Protecting sensitive information in the volatile memory from disclosure attacks

Stefanos Malliaros, ChristofoRos Ntantogian, Christos Xenakis
Department of Digital Systems
University of Piraeus
Piraeus, Greece
{stefmal, dadoyan, xenakis}@uniPi.gr

Abstract—The protection of the volatile memory data is an issue of crucial importance, since authentication credentials and cryptographic keys remain in the volatile memory. For this reason, the volatile memory has become a prime target for memory scrapers, which specifically target the volatile memory, in order to steal sensitive information, such as credit card numbers. This paper investigates security measures, to protect sensitive information in the volatile memory from disclosure attacks. Experimental analysis is performed to investigate whether the operating systems (Windows or Linux) perform data zeroization in the volatile memory. Results show that Windows kernel zeroize data after a process termination, while the Linux kernel does not. Next, we examine functions and software techniques in C/C++ programming language that can be used by developers to modify at process runtime the contents of the allocated blocks in the volatile memory. We have identified that only the Windows operating system provide a specific function named SecureZeroMemory that can reliably zeroize data. Finally, driven by the fact that malware scrapers primarily target web browsers, we examine whether it is feasible to extract authentication credentials from the volatile memory allocated by web browsers. The presented results show that in most cases we can successfully recover user authentication credentials from all the web browsers except when the user has closed the tab that used to access the website.

Keywords: Memory zeroization, Volatile memory, Operating Systems, memory management, Information disclosure.

I. INTRODUCTION

Recent research has shown that authentication credentials and cryptographic keys remain in the volatile memory and can be easily extracted [2]. For this reason, the volatile memory has become a prime target for malicious software. As a matter of fact, a new malware category has emerged named as memory scrapers, which specifically target the volatile memory, in order to steal sensitive information, such as credit card numbers [5]. To achieve this, memory scrapers use regular expression matches, in order to harvest credit card data from the volatile memory, and then the collected data are sent to a malicious server. The first known memory scraper, named StarDust targeted point of sale terminals and compromised nearly 20,000 credit cards in the US [7].

Due to the rise of memory scrapers [1], it is evident that the protection of volatile memory from disclosure attacks is of paramount importance. The majority of the related work in this area focuses on hardware-based solutions that encrypt data in volatile memory. Although these solutions can protect data in volatile memory, they raise deployment and performance issues, since they require modifications at the hardware level. As a result, these solutions are not considered to be generic. Apart from memory encryption solutions, there are also special devices, such as Hardware Security Modules [28] and FPGAs [30] that employ data zeroization to protect the volatile memory from disclosure attacks. The goal of data zeroization is to overwrite sensitive information in the volatile memory with zeroes [4]. On the other hand, it is not widely known whether the generic Operating Systems (OS) such as Windows or Linux provide protection in the volatile memory. Moreover, there are no specific guidelines that the developers should follow to protect sensitive data in the volatile memory of their applications.

In this paper we investigate techniques that can be applied at the software level either from the OS or the applications, in order to zeroize data in the volatile memory. More specifically, first, we investigate whether OS (Windows or Linux) provide built-in safeguards to protect against disclosure attacks from the volatile memory. Experimental results show that Windows kernel zeroize data after a process termination, while the Linux kernel does not. Next, we elaborate on software functions and techniques in C/C++ programming language that can be used by developers to protect the data in the volatile memory of their applications. We have identified that only the Windows OS provide a specific function named SecureZeroMemory that can reliably zeroize data. Finally, driven by the fact that malware scrapers primarily target web browsers, we examine whether it is feasible to extract authentication credentials from the volatile memory allocated by web browsers. The presented results show that in most cases we can successfully recover user authentication credentials from all the web browsers except when the user has closed the tab that used to access the website.

The rest of the paper is structured as follows. Section 2 presents the related work, while section 3 presents the protection of the volatile memory data at the software level. Section 4 evaluates the security of web browsers. Finally, the conclusions of this work are presented in section 5.

II. RELATED WORK

In this section, the related work is presented and elaborated which is composed of two parts. In the first part, previous works which show that sensitive information can be discovered in volatile memory are presented. In the second part, the previous works that protect volatile memory using encryption are elaborated.

Regarding previous works that focus on the retrieval of sensitive information in the volatile memory, Darren et al.
tried to recover data remnants from cloud storage applications including Dropbox [8], Skydrive [9], and Google Drive [10]. Similarly, in [11] the authors investigate the volatile memory of cloud services applications, such as Amazon S3, Dropbox, Google Docs and Evernote. In all the aforementioned publications, several artifacts were recovered such as authentication credentials, visited URLs, filenames and hashes. Apart from personal computers, sensitive information was also recovered from the volatile memory of Android devices using two different methods. More specifically, in the first method [12] the authors used the Linux Memory Extractor (LiME) kernel module [13] and a physical Samsung i9000 phone to dump the Android memory, whereas in the second technique [14] the Android emulator was used alongside with Dalvik Debug Monitor Server (DDMS) to acquire the memory data. In both cases, critical and secure applications, such as mobile banking and password managers, were examined and authentication credentials were recovered in plain text from the dumped memory.

Regarding memory encryption, the proposed solutions can be further classified into two categories: software-based and hardware-based. For software-based solutions, in [15], the authors propose a modified secure memory bus controlled by the OS, in which the encryption key is generated each time the system boots up. Peterson, in [16], modified the virtual memory manager of the Linux 2.6.24 kernel and partitioned the volatile memory into a plaintext and an encrypted segment. However, [17] shows that the memory maps, should be maintained in the plaintext segment; thus pointing the addresses to where the encrypted volatile data are stored. The second category of the proposed solutions for memory encryption is based on hardware modifications. In particular, several publications [18-23] [40] for single processor systems propose the addition of an encryption unit to cipher and decipher data from and to the volatile memory. Moreover, for multi-processor systems, [24] proposes a shared bus, containing a crypto engine, to coordinate and secure traffic between processors, while [25] [26] proposed the use of sequence numbers for the coordination between different processors. Lastly, in [27], the authors propose SecBus, a cryptographic coprocessor between the volatile memory and the main processor.

The main limitation of the proposed solutions that perform memory encryption has to do with the fact that hardware-based solutions require extensive changes in the current CPU architecture, while the software-based solutions require modifications at the OS kernel. In contrast to the relevant works, this paper investigates if the latest OS versions (Windows and Linux) provide built-in data zeroization methods as well as whether C/C++ developers can use existing software libraries and methods in order to perform data zeroization in their applications.

III. SOFTWARE LEVEL PROTECTION

A. Operating System level protection

Memory management is the procedure of administering the volatile memory at the system level. This is performed by the kernel of the Operating System (OS) with the support of a part of the central processing unit, named memory management unit. Allocation and deallocation requests are used in order to grant or revoke memory blocks to applications. Allocation is the procedure in which memory blocks are granted to applications, and are then used by them for handling the necessary data for their functionalities. On the other hand, deallocation is the procedure in which the applications free the memory blocks they do not longer need, making them available for other running or starting applications. It is important to note that the OS does not modify the allocated memory blocks, since this action could cause the running applications to crash. Subsequently, during the applications’ runtime, only the applications themselves are accountable of modifying their allocated memory blocks. In order to find out whether the OS performs data zeroization, we developed a testing application written in C programming language (see figure 1), that holds a secret value in the variable named as password. The aim of the experiments was to investigate how many instances of the password variable can be extracted from the volatile memory. More specifically, as shown in figure 1, the testing application defines the password variable at line 3, which is an array of type char and size length. Moreover, the stdin (e.g. keyboard input) is used to fill in the array of the password variable. For the experiments, three types of memory dumps were considered which are: A) Process: This memory dump includes only the memory blocks that are allocated to the executable of figure 1. B) All-Processes: This memory dump includes memory blocks allocated to all running user-mode processes in the OS. In this way we can find out whether the password variable of figure 1 can be extracted from other user-mode running processes; C) System: This memory dump contains the entire volatile memory including memory allocated not only to user-mode processes but also to the OS kernel, drivers, unallocated blocks. The technical methodology that we followed in order to obtain the memory dumps is as follows. To perform a Process dump in Linux, the GNU debugger (i.e. GDB) was used to dump the memory blocks of a process based on its PID. Similarly, the All-processes dump was performed using a script that feeds GDB with all the running PDIDs. The same methodology was followed in Windows. In particular, we used the Windows Powershell in order to list all the running PDIDs, and feed them to ProcDump [41] (i.e., a Windows utility which performs memory dumps of running processes). It is important to note that all the aforementioned memory dumps, were executed using root privileges both in Linux and Windows. To perform System dump, we used virtual machines, in order to dump the entire volatile memory of the system in an easy manner.

Moreover, two scenarios were considered. In the first scenario named as “Running process” we performed memory dumps (all three types) while the process of the

```c
1. void main() {
2.   static int length;
3.   char password[length];
4.   fgets(password, length * sizeof(char), stdin);
5.   sleep(120); } //suspend for 120 seconds
```

Figure 1. First testing application used to discover the total number of instances of the password variable in the volatile memory.
executable was running. This was achieved during the sleep function (see line 5 of figure 1), where the execution of the process was suspended and we were able to recover the memory dump. In the second scenario named as “After termination”, we performed the memory dumps immediately after the termination of the executable. Evidently, in this scenario, we performed only All-processes and System memory dumps, since Process dump cannot be performed after the termination of the executable. The experiments were conducted in Windows 7 and 10 and Ubuntu Linux 14.04, fully updated as of 15th of April 2016. In both versions of Windows, the compiler of Microsoft Visual Studio 2015 suite was used, while in Ubuntu Linux we used the latest version of the GCC compiler (i.e., v5.3).

The results of the experiments are summarized in Table 1. We can observe that in the “Running process” scenario in all three memory dump types for both Linux and Windows OS we were able to recover the value of the password variable. It is interesting to notice that in the All-processes memory dump type, the number of the instances of the password variable were the same as in the Process memory dump type (i.e., 1 time in Linux and 3 times in Windows). This means that apart from the process itself of the testing application (see figure 1), the other processes running in the system did not use the password variable. We can also observe that in the System memory dump, the number of recovered password instances increased (i.e., 9 times in Linux and 5 times in Windows). This result means that i) apart from the process of the testing application itself, the OS kernel stores also the value of the password variable and ii) the OS kernels stores in multiple memory regions the value of the password.

Regarding the results of the “After termination” scenario, we can observe an interesting outcome: for both All-processes and System dumps in Linux we were able of recovering the password variable (1 and 2 times respectively). On the other hand, in Windows we were not able to recover it. This result means that Windows kernel zeroize the deallocated blocks of a process immediately after its termination. On the other hand, the Linux kernel follows a different approach. That is, instead of zeroizing the deallocated memory blocks of a terminating process, it zeroizes the memory blocks right before their allocation [31]. Thus, in Linux, a malicious software that has access to the entire system memory can extract potentially sensitive information (such as authentication credentials) even from applications that were terminated, in case the related deallocated blocks have not been allocated to a new process. On the contrary, in Windows, a malicious software can extract information only from the memory blocks of running applications.

The above observation implies that Windows is more secure than Linux to memory disclosure attacks. To overcome this issue, we have identified that there is a Linux kernel patch, named as GRsecurity, which provide several security enhancements for the Linux kernel [32]. One of these enhancements enables the Linux kernel to zeroize the deallocated memory blocks after process termination by compiling the Linux kernel with the PAX_MEMORY_SANITIZE option that the GRsecurity provides. To this end, we repeated the experiments (using the testing application of figure 1) in Ubuntu 14.04 compiled with a kernel that has GRSecurity installed and the PAX_MEMORY_SANITIZE option enabled. We observed that this time we were not able of recovering instances of the password variable after the process termination. Based on the above discussion, we propose the use of GRsecurity (with the PAX_MEMORY_SANITIZE option enabled), in order to minimize information disclosure in volatile memory.

Despite the fact that GRSecurity may enable the kernel to perform data zeroization, it is not widely adopted in Linux Distributions. Even those that offer a GRSecurity patched kernel by default, many of them have not enabled the PAX_MEMORY_SANITIZE option. In total, we found six Linux distributions [33-34] [42-45] that come with a GRSecurity patched kernel and only three of them have the PAX_MEMORY_SANITIZE option enabled.

### B. Source code level protection

The previous results show that OS zeroize data only after the termination of the running process which means that during the runtime of a process, sensitive information can be extracted in its allocated memory blocks. In this section, we investigate functions and methods that developers can use in order to zeroize memory blocks during the runtime of their applications. We focus on C/C++ programming language, since it provides low-level memory manipulation. All experiments carried out in this section perform Process dump in a “Running-process” scenario.

First, we investigate for Windows OS, if there are special functions that can be used in order to zeroize data. More specifically, by including the windows.h header file in a C/C++ source code, a developer has the ability of using the macro SecureZeroMemory, which calls the function RtlSecureZeroMemory that guarantees to zeroize memory blocks, even if it is not subsequently written or accessed by the code [35]. We repeated the experiments.
performed in the previous section (i.e., as mentioned previously only Process dump in the "Running process" scenario) using the same testing application with the difference that at the end of the code we called the SecureZeroMemory macro. We observed that indeed the macro SecureZeroMemory replaced the contents of the password variable with zeroes. Thus, in Windows, developers should use the macro SecureZeroMemory to ensure that the memory blocks of their applications are zeroized.

On the other hand, for Linux OS, there is no similar C function that can be used to zeroize data in the volatile memory. To this end, we have used the function memset of the C programming language to manually try to zeroize memory blocks allocated to a process. In particular, we have used the testing application of figure 2, which is identical to the code of figure 1, with the difference that figure 2 includes in line 5, the command memset(password, '0', length). This command writes in the memory block, which is allocated for the value of the password variable, the 0 character as many times as indicated by the value of the length variable. This will result in the zeroization of the data of the array password. We repeated the experiments of the previous section and we observed that the memset function was not operating as we expected, since the value of the password variable was detected in the process dumps. After investigation, we identified that the memset function was not being called due to code optimization. The latter is the process in which a compiler tries to improve the generated executable code by making it consume fewer resources, such as CPU and Memory. This is performed by several techniques. One of these methods is to avoid compiling specific code which is not necessary for the execution flow. For this reason, in our experiments, the compiler skipped the calling of the memset function, because the new value of the password variable (i.e., the zeroized data) is not used after the memset function. Note that although the executable of figure 2 was compiled using GCC without optimization flags, the GCC compiler did perform optimization and did not include the memset function in the executable.

The above results raise the following question: "is it feasible to avoid optimization caused by the GCC compiler, in order to ensure that the memset function will be executed?" To answer this question we tried two different methods. In the first method we used the function memset_s. The latter has the same functionality as memset. The main difference between those two functions is that the memset_s cannot be optimized out by the compilers [38]. However, memset_s is included only in the currently last version of the standard of the C programming language (i.e., C11 [39]) in Annex K. Unfortunately, Annex K is not mandatory in C11 specifications and at the time of writing the paper, the latest version of the GCC compiler (i.e., v5.3) has not implemented the Annex K, and thus the developers have no way to use the memset_s function.

The second method that we attempted in order to avoid bypassing optimization was to write a testing application similar to the one described in [46] (see figure 3), which uses a function pointer of type volatile named memset_volatile, as defined at line 1. The declaration of a variable as volatile instructs the compiler not to optimize out functions that access the variable. This is due to the fact the volatile type is used mainly for buffers in communication with hardware devices or other applications. Based on this observation, we defined the function pointer named memset_volatile pointing to the function memset at line 1. At line 4, a pointer named password_heap is defined, which points to a block of memory of size length*sizeof(char). This block of memory is allocated using the malloc function, which is used for dynamic memory allocation during the application execution. In line 5, the user enters his password, and in line 6, the memory block allocated at line 4 is freed with the free command. It should be noted that the free command does not zeroize the data of the memory block it deallocates. Consequently, we used the memset_volatile function pointer to indirectly call the memset function. We repeated the experiments once again, using all the available optimization flags of the GCC compiler. In all cases we observed that the GCC compiler did not optimize the call to the memset function. Although the experiments showed that the data type volatile in C/C++ programming language prevents the optimization caused by the compilers, it should be noted that GCC compiler can arbitrary perform optimization even in volatile data types as mentioned in [47]. In any case, volatile function pointers can be used to increase the chances that the memset function will not be optimized out during compilation.

```
1. void main() {
2.   static int length;
3.   char password[length];
4.   fgets(password, length * sizeof(char), stdin);
5.   memset(password, '0', length);
6.   sleep(120); } // suspend for 120 seconds

Figure 2. Second testing application used to discover the total number of instances of the password variable in the volatile memory.
```

```
1. void *(*volatile memset_volatile)(void *, int, size_t) = memset;
2. void sensitive_function() {
3.   static int length;
4.   char *password_heap = malloc(length * sizeof(char));
5.   fgets(password_heap, n, stdin);
6.   memset_volatile(password_heap, 0, n * sizeof(char));
7.   free(password_heap)
8.   sleep(120); }

Figure 3. Third testing application used to discover the total number of instances of the password variable in the volatile memory.
```
IV. WEB BROWSER EXPERIMENTS

In this section we investigate whether web browsers perform data zeroization. The reason that we have selected web browsers is: i) their rendering engine is written in C++ programming language meaning that they can perform low-level memory management, and ii) malware scrapers primarily target web browsers to extract sensitive information from the volatile memory (such as authentication credentials). First, we present the methodology to obtain memory dumps and next we analyze numerical results.

A. Methodology

We have selected the most common web browsers to evaluate whether we can recover user login credentials (i.e., username and password) from their allocated memory blocks during their runtime. In particular, Microsoft Edge, Internet Explorer 11, Google Chrome 49.0.2623 and Mozilla Firefox 45.0.1 were investigated running in Windows 7 and Windows 10 OSes. In Ubuntu Linux 14.04, we evaluated Mozilla Firefox 45.0.1 and Google Chrome 49.0.2623. In the experiments we performed process-dumps (see section 3.1) at the end of four distinct steps, which are: i) after logging to the website using our authentication credentials; ii) after logging out of the website; iii) after browsing to a different webpage in the same tab of the web browser; and iv) after closing the browser tab that we used in the previous steps and browsing to a new page in a new tab. Furthermore, five different websites (i.e. LinkedIn, Twitter, Facebook, Pinterest, Instagram) were used that require user login. In the four process-dumps, we searched how many times we can recover the authentication credentials (both in ASCII [36] and Unicode [37] formats).

B. Results

Table 2 shows the average number of the recovered instances of the authentication credentials. In this table, the digit before the notation U indicates the number of occurrences of the username and the digit before the notation P indicates the number of occurrences of the passwords. For instance, the value 5U/2P indicates that we extracted 5 occurrences of the username and 2 occurrences the password. As shown in Table 2, in Firefox running in Windows we recovered 23 and 6 instances of the username and password respectively after the login process has been completed. Similarly, for Linux using the same web browser we recovered 38 and 8 times the username and password respectively. Google Chrome has exact the same behavior, since in Linux we recovered more occurrences of the authentication credentials than in Windows.

A generic observation is that the password variable is found fewer times in comparison to the username in all the performed experiments. This can be attributed to the fact that the username is displayed in the pages of the tested websites, while the password evidently is not shown anywhere in the website. Overall, we can observe that Microsoft Edge has the best performance, since the recovered instances of the authentication credentials are considerably fewer than the other web users. Moreover, Edge is the only browser that deleted all the instances of the authentication credentials after changing the web page without closing the tab. On the other hand, Firefox had the worst performance, since it had the most occurrences of the authentication credentials, both in Linux and Windows.

Finally, it is important to mention that the total occurrences of the authentication credentials decrease when we log out, change webpage and close the tab of the web browser. Subsequently, the authentication credentials are more likely to be disclosed right after the user has entered them into a website.

V. CONCLUSIONS

This paper investigates security measures that can be applied at the OS and the source code level to protect sensitive information in volatile memory from disclosure attacks. Based on our experimental analysis, we observed that Windows delete the data from deallocated memory blocks, while Linux does not. This can be solved using the GSecurity Linux kernel patch that enables the zeroization of deallocated memory blocks, using the `PAX_MEMORY_SANITIZE` option during the kernel compilation. At the source code level, the Windows developers may use the SecureZeroMemory function for manually modifying volatile memory data without facing any optimization issues. In Linux, we propose the use of `volatile` function pointers to ensure that the call to `memset` will not be optimized out. Lastly, the experiments performed in web browsers show that in most cases we can successfully recover user authentication credentials from
ACKNOWLEDGMENT
This research has been funded by the European Commission in part of the ReCRED project (Horizon 2020 Framework Programme of the European Union under GA number 653417).

REFERENCES


