

Lord of the x86 Rings: A Portable User Mode Privilege Separation Architecture on x86

Hojoon Lee  
GSIS, School of Computing, KAIST  
hojoon.lee@kaist.ac.kr  

Chihyun Song  
GSIS, School of Computing, KAIST  
ian0371@kaist.ac.kr  

Brent Byunghoon Kang  
GSIS, School of Computing, KAIST  
brentkang@kaist.ac.kr

ABSTRACT
Modern applications are increasingly advanced and complex, and inevitably contain exploitable software bugs despite the ongoing efforts. The applications today often involve processing of sensitive information. However, the lack of privilege separation within the user space leaves sensitive application secret such as cryptographic keys just as unprotected as a “hello world” string. Cutting-edge hardware-supported security features are being introduced. However, the features are often vendor-specific or lack compatibility with older generations of the processors. The situation leaves developers with no portable solution to incorporate protection for the sensitive application component.

We propose LOTRx86, a fundamental and portable approach for user space privilege separation. Our approach creates a more privileged user execution layer called PrivUser through harnessing the underused intermediate privilege levels on the x86 architecture. The PrivUser memory space, a set of pages within process address space that are inaccessible to user mode, is a safe place for application secrets and routines that access them. We implement the LOTRx86 ABI that exports the privcall interface to users to invoke secret handling routines in PrivUser. This way, sensitive application operations that involve the secrets are performed in a strictly controlled manner. The memory access control in our architecture is privilege-based, accessing the protected application secret only requires a change in the privilege, eliminating the need for costly remote procedure calls or change in address space.

We evaluated our platform by developing a proof-of-concept LOTRx86-enabled web server that employs our architecture to securely access its private key during SSL connection and thereby mitigating the HeartBleed vulnerability by design. We conducted a set of experiments including a performance measurement on the PoC on both Intel and AMD PCs, and confirmed that LOTRx86 incurs only a limited performance overhead.

1 INTRODUCTION
User applications today are prone to software attacks, and yet are often monolithically structured or lack privilege separation. As a result, adversaries who have successfully exploited a software vulnerability in an application can access to sensitive in-process code or data that are irrelevant to the exploited module or part of the application. Today’s applications often contain secrets that are too critical to reside in the memory along with the rest of the application contents, as we have witnessed in the incident of HeartBleed [18, 42].

The conventional software privilege model that coarsely divides the system privilege into only two levels (user-level and kernel-level) has failed to provide a fundamental solution that can support privilege separation in user applications. As a result, critical application secrets such as cryptographic information are essentially treated no differently than a “hello world” string in user memory space. When the control flow of a running user context is compromised, there is no access control left to prevent the hijacked context to access arbitrary memory addresses.

Many approaches have been introduced to mitigate the challenging issue within the boundaries of the existing application memory protection mechanisms provided by the operating system. A number of work proposed using the process abstraction as a unit of protection by separating a program into multiple processes [6, 27, 38]. The key idea is to utilize the process separation mechanism provided by the OS; these work achieve privilege separation by splitting a single program into multiple processes. However, this process-level separation incurs a significant overhead due to the cost of the inter-process communication (IPC) between the processes or address space switching that incur TLB flushes. Also, the coarse unit of separation still leaves a large attack surface for attackers. The direction has advanced through a plethora of works on the topic. One prominent aspect of the advancements is the granularity of protection. Thread-level protection schemes [5, 22, 41] have reduced the protection granularity compared to the process-level separation schemes while still suffering from performance overhead from page table modifications. Shreds presented fine-grained in-process memory protection using a memory partitioning feature that has long been present in ARM called Memory Domains [9]. However, the feature has been deprecated in the 64-bit execution mode of the ARM architecture (AARCH64).

In the more recent years, a number of processor architecture revisions and academic works have taken a more fundamental approach to provide in-process protection; Intel has introduced Software Guard Extensions (SGX) to its new x86 processors to protect sensitive application and code and data from the rest of the application as well as the possibly malicious kernel [4, 12]. Intel also offers hardware-assisted in-process memory safety and protection features [11, 13] and AMD has announced the plans to embed a similar feature to its future generations of x86 processors [17]. However, the support for the new processor features are fragmented; not only that the features are not inter-operable across processors from different vendors (Intel, AMD), they are also only available on the newer processors. Hypervisor-based application memory protection [8, 32] may serve to be a more portable solution, considering the widespread adoption of hypervisors nowadays. However, it is not reasonable for a developer to assume that her users are using a virtual machine.

The situation presents complications for developers who need to consider the portability as well as the security of the sensitive data her program processes. Therefore, we argue that there is a need
for an approach that provides a basis for an in-process privilege separation based on only the portable features of the processor. An in-process memory separation should not require a complete address space switching to access the protected memory or costly page table modifications.

In this paper, we propose a novel x86 user-mode privilege separation architecture called The Lord of the x86 Rings (LOTRx86) architecture. Our architecture proposes a drastically different, yet portable approach for user privilege separation on x86. While the existing approaches sought to retrofit the memory protection mechanisms within the boundaries of the OS kernel’s support, we propose the creation of a more privileged user layer called PrivUser that protects sensitive application code and data from the normal user mode. For this objective, LOTRx86 harnesses the underused x86 intermediate Rings (Ring1 and Ring2) with our unique design that satisfies security requirements that define a distinct privilege layer. The PrivUser memory space is a subset of a process memory space that is accessible to when the process context is in PrivUser mode but inaccessible when in user mode. In our architecture, user memory access control is privileged-based. Therefore, the architecture does not require costly run-time page table manipulations nor address space switching.

We also implemented the LOTRx86 ABI that exports the privcall interface that supports PrivUser layer invocation from user layer. To draw an analogy, the syscall interface is a controlled invocation of kernel services that involve kernel’s exclusive rights on sensitive system operations. In our architecture, PrivUser holds an exclusive right to application secrets and sensitive routines with a program, and user layer must invoke privcalls to enter PrivUser mode and perform sensitive operations involving the secrets in a strictly controlled way. Our architecture allows developers to protect applications secret within the PrivUser memory space and also write privcall routines that can securely process the application secret. We developed a kernel module that adds the support for the privcall ABI to the Linux kernel (lotr-kmod). In addition, we provide a library (liblotr) that provides the privcall interface to the user programs and C macros that enable declaration of privcall routines, a modified C library for the building the PrivUser side (lotr-libc), and a tool for building LOTRx86-enabled program (lotr-build).

We implemented a prototype of our architecture that is compatible with both Intel and AMD’s x86 processors. Based on our prototype, we developed a proof-of-concept LOTRx86-enabled web server. In our PoC, the web server’s private key is protected in the PrivUser memory space and the use of the key (e.g., sign a message with the key) is only allowed through our privcall interface. In our PoC web server, in-memory private key is inaccessible outside the privcall routines that are invoked securely, hence an arbitrary access to the key is automatically thwarted (i.e., HeartBleed). The evaluation of the PoC and other evaluations are conducted on both Intel and AMD PCs. We summarize the contributions of our LOTRx86 architecture as the following:

- We propose a portable privileged user mode architecture for sensitive application code and data protection that does not require address switching or run-time page table manipulation.
- We introduce the privcall ABI that allows user layer to invoke the privcall routines in a strictly controlled way. We also provide necessary software for building an LOTRx86-enabled software.
- We developed a PoC LOTRx86-enabled web server to demonstrate the protection of in-memory private key during SSL connection.

2 BACKGROUND: THE X86 PRIVILEGE ARCHITECTURE

The LOTRx86 architecture design leverages the x86 privilege structures in a unique way. Hence, it is necessary that we explain the x86 privilege system before we go further into the LOTRx86 architecture design. In this section, we briefly describe the x86 privilege concepts focusing on the topics that are closely related to this paper.

2.1 The Ring Privileges

Modern operating systems on the x86 architecture adopt the two privilege level model in which user programs run in Ring3 and kernel in Ring0. The x86 architecture, in fact, supports four privilege layers – Ring0 through Ring3 where Ring0 is the highest privilege on the system. The x86 architecture’s definition of privilege is closely tied to a feature called segmentation.

Segmentation divides virtual memory spaces into segments which are defined by a base address, a limit, and a Descriptor Privilege Level (DPL) that indicates the required privilege level for accessing the segment. A segment is defined by segment descriptor in either Global Descriptor Table (GDT) or Local Descriptor Table (LDT). The privilege of an executing context is defined by a 16-bit data structure called segment selector loaded in the code segment register (%cs). The segment selector contains an index to the code segment in the descriptor table, a bit field to signify which descriptor table it is referring to (GDT/LDT), and a 2-bit field to represent the Current Privilege Level (CPL). The CPL in %cs is synonymous to the context’s current Ring privilege number.

The privilege level (the Ring number) dictates an executing context’s permission to perform sensitive system operations and memory access. Notably, the execution of privileged instructions is only allowed to contexts running with Ring0 privilege. Also, the x86 paging only permits Ring0-2 to access supervisor pages.

2.2 Memory Protection

Operating systems use paging to manage memory access control, and the segmented memory model has long been an obsolete memory management technique. However, the paging-based flat memory model, which has become the standard memory management scheme, uses the Ring privilege levels for page access control. The x86 paging defines two-page access privilege: User and Supervisor. The Ring 3 can only access User pages while Ring 0-2 are allowed to access Supervisor pages. In general, the pages in the kernel memory space are mapped as Superuser such that they are protected.

1 Intel and AMD have introduced a CPU feature called Supervisor Mode Execution Prevention (SMEP) and Supervisor Mode Access Prevention (SMAP). SMEP prevents contexts in Ring 0-2 from executing code in User pages, effectively preventing ret2user style of attacks. SMAP prevents the kernel from accessing user pages as data [10, 24]
from user applications. Table 1 outlines the privileges of each Ring level.

**Algorithm 1 x86 callgate operation**

1. procedure \(CG:R_n \rightarrow R_m(\text{SEGSEL})\)
2. \(\text{DESC\_TBL} \leftarrow \text{SEGSEL}._i \oplus \text{LDT} : \text{GDT}\)
3. \(\text{CG} \leftarrow \text{DESC\_TBL}[\text{SEGSEL}._i]\)
4. if \(n > \text{CG}\).\text{RMPL}\) or \(n \leq m\) then
   5. return DENIED
6. end if
7. \(\text{Save}\!(\%\text{rip},\%\text{cs},\%\text{rsp},\%\text{ss})\) \(\triangleright\) Save caller context in temp space
8. \(\%\text{ss} \leftarrow \text{TSS}[m].\text{SS}\) \(\triangleright\) Load new context to be used in Ring \(m\)
9. \(\%\text{rsp} \leftarrow \text{TSS}[m].\text{RSP}\)
10. \(\%\text{cs} \leftarrow \text{CG}._\text{TargetCS}\) \(\triangleright\) Privilege Escalation: \(n \rightarrow m\)
11. \(\%\text{rip} \leftarrow \text{CG}._\text{TargetEntrance}\)
12. Push SavedSS
13. Push SavedRSP
14. Push SavedCS
15. Push SavedRIP
16. RESUME
17. end procedure

**Algorithm 2 x86 long return instruction (%lret)**

1. procedure \(\text{LONG RETURN}\)
2. \(\triangleright\) can only return to equal or lower privileges
3. if \(\text{Dest\_Priv} < \text{Current\_Priv}\) then
   4. return DENIED
5. end if
6. \(\%\text{rip} \leftarrow \text{Pop}(\)\()\) \(\triangleright\) target addr
7. \(\%\text{cs} \leftarrow \text{Pop}(\)\()\) \(\triangleright\) target ring privilege
8. \(\text{tempRSP} \leftarrow \text{Pop}(\)\()\)
9. \(\text{tempSS} \leftarrow \text{Pop}(\)\()\)
10. \(\%\text{rsp} \leftarrow \text{tempRSP}\)
11. \(\%\text{ss} \leftarrow \text{tempSS}\)
12. RESUME
13. end procedure

### 2.3 Moving Across Rings

The x86 architecture provides a number of mechanisms by which a running context can explicitly invoke privilege escalation for system services. While the privilege of the context is clearly specified in its %cs register, its contents cannot be directly altered (e.g., mov %eax, %cs) but indirectly with special instructions. The x86 provides special instructions that allow switching of the code segment as well as the program counter, namely the inter-segment control transfers instructions. For instance, the execution of the syscall instruction elevates the CPL of the context to Ring0 by loading the %cs with the kernel code segment. It also loads the PC register (%rip) with system call entrance point in the kernel. In modern operating system kernels, only the instructions that invoke system calls are frequently used. However, it is necessary that we explain the concepts and mechanisms of the inter-segment control transfer mechanisms that were introduced along with the four Ring system long before the instructions dedicated to invoking system calls.

**Privilege escalation.** Our design makes use of the callgate mechanism for privilege escalation, a feature present in all modern (since the introduction of the protected mode) x86 processors. A callgate descriptor can be defined at the descriptor tables to create an inter-privilege tunnel between the Rings. Specifically, it defines the target code segment, whose privilege will be referred to as the Target Privilege Level (TPL), a Target Addr, and a Required Minimum Privilege Level (RMPL). A context can pass through a call gate via a long call instruction\(^2\) that takes a segment selector as its operand. The long call instruction first performs privilege checks when it confirms that the operand given is a reference to a callgate. A callgate demands its caller’s CPL (the current Ring number) to be numerically equal to or less than (higher privilege) the callgate’s RMPL. Also, the caller’s CPL cannot be numerically less than the TPL of the callgate. In other words, a control transfer through a callgate does not allow privilege de-escalation. If these privilege checks fail, the context receives a general protection fault and is forced to terminate. If the privilege check is successful, the privilege of the context is escalated, and the program counter (%rip), as well as the stack pointer (%rsp), are loaded with the target address. A long call instruction results in privilege escalation if and only if it references a valid callgate that defines a privilege escalation and minimum privilege required to enter the callgate. Therefore a callgate is a controlled control transfer that facilitates privilege escalation. We provide a pseudocode that describes the set of operations performed at the callgate in Algorithm 1. Note that we denote a control flow transfer where a context executing in Ring\(n\) enters Ring\(m\) through a callgate using the following notation:

\[CG : R_n \rightarrow R_m, \text{where } n \leq \text{CG}.\text{RMPL} \text{ and } m \leq n\]

**Privilege de-escalation.** A context can return to its original privilege mode with a long call instruction\(^3\) after privilege escalation. A long return instruction restores the caller’s context that has been saved by the long call instruction as shown in Algorithm 2. It should be noted that a long return instruction only checks if the destination privilege level is numerically equal to or greater (lower privilege) by referencing the saved caller context. In fact, a long return instruction has no way of knowing if the saved context on the stack is indeed saved by the callgate. Hence, the long return instruction and similar return instructions such as iret can be thought of as privilege de-escalating control transfer instructions that pop the contents that are presumably saved registers. In this sense, a long return and its variants provide non-controlled control transfer mechanism that is used to de-escalate privileges. We denote this specific type of control transfer where privilege is de-escalated (or stays the same) from \(m\) to \(n\) as the following:

\[R_m \rightarrow R_n, \text{where } m \leq n\]

**Inter-bitness control transfer.** Inter-bitness control transfer is another type of an x86 control transfer that needs to be explained before we introduce our design. The x86-64 architecture provides 32-bit compatibility mode within the x86-64 (AMD’s amd64 or Intel’s IA-32e architecture). As with the privilege level, the bitness is also defined by the currently active code segment descriptor.

\(^2\)call\(\text{1}\) in AT&T syntax and “call far” in Intel syntax
\(^3\)iret in AT&T syntax and “retf” in Intel syntax

| Privileged instruction | Ring0 | Ring1 | Ring2 | Ring3 |
|------------------------|-------|-------|-------|-------|
| Supervisor page access | ✓     | ×     | ×     | ×     |
| Inter-bitness control transfer | | | | |
When a context is executing in a code segment whose descriptor has the L flag set, the processor operates in the 64-bit instruction architecture (e.g., registers are 64-bit, and 64-bit instructions are used). Otherwise, the context executes as if the processor is an x86-32 architecture processor. The bitness switching, although it changes the processor (current CPU core) execution mode, is no different than any other inter-segment control transfers with one exception: a ccall gate cannot target a 32-bit code segment. This is a perk that came with the introduction of the x86-64 implementation. In summary, we denote 32-bit code segments with a x32 suffix as the following:

\[ R_{n,x32} \rightarrow R_m \]

3 ATTACK MODEL AND SECURITY GUARANTEES

3.1 Attack model

We assume that the adversary is either an outside entity or a non-administrator user (i.e., no access to root account) who seeks to extract sensitive application code or data. The adversary may have an exploitable vulnerability in the victimized application that could lead to arbitrary code execution and direct access to application secret. We assume such vulnerabilities are present when the app has fully initialized and is servicing its user. However, we presume that the program is safe from the adversary during the initialization phase of the application. We also assume a non-compromised kernel that can support the LOTRx86 architecture. Our design requires the presence of a kernel module that depends on kernel capabilities such as marking memory regions supervisor or installing custom segment descriptors. Also, our design includes Enter/Exit gates that facilitate the control transfer between the PrivUser and normal user mode. The gates amount to about 50 lines of assembly code and we assume that they are verifiable and absent of vulnerabilities.

3.2 Security guarantees

Our work focus on providing developers with an underlying architecture, a new user privilege layer, which can be leveraged to protect application secrets and also program routines that access secrets securely. Using our architecture, we guarantee that a context in normal user mode cannot directly access a region protected (as a part of the PrivUser memory space) even in the presence of vulnerabilities. The adversary cannot jump into an arbitrary location in the PrivUser memory space to leak secrets since LOTRx86 leverages the x86 privilege structures to allow only controlled invocation of routines that handle sensitive information.

On the other hand, we do not focus on the security of the code that executes in our PrivUser mode. We also argue that protection of application secret in the presence of a vulnerability in the trusted code base (PrivUser code in our case) is an unrealistic security objective for any privilege separation scheme or even hardware-based Trusted Execution Environments [12, 31]. For instance, a recent work [29] proved that vulnerabilities inside SGX could be used to disclose protected application secrets. However, we do guarantee that the PrivUser layer is architecturally confined to its privilege that it cannot modify kernel memory nor infringe upon the kernel (Ring0) privileges even in the presence of a vulnerability in the PrivUser code. As we will explain in the coming section (section 4), this is a pivotal part of our architecture design. The privilege structures and gates that exactly achieve this security guarantee is one of the key contributions of this paper.

4 LOTRX86 DESIGN

The primary goal of the LOTRx86 design is to establish a new user memory protection and access mechanism through the introduction of new user mode privilege called PrivUser. Our design eliminates the necessity for page table switching or manipulation; the access to the protected memory regions is granted based on the privilege. Our architecture approaches the problem of application memory safety at a fundamental level. Instead of leveraging the existing OS-supported protection mechanisms, we create another privilege distinction within the user program execution model that resembles the user vs. kernel privilege.

Developers can write sensitive memory handling routines into PrivUser layer, then simply place a privcall in place to invoke a routine that she defined. Below is the privcall interface that we provide to developers:

\[ \text{privcall}(\text{PRIVCALL_NR}, \ldots); \]

The privcall interface and its ABI is modeled after the Linux kernel's system call interface. The routine in PrivUser is identified with a number (e.g., PRIV_USEPKEY=3). For developers who have experience in POSIX system programming, using the privcalls to perform operations that involve application secret is intuitive.

Privileged-based memory access control. Our approach introduces a privilege-based memory access control, and it offers clear advantages over the existing process and thread level approaches. The cost of the remote procedure calls for bridging two independent processes, or the cost of page table manipulation is eliminated. In our architecture, the memory access permissions do not change when the application secret needs to be accessed. Instead, the privilege of the the execution mode is elevated to obtain access to the protected memory.

Secure invocation. privcall is a single control transfer instruction (lcall), by which a context enters PrivUser mode through the LOTRx86 Enter gate and returns upon finishing the privcall routine. Due to this design, the adversary cannot jump into an arbitrary location with the PrivUser privilege. Therefore, our architecture does not experience the security complications inherent to enable and disable models [9].

Portability. LOTRx86 does not rely on new processor features for memory protection [11–13, 17]. Instead, we re-purpose the underused privilege layers to implement PrivUser. Hence, our architecture is compatible across all generations of x86-64 processors. As we will present in section 6, we evaluated our architecture and a PoC on both Intel and AMD's x86-64 processors.

Flexible application privilege separation.

4.1 Establishing PrivUser memory space

We face formidable challenges in the process of establishing the PrivUser layer. Our design creates a distinct execution mode (PrivUser execution mode) and its address space (PrivUser address space) for
PrivUser layer. However, the resulting PrivUser layer must be intermediate, meaning that its address space should not be accessible by a user mode context, and at the same time, PrivUser execution mode must not be able to access the kernel address space. However, the x86 paging architecture provides only two memory privilege distinction: U-pages and S-pages. The memory segmentation feature that existed in x86-32 is deprecated in x86-64, eliminating an additional memory access control mechanism to paging.

In summary, PrivUser layer must satisfy the two fundamental memory access security requirements (M-SR1 and M-SR2) to function as an intermediate layer.

M-SR1. User mode must not be able to access PrivUser memory space

M-SR2. PrivUser mode must not be able to access kernel memory space

Satisfying M-SR1. We satisfy M-SR1 by mapping all pages that belong to PrivUser as S-pages to protect a user mode context from accessing PrivUser code and data. As a result, PrivUser memory space that is mapped as S-page is accessible to PrivUser mode, but not to user mode. Now, we see that we are already using both of the two privilege distinction recognized by the paging system, and we are unable to protect the kernel from PrivUser mode.

Solution for M-SR2. LOTRx86 adapts a scheme that temporarily enables segmentation when a certain code segment is in use; we enforce PrivUser mode to be a segmentation-enforced execution mode by defining it as a 32-bit segmentation-enabled code segment as shown in Table 2. This way, entering PrivUser mode changes not only the currently active code segment but also the bitness of the execution mode. That is, when user mode enters PrivUser mode through privcall the execution mode is set to the 32-bit compatibility mode. As a result, we can enforce segmentation to set boundaries to the powerful PrivUser mode (Ring2) that are capable of accessing S-pages. The resulting memory access map of the three execution mode is illustrated in Figure 1a. With our design, the PrivUser memory space serves as a functionally intermediate memory space for PrivUser.

Remaining challenge. However, we found that satisfying MSR2 is a non-trivial issue. The segmentation enforcement alone is not sufficient for ensuring M-SR2. As mentioned in subsection 3.2, we must guarantee that the PrivUser layer has architecturally well-defined memory boundary against kernel; we must ensure that all memory access under any circumstances should not be able to affect kernel. The challenge is that we must carefully inspect all possible inter-segment control transfer from PrivUser mode, to verify that PrivUser mode cannot enter a state where it can access kernel memory.

4.2 Inescapable segmentation enforcement

LOTRx86 needs to guarantee that PrivUser mode is architecturally confined. Hence, we need to ensure it cannot escape the segmentation to access kernel memory. More specifically, we need to ensure that no non-controlled (i.e., not through callgate) inter-segment control transfer paths out of the PrivUser mode arrives in a segment that is 1. has a Ring privilege numerically less than 3 (can access S-pages), 2. and is a 64-bit segment (no segmentation is enforced). We denote this control transfer security requirement for the enforcement of inescapable segmentation as CT-SR, and we explain how our privilege definitions (i.e., entries in LDT) and our gate structure (Figure 1b) satisfy the above requirement.

CT-SR. If $R_{x32} \rightarrow R_{e \times 64}$ where $e < 3$: there must be no possible non-controlled control transfer from PrivUser mode ($R_{x32}$) to a 64-bit Ring privilege $e$ that is capable of accessing S-page access privilege
**Hardware constraint and gate mode.** Along with the CT-SR, there is an x86-64 specific perk that has been proved to be a constraint in our design. The x86-64 mode (both 64-bit mode and the 32-bit compatibility mode) only supports a 64-bit mode callgate which is an extended version of its counterpart that existed in x86-32. Specifically, it does not allow the target code segment of a callgate to be a 32-bit segment. This implies that an inter-bitness control transfer through callgate is not supported in both ways; while \( CG : R_{n,x32} \rightarrow R_m \) is possible, but \( CG : R_m \rightarrow R_{n,x32} \) is an invalid callgate definition. Due to this constraint \( C \), a privilege escalation and a switch to the 32-bit mode cannot be achieved in a single callgate transfer. Therefore, we see that we need a separate 64-bit Gate mode segment to elevate privilege, then enter the 32-bit Priv\( \text{User} \) mode. However, there exists a more important reason for the existence of the 64-bit Gate mode and that its privilege \( R_g \) must be higher than that of Priv\( \text{User} \) mode \( (R_p) \).

\[ C. \ CG : R_{n} \rightarrow R_{m,x32} : \text{callgate cannot target a 32-bit code segment} \]

**Inspecting non-controlled control transfer routes.** As we explained in section 2, an uncontrolled inter-segment control transfer can be made to jump to a less privileged code segment without any security checks. Therefore we must rigorously verify all possible non-controlled transfers from \( R_{p,x32} \) to all Ring levels e that are \( e \geq 2 \) (Ring privilege levels that are numerically equal or greater, meaning equal or lower privilege). First of all, we must make sure that a context in Priv\( \text{User} \) mode cannot arbitrarily jump into an arbitrary place in Gate mode. In order to prevent a non-controlled control transfer \( R_p \rightarrow R_g \), we realize that the gate mode privilege must be higher (numerically lower in terms of Ring number). Hence the following property must hold in our design:

\[ \text{P1.} \ g < p \ (R_g \text{ is higher in privilege than } R_p) : \text{privilege of Gate mode must be higher than that of Priv\( \text{User} \) mode} \]

The second possible escape route is to perform a same-privilege inter-bitness (32bit \( \rightarrow \) 64bit) inter-segment control flow. We prevent such route by intentionally not defining a 64-bit code segment for the Ring level 2. A Ring privilege level in the x86 architecture come into existence when it is defined in the descriptor table, and a context loads the segment selector that points to the code segment through inter-segment control flow instruction. Hence, a Ring level that is not defined in the descriptor tables, does not exist within the system. Hence, by only defining a 32-bit code segment for Ring2, Ring2 becomes a 32-bit only, segmentation enforced Ring level in our system definition. We denote this property of our privilege structure design as the following:

\[ \text{P2.} \ \not\exists R_{p,x64} : \text{64-bit counter part of Priv\( \text{User} \) mode segment must not exist} \]

Our privilege definitions and gate structures (Table 2 and Figure 1b) meet the constraint C. C is satisfied by the Enter gate in Gate mode. A privcall first enters Gate mode through the CG1 into the Enter Gate (Table 2). At the gate mode, we load the stack with the following arguments: [Priv\( \text{User} \) entry point, Priv\( \text{User} \) code segment selector, Priv\( \text{User} \) stack address, Priv\( \text{User} \) stack segment selector], and then perform a far return lret to enter Priv\( \text{User} \) mode. While this control transfer is made through a non-controlled control transfer instruction, the Enter gate consisting of about 30 lines of assembly instructions are guaranteed to be executed from the beginning by the CG1. In other words, we chain a non-controlled control transfer with a controlled control transfer (CG1) to guarantee its correct execution. Our design also satisfies CT-SR by maintaining the required properties P1 and P2. We chose Ring1 as the privilege level for Gate mode while enforcing the segmentation on all Priv\( \text{User} \) mode execution by defining only 32-bit segmentation-enforced code segment for the Ring level 2. By meeting CT-SR, we complete our solution for the establishment of the Priv\( \text{User} \) memory space that satisfies both M-SR1 and M-SR2; the Priv\( \text{User} \) memory space is protected from context running in the user mode, while Priv\( \text{User} \) mode is architecturally bound to its memory space that it cannot access kernel memory under all circumstances.

## 5 PROTOTYPE IMPLEMENTATION

In this section, we explain the prototype of our LOTRx86 architecture in detail. Our prototype implementation consists of the following components:

**lotr-kmod.** We built a Linux kernel module that communicates with the host process (LOTRx86 enabled process). The module creates a virtual device interface at /dev/lotr, and an LOTRx86 enabled program communicates with our kernel module with the ioctl interface. The kernel module builds the Priv\( \text{User} \)-space for the program when requested.

**liblotr.** The user library liblotr allows developers to the use of our architecture in the host program, isolate the application secrets, and implement privcalls that securely access the secrets. A developer can initialize the Priv\( \text{User} \)-space and utilize the privcall interface through our user library. The library also includes tools and scripts for building the executable that runs in the Priv\( \text{User} \)-space.

**lotr-lib.** We provide a modified version of the musl [1] lib for building the Priv\( \text{User} \) executable. We modified the heap memory manager such that only 5-pages are allocated to the heap managers used in the Priv\( \text{User} \) mode. In this way, we prevent the leakage of the application secret and the by-products of its processing to the user space.

**lotr-build.** lotr-build is a collection of compilation scripts and tools that help developers in compiling the Priv\( \text{User} \) portion of their application and incorporating it into the host application. We further explain this procedure later.

### 5.1 Priv\( \text{User} \) mode Initialization

The lotr-kmod kernel module initializes the LOTRx86 infrastructure such as the Gate-mode, Priv\( \text{User} \) mode and control transfer structures for the host process. The host application is required to call init\_lotr\&(req) function from liblotr with an argument of the struct init\_request type during its initialization. The request structure contains the addresses and sizes of Priv\( \text{User} \) components that lotr-kmod need in its initialization routine. Such information includes the range of Priv\( \text{User} \) code segment, data segment, the entry point for the Priv\( \text{User} \)-space, pages to be used as a stack in Priv\( \text{User} \), and so forth. The addresses of the segments
are available through the symbols generated by our build tools during the compile-time, while the stack is allocated through mmap in liblotr. Additionally, lotr-kmod contains the Enter gate and Exit gate that are loaded into the kernel memory upon module load.

The lotr-kmod kernel module creates an LDT for the host process and writes the segment and callgate descriptors that are used for the Gate-mode and PrivUser mode. Unlike the GDT, an LDT is referenced on a per-process basis; an LDT can be created for each process, and the register that points to the currently active LDT called ldr is updated in each context switch. For this reason, the LOTRx86 descriptors can only be referenced by the host process that explicitly requested the initialization of the LOTRx86 infrastructure. lotr-kmod creates the descriptor segments listed in Table 2. A set of Ring1 code and data segments are used for the Gate-mode, and Ring2-32bit segment descriptors are loaded as a context enters the PrivUser mode.

The initialization also sets the Gate-mode stack to be loaded at the Enter callgate. As briefly explained in section 2, the x86 callgate mechanism finds the address of the new stack for the control transfer at the TSS structure. The TSS structure holds the addresses of each Ring levels. In our case, we use two callgates, CG(R3→R1) and CG(R2→R1), that both require a stack for Ring1. Hence, we allocate stack space and record the top of the stack in the Ring1 stack field of the TSS (TSS.SP1).

Another important task carried out during the initialization (in lotr-kmod) is marking the pages that belong to the PrivUser space Supervisor pages. The kernel module walks the page tables and marks PrivUser pages Supervisor by clearing the User bit in the page table entry. All pages that are marked Supervisors are maintained in a linked list so they can be reverted or freed when necessary as the host process terminates.

When all necessary initialization procedures are finished, the kernel module creates a lock for the host process based on its PID. From this point on, lotr-kmod ignores additional initialization request delivered via the ioctl requests from the host to thwart any possible attempt to compromise the PrivUser-space.

### 5.2 LOTRx86 ABI

The privcall interface of the LOTRx86 is almost identical to the syscall interface; privcall follows the x86-64 System V AMD64 ABI system call convention [14]. That is, we use %rax, %rdi, %rsi, %rdx, %r10, %r8, %r9 registers for passing arguments to the PrivUser mode, and the return value is stored in %rax. Underneath the surface, however, our unique design enables establishment and secure use of the PrivUser-space. From here on, we explain each stage of

| Type          | Priv.       |
|---------------|-------------|
| Gate-mode CS  | Code Segment| Ring1 |
| Gate-mode DS  | Data Segment| Ring1 |
| PrivUser mode CS | Code Segment| Ring2-x32 |
| PrivUser mode DS | Data Segment| Ring2-x32 |
| CG1           | CG(R3→R1)   | CPL ≤ 3 |
| CG2           | CG(R2→R1)   | CPL ≤ 2 |

**Table 2: LOTRx86 LDT descriptors: by defining segment and callgate descriptors in LDT, LOTRx86 creates Gate-mode and PrivUser mode for a process**

The control flow transfers in the ABI – starting from a privcall and its return to its caller.

**privcall interface.** A privcall (NR_PRIVCALL, ...) consists of layers of macros that handle a variable number of arguments and place them in the argument registers in order. After the arguments are placed according to the x86-64 syscall ABI, a long call (lcall) instruction is executed with a segment selector that points to the Enter callgate as an argument. Upon the executing of lcall, the execution continues at the Enter gate with a privilege of Ring1.

**Enter gate.** The LOTRx86 Enter gate plays a pivotal role in safeguarding the user mode context that invoked a privcall into the PrivUser mode. Figure 2 is simplified pseudo assembly code of the implementation. The Enter gate is written in assembly code and is about 30 instructions that carry out three main operations.

First, the Enter gate checks the saved $cs in the gate stack. At this point, the ring privilege has been escalated to that of the Gate mode (Ring1), stack pointer now points to Gate mode stack, and the caller context is saved in the new stack. (for detailed x86 callgate operation, revisit Algorithm 1 in section 2). The least significant 2 bits of the saved $cs ($cs[10]) indicates the caller’s Ring privilege. By ensuring the value to be 3, we prevent PrivUser mode from entering the Enter gate for possibly malicious intent.

Then the gate saves the user mode caller context in the Gate mode stack. Note that the x86 long call instruction has context saving feature built in. However, since we use the Ring1 for both

**Figure 2: Simplified pseudo assembly code of LOTRx86 Enter gates**
Enter gate and Exit gate, the saved context is overwritten when the context returns from PrivUser mode back to the Exit gate. Therefore, we found that it is necessary to perform a manual context saving of the four registers (%RIP, %CS, %RSP, %SS) in the beginning of our Enter gate as shown in the code block (b) in Figure 2.

The second operation (code block (c) in Figure 2) illustrates the transforming of the privcall arguments that follow the x86-64 calling convention into that of the PrivUser mode ABI; the in-register arguments must be transferred to the PrivUser mode stack as preparation before entering the PrivUser mode. Unlike the conventional x86-32 ABI, we use the 64-bit arguments in the PrivUser mode by default. The fact that the PrivUser mode runs in 32-bit compatibility mode but uses 64-bit length arguments is a peculiar characteristic of our design, and the LOTRx86 Enter gate resolves the calling convention discrepancy.

The last operation (code block (d)) performed in at the Enter gate is to transfer the control flow into the entry point of PrivUser mode. We push the entry point address, the address of the PrivUser mode stack that contains the arguments passed on by the privcall in the user mode at this point, and their segments (%cs and %ss) on to the current (Gate mode stack). Then, we execute the lret instruction to enter the PrivUser mode.

PrivUser entry point. The PrivUser mode entry point first performs a bound check on the %eax that contains the privcall number (i.e., \(1 \leq \text{nr\_privcall} \leq \text{MAX\_PRIVCALL}\)). The pointers to the predefined privcall routines are arranged in the Privuser Call Table (PCT) whose role is identical to the system call table in the Linux kernel. This mechanism prevents a maliciously crafted privcall from calling an arbitrary memory address. If the check is valid then the entry point calls the wrapper function for the privcall routine that corresponds to the number is invoked.

PrivUser routine. The developers can define a privcall routine through the PRIVCALL\_DEFINE(func_name,...) macro. The macro creates and exports a wrapper function that calls the main routine that corresponds to the number is invoked. First is the declaration macro #PRIVCALL\_DEFINE. The macro takes the name of the function as the first argument and up to six arguments. The type and the name of the arguments must be entered as if they are separate argument (e.g., (int, mynumber)). This is because PRIVCALL\_DEFINE generates a wrapper function that casts the ABI-defined arguments into the argument's type. We refrain from further explaining the details of the macro since it is almost identical to the kernel's SYSCALL\_DEFINE macro.

Compiling with lotr-libc. We provide gcc-lotr which is a wrapper to the gcc compiler. gcc-lotr links the user's PrivUser code with lotr-lIBC instead of the default glibc (32bit). lotr-lIBC is a modified version of musl-libc. We modified the malloc function in themusl-libc so that it manages a memory block from the PrivUser memory space 5-pages. This is to prevent the by-products or the application itself from being placed in a memory region accessible to the normal user mode. Additionally, we implemented a function that initializes process Thread Local Storage (TLS) that can be called from liblotr's init\_lotr() function. The initialization of a process TLS is performed by the libc library before the program's main() is executed. Therefore, it is necessary to implement a separate function to initialize the TLS for LOTRx86.

Argument passing. liblotr provides an argument page that is always allocated within the 32-bit address space. Developers can first copy the argument into the argument page then pass the reference to PrivUser mode. Alternatively, developers can use LOTR\_SECRET keyword to global variables to force them to be placed in a data section called .lotr\_secret which is loaded into the PrivUser memory space.

Building final executable. Compiling the PrivUser code with our build tools yields two files: a header file in which privcall numbers are defined, and a LOTRx86 object in .lotr extension. The header file lists the assigned number for each declared privcall routines, and the .lotr file is an object file ready to be linked to the main program. Our build tool compiles the PrivUser code in x86-32 code. However, we copy the sections of the 32-bit object into a new 64-bit ELF object format so that it can be linked into the main program. The PrivUser build tools also strip all the symbols to prevent symbol collision between the 32-bit libc (lotr-lIBC) and the 64-bit libc used in the main program, then it generates a symbol table that includes addresses of the PrivUser object sections and most importantly, the PrivUser entry point. The main program is built with our linker script (LOTR\_linkerscript) that loads the symbol table generated during PrivUser build. When the main program launches, init\_lotr() fetches the symbols and transfers
them to 1otr-mod, and the kernel module marks the memory pages that belong to the PrivUser memory space S-pages.

5.4 Kernel changes
The LOTRx86 prototype is implemented as a kernel module. However, we also made minor but necessary modifications to the Linux kernel. First of all, we made sure that system calls (e.g., mprotect) that alter the memory permissions of the user memory space ignore the request when the affected region includes PrivUser-memory. This is achieved by simply placing a "if-then-return -ERR" statement for the case where the address belongs to the user-space but the page is an S-page. We made a similar change to the munlock system call so that PrivUser’s P-pages are excluded from possible memory swap-outs.

6 PROOF-OF-CONCEPTS AND EVALUATION
To show the feasibility and efficiency of the LOTRx86 architecture approach, we develop a proof-of-concept (PoC) by incorporating our architecture into the Nginx web server [23] as well as the Libressl [36] that is used by the web browser to support SSL. We modified the parts of the web server to protect the in-memory private key in the PrivUser memory space and only allow accesses to the key through our privcall interface. In this section, we present the results from a set of microbenchmarks that we performed to measure the latency induced by a privcall. Then we compare the performance of the PoC web server whose private key is protected with its original version. Our experiments are conducted on both Intel and AMD to show that our approach to show that our approach is portable. It should be noted that the PoC and all examples are compiled as 64-bit programs. The specifications of the two x86 machines are as follows:

Intel-based PC: i7-4770 @ 3.40GHz, 4 cores, 16GB RAM
AMD-based PC: Ryzen 7 1800X @ 3.60GHZ, 8 cores, 32GB RAM
OS on both PC: Ubuntu 16.04 LTS, kernel version 4.13

6.1 Microbenchmarks
The privcall allows developers to invoke routines that access application secrets in the PrivUser layer. A certain amount of added latency is inevitable since we perform a chain of control transfers to securely invoke the privcall routines. In this experiment, we compare the latency of an empty privcall against the commonly used library calls to show that the added overhead is indeed a reasonable trade-off for the protection of application secrets. We also conducted the microbenchmark in two varying setups. In the first setup, we built executables that makes a single invocation of each call (privcalls, library calls), and we produced the results by executing the executables 1000 times. In the second setup, we measure the latency of a 1000 consecutive invocations of each calls using a for loop. These two setups represent the two situations where privcall is infrequently called and frequently called.

Figure 4a shows the experiment results on the Intel PC while Figure 4b shows the results from the same experiment on the AMD PC. The latency incurred by a privcall proves to be at a reasonable level. The single invocation performance is on par with the most basic library calls such as ioctl or gettid, consuming about 1500-2000 cycles on both PCs. It is noticeable that the latency of a privcall does not improve drastically as some of the other calls such as gettid whose number of cycles has dropped from 2148 to 404 on Intel PC. As to this result, we surmise that the control flow transfer chain used in our architecture affect the caching behavior of the processor negatively. Also, the libc and kernel’s system call invocation have been extremely well optimized for a long period of time. Hence, we plan to investigate possible optimizations that can be applied to LOTRx86 in future. However, LOTRx86 is the only portable solution on the x86 architecture that achieves in-process memory protection. Also it is the most efficient solution among portable solutions. We present a comparison of LOTRx86 against the traditional memory protection techniques to support our claim. Moreover, we argue that the performance overhead of LOTRx86 is at a reasonable level in an application scale. We present a macro benchmark results using our PoC (Nginx with LOTRx86).

6.2 Comparison with traditional memory protection techniques
We implemented a simple demonstration in-process memory protection using LOTRx86, page table manipulation technique implemented with mmap and mprotect, and a socket-based remote procedure call mechanism (from <rpc/rpc.h>.

Test program. Our simple program first load a password from a file into the protected memory region, then it receives an input from user via stdin to compare it against the protected password. In more detail, we implemented two functions load_password and check_password using the three protection mechanisms to evaluate their performance overhead. For page table based method,
we use mprotect to set the page that contains the load_password and check_password and the page dedicated for storing the loaded password to PROT_NONE. In case of the RPC mechanism we simply place the two measured functions and the password-storing buffer in a different process and make RPCs to execute the functions remotely.

Performance overhead comparison. The measurements for the execution time of the two functions, implemented with three different mechanisms, are illustrated in Figure 4. We averaged the results from 1000 trials (the y-axis is in log scale). The results show that LOTRx86 proves to be much faster than the two traditional methods by a large margin. On Intel PC, LOTRx86 greatly reduces the execution time (33051 cycles) of load_password by 83.11% (195661 cycles) and by 97.93% (1600291 cycles), compared to mmap- and RPC-based implementation, respectively. This is because LOTRx86 does not require page table modifications that may cause system-wide performance overhead, nor the cost of communication with an external entity.

Figure 4: Micro-benchmark privcall vs common library calls (a,b) and comparison against traditional memory protection methods (c,d)

| Isolation technique | #Cycles |
|---------------------|---------|
| LOTRx86             | 10⁴     |
| mmap                | 10⁵     |
| RPC                 | 10⁶     |

6.3 LOTRx86-enabled web server

To develop a proof-of-concept LOTRx86-enabled web server, we made changes to LibreSSL and the Nginx web browser. Specifically, we replaced parts of the software that accesses private keys with a privcall routine that performs the equivalent task. In the resulting web server’s process address space, the private key always resides in the PrivUser memory space. Therefore, any arbitrary memory access (e.g., buffer over-read in HeartBleed) are thwarted. Only through the pre-defined privcall routines, the web server can perform operations that involve the private key.

Implementation. During its initialization, Nginx loads the private key through a function called SSL_CTX_use_PrivateKey_file. This function performs a series of operations to read the private key then parsing the contents into a ASN1 structure, then the function eventually produces an RSA structure that is used by LibreSSL during SSL connections. We re-implemented the function using privcalls. In our version of the function, the opening of the file and loading its contents into memory are performed in PrivUser mode the structures that contain the private key or its processed forms, are stored in the PrivUser memory space. For passing arguments, we created a custom C structure that contains the necessary information that needs to be passed via privcalls. Once the private key is converted into a RSA data structure, it is stored safely in the PrivUser memory space until it needs to be accessed during the handshake stage in an SSL connection. During the handshake, server digitally signs a message using the private key to authenticate itself to its client. We modified the RSA_sign() such that it makes privcalls to request operations involving the
RSA structure. In more detail, we copy the message to be signed in the argument page shared between user mode and PrivUser mode that is designated by liblotr.

**Performance measurements.** We used the ab apache benchmark tool to perform a benchmark similar to the one performed in [32], a work that leverages hypervisor to achieve a similar objective to LOTRx86. Using the tool, we make 1000 KeepAlive requests to the server, then the server responds by sending a file back to the client. In the benchmark, we measured the average execution time from the socket connection to the last response from the server; The we varied the size of the requested file size to represent different configurations (we used {1k, 5k, 20k, 50k, 100k, 500k} and [32] uses {5k,20k,50k}). The results are shown in Figure 5a and Figure 5b. Note that due to the difference of CPU performance, the range of y-axes is different.

The additional performance overhead due to LOTRx86 mainly comes from the execution mode transition (user mode to PrivUser mode). A total of three privcall invocations are made in opening and loading the contents of the private key file into a buffer in PrivUser memory space, and a single privcall to sign the message using the private key. In case of 5K requested file size, a rather extreme case, LOTRx86 adds about 35% on Intel processor and 40% on AMD. However, the overhead becomes relatively irrelevant as the file size increases: 20K: 29.36% (Intel), 16.07% (AMD), 500K: 6.25% (Intel) 0% (AMD). This particular experiment tests the feasibility of LOTRx86 in latency-critical tasks, and the results show that our approach is feasible even in such cases.

![Figure 5: SSL KeepAlive response latency with varying file sizes on LOTRx86-enabled Nginx](image)

**7 RELATED WORK**

The LOTRx86 architecture proposes a method for construction of a privileged userspace is capable of protecting designated set of user application code and data. The new architecture involves the underused intermediate rings and segmentation-based memory bounds. In this section we discuss previous work on user-space memory protection and system privilege restructuring methods for system fortification. Our architecture creates a virtual intermediate memory privilege only using the existing features in the x86 architecture. There have been a number of approaches that adapted varying memory protection techniques.

**Alternative Privilege Models.** Nested kernel [16] introduced a concept of inner-kernel that takes control of the hardware MMU by deprivileging the original kernel by disabling a subset of its Ring0 power. Nested Kernel exports virtual MMU interface that allows the deprivileged kernel to request sensitive memory management operation explicitly.

The x86-32 hypervisor implementation before the introduction of Intel's hardware-assisted virtualization features such as Intel's VT-x and AMD's SVM [2, 24], made use of the intermediate Rings and segmentation to achieve virtualization. Hypervisor implementations [3, 7, 15] depriviledged the operating system kernel by making them run in the intermediate Rings then enforced segmentation to protect the in-memory hypervisor. LOTRx86 not only explores the use of the intermediate Rings on 64-bit operating systems but also for a very different purpose. Our architecture inhabits the abandoned Rings into more privileged user mode in which developers can place their sensitive application code and data.

**Use of processor features.** The segmentation feature present in x86-32 had been employed in a number of approaches for application memory protection [20, 44](and as aforementioned in early hypervisor implementations). Both Native Client(NaCl) and Vx32 provides a safe is a user-level sandbox that enables safe execution of guest plug-ins to the host program. In this regard, the attack model of these work differ from that of LOTRx86. LOTRx86 assumes that the host application is untrusted and we place sensitive code and data in the PrivUser space. Also, The Nacl sandbox has adopted SFI to compensate for the lack of segmentation in x86-64 [39]. LOTRx86 enables inescapable segmentation enforcement to construct PrivUser mode in x86-64.

Processor architectures have been extended to support user-space memory protection. Intel has recently introduced Software Guard eXtensions (SGX) that creates an enclave which a predefined set of code and data can be protected [12]. Intel also ships memory bound checking functionality called MPX [13], which provides hardware assist for bound checking that software fault isolation approaches advocated. Intel has also disclosed plans for domain-based memory partitioning that resembles the memory domain feature in the ARM architecture. More recently AMD has disclosed
a white paper describing the upcoming memory encryption feature to be added to its x86 processors [17].

The fragmented hardware support for application-level memory protection served as a pivotal motivation for our approach. Our design does not rely on any specific hardware feature and preserves portability. However, there is a difference in attack model. SGX distrusts kernel and the protected user memory within its enclave stays intact even under kernel compromise. On the other hand, LOTRx86 is incapable of operating in a trustworthy way when the kernel is compromised. Nevertheless, we argue that our work presents an unique approach that achieves in-process memory protection while preserving portability.

Process/thread level partitioning. Early work on application privilege separation focused on restructuring a program into separate processes [6, 27, 38, 44]. In essence, placing program components into process-level partitions aims to achieve complete address space separation. The approach presents a number of disadvantages. First, the approach inevitably involves an Inter-Process Communication (IPC) mechanism to establish a communication channel between the partitions so that they can remotely invoke functions (i.e., Remote Procedure Calls (RPC)) in other partitions and pass arguments as necessary.

More recent work used threads as a unit of separation compartment that prevents leakage of sensitive memory [5, 22, 26, 41]. Chen et al. [9] pointed out that even the thread compartments are still too coarse-grained, introduce a high performance overhead due to page table switches, and requires developers to make structural changes to their program. Their work Shreds takes advantage of the memory domain feature on the ARM architecture to create a secure code block within the program. However, the Domain-based memory partitioning is only available has been deprecated on AARCH64.

Unlike the approaches that retrofit the OS protection mechanisms, our approach makes a fundamental change in privilege architecture. Another significant difference is that our approach does not require address space switch nor run-time page table modifications. In our design, the privilege is what changes when granted access to the protected memory through privcall; the page tables that map the protected memory region as S-pages are intact whether the running context is in the normal user mode or in PrivUser mode.

Software-based fault isolation (SFI). Software-based fault isolation techniques [19, 20, 33, 35, 39, 40, 44] employ software techniques such as compilers or instrumentation to create logical fault domains within the same address space often to contain code execution or memory access. SFI is often used to partition an untrusted module into a sandbox to protect the host program [20, 35, 39, 44]. The attack surface that LOTRx86 covers is the exact opposite, or inverse, of their approach; our work seeks to protect a sensitive and more privileged user code and data from rest of the program.

Hypervisor-based Approaches. A number of works have leveraged hypervisors to protect applications in virtualized systems. memory [8, 21, 28, 30, 32, 34, 43]. Hypervisor-based approaches leverage hypervisor-controlled page tables and other hypervisor control over the virtualized system to ensure trustworthiness of applications and system services. Similar to SGX, hypervisor-based approaches are designed on the premise that kernel is vulnerable or possibly malicious. Additionally, these works assume the presence of a hypervisor on the system.

8 LIMITATIONS AND FUTURE WORK

LOTRx86 proposes a novel approach to application memory protection. However, the architecture is still at its infancy. We describe the limitations of the current prototype and discuss issues that needs to be addressed.

Argument passing. While the LOTRx86 supports passing of 64-bit arguments to PrivUser, PrivUser mode is confined to a 32-bit address space. Hence, PrivUser mode is incapable of accessing 64-bit pointers to further reference the members of an object. However, passing a complex object with multiple pointer members to a trusted execution mode is generally discouraged [25]. Such practices will expose the trusted execution mode to a large attack surface. We advise developers to clearly define the necessary arguments with known fixed sizes for a privcall routine.

Further optimization. We believe there is a room for further optimizations for the LOTRx86 architecture. However, finding resourceful optimization guides for using the intermediate Rings were absent due to its rare usage in modern operating systems. We plan to investigate further to find ways to improve the performance of our architecture.

SMEM/SMAP. Intel’s SMEP and SMAP [24] prevents supervisor mode (Ring0-2) to access or execute U-pages. SMEP does not affect LOTRx86 since PrivUser does not execute any code in U-pages. However, SMAP prevents PrivUser mode from accessing the argument page shared with user mode. One possible solution is to implement a system call or an ioctl call that toggles the SMAP enforcement such that PrivUser mode can fetch data from the shared page. Note that kernel’s copy_from_user API also temporarily disables SMAP to copy from user-supplied pointer to the argument.

ASLR in PrivUser memory space. We currently build PrivUser executable statically, then copy its sections into a 64-bit object that can be linked to the main program executable. While the only a small and easily verifiable amount of code should reside in PrivUser in principle, ASLR support in PrivUser memory space is one of future work.

9 CONCLUSION

We presented LOTRx86, a novel approach that establishes a new user privilege layer called PrivUser that protects and safeguards secure access to application secrets. The new PrivUser memory space is protected from user mode access. We introduced the privcall interface that provides user mode a controlled invocation mechanism of the PrivUser routines to securely perform operations involving application secrets. Our design introduced a unique privilege and control transfer structures that establish a new user mode privilege. We also explained how our design satisfies the security requirements for the PrivUser layer to have a distinct execution mode and memory space. In our evaluation, we showed that the latency added by a privcall is on par with frequently used C function calls such as ioctl and malloc. We also implemented and evaluated the LOTRx86-enabled Nginx web server that securely accesses its private key through the privcall interface. Using the Apache
ab server benchmark mark tool, we measured the average keep-alive response time of the server to find the average overhead incurred by LOTTrx86 in various response file size. The average overhead is limited to 30.40% on the Intel processor and 20.19% on the AMD processor.

REFERENCES

[1] 2018. musl libc. https://www.musl-libc.org. (2018). Last accessed Jan 23, 2018.

[2] AMD 2013. AMD64 Architecture Programmer’s Manual. AMD.

[3] Paul Barham, Boris Dragovic, Ken Fraser, Steven Hand, Tim Harris, Alex Ho, Rolf Neugebauer, Ian Pratt, and Andrew Warfield. 2003. Xen and the art of virtualization. In Proceedings of the nineteenth ACM symposium on Operating systems principles (SOSP ’03). ACM, New York, NY, USA, 164–177. https://doi.org/10.1145/845445.845462

[4] Andrew Baumann, Marcio Feinado, and Galen Hunt. 2015. Shielding Applications from an Untrusted Cloud with Haven. ACM Trans. Comput. Syst. 33, 3, Article 8 (Aug. 2015), 26 pages. https://doi.org/10.1145/2799647

[5] Andrea Bittau, Petr Marchenko, Mark Handley, and Brad Karp. 2008. Wedge: Splitting Applications into Reduced-privilege Compartments. In Proceedings of the 5th USENIX Symposium on Networked Systems Design and Implementation (NSDI’08). USENIX Association, Berkeley, CA, USA, 309–322. http://dl.acm.org/citation.cfm?id=1385693.1387611

[6] David Brumley and Dawn Song. 2004. Privatron: Automatically Partitioning Programs for Privilege Separation. In Proceedings of the 13th Conference on USENIX Security Symposium - Volume 13 (SYM’04). USENIX Association, Berkeley, CA, USA, 5–5. http://dl.acm.org/citation.cfm?id=1251375.1251380

[7] Edouard Bugnion, Scott Devine, Mender Rosenblum, Jeremy Sugerman, and Edward Y. Wang. 2012. Bringing Virtualization to the X86 Architecture with the Original VMware Workstation. ACM Trans. Comput. Syst. 30, 4, Article 12 (Nov. 2012), 51 pages. https://doi.org/10.1145/2283553.2283554

[8] Xiaoxion Chen, Tal Garfinkel, E. Christopher Lewis, Pratap Subramaniam, Carl A. Waldspurger, Dan Boneh, Jeffrey Dvoskin, and Dan R.K. Ports. 2008. Overshadow: A Virtualization-based Approach to Retricting Protection in Commodity Operating Systems. SIGPLAN Not. 43, 3 (March 2008), 2–13. https://doi.org/10.1145/1353536.1346284

[9] Y. Chen, S. Reymondjohnson, Z. Sun, and L. Lu. 2016. Shreds: Fine-Grained Execution Units with Private Memory. In 2016 IEEE Symposium on Security and Privacy (SP). 56–71. https://doi.org/10.1109/SP.2016.12

[10] Jonathan Corbet. 2012. Supervisor mode access prevention. https://lwn.net/Articles/643797/(2015).

[11] Jonathan Corbet. 2012. Supervisor mode access prevention. https://lwn.net/Articles/643797/(2015).

[12] Linux Kernel Organization. 2018. The Linux Kernel Archives. https://www.kernel.org. (2018). Last accessed April 2, 2018.

[13] Greg Morrisett, Gang Tan, Joseph Tassarotti, Jean-Baptiste Tristan, and Edward Y. Wang. 2012. Bringing Virtualization to the x86 Architecture with the Original VMware Workstation. ACM Trans. Comput. Syst. 30, 4, Article 12 (Nov. 2012), 51 pages. https://doi.org/10.1145/2283553.2283554

[14] ARM Limited. 2009. Building a Secure System using TrustZone® Technology. http://infocenter.arm.com/help/topic/com.arm.doc.prd29-genc-009492c_trustzone_security_whitepaper.pdf. (2009).

[15] Yutao Liu, Tianyu Zhou, Kexin Chen, Haibo Chen, and Yubin Xia. 2015. Thwarting Memory Disclosure with Efficient Hypervisor-enforced Intra-domain Isolation. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS ’15). ACM, New York, NY, USA, 1607–1619. https://doi.org/10.1145/2810103.2813690

[16] Stephen McCamant and Greg Morrisett. 2006. Evaluating SFI for a CISC Architecture. In Proceedings of the 15th Conference on USENIX Security Symposium - Volume 15 (USENIX-S06). USENIX Association, Berkeley, CA, USA, Article 15. http://dl.acm.org/citation.cfm?id=1267336.1267351

[17] J. M. McCune, Y. Li, N. Qu, Z. Zhou, A. Datta, V. Gilgor, and A. Perrig. 2010. TrustVisor: Efficient TCB Reduction and Attestation. In 2010 IEEE Symposium on Security and Privacy. 143–158. https://doi.org/10.1109/SP.2010.17

[18] Greg Morrisett, Gang Tan, Joseph Tassarotti, Jean-Baptiste Tristan, and Edward G. W. 2012. RockSalt: Better, Faster, Stronger SFI for the x86. In Proceedings of the 4th Conference on Internet Measurement Conference (IMC ’12). ACM, New York, NY, USA, 395–404. https://doi.org/10.1145/2254604.2254611

[19] OpenBSD. 2017. Libressl. http://www.libressl.org. (2017). Last accessed Feb 27, 2018.

[20] Linux Kernel Organization. 2018. The Linux Kernel Archives. https://www. kernel.org. (2018). Last accessed April 2, 2018.

[21] Niels Provos, Markus Friedl, and Peter Honeyman. 2003. Preventing Privilege Escalation. In Proceedings of the 2003 USENIX Security Symposium - Volume 2003 (USENIX Security Symposium - Volume 12 (SYM’03). USENIX Association, Berkeley, CA, USA, 16–16. http://dl.acm.org/citation.cfm?id=1251353.1251369

[22] Robert Wahle, Steven Lacco, Thomas E. Anderson, and Susan L. Graham. 1993. Efficient Software-based Fault Isolation. In Proceedings of the Fourteenth ACM Symposium on Operating Systems Principles (SOSP ’93). ACM, New York, NY, USA, 283–296. https://doi.org/10.1145/164619.164635

[23] Jun Wang, Xi Xiong, and Peng Liu. 2015. Between Mutual Trust and Mutual Distrust: Practical Fine-grained Privilege Separation in Multithreaded Applications. In Proceedings of the 2015 Conference on USENIX Annual Technical Conference (USENIX ATC ’15). USENIX Association, Berkeley, CA, USA, 361–373.

[24] Youngjin Kwon, Alan M. Dunn, Michael Z. Lee, Owen S. Hofmann, Yuanzhong Xu, and Emmitt Mitchel. 2016. Seqo: Pervasive Trusted Metadata for Efficiently Verified Untrusted Systems Service. SIGOPS Oper. Syst. Rev. 50, 2 (March 2016), 277–290. https://doi.org/10.1145/2946801.2873732

[25] Fayeohny Lue, Jinsoon Jang, Noh Young, Kyaw, Yesuel Choi, Changho Choi, Taesoo Kim, Marcus Feinado, and Brent ByungHoon Kang. 2017. Hacking in Darkness: Return-oriented Programming against Secure Enclaves. In 26th USENIX Security Symposium (USENIX Security 17). USENIX Association, Vancouver, BC, 525–539. https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/lee-fayeohny
[42] D. A. Wheeler. 2014. Preventing Heartbleed. Computer 47, 8 (Aug 2014), 80–83. https://doi.org/10.1109/MC.2014.217

[43] Jisoo Yang and Kang G. Shin. 2008. Using Hypervisor to Provide Data Secrecy for User Applications on a Per-page Basis. In Proceedings of the Fourth ACM SIGPLAN/SIGOPS International Conference on Virtual Execution Environments (VEE ’08). ACM, New York, NY, USA, 71–80. https://doi.org/10.1145/1346256.

[44] Bennet Yee, David Sehr, Gregory Dardyk, J. Bradley Chen, Robert Muth, Tavis Ormandy, Shiki Okasaka, Neha Narula, and Nicholas Fullagar. 2009. Native Client: A Sandbox for Portable, Untrusted x86 Native Code. In Proceedings of the 2009 30th IEEE Symposium on Security and Privacy (SP ’09). IEEE Computer Society, Washington, DC, USA, 79–93. https://doi.org/10.1109/SP.2009.25