compared to an equivalent software-only CFI scheme.

One such example is Pointer Authentication (Rutland 2017), an Armv8.3 feature, supported only in AArch64 state, that can be used to mitigate code reuse attacks. Pointer Authentication introduces instructions that generate a pointer *signature*, called a Pointer Authentication Code (PAC), based on a key and a modifier. It also introduces matching instructions to authenticate this signature. Incorrect authentication leads to an unusable pointer, that will cause a fault when used.[7] The key is not directly accessible by user space software.

Pointers are stored as 64-bit values, but they don't need all of these bits to describe the available address space, so a number of bits in the top of each pointer are unused. The unused bits must be all ones or all zeros, so we refer to them as extension bits. Pointer Authentication Codes are stored in those unused extension bits of a pointer. The exact number of PAC bits depends on the number of unused pointer bits, which varies based on the configuration of the virtual address space size.[8]

Clang and GCC both use Pointer Authentication for return address signing, when compiling with the `-mbranch-protection=pac-ret` flag. When compiling with Clang using this flag, the `main` function from the earlier stack buffer overflow example looks like:

```
main:
    cmp w0, #2
    b.lt    .LBB1_2
    paciasp
    stp x29, x30, [sp, #-16]!
    ldr x0, [x1, #8]
    mov x29, sp
    bl  copy_and_print
    ldp x29, x30, [sp], #16
    autiasp
.LBB1_2:
    mov w0, wzr
    ret
```

Notice the `paciasp` and `autiasp` instructions: `paciasp` computes a PAC for the return address in the link register (`x30`), based on the current value of the stack pointer (`sp`) and a key. This PAC is inserted in the extension bits of

[7] With the FPAC extension, a fault is raised at incorrect authentication.

[8] If the Top-Byte-Ignore (TBI) feature is enabled, the top byte of pointers is ignored when performing memory accesses. This restricts the number of available PAC bits.

the pointer. We then store this signed version of the link register on the stack. Before returning, we load the signed return address from the stack, we execute `autiasp`, which verifies the PAC stored in the return address, again based on the value of the key and the value of the stack pointer (which at this point will be the same as when we signed the return address). If the PAC is correct, which will be the case in normal execution, the extension bits of the address are restored, so that the address can be used in the `ret` instruction. However, if the stored return address has been overwritten with an address with an incorrect PAC, the upper bits will be corrupted so that subsequent uses of the address (such as in the `ret` instruction) will result in a fault.

By making sure we don't store any return addresses without a PAC, we can significantly reduce the effectiveness of ROP attacks: since the secret key is not retrievable by an attacker, an attacker cannot calculate the correct PAC for a given address and modifier, and is restricted to guessing it. The probability of success when guessing a PAC depends on the exact number of PAC bits available in a given system configuration. However, authenticated pointers are vulnerable to pointer substitution attacks, where a pointer that has been signed with a given modifier is replaced with a different pointer that has also been signed with the same modifier.

Another backward-edge CFI scheme that uses Pointer Authentication instructions is PACStack (Liljestrand et al. 2021), which chains together PACs in order to include the full context (all of the previous return addresses in the call stack) when signing a return address.

Pointer Authentication can also be used more widely, for example to implement a forward-edge CFI scheme, as is done in the arm64e ABI (McCall and Bougacha 2019). The Pointer Authentication instructions, however, are generic enough to also be useful in implementing more general memory safety measures, beyond CFI.

Add more references to relevant research #166

Mention more Pointer Authentication uses in later section, and add link here #167

2.5.2.4 BTI

Branch Target Identification (BTI) , introduced in Armv8.5, offers coarse-grained forward-edge protection. With BTI, the locations that are targets of indirect branches have to be marked with a new instruction, `BTI`. There are four different types of BTI instructions that permit different types of indirect branches (indirect jump, indirect call, both, or none). An indirect branch to a non-BTI instruction or the wrong type of BTI instruction will raise a Branch Target Exception.

Both Clang and GCC support generating BTI instructions, with the `-mbranch-protection=bti` flag, or, to enable both BTI and return address signing with Pointer Authentication, `-mbranch-protection=standard`.

Two aspects of BTI can simplify its deployment: individual pages can be marked as guarded or unguarded, with BTI checks as described above only applying to indirect branches targeting guarded pages. In addition to this, the BTI instruction has been assigned to the hint space, therefore it will be executed as a no-op in cores that do not support BTI, aiding its adoption.

2.5.2.5 CFI implementation pitfalls

When implementing CFI measures like the ones described here, it is important to be aware of known weaknesses that affect similar schemes. (Conti et al. 2015) describes how CFI implementations can suffer when certain registers are spilled on the stack, where they could be controlled by an attacker. For example, if a register that contains a function pointer that has just been validated gets spilled, the check can effectively be bypassed by overwriting the spilled pointer.

Having discussed various mitigations against code reuse attacks, it's time to turn our attention to a different type of attacks, which do not try to overwrite code pointers: attacks against non-control data, which will be the topic of the next section.

2.6 Non-control data attacks

In the previous sections, we have focused on subverting control flow by overwriting control data, which are used to change the value of the program counter, such as return addresses and function pointers. Since these types of attacks are prominent, many mitigations have been designed with the goal of maintaining control-flow integrity. Non-control data attacks, also known as data-only attacks, can completely bypass these mitigations, since the data they modify is not the control data that these mitigations protect.

Non-control data attacks can range from very simple attacks targeting a single piece of data to very elaborate attacks with very high expressiveness (Beer and Groß 2021). A very simple example may look something like this:

```
// Returns zero for failure, non-zero for success.
int authenticate() {
  int authenticated = 0;
  char passphrase[10];
  if (fgets(passphrase, 20, stdin)) {     // buffer overflow
    if (!strcmp(passphrase, "secret\n")) {
      authenticated = 1;
    }
  }
  return authenticated;
}
```

The example shows a simplified[9] function that reads a passphrase from a user, compares it with a known value and sets an integer stack variable to indicate whether "authentication" was successful or not. The function contains a very obvious buffer overflow, as the string length limit passed to `fgets` does not match the buffer size.

Figure 2.4 shows the stack frame layout for this function when the code is compiled for AArch64 with Clang 10.0[10]. As the figure shows, an overflow of

[9] This is obviously not a realistic example of how authentication should be done, but simply serves to illustrate how a buffer overflow into a non-control variable can have serious security consequences.

[10] The stack frame layout may be significantly different for other architectures and compilers.

`passphrase` will overwrite `authenticated`, setting it to a non-zero value, even though the passphrase was incorrect. The `authenticate` function will then return a non-zero value, incorrectly indicating authentication success.

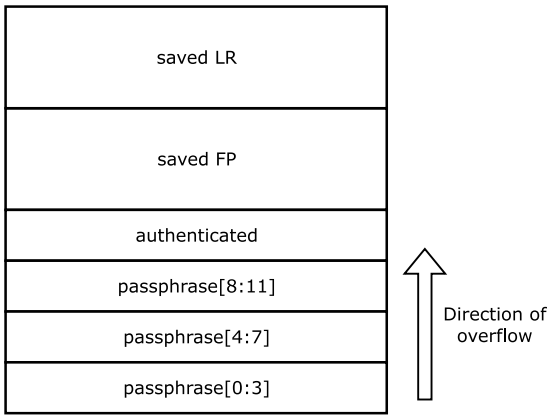


Figure 2.4: Stack frame for `authenticate`

For many more simple examples of data-only attacks that can occur in real applications, see (Chen et al. 2005). Although this makes it clear that data-only attacks are a real issue, it leaves open a very important question: what are the limits of such attacks? It is tempting to assume that data-only attacks are somehow inherently limited, however it has been demonstrated in (Hu et al. 2016) that they can, in fact, be very expressive. (Hu et al. 2016) describes Data-Oriented Programming (DOP), a general method for building data-only attacks against a vulnerable program, starting from a known memory error in the program[11].

The authors of (Hu et al. 2016) describe a small language called MINDOP, with a virtual instruction set and virtual registers. The virtual registers of MINDOP correspond to memory locations. The MINDOP instructions correspond to operations on these virtual registers, for example loading a value into a virtual register, storing a value from a virtual register, arithmetic operations and even conditional and unconditional jumps. The authors show how to identify gadgets in the code that implement the various MINDOP instructions and are reachable from memory errors, and how those gadgets can be stitched together with the help of dispatcher gadgets, the role of which is specifically to chain gadgets together.

Stitching gadgets together is simpler for interactive attacks, where the attacker can keep providing malicious input to trigger the initial memory error and a certain chain of gadgets, as many times as needed. For non-interactive attacks,

[11] The authors describe how DOP gadgets can be chained to simulate a Turing machine, making DOP attacks Turing-complete (it's not possible to simulate the infinite tape of a Turing machine on any actual hardware, of course). Turing-completeness is not, however, a particularly useful measure of exploitability, as explained in (Flake 2018). Many applications offer their users the ability to perform arbitrary computation, for example JavaScript engines, and those capabilities can be useful to an attacker, but performing a computation without affecting normal program behavior does not constitute "exploitation".

the MINDOP jump operations are required as well, used in conjunction with a memory location that provides a virtual program counter.

The process of creating a DOP attack is not so simple and not fully automated. Related literature (Ispoglou et al. 2018) focuses on automating data-only attacks.

When reading write-ups on recent security issues, instead of terminology related to data-oriented gadgets, you are more likely to encounter the term "primitive", which has been described in an earlier section. These concepts are related: an arbitrary read primitive, for example, can be produced by chaining a (possibly large) number of DOP gadgets. Talking about primitives offers a nicer level of abstraction, as it tends to be simpler to reason in terms of higher-level operations instead of many small pieces of code that need to be stitched together to perform the operations.

To summarize, data-only attacks are a significant concern. As most of the mitigation techniques we have seen so far are control-flow oriented, they are by design inadequate to protect against this different type of attacks. In the next section, we will look at what we can do to address them at their source: memory errors.

2.7 Preventing and detecting memory errors

We have so far discussed how languages that are not memory safe, like C and C++, are vulnerable to memory errors and therefore exploitation. In this section, we will discuss tools that are available to C/C++ programmers to help them detect vulnerabilities that can lead to memory errors.

2.7.1 Sanitizers

Sanitizers are tools that detect bugs during program execution. Sanitizers usually have two components: a compiler instrumentation part that introduces the new checks, and a runtime library part. They are often too expensive to run in production mode, as they tend to increase execution time and memory usage. They are commonly used during testing of an application, frequently in combination with fuzzers[12].

A very popular sanitizer is Address Sanitizer (ASan). It aims to detect various memory errors. These include out-of-bounds accesses, use-after-free, double-free and invalid free[13]. There are Address Sanitizer implementations for both GCC and Clang, but we will focus on the Clang implementation here.

ASan uses shadow memory to keep track of the state of the application's memory. Each byte of shadow memory records information on 8 bytes of the application's memory. It represents how many of the 8 bytes are addressable. When none of the bytes are addressable, it encodes additional details (whether the 8 bytes are out-of-bounds stack, out-of-bounds heap, freed memory, and so on). Requiring one byte of shadow memory for every 8 bytes of application memory means that ASan needs to reserve one-eighth of the application's virtual address space

[12] Fuzzing is a powerful testing technique that relies on automatically generating large amounts of random inputs to the program under test.

[13] ASan also includes a memory leak detector .

(Serebryany et al. 2012). Shadow memory is allocated in one contiguous chunk, which keeps mapping application memory to shadow memory simple.

ASan's runtime library replaces memory allocation functions like `malloc` and `free` with its own specialized versions. `malloc` introduces redzones before and after each allocation, which are marked as unaddressable. `free` marks the entire allocation as unaddressable and places it in quarantine, so that it doesn't get reallocated for a while (in a FIFO basis). This allows for detecting use-after-free. The runtime library also handles management of the shadow memory.

ASan's code instrumentation in the compiler introduces redzones around each stack array allocation, and around globals. It then instruments loads and stores to check whether the accessed memory is addressable, based on the information stored in the shadow memory, and reports an error if unaddressable memory is accessed.

ASan doesn't produce false positives and is easy to use. It requires compiling and linking a program with the `-fsanitize=address` option. It is used in practice for testing large projects. There is a similar tool for dynamic memory error detection in the Linux kernel, KASAN.

ASan's biggest drawback is its high runtime overhead and memory usage, due to the quarantine, redzones and shadow memory. Hardware-assisted AddressSanitizer (HWASAN) works similarly to ASan, but with partial hardware assistance can result in lower memory overheads, at the cost of being less portable.

On AArch64, HWASAN uses Top-Byte Ignore (TBI). When TBI is enabled, the top byte of a pointer is ignored when performing a memory access, allowing software to use that top byte to store metadata, without affecting execution. Each allocation is aligned to 16 bytes and each 16-byte chunk of memory (called "granule") is randomly assigned an 8-bit tag. The tag is stored in shadow memory and is also placed in the top byte of the pointer to the object. Memory loads and stores are then instrumented to check that the tag stored in the pointer matches the tag stored in memory, and report an error when a mismatch happens.

Add diagram to demonstrate how HWASAN works #168

For granules shorter than 16 bytes, the value stored in shadow memory is not the actual tag, but the length of the granule. The actual tag is stored at the last byte of the granule itself. For tags in shadow memory with values between 1 and 15, HWASAN checks that the access is within the bounds of the granule and the pointer tag matches the tag stored at the last byte of the granule.

HWASAN is also easy to use, and simply requires compiling and linking an application with the `-fsanitize=hwaddress` flag.

MemTagSanitizer goes one step further and uses the Armv8.5-A Memory Tagging Extension (MTE). With MTE, the tag checking is done automatically by hardware, and an exception is raised on mismatch. MTE's granule size is 16 bits, whereas tags are 4-bit.

Consider adding a whole section on MTE and its applications #169

UndefinedBehaviorSanitizer (UBSan) detects undefined behavior during program execution, for example array out-of-bounds accesses for statically determined array bounds, null pointer dereference, signed integer overflow and various kinds of integer conversions that result in data loss. Although some of these checks are not directly related to memory errors, these kinds of errors can lead to

incorrect pointer arithmetic, incorrect allocation sizes, and other issues that lead to memory errors, so it is important to detect them and address them.

UBSan's documentation describes the full list of available checks. The majority of these checks are enabled with the `-fsanitize=undefined` flag, but there are also other useful groupings of checks, for example `-fsanitize=integer` for checks related to integer conversions and arithmetic.

There are many other sanitizers, more than can reasonably be covered in this section. For the interested reader, we list a few more:

- MemorySanitizer: detects uninitialized reads.
- ThreadSanitizer: detects data races.
- GWP-ASan: detects use-after-free and heap buffer overflows, with low overhead that makes it suitable for production environments. It performs checks only on a sample of allocations.

Describe other mechanisms for detecting memory errors, both software-based (static analysis, library and buffer hardening) and hardware-based, e.g. PAuth-based pointer integrity schemes, MTE etc #170

2.8 JIT compiler vulnerabilities

Compiler correctness is obviously very important, as miscompilation creates buggy programs even when the source code has no bugs. What might be less obvious is that these bugs can have security implications. For example, they can introduce memory safety errors in languages that are otherwise memory safe. In some cases, a bug might leave most programs unaffected and not cause security issues in practice before it is detected and fixed. This is, of course, assuming that the bug has not been intentionally injected in the compiler.

Compiler bugs are an interesting source of security issues for just-in-time (JIT) compilers[14]. JIT compilation is often used in programs that receive source code as input during program execution, for example in web browsers, for executing JavaScript code included in web pages. In this context, the input to the JIT compiler comes from arbitrary websites and is therefore untrusted. Bugs in such JIT compilers can lead to compromise of the whole program (here, the browser) if a malicious input (e.g. coming from a malicious website) deliberately triggers miscompilation in order to break memory safety of the language being implemented.

For this section, we focus on JavaScript, which is a dynamically typed, memory safe language, but the concerns we discuss also apply to other languages that are compiled dynamically.

Without statically known types, in order to optimize JavaScript code, JavaScript engines resort to type profiling (Pizlo 2020), recording the types encountered while executing code. These types are then used during optimization, which

[14] JIT compilers compile code during execution of a program, as opposed to the more traditional compilation where code is compiled before the program is executed.

speculates that the same types will be encountered in future runs of the code, and inserts checks to validate that these assumptions about types still hold. When a check fails, the optimized code is replaced by unoptimized code that can handle all types, a process known as deoptimization or on-stack replacement (OSR). Deoptimization makes sure that the state of the deoptimized function is recreated correctly for the point of execution where the type check failed.

For example, a function such as:

```
function foo(x, y) {
  return x + y;
}
```

will return a number when `x` and `y` are numbers, but a string when either is a string. An optimizing compiler can use the results of profiling to generate optimized code. For example, when both arguments are integers during profiling, it can generate code that looks like this in pseudocode:

```
foo:
  if x not integer, deoptimize
  if y not integer, deoptimize
  result = x + y
  if overflowed, deoptimize
  return result
```

You may be wondering how the type checks are implemented, and this is closely related to the representation of values in a JavaScript engine (Wingo 2011). In short, JavaScript engines use specific bit patterns to indicate whether a value should be interpreted as a pointer, or as an integer or floating-point value. For example, the V8 JavaScript engine uses the least significant bit to denote that a value is a pointer, otherwise it is a small integer (which needs to be shifted down to access its value). Pointers then point to objects that contain a hidden class member which is used for type checking.

In addition to the values for which typing information is gathered during profiling, optimizing JavaScript compilers propagate the profiled types to dependent values. For example if a value `x` is expected to be a string, and we check this assumption, then `x + 1` will also be a string (and no additional check is needed in this case). In addition to simple type propagation, they usually perform range analysis to determine as precise a range for a value as possible, which is useful for bounds check elimination.

Bounds check elimination (BCE) is a common optimization in languages that perform bounds checks on array accesses to ensure every accessed index is within the bounds of the array. BCE gets rid of bounds checks when they are proven to be redundant, e.g. when the array access uses a constant index that's known to be smaller than the length of the array. See here for details on how out-of-bounds array accesses behave in JavaScript.

Range analysis is a good example of an analysis where a JIT compiler bug can introduce a vulnerability. Incorrect range analysis results can be used by bounds check elimination to incorrectly eliminate bounds checks that should actually have been maintained in the optimized code. For example, for the following function:

```
function foo(x) {
  y = bar(x);
  var a = [0, 1, 2];
  return a[y];
}
```

If range analysis decides that the value of `y` is in the range `[0, 2]`, but in reality the value is in the range `[0, 3]`, the bounds check for the access `a[y]` can be eliminated incorrectly, assuming the access is in-bounds. (Glazunov 2021) lists a few examples of similar hypothetical vulnerabilities, along with examples of vulnerabilities of this type that affected widely-used JavaScript engines.

The type of bug described above provides an attacker with a limited read or write primitive, as a linear overflow of the array allocation occurs. The attacker can then build on this primitive to get to an arbitrary read/write primitive. As JIT compilers generate executable code at runtime, they often use memory that is writable and executable at the same time. Such memory is very useful to attackers, who can use an arbitrary write primitive to copy their payload into this code memory, and then jump to it. Writable and executable memory, therefore, makes JITs lucrative targets for attackers.

Bugs related to range analysis are just one of the common types of bugs encountered in a JavaScript engine. (Groß and Burnett 2022) lists some other common types of bugs that result in violations of temporal and spatial memory safety, as well as type safety, in JavaScript engines.

How can we defend against such vulnerabilities? There are several complementary approaches, for example:

1. Use fuzzing to discover compiler bugs. For JavaScript, a useful fuzzing tool is Fuzzilli.
2. Be more conservative when it comes to error-prone compiler optimizations such as bounds check elimination. For example, the V8 JavaScript engine has introduced hardening of bounds checks against typer bugs.[15]
3. Instead of trying to prevent compiler (and other) bugs, assume they will be present and introduce mitigations that prevent attackers from building arbitrary read/write primitives on top of the initial limited primitives that bugs provide. For example, for 64-bit architectures, V8 implements a sandbox, built on top of pointer compression . With pointer compression, pointers are represented by 32-bit indices off a base pointer instead of as full 64-bit values. By making sure that all pointers inside the sandbox (where the JavaScript heap is located) are compressed, and that compressed pointers always point inside the sandbox, a limited primitive that allows overwriting memory within the sandbox cannot be used to build an arbitrary read/write primitive by overwriting pointer values.
4. Preventing code memory from being executable and writable at the same time is also desirable. This is known as W^X . A naive implementation of W^X that simply switches memory permissions based on page tables temporarily is not enough to prevent attackers from writing to code memory (Song et al. 2015), when multiple threads are involved. A more effective solution would use a separate compilation process, which is the only process

[15] This naturally leads to attempts to bypass the hardening too (Fetiveau 2019).

that has write access to the JIT's code memory. Alternatively, some architectures provide special features that can restrict page-based memory permissions from userspace, effectively allowing permissions to be different for different threads. Such features can also be of use in implementing WˆX. For AArch64, this feature is called permission overlays.

In this section, we have discussed JIT compiler security and described JavaScript compiler bugs that lead to vulnerabilities. Although we haven't focused on the details of JavaScript exploitation, an interested reader could take a look at (saelo 2021b) and (saelo 2021a).

Chapter 3

Covert channels and side-channels

A large class of attacks make use of so-called side-channels, which are attacks making use of so-called covert channels. Side-channels and covert channels are communication channels between two entities in a system, where the entities should not be able to communicate that way.

A **covert channel** is such a channel where both entities intend to communicate through the channel. A **side-channel** is a such a channel where one end is the victim of an attack using the channel.

In other words, the difference between a covert channel and a side-channel is whether both entities intend to communicate. If one entity does not intend to communicate, but the other entity nonetheless extracts some data from the first, it is called a side-channel attack. The entity not intending to communicate is called the **victim**. The other entity is sometimes called the **spy**.

As we focus on attacks in this book, we'll mostly use the term side-channels in the rest of this chapter.

3.1 Timing side-channels

An implementation of a cryptographic algorithm can leak information about the data it processes if its run time is influenced by the value of the processed data. Attacks making use of this are called timing attacks.

The main mitigation against such attacks consists of carefully implementing the algorithm such that the execution time remains independent of the processed data. This can be done by making sure that both:

a) The control flow, i.e. the trace of instructions executed, does not change depending on the processed data. This guarantees that every time the algorithm runs, exactly the same sequence of instructions is executed, independent of the processed data.

b) The instructions used to implement the algorithm are from the subset of instructions for which the execution time is known to not depend on the data values it processes.

   For example, in the Arm architecture, the Armv8.4-A DIT extension guarantees that execution time is data-independent for a subset of the AArch64 instructions.

   By ensuring that the extension is enabled and only instructions in the subset are used, data-independent execution time is guaranteed.

At the moment, we do not know of a compiler implementation that actively helps to guarantee both (a) and (b).

Using compiler techniques to transform a function such that it respects property (a) is an active research area. (Wu et al. 2018) provides a method to convert a program such that it respects property (a), albeit by potentially introducing unsafe memory accesses. (Soares and Pereira 2021) improves on that result by not introducing unsafe memory accesses, albeit by potentially needing to change the interface of the transformed function.


Also discuss the techniques implemented in the Constatine compiler #172

A great reference giving practical advice on how to achieve (a), (b) and more security hardening properties specific for cryptographic kernels is found in (Pornin 2018).

As discussed in (Pornin 2018), when implementing cryptographic algorithms, you also need to keep cache side-channel attacks in mind, which are discussed in the section on cache side-channel attacks.

3.2 Cache side-channels

Caches are used in almost every computing system. They are small and much faster memories than the main memory. They aim to automatically keep frequently used data accessed by programs, so that average memory access time improves. Various techniques exist where a covert communication can happen between processes that share a cache, without the processes having rights to read or write to the same memory locations. To understand how these techniques work, one needs to understand typical organization and operation of a cache.

3.2.1 Typical CPU cache architecture

There is a wide variety in CPU cache micro-architecture details, but the main characteristics that are important to set up a covert channel tend to be similar across most popular implementations.

Caches are small and much faster memories than the main memory that aim to keep a copy of the data at the most frequently accessed main memory addresses. The set of addresses that are used most frequently changes quickly over time as a program executes. Therefore, the addresses that are present in CPU caches also evolve quickly over time. The content of the cache may change with every executed read or write instruction.

On every read and write instruction, the cache micro-architecture looks up if the data for the requested address happens to be present in the cache. If it is, the CPU can continue executing quickly; if not, dependent operations will have to wait until the data returns from the much slower main memory. A typical access time is 3 to 5 CPU cycles for the fastest cache on a CPU versus hundreds of cycles for a main memory access. When data is present in the cache for a read or write, it is said to be a cache hit. Otherwise, it's called a cache miss.

Most systems have multiple levels of cache, each with a different trade-off between cache size and access time. Some typical characteristics might be:

- L1 (level 1) cache, 32kB in size, with an access time of 4 cycles.
- L2 cache, 256Kb in size, with an access time of 10 cycles.
- L3 cache, 16MB in size, with an access time of 40 cycles.
- Main memory, gigabytes in size, with an access time of more than 100 cycles.

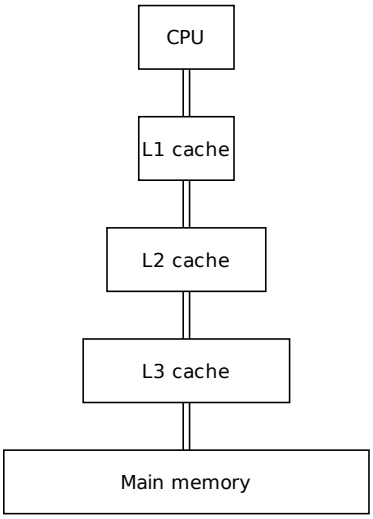


Illustration of cache levels in a typical system

If data is not already present in a cache layer, it is typically stored there after it has been fetched from a slower cache level or main memory. This is often a good decision to make as there's a high likelihood the same address will be accessed by the program soon after. This high likelihood is known as the principle of locality.

Data is stored and transferred between cache levels in blocks of aligned memory. Such a block is called a cache block or cache line. Typical sizes are 32, 64 or 128 bytes per cache line.

When data that wasn't previously in the cache needs to be stored in the cache, most of the time, room has to be made for it by removing, or evicting, some other address/data from it. How that choice gets made is decided by the cache replacement policy. Popular replacement algorithms are Least Recently Used (LRU), Random and pseudo-LRU. As the names suggest, LRU evicts the cache line that is least recently used; random picks a random cache line; and

pseudo-LRU approximates choosing the least recently used line.

If a cache line can be stored in all locations available in the cache, the cache is fully-associative. Most caches are however not fully-associative, as it's too costly to implement. Instead, most caches are set-associative. In an N-way set-associative cache, a specific line can only be stored in one of N cache locations. For example, if a line can potentially be stored in one of 2 locations, the cache is said to be 2-way set-associative. If it can be stored in one of 4 locations, it's called 4-way set-associative, and so on. When an address can only be stored in one location in the cache, it is said to be direct-mapped, rather than 1-way set-associative. Typical organizations are direct-mapped, 2-way, 4-way, 8-way, 16-way or 32-way set-associative.

The set of cache locations that a particular cache line can be stored at is called a cache set.

3.2.1.1 Indexing in a set-associative cache

For some cache covert channels, it is essential to know exactly how a memory address maps to a specific cache set.

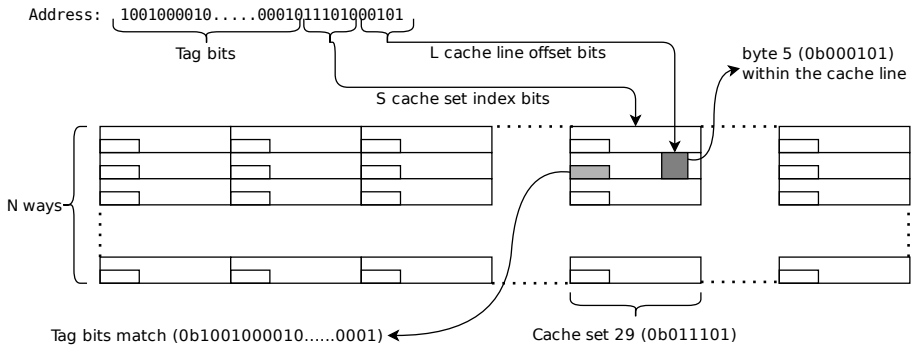


Figure 3.1: Illustration of indexing into a set-associative cache. In this example: $L = 6$ bits, hence the cache line size is $2^6 = 64$ bytes. $S = 5$ bits, so there are $2^5 = 32$ cache sets. N can be independent of address bits used to index the cache. If we assume $N = 12$ for a 12-way set-associative cache, the total cache size is $N * 2^L * 2^S = 12 * 64 * 32 = 24$KB.

Specific bits in the memory address are used for different cache indexing purposes, as illustrated in fig. 3.1. The least-significant L bits, where 2^L is the cache line size, are used to compute an address's offset within a cache line. The next S bits, where 2^S is the number of cache sets, are used to determine which cache set an address maps to. The remaining top bits are "tag bits". They are stored alongside a line in the cache so later operations can detect which specific memory address is replicated in that cache line.

For direct-mapped and fully-associative caches, the mapping of an address to cache locations also works as described above. In fully-associative caches the number of cache sets is 1, so S=0.

Also explain cache coherency ? #173

Also say something about TLBs and prefetching? #174

3.2.2 Operation of cache side-channels

Cache side-channels typically work by the spy determining whether a memory access was a cache hit or a cache miss. From that information, in specific situations, it may be able to deduce bits of data that only the victim has access to.

Let's illustrate this with describing a few well-known cache side-channels:

3.2.2.1 Flush+Reload

In a so-called Flush+Reload attack(Yarom and Falkner 2014), the spy process shares memory with the victim process. The attack works in 3 steps:

1. The Flush step: The spy flushes a specific address from the cache.
2. The spy waits for some time to give the victim time to potentially access that address, resulting in bringing it back into the cache.
3. The Reload step: The spy accesses the address and measures the access time. A short access time means the address is in the cache; a long access time means it's not in the cache. In other words, a short access time means that in step 2 the victim accessed the address; a long access time means it did not access the address.

Should there be a more elaborate example with code that demonstrates in more detail how a flush+reload attack works? #175

Knowing if a victim accessed a specific address can leak sensitive information. Such as when accessing a specific array element depends on whether a specific bit is set in secret data. For example, (Yarom and Falkner 2014) demonstrates that a Flush+Reload attack can be used to leak GnuPG private keys.

3.2.2.2 Prime+Probe

In a Prime+Probe attack, there is no need for memory to be shared between victim and spy. The attack works in 3 steps:

1. The Prime step: The spy fills one or more cache sets with its data, for example, by accessing data that maps to those cache sets.
2. The spy waits for some time to let the victim potentially access data that maps to those same cache sets.
3. The Probe step: The spy accesses that same data as in the prime step. Measuring the time it takes to load the data, it can derive how many cache

lines the victim evicted from each cache set in step 2, and from that derive information about addresses that the victim accessed.

(Osvik, Shamir, and Tromer 2005) which first documented this technique in 2005 demonstrates extracting AES keys in just a few milliseconds using Prime+Probe.

3.2.2.3 General schema for cache covert channels

An attentive reader may have noticed that the concrete named attacks above follow a similar 3-step pattern. Indeed, (Weber et al. 2021) describes this general pattern and uses it to automatically discover more side-channels that follow this 3-step pattern. They describe the general pattern as being:

1. An instruction sequence that resets the inner CPU state (*reset sequence*).
2. An instruction sequence that triggers a state change (*trigger sequence*).
3. An instruction sequence that leaks the inner state (*measurement sequence*).

Other cache-based side channel attacks following this general 3-step approach include: Flush+Flush(Gruss, Maurice, Wagner, et al. 2016), Flush+Prefetch(Gruss, Maurice, Fogh, et al. 2016), Evict+Reload(Percival 2005), Evict+Time(Osvik, Shamir, and Tromer 2005), Reload+Refresh(Briongos et al. 2020), Collide+Probe(Lipp et al. 2020), etc.

3.2.3 Mitigating cache side-channel attacks

As described in (Su and Zeng 2021), 3 conditions need to be met for a cache-based side-channel attack to succeed:

1. There is a mapping between a state change in the cache and sensitive information in the victim program.
2. The spy program needs to run on a CPU that shares the targeted cache level with the CPU the victim program runs on.
3. The spy program can infer a cache status change caused by the victim program through its own cache status.

Mitigations against cache side-channel attacks can be categorized according to which of the 3 conditions above they aim to prevent from happening:

3.2.3.1 Mitigations de-correlating cache state change with sensitive information in the victim program

A typical example of when a cache state change could be correlated to sensitive information is when a program uses secret information to index into an array. An attacker could derive bits of the secret information by observing which cache line was fetched.

Especially in crypto kernels, indexing into an array using a secret value is generally avoided. An alternative mitigation is to always access all array indices, independent of the secret value, e.g. as done in commit 46fbe375 to the PuTTY project, which contains this comment:

```
* Side-channel considerations: the exponent is secret, so
* actually doing a single table lookup by using a chunk of
```

```
* exponent bits as an array index would be an obvious leak of
* secret information into the cache. So instead, in each
* iteration, we read _all_ the table entries, and do a sequence
* of mp_select operations to leave just the one we wanted in the
* variable
```

3.2.3.2 Mitigations disallowing spy programs to share the cache with the victim program

If the victim and the spy do not share a common channel – in this case a cache level – then a side channel cannot be created.

One way to achieve this is to only allow one program to run at the same time, and when a context switch does happen, to clear all cache content. Obviously, this has a huge performance impact, especially in systems with multiple cores and with large caches. Therefore, a wide variety of mitigations have been proposed that aim to make attacks somewhat harder without losing too much system efficiency. (Mushtaq et al. 2020) and (Su and Zeng 2021) summarize dozens of proposals and implementations – too many to try to describe them all here.

One popular such mitigation is disabling cpu multithreading. For example, Azure suggests that users who run untrusted code should consider disabling cpu multithreading. The linux kernel's core scheduling documentation also states mutually untrusted code should not run on the same core concurrently. It implements a scheduler that takes into account which processes are mutually-trusting and only allows those to run simultaneously on the same core.

One could argue that site isolation as implemented in many web browsers is a mitigation that also falls into this category. Site isolation is described in more detail in its own section.

3.2.3.3 Mitigations disabling the spy program to infer a cache status change in the victim program through its own cache status

In some contexts, the resolution of the smallest time increment measurable by the spy program can be reduced so much that it becomes much harder to distinguish between a cache hit and a cache miss. Injecting noise and jitter into the timer also makes it harder to distinguish between a cache hit and cache miss. This is one of the mitigations in javascript engines against Spectre attacks. For more information see this v8 blog post or this Firefox documentation of the performance.now() method.

Note that this is not always a perfect mitigation - there are often surprising ways that an attacker can get a fine-grained enough timer or use statistical methods to be able to detect the difference between a cache hit or miss. One extreme example is NetSpectre (Schwarz et al. 2019) where the difference between cache hit and cache miss is measured over a network, by statistically analyzing delays on network packet responses. Furthermore, (Schwarz et al. 2017) demonstrates how to construct high-resolution timers in various indirect ways in all browsers that have removed explicit fine-grained timers.

Another possibility is to clear the cache between times when the victim runs and the spy runs. This is probably going to incur quite a bit of performance

overhead, and may also not always possible e.g. when victim and spy are running at the same time on 2 CPUs sharing a cache level.

3.3 Resource contention channels

3.4 Channels making use of aliasing in branch predictors and other predictors

Should we also discuss more "covert" channels here such as power analysis, etc? #176

3.5 Transient execution attacks

3.5.1 Transient execution

CPUs execute sequences of instructions. There often are dependencies between instructions in the sequence. That means that the outcome of one instruction influences the execution of a later instruction.

Apart from the smallest micro-controllers, all CPUs execute multiple instructions in parallel. Sometimes even multiple hundreds of them at the same time, all in various stages of execution. In other words, instructions start executing while potentially hundreds of previous instructions haven't produced their results yet. How can a CPU achieve this when the output of a previous instruction, which might not have fully executed yet, and hence whose output may not yet be ready, may affect the execution of that later instruction?

In other words, there may be a dependency between an instruction that has not finished yet and a later instruction that the CPU also already started executing. There are various kinds of dependencies. One kind is *control dependencies*, where whether the later instruction should be executed at all is dependent on the outcome of the earlier instruction. Other kinds are *true data dependencies*, *anti-dependencies* and *output dependencies*. More details about these kinds of dependencies can be found on the wikipedia page about them.

CPUs overcome parallel execution limitations imposed by dependencies by making massive numbers of *predictions*. For example, most CPUs predict whether conditional branches are taken or not, which is making a prediction on control dependencies. Another example is a CPU making a prediction on whether a load accesses the same memory address as a preceding store. If they do not access the same memory locations, the load can run in parallel with the store, as there is no data dependency between them. If they do access overlapping memory locations, there is a dependency and the store should complete before the load can start executing.

Starting to execute later instructions before all of their dependencies have been resolved, based on the predictions, is called *speculation*.

Let's illustrate that with the following example The following C code

```
long abs(long a) {
  if (a>=0)
    return a;
  else
    return -a;
  }
```

can be translated to the following AArch64 assembly code:

```
        cmp     x0, #0
        b.ge    Lbb2
Lbb1:
        neg     x0, x0
Lbb2:
        ret
```

The `b.ge` instruction is a conditional branch instruction. It computes whether the next instruction should be the one immediately after, or the one pointed to by label `Lbb2`. In case it's the instruction immediately after, the branch is said to not be taken. Instead, if it's the instruction pointed to be label `Lbb2`, the branch is said to be taken. When the condition `.ge` (greater or equal) is true, the branch is taken. That condition is defined or set by the previous instruction, the `cmp x0, #0` instruction, which compares the value in register `x0` with 0. Therefore, there is a dependency between the `cmp` instruction and the `b.ge` instruction. To overcome this dependency, and be able to execute the `cmp`, `b.ge` and potentially more instructions in parallel, the CPU predicts the outcome of the branch instruction. In other words, it predicts whether the branch is taken or not. The CPU will pick up either the `neg` or the `ret` instruction to start executing next. This is called *speculation*, as the CPU *speculatively executes* either instruction `neg`, or `ret`.

Show a second example of cpu speculation that is not based on branch prediction. #177

Of course, as with all predictions, the CPU gets the prediction wrong from time to time. In that case, all changes to the system state that affect the correct execution of the program need to be undone. In the above example, if the branch should have been taken, but the CPU predicted it to not be taken, the `neg` instruction is executed incorrectly and changes the value in register `x0`. After discovering the branch was mis-predicted, the CPU would have to restore the correct, non-negated, value in register `x0`.

Any instructions that are executed under so-called *mis-speculation*, are called *transient instructions*.

The paragraph above says "*the system state that affects the correct execution of the program, needs to be undone*". There is a lot of system state that does not affect the correct execution of a program. And the changes to such system state

by transient instructions is often not undone.

For example, a transient load instruction can fetch a value into the cache that was not there before. By bringing that value in the cache, it could have evicted another value from the cache. Whether a value is present in the cache does not influence the correct execution of a program; it merely influences its execution speed. Therefore, the effect of transient execution on the content of the cache is typically not undone when detecting mis-speculation.

Sometimes, it is said that the *architectural effects* of transient instructions need to be undone, but the *micro-architectural effects* do not need to be undone.

The above explanation describes architectural effects as changes in system state that need to be undone after detecting mis-speculation. In reality, most systems will implement techniques that keep all state changes in micro-architectural buffers until it is clear that all predictions made to execute that instruction were correct. At that point the micro-architectural state is *committed* to become architectural state. In that way, mis-predictions naturally do not affect architectural state.

Transient execution attacks are a category of side-channel attacks that use the micro-architectural side-effects of transient execution as a side channel.

Write sections on specific transient execution attacks such as Spectre and Meltdown. #178

Could we find a good reference that explains micro-architectural versus architectural state in more detail? Is "Computer Architecture: A Quantitative Approach" the best reference available?

3.5.1.1 Site isolation

Write section on site isolation as a SpectreV1 mitigation #179

3.6 Physical access side-channel attacks

Chapter 4

Supply chain attacks

A software *supply chain attack* occurs when an attacker interferes with the software development or distribution processes with the intention to impact users of that software.

Supply chain attacks and their possible mitigations are not specific to compilers. However, compilers are an attractive target for attack because they are widely deployed to developers, in continuous integration systems and as JITs. Also, an infected compiler has the possibility to make a much larger impact if it can silently spread the infection to other software created with or run using it.

This chapter explores the history of supply chain attacks that involve compilers and what can be done to prevent them.

4.1 History of supply chain attacks

As far back as 1974 Karger & Schell theorized about an attack on the Multics operating system via the PL/I compiler (Paul A. and Roger R. 1974). In this attack, a trap door is inserted into the compiler, which then injects malicious code into generated object code. Furthermore, the trap door could be designed to reinsert itself into the compiler binary so that future compilers are silently infected without needing changes to their source code. This attack method was subsequently popularized by Ken Thompson in his 1984 ACM Turing Award acceptance speech *Reflections on Trusting Trust* (Thompson 1984).

If these cases seem far-fetched then consider that there have been several real examples of supply chain attacks on development tools.

Induc is a family of viruses that infects a pre-compiled library in the Delphi toolchain with malicious code (Gostev 2009). When Delphi compiles a project the malicious library is included into the resulting executable, thus enabling the virus to spread. The virus was first detected in 2009 and was circulating undetected for at least a year beforehand. Several popular applications are known to have been infected, including a chat client and a media player. Overall, in excess of a hundred thousand infected computers were detected world-wide by anti-virus solutions.

XcodeGhost is the name given to malware first detected in 2015 that infected thousands of iOS applications (Cox 2015). The source of the infection was tracked down to a trojanized version of Xcode tools. The malware exists in an extra object file within the Xcode tools and is silently linked into each application as it is built. File sharing sites were used to spread the trojanized Xcode tools to unwitting developers.

A trojanized linker was found to be involved in a supply chain attack discovered in 2017 named ShadowPad (Greenberg 2019). Some instances of the attack were perpetrated using a trojanized Visual Studio linker that silently incorporates a malicious library into applications as they are built. Related attacks named CCleaner and ShadowHammer used the same approach of a trojanized linker to infect built applications. Infected applications from these attacks were distributed to millions of users world-wide.

These cases highlight that attacks on compilers, and especially linkers and libraries, are a viable route to silently infect many other applications, and there is no doubt that there will be more such attacks in the future. Let us now explore what we can do about these.

Explain how these vulnerabilities arise and how to mitigate them. #180

Chapter 5

Physical attacks

This chapter should probably be moved under section 'Physical access side-channel attacks' higher-up #181

5.1 Overview

There are many types of physical attacks – these attack methods focus on one or multiple physical properties of systems (e.g. CPU, GPU, crypto hardware), and can either be

- Passive – just monitoring physical quantities (e.g. side channel information leakage), or
- Active – modification of physical quantities, for example, by
  - changing the operation conditions of the system so that the circuit operates outside its specifications (e.g. by changing temperature, or by applying glitches to supply voltage/clock source)
  - injecting faults to the system (e.g. altering the electrical state of the system using Electromagnetic pulse injection, or laser beam)
  - physically modifying the system/chip

In the rest of this section, we will focus on a subset of physical attacks:

- Side channel information leakage
- Physical attack using glitches

These two forms of attacks can be carried out using low cost hardware, and have been widely demonstrated by researchers on SoCs or microcontrollers developed for IoT (Internet-of-Things) applications.

5.2 Physical access side-channel attacks

If an attacker has physical access to a device, even without debug access, the attacker can collect side channel information about the program execution on a

processor. If the processor is used to handle cryptographic operations, the side channel information can be used to deduce the crypto key(s) or the data being processed. Please note that some forms of physical attacks (e.g. fault injection attacks like rowhammer and voltjockey) do not require physical access, but those attacks are not covered in this section.

5.2.1 How is information leaked?

The most common physical access side channel attack method is to capture the voltage or current consumption of the device during its operation. Every time a flip-flop toggles, the switching activity results in a small current spike. Even though there is capacitance on the power distribution connections inside the chip, the toggling of registers (composed of flip-flops) still results in variations in the power supply current, which can be observed easily. Because the connections for delivering power (at the power supply, printed circuit board, on chip packages as well as on the silicon dies) also contain resistance, the variation of electrical current in the chip's power supply also results in variations in the power supply voltage. Again, this can be observed easily if the attacker has physical access to the device. If an attacker has access to data acquisition equipment that can record the current/voltage/power patterns, he/she can record the "power signature" for different crypto operations, including the power signature using different data inputs. By applying analysis techniques like differential power analysis, the attacker can extract the information being processed. One additional form of side-channel leakage is electro-magnetic radiation. Because the processor's clock frequency is usually in the radio frequency (RF) range, the wires on the die and the tracks on the PCB become small antennas, and the ripples in the processor's voltage/current results in radio frequency signals. Although the RF power radiated can be tiny, it still means that an attacker can observe the side-channel leakage if he/she is in close proximity from the device and has the right equipment to amplify and record the RF power signature. However, the risk of such attack can be reduced by reducing the radiation energy level using:

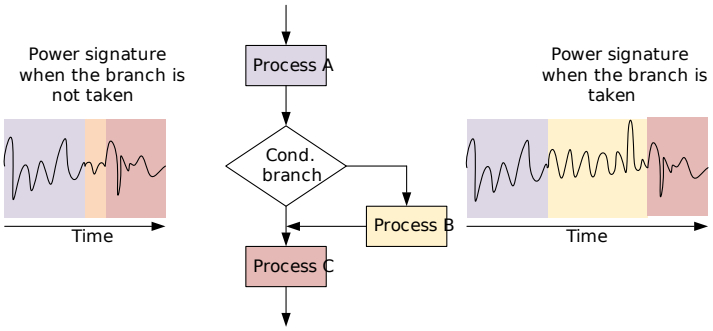
- Shielding around the device, including ground plate on the circuit board.
- Coupling capacitors on power supply tracks on the printed circuit board.

Generally, such an attack requires knowledge of radio circuit techniques and the result can be affected by other factors. For example, in normal environments there are many other source of RF noises that affects the accuracy of signal measurement. In a "noisy" environment, the RF signals from various wireless communication gadgets nearby might drown out the signals from the device being monitored.

5.2.2 Side channel leakage at instruction level

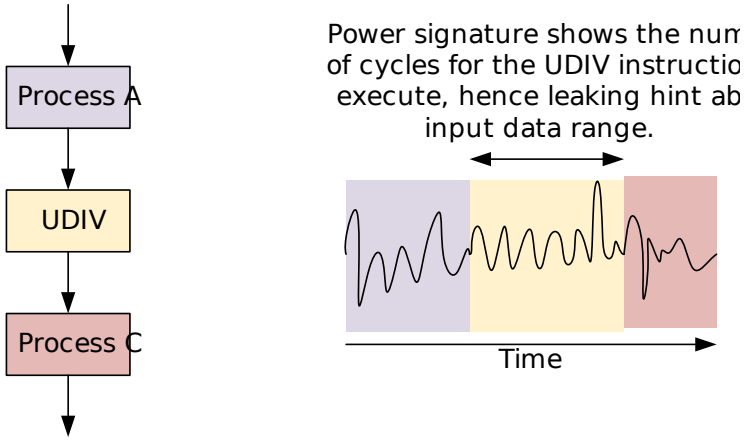
Instruction executions can result in various forms of side channel leakage:

Cycle timing resulting from conditional branch – A code sequence containing a conditional branch could result in observable side channel leakage. For example, if the power signatures of several code segments are easily recognizable (process A, B and C in the following diagram), it is possible to detect if the conditional branch was taken or not.



leakage of conditional branch

Cycle timing resulting from specific data values – The execution cycles of some instructions can be dependent on the values of input data, resulting in timing side-channel leakage. E.g., the integer divide instruction in Arm Cortex-M processors.



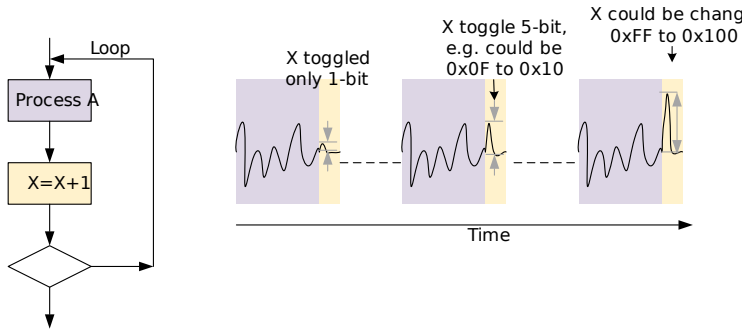
leakage of execution cycle

Power variation due to value changes – The power spikes in the power signature are often dependent on a combination of how many bits are set and how many bits have toggled in the register(s) — the so-called Hamming weight and Hamming distance, so the amplitude of the spike could be used to guess the register value in that clock cycle. The power spikes can be caused by a combination of

- Logic switching due to the operations of an instruction (e.g. power consumed by a single cycle multiplier can be much higher than the power used by a Boolean logic function), and
- Logic switching due to changes in data values in the register bank and data paths.

The switching activities are dependent on preceding and next operations. If the power signature of the codes around a specific instruction is recognizable, then

the data value being process could be guessed.



leakage of number of bit toggled

In some SoC or microcontroller implementations, the power spike effect of the operations can be much higher than the effect of data value changes in the register banks. In such case the program execution flow can be observed, and as a result, might also indirectly leak information about the data that it is processing.

5.2.3 Countermeasures

For normal embedded devices that don't have physical protection features, there is a much higher chance that power/voltage/radiation side channels can result in information leakage. However, some aspects of timing signature leakage could be reduced:

- Using data processing instruction with data independent timing for cryptographic operations. In recent Arm architectures (including Armv8-A and Armv8-M), some instructions are architecturally defined as DIT (Data Independent Timing).
- For conditional branches where the condition is dependent on secret data, use table branch instead might help reduce timing base leakage (both paths result in a branch). It is not necessary to replace all conditional branch. For example, many loop counters in crypto operation can be independent to the crypto key or input data values, so there is no need to change those loops.

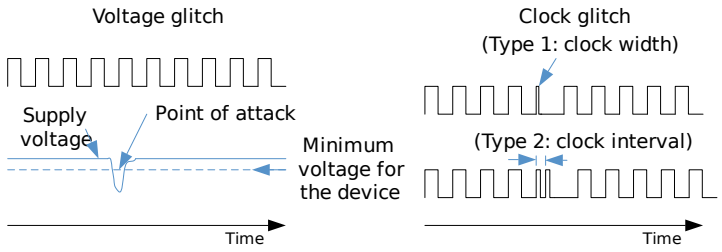
There is overlap with section timing-side-channels. How to best consolidate that? #182

There are additional software techniques to mitigate power leakage. One of the most well-known techniques is masking (e.g. Boolean, multiplicative, affine). When applying software mitigation, software developers need to check that optimizations carried out by compilers (C/C++) do not impact the mitigation, as compilers can be very smart and undo the masking in order to perform faster operations (or reducing code size).

5.3 Fault injection attacks

5.3.1 Common forms of Fault injection attacks

If an attacker has physical access to a device, they can also choose to use physical attacks to modify the behavior of the software, for example, prevent the software from setting up certain security features during the device's initialization sequence. The two most common forms of such attacks are voltage glitching and clock glitching.



common fault injection attacks

- Voltage glitch attack
  - Using a programmable power supply that can switch the voltage level rapidly, it is possible to reduce/increase the power supply voltage of a chip at specific clock cycle of the software execution. In some case, a precise voltage drop can cause a processor to "skip" an instruction, for example, the write to memory or a hardware register might not be taken. Or if a write has taken place, the actual write value used could be changed by the voltage glitch.
- Clock glitch attack
  - Using a clock switching circuit, it is possible to reduce the width of a clock pulse, or the interval between two clock pulses so that some of the hardware registers are not updated correctly at certain clock edge(s). Similar to voltage glitch, this can make the hardware seems to be skipping an instruction.

Such voltage/clock glitch attack could affect multiple parts in the processors, but sometimes the impact might not lead to any visible error in the operation, leaving the only effect that the processor skipping a memory/register write, or writing an incorrect value. Potentially, a glitch attack could result in other observable effect (e.g. register reset, bit toggle). The analysis of fault injection methods (and their physical effect) and the observable effects at the program or instruction execution level are often referred to as fault models, where one can say that a specific fault injection behaves as an instruction skip, etc. More details about the concept of fault models can be found in the paper "Fault Attacks on Secure Embedded Software: Threats, Design and Evaluation" (https://arxiv.org/pdf/2003.10513.pdf), where a good illustration of the concept is shown in figure 1 of that paper.

Make the above reference to a paper use bibtex. #159

Using glitching methods, there are several common ways of attacking a system. For example:

- Skipping an instruction during setup sequence for security features – e.g. skipping the write to the MPU (Memory Protection Unit)/Security Attribution Unit (SAU) so that the MPU/SAU is not enabled.
- Skipping an instruction after a security authentication that branch to an error handling code. As the branch is not taken, the code can continue to operate even a security authentication has failed.
- Causing an incorrect value to be written in a memory or hardware. E.g. When writing a crypto key to a crypto accelerator, forcing the key value written to be zero (caused to low voltage on bus hardware).

Example: Attack on TrustZone for Armv8-M: https://www.youtube.com/watch?v=4u6BAH8mEDw

There are other forms of physical attacks, but most of them requires significant effort or cost (e.g. cut open the chip package can carry out fault injection or readout secret data on chip).

5.3.2 Countermeasures

Ideally, system designers can use hardware (SoCs or microcontroller) that support protection against fault injection. For example, a hardware circuit can include redundancy logic (spatial and temporal). In addition, software developers can make such attack harder by adding checks after the write operations. When applying software mitigation, software developers need to check that optimizations carried out by compilers (C/C++) do not impact the mitigation.

Chapter 6

Other security topics relevant for compiler developers

Write chapter with other security topics.

Write section on securely clearing memory in C/C++ and undefined behaviour. #183

Appendix: contribution guidelines

If you'd like to start contributing to this book: please do, we're looking forward to your contributions!

The project lives on github at https://github.com/llsoftsec/llsoftsecbook. We also have a Discord server where you can have an interactive chat with us at https://discord.gg/Bm55Z9Ppgn.

We use github issues as our issue tracker and use github pull requests to accept edits, changes, additions and more.

If you'd like to contribute, but are not sure where to start, the list of open issues labeled with "good first issue" may give you inspiration of things to contribute. Please, also don't be shy to reach out to us on Discord.

We follow the Contributor Covenant Code of Conduct in this project.

For more details on how to write text for the book, please read contributing.md. If after reading that, you think some specific aspects could be explained better, please do let us know by raising an issue.

Index

Todo list

References

Aleph One. 1996. "Smashing the Stack for Fun and Profit." 1996. http://www.phrack.org/issues/49/14.html#article.

Beer, Ian. 2020. "An iOS Zero-Click Radio Proximity Exploit Odyssey." 2020. https://googleprojectzero.blogspot.com/2020/12/an-ios-zero-click-radio-proximity.html.

Beer, Ian, and Samuel Groß. 2021. "A Deep Dive into an Nso Zero-Click iMessage Exploit: Remote Code Execution." 2021. https://googleprojectzero.blogspot.com/2021/12/a-deep-dive-into-nso-zero-click.html.

Bletsch, Tyler, Xuxian Jiang, Vince W. Freeh, and Zhenkai Liang. 2011. "Jump-Oriented Programming: A New Class of Code-Reuse Attack." In *Proceedings of the 6th Acm Symposium on Information, Computer and Communications Security*, 30–40. ASIACCS '11. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1966913.1966919.

Bosman, Erik, and Herbert Bos. 2014. "Framing Signals - a Return to Portable Shellcode." In *2014 Ieee Symposium on Security and Privacy*, 243–58. https://doi.org/10.1109/SP.2014.23.

Briongos, Samira, Pedro Malagon, Jose M. Moya, and Thomas Eisenbarth. 2020. "RELOAD+REFRESH: Abusing Cache Replacement Policies to Perform Stealthy Cache Attacks." In *29th Usenix Security Symposium (Usenix Security 20)*. https://www.usenix.org/conference/usenixsecurity20/presentation/briongos.

Burow, Nathan, Scott A. Carr, Joseph Nash, Per Larsen, Michael Franz, Stefan Brunthaler, and Mathias Payer. 2017. "Control-Flow Integrity: Precision, Security, and Performance." *ACM Comput. Surv.* 50 (1). https://doi.org/10.1145/3054924.

Chen, Shuo, Jun Xu, Emre Can Sezer, Prachi Gauriar, and Ravishankar K Iyer. 2005. "Non-Control-Data Attacks Are Realistic Threats." In *USENIX Security Symposium*, 5:146.

Conti, Mauro, Stephen Crane, Lucas Davi, Michael Franz, Per Larsen, Marco Negro, Christopher Liebchen, Mohaned Qunaibit, and Ahmad-Reza Sadeghi. 2015. "Losing Control: On the Effectiveness of Control-Flow Integrity Under Stack Attacks." In *Proceedings of the 22nd Acm Sigsac Conference on Computer and Communications Security*, 952–63. CCS '15. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2810103.2813671.

Cox, Joseph. 2015. "Hack Brief: Malware Sneaks into the Chinese iOS App Store." *WIRED*. https://www.wired.com/2015/09/hack-brief-malware-sneaks-chinese-ios-app-store/.

Dowd, Mark, John McDonald, and Justin Schuh. 2006. *The Art of Software Security Assessment: Identifying and Preventing Software Vulnerabilities*. Addison-Wesley Professional.

Dullien, Thomas. 2020. "Weird Machines, Exploitability, and Provable Unexploitability." *IEEE Transactions on Emerging Topics in Computing* 8 (2): 391–403. https://doi.org/10.1109/TETC.2017.2785300.

Fetiveau, Jeremy. 2019. "Circumventing Chrome's Hardening of Typer Bugs." 2019. https://doar-e.github.io/blog/2019/05/09/circumventing-chromes-hardening-of-typer-bugs/.

Flake, Thomas Dullien/Halvar. 2018. "Turing Completeness, Weird Machines, Twitter, and Muddled Terminology." 2018. http://addxorrol.blogspot.com/2018/10/turing-completeness-weird-machines.html.

Glazunov, Sergei. 2021. "In-the-Wild Series: Chrome Infinity Bug." 2021. https://googleprojectzero.blogspot.com/2021/01/in-wild-series-chrome-infinity-bug.html.

Gostev, Alexander. 2009. "A Short History of Induc." 2009. https://securelist.com/a-short-history-of-induc/30555/.

Greenberg, Andy. 2019. "Supply Chain Hackers Snuck Malware into Videogames." *WIRED*. https://www.wired.com/story/supply-chain-hackers-videogames-asus-ccleaner/.

Groß, Samuel. 2020. "JITSploitation I: A Jit Bug." 2020. https://googleprojectzero.blogspot.com/2020/09/jitsploitation-one.html.

Groß, Samuel, and Amy Burnett. 2022. "Attacking Javascript Engines in 2022." 2022. https://saelo.github.io/presentations/offensivecon_22_attacking_javascript_engines.pdf.

Gruss, Daniel, Clémentine Maurice, Anders Fogh, Moritz Lipp, and Stefan Mangard. 2016. "Prefetch Side-Channel Attacks: Bypassing Smap and Kernel Aslr." In *Proceedings of the 2016 Acm Sigsac Conference on Computer and Communications Security*. CCS '16. https://doi.org/10.1145/2976749.2978356.

Gruss, Daniel, Clémentine Maurice, Klaus Wagner, and Stefan Mangard. 2016. "Flush+Flush: A Fast and Stealthy Cache Attack." In *Proceedings of the 13th International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment - Volume 9721*, 279–99. DIMVA 2016. https://doi.org/10.1007/978-3-319-40667-1_14.

Hicks, Michael. 2014. "What Is Memory Safety?" 2014. http://www.pl-enthusiast.net/2014/07/21/memory-safety/.

Hu, Hong, Shweta Shinde, Sendroiu Adrian, Zheng Leong Chua, Prateek Saxena, and Zhenkai Liang. 2016. "Data-Oriented Programming: On the Expressiveness of Non-Control Data Attacks." In *2016 Ieee Symposium on Security and Privacy (Sp)*, 969–86. https://doi.org/10.1109/SP.2016.62.

Ispoglou, Kyriakos K., Bader AlBassam, Trent Jaeger, and Mathias Payer. 2018. "Block Oriented Programming: Automating Data-Only Attacks." In *Proceedings of the 2018 Acm Sigsac Conference on Computer and Communications Security*, 1868–82. CCS '18. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3243734.3243739.

Liljestrand, Hans, Thomas Nyman, Lachlan J Gunn, Jan-Erik Ekberg, and N Asokan. 2021. "{PACStack}: An Authenticated Call Stack." In *30th Usenix Security Symposium (Usenix Security 21)*, 357–74.

Lipp, Moritz, Vedad Hadzic, Michael Schwarz, Arthur Perais, Clementine Lucie Noemie Maurice, and Daniel Gruß. 2020. "Take a Way: Exploring the Security Implications of Amd's Cache Way Predictors." In *Proceedings of the 15th Acm Asia Conference on Computer and Communications Security, Asia Ccs 2020*. https://doi.org/10.1145/3320269.3384746.

McCall, John, and Ahmed Bougacha. 2019. "Arm64e: An Abi for Pointer Authentication." 2019. https://llvm.org/devmtg/2019-10/slides/McCall-Bougacha-arm64e.pdf.

Miller, Matt. n.d. "Modeling the Exploitation and Mitigation of Memory Safety Vulnerabilities." Breakpoint 2012. https://github.com/Microsoft/MSRC-Security-Research/blob/master/presentations/2012_10_Breakpoint/BreakPoint2012_Miller_Modeling_the_exploitation_and_mitigation_of_memory_safety_vulnerabilities.pdf.

Mushtaq, Maria, Muhammad Asim Mukhtar, Vianney Lapotre, Muhammad Khurram Bhatti, and Guy Gogniat. 2020. "Winter Is Here! A Decade of Cache-Based Side-Channel Attacks, Detection & Mitigation for RSA." *Information Systems*, September. https://doi.org/10.1016/j.is.2020.101524.

Osvik, Dag Arne, Adi Shamir, and Eran Tromer. 2005. "Cache Attacks and Countermeasures: The Case of Aes." In *IACR Cryptology ePrint Archive*. https://doi.org/10.1007/11605805_1.

Paul A., Karger, and Schell Roger R. 1974. "MULTICS Security Evaluation: VULNERABILITY Analysis," 52. https://csrc.nist.gov/csrc/media/publications/conference-paper/1998/10/08/proceedings-of-the-21st-nissc-1998/documents/early-cs-papers/karg74.pdf.

Percival, Colin. 2005. "Cache Missing for Fun and Profit." BSDCan. https://eprint.iacr.org/2005/271.

Pizlo, Filip. 2020. "Speculation in Javascriptcore." 2020. https://webkit.org/blog/10308/speculation-in-javascriptcore/.

Pornin, Thomas. 2018. "Why Constant-Time Crypto?" 2018. https://www.bearssl.org/constanttime.html.

Rutland, Mark. 2017. "ARMv8.3 Pointer Authentication." 2017. https://events.static.linuxfound.org/sites/events/files/slides/slides_23.pdf.

saelo. 2021a. "Attacking Javascript Engines: A Case Study of Javascriptcore and Cve-2016-4622." 2021. http://www.phrack.org/issues/70/3.html.

———. 2021b. “Exploiting Logic Bugs in Javascript Jit Engines.” 2021. http://www.phrack.org/issues/70/9.html.

Schuster, F., T. Tendyck, C. Liebchen, L. Davi, A. Sadeghi, and T. Holz. 2015. “Counterfeit Object-Oriented Programming: On the Difficulty of Preventing Code Reuse Attacks in C++ Applications.” In *2015 Ieee Symposium on Security and Privacy (Sp)*, 745–62. Los Alamitos, CA, USA: IEEE Computer Society. https://doi.org/10.1109/SP.2015.51.

Schwarz, Michael, Clémentine Maurice, Daniel Gruss, and Stefan Mangard. 2017. “Fantastic Timers and Where to Find Them: High-Resolution Microarchitectural Attacks in JavaScript.” In *Financial Cryptography and Data Security*, 247–67. Springer International Publishing.

Schwarz, Michael, Martin Schwarzl, Moritz Lipp, Jon Masters, and Daniel Gruss. 2019. “NetSpectre: Read Arbitrary Memory over Network.” In *Computer Security - ESORICS 2019 - 24th European Symposium on Research in Computer Security, Luxembourg, September 23-27, 2019, Proceedings, Part I*, 11735:279–99. Lecture Notes in Computer Science. Springer.

Seacord, Robert C. 2013. *Secure Coding in c and C++*. 2nd ed. Addison-Wesley Professional.

Serebryany, Konstantin, Derek Bruening, Alexander Potapenko, and Dmitriy Vyukov. 2012. “{AddressSanitizer}: A Fast Address Sanity Checker.” In *2012 Usenix Annual Technical Conference (Usenix Atc 12)*, 309–18.

Serna, Fermin J. 2012. “The Info Leak Era on Software Exploitation.” 2012. https://www.youtube.com/watch?v=VgWoPa8Whmc.

Shacham, Hovav. 2007. “The Geometry of Innocent Flesh on the Bone: Return-into-Libc Without Function Calls (on the X86).” In *Proceedings of the 14th Acm Conference on Computer and Communications Security*, 552–61. CCS ’07. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1315245.1315313.

Sharma, Siddharth. 2014. 2014. https://www.redhat.com/en/blog/enhance-application-security-fortifysource.

Soares, Luigi, and Fernando Magno Quintan Pereira. 2021. “Memory-Safe Elimination of Side Channels.” In *2021 IEEE/ACM International Symposium on Code Generation and Optimization (CGO)*. IEEE. https://doi.org/10.1109/cgo51591.2021.9370305.

Solar Designer. 1997. “Getting Around Non-Executable Stack (and Fix).” 1997. https://seclists.org/bugtraq/1997/Aug/63.

Song, Chengyu, Chao Zhang, Tielei Wang, Wenke Lee, and David Melski. 2015. “Exploiting and Protecting Dynamic Code Generation.” In *NDSS*.

Su, Chao, and Qingkai Zeng. 2021. “Survey of CPU Cache-Based Side-Channel Attacks: Systematic Analysis, Security Models, and Countermeasures.” *Security and Communication Networks*, June. https://doi.org/10.1155/2021/5559552.

Thompson, Ken. 1984. “Reflections on Trusting Trust.” https://www.cs.cmu.edu/~rdriley/487/papers/Thompson_1984_ReflectionsonTrustingTrust.pdf.

Weber, Daniel, Ahmad Ibrahim, Hamed Nemati, Michael Schwarz, and Christian Rossow. 2021. "Osiris: Automated Discovery of Microarchitectural Side Channels." In *USENIX Security'21*. https://arxiv.org/abs/2106.03470.

Wingo, Any. 2011. "Value Representation in Javascript Implementations." 2011. https://wingolog.org/archives/2011/05/18/value-representation-in-javascript-implementations.

Wu, Meng, Shengjian Guo, Patrick Schaumont, and Chao Wang. 2018. "Eliminating Timing Side-Channel Leaks Using Program Repair." In *Proceedings of the 27th ACM SIGSOFT International Symposium on Software Testing and Analysis*. ACM. https://doi.org/10.1145/3213846.3213851.

Yarom, Yuval, and Katrina Falkner. 2014. "FLUSH+RELOAD: A High Resolution, Low Noise, L3 Cache Side-Channel Attack." In *23rd Usenix Security Symposium (Usenix Security 14)*, 719–32. San Diego, CA: USENIX Association. https://www.usenix.org/conference/usenixsecurity14/technical-sessions/presentation/yarom.