Lock-in-Pop: Securing Privileged Operating System Kernels by Keeping on the Beaten Path

Yiwen Li, Brendan Dolan-Gavitt, Sam Weber, and Justin Cappos, ***New York University***

https://www.usenix.org/conference/atc17/technical-sessions/presentation/li-yiwen

This paper is included in the Proceedings of the 2017 USENIX Annual Technical Conference (USENIX ATC '17).

July 12–14, 2017 • Santa Clara, CA, USA

ISBN 978-1-931971-38-6

Open access to the Proceedings of the 2017 USENIX Annual Technical Conference is sponsored by USENIX.

Lock-in-Pop: Securing Privileged Operating System Kernels by Keeping on the Beaten Path

Yiwen Li Brendan Dolan-Gavitt Sam Weber Justin Cappos
New York University

Abstract

Virtual machines (VMs) that try to isolate untrusted code are widely used in practice. However, it is often possible to trigger zero-day flaws in the host Operating System (OS) from inside of such virtualized systems. In this paper, we propose a new security metric showing strong correlation between "popular paths" and kernel vulnerabilities. We verify that the OS kernel paths accessed by popular applications in everyday use contain significantly fewer security bugs than less-used paths. We then demonstrate that this observation is useful in practice by building a prototype system which ***locks*** an application into using only ***popular*** OS kernel paths. By doing so, we demonstrate that we can prevent the triggering of zero-day kernel bugs significantly better than three other competing approaches, and argue that this is a practical approach to secure system design.

1 Introduction

The number of attacks involving the exploitation of zero-day vulnerabilities more than doubled from 2014 to 2015 [52]. Skilled hackers can find a security flaw in a system and use it to hold the system's users hostage, e.g., by gaining root access and compromising the host [25]. Similarly, zero-day vulnerabilities can be exploited [17] or their presence not be acknowledged [30] by government agencies, thus rendering millions of devices vulnerable.

In theory, running a program in an operating-system-level virtual machine (OSVM) like Docker [15] or LXC [28] should prevent bugs in the host OS kernel from triggering. However, the isolation provided by such systems is not the whole answer and faces some significant drawbacks. To be effective, the OSVM's software must not contain any bugs that could allow the program to escape the machine's containment and interact directly with the host OS. Unfortunately, these issues are very common in OSVMs, with 14 CVE vulnerabilities confirmed for Docker [14] since 2014. The large amount of complex code needed to run such a system increases the odds that flaws will be present, and, in turn, that tens of millions of user machines could be at risk [25]. Furthermore, isolation will not work if a malicious program can access even a small portion of the host OS's kernel that contains a zero-day flaw [12]. Both of these drawbacks reveal the key underlying weakness in designing OSVM systems – a lack of information as to which parts of the host kernel can be safely exported to user programs.

Several attempts have been made to find a reliable metric to pinpoint where bugs are most likely to be in kernel code. A number of previous studies have suggested that older code may be less vulnerable than new code [32] or that certain parts (such as device drivers) of the kernel [10] may be more bug-prone than others. To these hypotheses, we add a new security metric idea, called "popular paths." Positing that bugs in the popular paths, associated with frequently-used programs, are more likely to be found in software testing because of the numerous times they are executed by diverse pieces of software, we propose that kernel code found in these paths would have less chance of containing bugs than code in less-used parts of the kernel. We perform a quantitative analysis of resilience to flaws in two versions of the Linux kernel (version 3.13.0 and version 3.14.1), and find that only about 3% of the bugs are present in popular code paths, despite these paths accounting for about one-third of the total reachable kernel code. When we test our "popular paths" metric against the two aforementioned "code age" and "device drivers" metrics, we find our "popular paths" metric is much more effective (Section 3.2).

This key information inspired the idea that if we could design virtual machines that use only "popular kernel paths," a strategy we have dubbed *Lock-in-Pop*, it would greatly increase resilience to zero-day bugs in the host OS kernel. Yet using such a design scheme creates a few challenges that would need to be overcome. These include:

- It might not be possible in real-life codebases to completely avoid "unpopular paths." If other applications, or future versions of applications we tested, frequently require the use of "unpopular paths," would this make our metric untenable? (Section 4.2)

- The exploits that adversaries use change over time. Could our observation that "popular paths" are safer be only an artifact of when we did our measurements, and not be predictive of future exploits? (Section 3.2)

- Lastly, can developers make use of this observation in a practical setting? That is, is it feasible for developers to actively try to avoid unpopular code paths? (Section 4.3)

While we address some of these challenges in developing the *Lock-in-Pop* design, we want to test how well a system could function if it forced applications to use only popular kernel paths. To conduct these tests, we built a prototype system, called Lind. For Lind, we pick two key components – Google's Native Client (NaCl) [51] and Seattle's Repy [8]. NaCl serves as a computational module that isolates binaries, providing memory safety for legacy programs running in our OSVM. It also passes system calls invoked by the program to the operating system interface, called SafePOSIX. SafePOSIX re-creates the broader POSIX functionalities needed by applications, while being contained within the Repy sandbox. An API in the sandbox only allows access to popular kernel paths, while the small (8K LOC) sandbox kernel of Repy isolates flaws in SafePOSIX to prevent them from directly accessing the host OS kernel.

To test the effectiveness of Lind and our "popular paths" metric, we replicated 35 kernel bugs discovered in Linux kernel version 3.14.1. We attempted to trigger those bugs in Lind and three other virtualized environments, including Docker [15], LXC [28], and Graphene [43]. In this study, our evaluation was focused on comparing operating-system-level virtualization containers, such as Docker and LXC, and library OSes, such as Graphene. We excluded bare-metal hypervisors [4, 46], hardware-based virtualization [3, 22] and full virtualization virtual machines, such as VirtualBox [45], VMWare Workstation [47], and QEMU [37]. While our "popular paths" metric may potentially apply to those systems, a direct comparison is not possible since they have different ways of accessing hardware resources, and would require different measurement approaches.

Our results show that applications in Lind are substantially less likely to trigger kernel bugs. By doing so, we demonstrate that forcing an application to use only popular OS paths can be an effective and practical method to improve system security. Armed with this knowledge, the *Lock-in-Pop* principle can be adapted to incorporate other OSVM design configurations.

In summary, the main contributions of this paper are as follows:

- We propose a quantitative metric that evaluates security at the line-of-code level, and verify our hypothesis that "popular paths" have significantly fewer security bugs than other paths.
- Based on the "popular paths" metric, we develop a new design scheme called *Lock-in-Pop* that accesses only popular code paths through a very small trusted computing base. The need for complex functionality is addressed by re-creating riskier system calls in a memory-safe programming language within a secure sandbox.
- To demonstrate the practicality of the "popular paths" metric, we build a prototype virtual machine, Lind, using the *Lock-in-Pop* design, and test its effectiveness against three other virtual machines. We find that Lind exposes 8-12x fewer zero-day kernel bugs.

2 Goals and Threat Model

In this section, we define the scope of our efforts. We also briefly note why this study does not evaluate a few existing design schemes.

Goals. Ultimately, our goal is to help designers create systems that allow untrusted programs to run on unpatched and vulnerable host OSes without triggering vulnerabilities that attackers could exploit. Developing effective defenses for the host OS kernel is essential as kernel code can expose privileged access to attackers that could lead to a system takeover.

Our hypothesis is that OS kernel code paths that are frequently used receive more attention and therefore are less likely to contain security vulnerabilities. Our approach will be to test this hypothesis and explore the feasibility of building more secure virtualization systems, such as guest OSVMs, system call interposition modules, and library OSes, by forcing untrusted applications to stay on popular kernel code paths.

Threat model. When an attack attempt is staged on a host OS in a virtualization system, the exploit can be done either directly or indirectly. In a direct exploit, the attacker accesses a vulnerable portion of the host OS's kernel using crafted attack code. In an indirect exploit, the attacker first takes advantage of a vulnerability in the virtualization system itself (for example, a buffer overflow vulnerability) to escape the VM's containment. Once past the containment, the attacker can run arbitrary code in the host OS. The secure virtualization system design we propose in Section 4 can prevent both types of attacks effectively.

Based on the goals mentioned above, we make the following assumptions about the potential threats our system could face:

- The attacker possesses knowledge of one or more unpatched vulnerabilities in the host OS.
- The attacker can execute any code in the secure virtualization system.
- If the attack program can trigger a vulnerability in any privileged code, whether in the host OS or the secure virtualization system, the attacker is then considered successful in compromising the system.

3 Developing a Quantitative Metric for Evaluating Kernel Security

If we knew which lines of code in the kernel are likely to contain zero-day bugs, we could try to avoid using them in an OSVM. In this section, we formulate and test a quantitative evaluation metric that can indicate which lines of code are likely to contain bugs. This metric is based on the idea that kernel paths executed by popular applications during everyday use are less likely to contain security flaws. The rationale is that these code paths are well-tested due to their constant use, and thus fewer bugs can go undetected. Our initial tests yielded promising results. Additionally, when tested against two earlier strategies for predicting bug locations in the OS kernel, our metric compared favorably.

3.1 Experimental Setup

We used two different versions of the Linux kernel in our study. Since our findings for these versions are quantitatively and qualitatively similar, we report the results for 3.13.0 in this section and use 3.14.1 in Section 5. To trace the kernel, we used `gcov` [19], a standard program profiling tool in the GCC suite. The tool indicates which lines of kernel code are executed when an application runs.

Popular kernel paths. To capture the popular kernel paths, we used two strategies concurrently. First, we attempted to capture the normal usage behavior of popular applications. To do this, two students used applications from the 50 most popular packages in Debian 7.0 (omitting libraries, which are automatically included by packages that depend on them) according to the Debian Popularity Contest [1], which tracks the usage of Debian packages on an opt-in basis. Each student used 25 applications for their tasks (e.g., writing, spell checking, printing in a text editor, or using an image processing program). These tests were completed over 20 hours of total use over 5 calendar days.

The second strategy was to capture the total range of applications an individual computer user might regularly access. The students used the workstation as their desktop machine for a one-week period. They did their homework, developed software, communicated with friends and family, and so on, using this system. Software was installed as needed. From these two strategies, we obtained a profile of the lines of kernel code that defined our popular kernel paths. We make these traces publicly available to other researchers [24], so they may analyze or replicate our results.

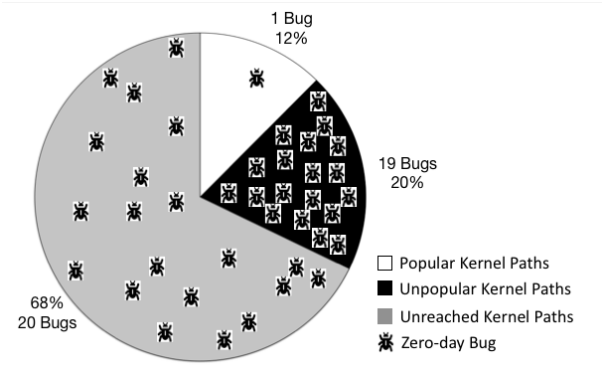

Figure 1: Percentage of different kernel areas that were reached during LTP and Trinity system call fuzzing experiments, with the zero-day kernel bugs identified in each area.

Reachable kernel paths. There are certain paths in the kernel, such as unloaded drivers, that are unreachable and unused. To determine which paths are unreachable, we used two techniques. First, we performed system call fuzzing with the Trinity system call fuzz tester [42]. Second, we used the Linux Test Project (LTP) [26], a test suite written with detailed kernel knowledge.

Locating bugs. Having identified the kernel paths used in popular applications, we then investigated how bugs are distributed among these paths. We collected a list of severe kernel bugs from the National Vulnerability Database [31]. For each bug, we found the patch that fixed the problem and identified which lines of kernel code were modified to remove it. For the purpose of this study, a user program that can execute a line of kernel code changed by such a patch is considered to have the *potential to exploit that flaw*. Note that it is possible that, in some situations, this will over-estimate the exploitation potential because reaching the lines of kernel code where a bug exists does not necessarily imply a reliable, repeatable capability to exploit the bug.

3.2 Results and Analysis

Bug distribution. The experimental results from Section 3.1 show that only one of the 40 kernel bugs tested for was found among the popular paths, even though these paths make up 12.4% of the kernel (Figure 1).

To test the significance of these results, we performed a power analysis. We assume that kernel bugs appear at

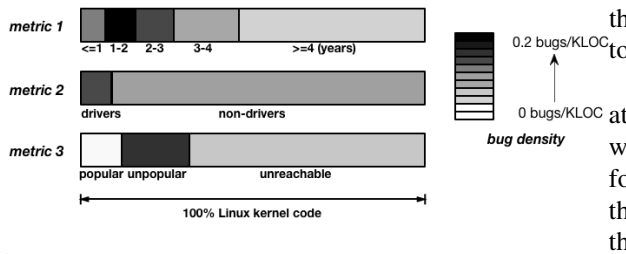

Figure 2: Bug density comparison among three metrics.

an average rate proportional to the number of lines of kernel code. Therefore, consistent with prior research [29], the rate of defect occurrence per LOC follows a Poisson distribution [35]. The premise we tested is that bugs occur at different rates in different parts of the kernel, i.e., that the less popular kernel portion has more bugs.

We first divided the kernel into two sets, A and B, where bugs occur at rates λ_A and λ_B, and $\lambda_A \neq \lambda_B$. In this test, A represents the popular paths in the kernel, while B addresses the less commonly-used paths. Given the null hypothesis that the rate of defect occurrences is the *same* in set A and B (or bugs in A and B are drawn from the same Poisson distribution), we used the Uniformly Most Powerful Unbiased (UMPU) test [39] to compare unequal-sized code blocks. At a significance level of $\alpha = 0.01$, the test was significant at $p = 0.0015$, rejecting the null hypothesis. The test also reported a 95% confidence that $\lambda_A/\lambda_B \in [0.002, 0.525]$. This indicates that the ratio between the bug rates is well below 1. Since B has a bug rate much larger than that of A, this result shows that popular paths have a much lower bug rate than unpopular ones.

Comparison with other security metrics. Ozment, et al. [32] demonstrated that older code in the Berkeley Software Distribution (BSD) [7] kernel tended to have fewer bugs (metric 1). To test Ozment's metric using our Linux bug dataset, we separated the code into five different age groups. Our results (Figure 2) showed a substantial number of bugs located in each group, and not just in the newer code. Therefore, buggy code in the Linux kernel cannot be identified simply by this age-based metric. In addition, this metric would seem to have limited use for designing a secure virtualization system, as no system could run very long exclusively on old code.

Another metric, reported by Chou, et al. [10], showed that certain parts of the kernel, particularly device drivers, were more vulnerable than others (metric 2). Applying this metric on our dataset, we found that the driver code in our version of the Linux kernel accounted for only 8.9% of the total codebase, and contained just 4 out of the 40 bugs (Figure 2). One reason for this is that after Chou's study was published system designers focused efforts on improving driver code. Palix [33] found that `drivers` now has a lower fault rate than other directories, such as `arch` and `fs`.

Additionally, there are other security metrics that operate at a coarser granularity, e.g., the file level. However, when our kernel tests were run at a file granularity, we found that even popular programs used parts of 32 files that contained flaws. Yet, only one bug was triggered by those programs. In addition, common programs tested at this level also executed 36 functions that were later patched to fix security flaws, indicating the need to localize bugs at a finer granularity.

To summarize, our results demonstrate that previously proposed security metrics show only weak correlation between the occurrence of bugs and the type of code they target. In contrast, our metric (metric 3) provides an effective and statistically significant means for predicting where in the kernel exploitable flaws will likely be found. For the remainder of the paper, we will focus on using our "popular paths" metric to design and build secure virtualization systems.

4 A New Design for Secure Virtualization Systems

In the previous section we have shown that "popular paths" correlate in a statistically significant manner with security. Next, we want to demonstrate that our "popular paths" metric is useful in practice for designing secure virtualization systems. We first briefly discuss the limitations faced by existing methods, due to the lack of a good security metric. We then discuss our new design scheme named *Lock-in-Pop*, which follows our metric by accessing only popular code paths.

4.1 Previous Attempts and Their Limitations

System call interposition (SCI). SCI systems [20, 48] filter system calls to mediate requests from untrusted user code instead of allowing them to go directly to the kernel. The filter checks a predefined security policy to decide which system calls are allowed to pass to the underlying kernel, and which ones must be stopped.

This design is limited by its overly complicated approach to policy decisions and implementation. To make a policy decision, the system needs to obtain and interpret the OS state (e.g., permissions, user groups, register flags) associated with the programs it is monitoring. The complexity of OS states makes this process difficult and can lead to inaccurate policy decisions.

Functionality re-creation. Systems such as Drawbridge [36], Bascule [5], and Graphene [43] can provide richer functionality and run more complex programs than most systems built with SCI alone because they have their own interfaces and libraries. We label such a design as "functionality re-creation."

The key to this design is to not fully rely on the underlying kernel for system functions, but to re-create its own

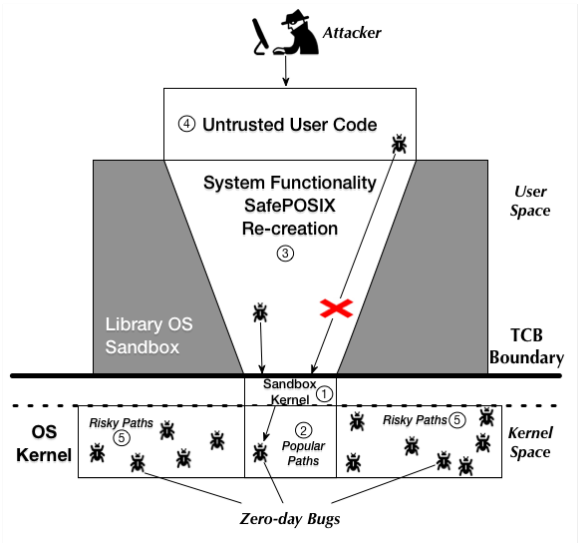

Figure 3: *Lock-in-Pop* design ensures safe execution of untrusted user code despite existing potential zero-day bugs in the OS kernel.

system functionality. When it has to access resources, like memory, CPU, and disk storage, the system accesses the kernel directly with its underlying TCB code.

Functionality re-creation provides a more realistic solution to building virtualization systems than earlier efforts. However, functionality re-creation has two pitfalls: first, if the re-created functionality resides in the TCB of the virtualization system, then vulnerabilities there can expose the host OS to attack as well. For example, hundreds of vulnerabilities have been reported in existing virtualization systems, such as QEMU and VMWare, over the past ten years [31].

Second, functionality re-creation may assume that the underlying host kernel is correct. As we have seen, this assumption is often incorrect: host kernels may have bugs in their implementation that leave them vulnerable to attack. Thus, to provide the greatest assurance that the host kernel will not be exposed to malicious user programs, a secure functionality re-creation design should try to deliberately avoid kernel paths that are likely to contain flaws. We discuss this approach in detail next.

4.2 Lock-in-Pop: Staying on the Beaten Path

Recall that we want to show that the "popular paths" metric can be used in practice. We do so by devising a design in which all code, including the complex part of the operating system interface, accesses only popular kernel paths through a small TCB. As it "locks" all functionality requests into only the "popular paths," we call this design *Lock-in-Pop*.

At the lowest level of the design (interfacing with the host OS) is the sandbox kernel (① in Figure 3). The sandbox kernel's main role is to ensure that only popular paths (② in Figure 3) of the host OS's kernel can be accessed. The sandbox kernel could thus function as a very granular system call filter, or as the core of a programming language sandbox. Note that the functionality provided by the sandbox kernel is (intentionally) much less than what an application needs. For example, an application may store files in directories and set permissions on those files. The sandbox kernel may provide a much simpler abstraction (e.g., a block storage abstraction), so long as the strictly needed functionality (e.g., persistent storage) is provided.

Constructing the sandbox kernel is not dependent on any specific technique or programming language. Instead, the sandbox kernel follows a central design principle to include only simple and necessary system calls with basic flags, which can be checked to verify that only "popular paths" are used. The sandbox kernel should start with building-block functions to first form a minimum set of system calls. To give one example, for network programs, opening a TCP connection would be considered an essential function. We can verify that the lines of kernel code that correspond to opening TCP sockets, such as lines in `void tcp_init_sock(struct sock *sk)`, are included in the "popular paths" for that system, and so decide to include the `open_tcp_connection()` function in the sandbox kernel. Examples of other necessary functions are `file.open`, `file.close`, `file.read`, and `file.write` for filesystem functions, and `create_thread`, `create_lock`, `lock.acquire`, and `lock.release` for threading functions.

In order to make security our priority, the designed sandbox kernel should only use a subset of the "popular paths." For systems where security is not as critical, trade-offs can certainly be made to include some "unpopular paths" to accommodate applications. Further discussion of this trade-off is beyond the scope of this paper, though we acknowledge it is an issue that should be addressed as *Lock-in-Pop* is deployed. While restricting the system call interface is a big hammer for limiting access to "popular paths" in the kernel, we believe that this is the best choice available, given that we do not want to require modification to the kernel, and would like to allow users to easily run their applications without much extra effort.

The application is provided more complex functionality via the SafePOSIX re-creation (③ in Figure 3). SafePOSIX has the needed complexity to build more convenient higher-level abstractions using the basic functionality provided by the sandbox kernel. The SafePOSIX re-creation is itself isolated within a library OS sandbox, which forces all system calls to go through the sandbox kernel. So long as this is performed, all calls from the SafePOSIX re-creation will only touch the permitted

(popular) kernel paths in the underlying host OS.

Similarly, untrusted user code (④ in Figure 3) also must be restricted in the way in which it performs system calls. System calls must go through the SafePOSIX re-creation, into the sandbox kernel, and then to the host OS. This is done because if user code could directly make system calls, it could access any desired path in the host OS's kernel, and thus exploit bugs within it.

Note that it is expected that bugs will occur in many components, including both the non-popular (risky) kernel paths (⑤ in Figure 3), and in the SafePOSIX re-creation. Even the user program will be buggy or perhaps explicitly malicious (created by attackers). Since the remaining components (① and ② in Figure 3) are small and can be thoroughly tested, this leads to a lower risk of compromise.

4.3 Implementation of Lock-in-Pop

To test the practicality of the "popular paths" metric and our *Lock-in-Pop* design, we implement a prototype virtual machine called Lind.[1] The purpose of building the Lind prototype is to demonstrate that our "popular paths" metric is practical, and that developers can build secure systems using it. Lind is divided into a *computational module* that enforces software fault isolation (SFI) and a *SafePOSIX module* that safely re-creates the OS functionality needed by user applications. We use a slightly modified version of Native Client (NaCl) [51] for the computational module; SafePOSIX is implemented using Restricted Python (Repy) [8] and supports complex user applications without exposing potentially risky kernel paths.

In this section we provide a brief description of these components and how they were integrated into Lind, followed by an example of how the system works.

4.3.1 Primary Components

Native Client. We use NaCl to isolate the computation of the user application from the kernel. NaCl allows Lind to work on most types of legacy code. It compiles the programs to produce a binary with software fault isolation. This prevents applications from performing system calls or executing arbitrary instructions. Instead, the application will call into a small, privileged part of NaCl that forwards system calls. In NaCl's original implementation, these calls would usually be forwarded to the host OS kernel. In Lind, we modified NaCl to instead forward these calls to our SafePOSIX re-creation (described in detail below).

Repy Sandbox. To build an API that can access the safe parts of the underlying kernel while still supporting existing applications, we need two things. First, we need a restricted sandbox kernel that only allows access to popular kernel paths. We used Seattle's Repy [8] sandbox to perform this task. Second, we have to provide complex system functions to user programs. For this task we created SafePOSIX, which implements the widely accepted standard POSIX interface on top of Repy.

Because the sandbox kernel is the only code that will be in direct contact with host system calls, it should be small (to make it easy to audit), while providing primitives that can be used to build more complex functionality. We used Seattle's Repy system API due to its tiny (around 8K LOC) sandbox kernel and its minimal set of system call APIs needed to build general computational functionality. Repy allows access only to the popular portions of the OS kernel through 33 basic API functions, including 13 network functions, 6 file functions, 6 threading functions, and 8 miscellaneous functions (Table 1) [8, 38].

Repy is only one possible implementation of the sandbox kernel built for our *Lock-in-Pop* design. It was chosen because it starts with basic building-block functions and tries to be conservative in what underlying kernel functionality it uses. Repy was designed and implemented before our "popular paths" study, and so it was not a perfect match, but it we experimentally verified that it uses a subset of the "popular paths." As reported in our evaluation (Section 5.3), Repy accessed a subset (around 70% to 80%) of the "popular paths."

Our current implementation does not end up using all of the "popular paths." It is certainly safe to use fewer paths than are available, but it is possible that we are missing out on some performance or compatibility gains. As we extend our prototype, the "popular path" metric will allow us to check whether new APIs we add expose potentially unsafe kernel code to applications in the sandbox.

4.3.2 Enhanced Safety in Call Handling with SafePOSIX Re-creation

The full kernel interface is extremely rich and hard to protect. The *Lock-in-Pop* design used to build Lind provides enhances safety protection through both isolation and a POSIX interface (SafePOSIX). The latter re-creates risky system calls to provide full-featured API for legacy applications, with minimal impact on the kernel.

In Lind, a system call issued from user code is received by NaCl, and then redirected to SafePOSIX. To service a system call in NaCl, a server routine in Lind marshals its arguments into a text string, and sends the call and the arguments to SafePOSIX. The SafePOSIX re-creation services the system call request, marshals the result, and returns it back to NaCl. Eventually, the result is returned as the appropriate native type to the calling program.

SafePOSIX is safe because of two design principles. First, its re-creation only relies on a small set of basic Repy functions (Table 1). Therefore, the interac-

[1] Lind is an old English word for a lightweight, but still strong shield constructed from two layers of linden wood.

Repy Function	Available System Calls
Networking	*gethostbyname, openconnection, getmyip, socket.send, socket.receive, socket.close, listenforconnection, tcpserversocket.getconnection, tcpserversocket.close, sendmessage, listenformessage, udpserversocket.getmessage, and udpserversocket.close.*
File System I/O Operations	*openfile(filename, create), file.close(), file.readat(size limit, offset), file.writeat(data, offset), listfiles(), and removefile(filename).*
Threading	*createlock, sleep, lock.acquire, lock.release, createthread, and getthreadname.*
Miscellaneous Functions	*getruntime, randombytes, log, exitall, createvirtualnamespace, virtualnamespace.evaluate, getresources, and getlasterror.*

Table 1: Repy sandbox kernel functions that support Lind's SafePOSIX re-creation.

tion with the host OS kernel is strictly controlled. Second, the SafePOSIX re-creation is run within the Repy programming language sandbox, which properly isolates any bugs inside SafePOSIX itself.

5 Evaluation

To demonstrate that our "popular paths" metric is useful and practical, we used our Lind prototype as a testing tool. We compared Lind against three existing virtualization systems – Docker, LXC, and Graphene. We chose these three systems because they currently represent the most widely-used VM design models for securing the OS kernel. LXC is a well-known container designed specifically for the Linux kernel. Docker is a widely-used container that wraps an application in a self-contained filesystem, while Graphene is an open source library OS designed to run an application in a virtual machine environment. Lastly, we also tested Native Linux to serve as a baseline for comparison. Our tests were designed to answer four fundamental questions:

How does Lind compare to other virtualization systems in protecting against zero-day Linux kernel bugs? (Section 5.1)

How much of the underlying kernel code is exposed, and is thus vulnerable in different virtualization systems? (Section 5.2)

If Lind's SafePOSIX construction has bugs, how severe an impact would this vulnerability have? (Section 5.3)

In the Lind prototype, what would be the expected performance overhead in real-world applications? Can developers make use of the "popular paths" metric to develop practical systems? (Section 5.4)

5.1 Linux Kernel Bug Test and Evaluation

Setup. To evaluate how well each virtualization system protects the Linux kernel against reported zero-day bugs, we examined a list of 69 historical bugs that had been identified and patched in versions 3.13.0 and 3.14.1 of the Linux kernel [13]. By consulting the National Vulnerability Database (NVD) [31], we obtained a list of all CVEs [11] that were known to exist in these Linux kernel versions as of September 2015; we found 69 such vulnerabilities. By analyzing security patches for those bugs, we were able to identify the lines of code in the kernel that correspond to each one.

In the following evaluation, we assume that a bug is potentially triggerable if the lines of code that were changed in the patch are reached (i.e., the same metric described in Section 3). This measure may overestimate potential danger posed by a system since simply reaching the buggy code does not mean that guest code actually has enough control to exploit the bug. However, this overestimate should apply equally to all of the systems we tested, which means it is still a useful method of comparison.

Next, we sought out proof-of-concept code that could trigger each bug. We were able to obtain or create code to trigger nine out of the 69 bugs [16]. For the rest, we used the Trinity system call fuzzer [42] on Linux 3.14.1 (referred to as "Native" Linux in Table 2). By comparing the code reached during fuzzing with the lines of code affected by security patches, we were able to identify an additional 26 bugs that could be triggered. All together, we identified a total of 35 bugs that we were able to trigger from user space, and these formed our final dataset for the evaluation.

We then evaluated the protection afforded by four virtualization systems (including Lind) by attempting to trigger the 35 bugs from inside each one. The host system for each test ran a version of Linux 3.14.1 with gcov instrumentation enabled. For the nine bugs that we could trigger directly, we ran the proof-of-concept exploit inside the guest. For the other 26, we ran the Trinity fuzzer inside the guest, exercising each system call 1,000,000 times with random inputs. Finally, we checked whether the lines of code containing each bug were reached in the host kernel, indicating that the guest could have triggered the bug.

Results. We found that a substantial number of bugs could be triggered in existing virtualization systems, as shown in Table 2. All (100%) bugs were triggered in Native Linux, while the other programs had lower rates:

8/35 (22.9%) in Docker, 12/35 (34.3%) in LXC, and 8/35 (22.9%) bugs in Graphene. Only 1 out of 35 bugs (2.9%) was triggered in Lind.

When we take a closer look at the results, we can see that these outcomes have a lot to do with the design principles of the virtualization systems and the way in which they handle system call requests. Graphene [43] is a library OS that relies heavily on the Linux kernel to handle system calls. Graphene's Linux library implements the Linux system calls using a variant of the Drawbridge [36] ABI, which has 43 functions. Those ABI functions are provided by the Platform Adaptation Layer (PAL), implemented using 50 calls to the kernel. It turns out that 8 vulnerabilities in our test were triggered by PAL's 50 system calls. By contrast, Lind only relies on 33 system calls, which significantly reduces risk and avoids 7 out of the 8 bugs.

Graphene supports many complex and risky system calls, such as `execve`, `msgsnd`, and `futex`, that reached the risky (unpopular) portion of the kernel and eventually led to kernel bugs. In addition, for many basic and frequently-used system calls like `open` and `read`, Graphene allows rarely-used flags and arguments to be passed down to the kernel, which triggered bugs in the unpopular paths. In Lind, all system calls only allow a restricted set of simple and frequently-used flags and arguments. One example from our test result is that Graphene allows `O_TMPFILE` flag to be passed to the `path_openat()` system call. This reached risky lines of code inside `fs/namei.c` in the kernel, and eventually triggered bug CVE-2015-5706. The same bug was triggered in the same way inside Docker and LXC, but was successfully prevented by Lind, due to its strict control of flags and arguments. In fact, the design of Graphene requires extensive interaction with the host kernel and, hence, has many risks. The developers of Graphene manually conducted an analysis of 291 Linux vulnerabilities from 2011 to 2013, and found out that Graphene's design can not prevent 144 of those vulnerabilities.

LXC [28] is an operating-system-level virtualization container that uses Linux kernel features to achieve containment. Docker [15] is a Linux container that runs on top of LXC. The two containers have very similar design features that both rely directly on the Linux kernel to handle system call requests. Since system calls inside Docker are passed down to LXC and then into the kernel, we found out that all 8 kernel vulnerabilities triggered inside Docker were also triggered with LXC. In addition, LXC interacts with the kernel via its `liblxc` library component, which triggered the extra 4 bugs.

It should be noted that although the design of Lind only accesses popular paths in the kernel and implements SafePOSIX inside of a sandbox, there are a few fundamental building blocks for which Lind must rely on the kernel. For example, `mmap` and `threads` cannot be recreated inside SafePOSIX without interaction with the kernel, since there have to be some basic operations to access the hardware. Therefore, Lind passes `mmap` and `threads` directly to the kernel, and any vulnerabilities related to them are unavoidable. CVE-2014-4171 is a bug triggered by `mmap` inside Lind. It was also triggered inside Docker, LXC, and Graphene, indicating that those systems rely on the kernel to perform `mmap` operations as well.

Our initial results suggest that bugs are usually triggered by extensive interaction with the unpopular paths in the kernel through complex system calls, or basic system calls with complicated or rarely used flags. The *Lock-in-Pop* design, and thus Lind, provides strictly controlled access to the kernel, and so poses the least risk.

Vulnerability	Native Linux	Docker	LXC	Graphene	Lind
CVE-2015-5706	✓	✓	✓	✓	✗
CVE-2015-0239	✓	✗	✓	✗	✗
CVE-2014-9584	✓	✗	✗	✗	✗
CVE-2014-9529	✓	✗	✓	✗	✗
CVE-2014-9322	✓	✓	✓	✓	✗
CVE-2014-9090	✓	✗	✗	✗	✗
CVE-2014-8989	✓	✓	✓	✓	✗
CVE-2014-8559	✓	✗	✗	✗	✗
CVE-2014-8369	✓	✗	✗	✗	✗
CVE-2014-8160	✓	✗	✓	✗	✗
CVE-2014-8134	✓	✗	✓	✓	✗
CVE-2014-8133	✓	✗	✗	✗	✗
CVE-2014-8086	✓	✓	✓	✗	✗
CVE-2014-7975	✓	✗	✗	✗	✗
CVE-2014-7970	✓	✗	✗	✗	✗
CVE-2014-7842	✓	✗	✗	✗	✗
CVE-2014-7826	✓	✗	✗	✓	✗
CVE-2014-7825	✓	✗	✗	✓	✗
CVE-2014-7283	✓	✗	✗	✗	✗
CVE-2014-5207	✓	✗	✗	✗	✗
CVE-2014-5206	✓	✓	✓	✗	✗
CVE-2014-5045	✓	✗	✗	✗	✗
CVE-2014-4943	✓	✗	✗	✗	✗
CVE-2014-4667	✓	✗	✗	✓	✗
CVE-2014-4508	✓	✗	✗	✗	✗
CVE-2014-4171	✓	✓	✓	✓	✓
CVE-2014-4157	✓	✗	✗	✗	✗
CVE-2014-4014	✓	✓	✓	✗	✗
CVE-2014-3940	✓	✓	✓	✗	✗
CVE-2014-3917	✓	✗	✗	✗	✗
CVE-2014-3153	✓	✗	✗	✗	✗
CVE-2014-3144	✓	✗	✗	✗	✗
CVE-2014-3122	✓	✗	✗	✗	✗
CVE-2014-2851	✓	✗	✗	✗	✗
CVE-2014-0206	✓	✗	✗	✗	✗
Vulnerabilities Triggered	**35/35 (100%)**	**8/35 (22.9%)**	**12/35 (34.3%)**	**8/35 (22.9%)**	**1/35 (2.9%)**

Table 2: Linux kernel bugs, and vulnerabilities in different virtualization systems (✓: vulnerability triggered; ✗: vulnerability not triggered).

5.2 Comparison of Kernel Code Exposure in Different Virtualization Systems

Setup. To determine how much of the underlying kernel can be executed and exposed in each system, we conducted system call fuzzing with Trinity (similar to our approach in Section 3) to obtain kernel traces. This helps us understand the potential risks a virtualization

Virtualization system	# of Bugs	Kernel trace (LOC)		
		Total coverage	In popular paths	In risky paths
LXC	12	127.3K	70.9K	56.4K
Docker	8	119.0K	69.5K	49.5K
Graphene	8	95.5K	62.2K	33.3K
Lind	1	70.3K	70.3K	0

Table 3: Reachable kernel trace analysis for different virtualization systems.

system may pose based upon how much access it allows to the kernel code. All experiments were conducted under Linux kernel 3.14.1.

Results. We obtained the total reachable kernel trace for each tested system, and further analyzed the components of those traces. These results, shown in Table 3, affirm that Lind accessed the least amount of code in the OS kernel. More importantly, all the kernel code it did access was in the popular kernel paths, which contain fewer bugs (Section 3.2). A large portion of the kernel paths accessed by Lind lie in `fs/` and perform file system operations. To restrict file system calls to popular paths, Lind allows only basic calls, like `open()`, `close()`, `read()`, `write()`, `mkdir()`, and `rmdir()`, and permits only commonly-used flags like `O_CREAT`, `O_EXCL`, `O_APPEND`, `O_TRUNC`, `O_RDONLY`, `O_WRONLY`, and `O_RDWR` for `open()`.

The other virtualization systems all accessed a substantial number of code paths in the kernel, and they all accessed a larger section from the unpopular paths. This is because they rely on the underlying host kernel to implement complex functionality. Therefore, they are more dependent on complex system calls, and allow extensive use of complicated flags. For example, Graphene's system call API supports multiple processes via `fork()` and signals, and therefore accesses many risky lines of code. For basic and frequently-used system calls like `open`, Graphene allows rarely-used flags, such as `O_TMPFILE` and `O_NONBLOCK` to pass down to the kernel, thus reaching risky lines in the kernel that could lead to bugs. By default, Docker and LXC do not wrap or filter system calls made by applications running in a container. Thus, programs have access to basically all the system calls, and rarely used flags, such as `O_TMPFILE`, `O_NONBLOCK`, and `O_DSYNC`. Again, this means they can reach risky lines of code in the kernel.

To summarize, our analysis suggests that Lind triggers the fewest kernel bugs because it has better control over the portions of the OS kernel accessed by applications.

5.3 Impact of Potential Vulnerabilities in Lind's SafePOSIX Re-creation

Setup. To understand the potential security risks if Lind's SafePOSIX re-creation has vulnerabilities, we conducted system call fuzzing with Trinity to obtain the reachable kernel trace in Linux kernel 3.14.1. The goal is to see how much of the kernel is exposed to SafePOSIX. Since our SafePOSIX runs inside the Repy sandbox kernel, fuzzing it suffices to determine the portion of the kernel reachable from inside the sandbox.

Virtualization system	# of Bugs	Kernel trace (LOC)		
		Total coverage	In popular paths	In risky paths
Lind	1	70.3K	70.3K	0
Repy	1	74.4K	74.4K	0

Table 4: Reachable kernel trace analysis for Repy.

Results. The results are shown in Table 4. The trace of Repy is slightly larger (5.8%) than that of Lind. This larger design does not allow attackers or bugs to access the risky paths in the OS kernel, and it leaves open only a small number of additional popular paths. These are added because some functions in Repy have more capabilities for message sending and network connection than Lind's system call interface. For example, in Repy, the `sendmessage()` and `openconnection()` functions could reach more lines of code when fuzzed. However, the kernel trace of Repy still lies completely within the popular paths that contain fewer kernel bugs. Thus, the Repy sandbox kernel has only a very slim chance of triggering OS kernel bugs.

Since it is the direct point of contact with the OS kernel, in theory, the Repy sandbox kernel could be a weakness in the overall security coverage provided by Lind. Nevertheless, the results above show that, even if it has a bug or failure, the Repy kernel should not substantially increase the risk of triggering bugs.

5.4 Practicality Evaluation

The purpose of our practicality evaluation is to show that the "popular paths" metric is practical in building real-world systems. Overhead is expected. We have not optimized our Lind prototype to try to improve performance, since that is not our main purpose for building the prototype.

Setup. We ran a few programs of different types to understand Lind's performance impact. All applications ran unaltered and correctly in Lind. To run the applications, it was sufficient to just recompile the unmodified source code using NaCl's compiler and Lind's `glibc` to call into SafePOSIX.

To measure Lind's runtime performance overhead compared to Native Linux when running real-world applications, we first compiled and ran six widely-used legacy applications: a prime number calculator Primes 1.0, GNU Grep 2.9, GNU Wget 1.13, GNU Coreutils 8.9, GNU Netcat 0.7.1, and K&R Cat. We also ran more extensive benchmarks on two large legacy applications, Tor 0.2.3 and Apache 2.0.64, in Lind. We used Tor's built-in benchmark program and Apache's benchmarking tool `ab` to perform basic testing operations and record the execu-

Application	Native Code	Lind	Impact
Primes	10000 ms	10600 ms	1.06x
GNU Grep	65 ms	260 ms	4.00x
GNU Wget	25 ms	96 ms	3.84x
GNU Coreutils	275 ms	920 ms	3.35x
GNU Netcat	780 ms	2180 ms	2.79x
K&R Cat	20 ms	125 ms	6.25x

Table 5: Execution time performance results for six real-world applications: Native Linux vs. Lind.

Benchmark	Native Code	Lind	Impact
Digest Tests:			
Set	54.80 nsec/element	176.86 nsec/element	3.22x
Get	42.30 nsec/element	134.38 nsec/element	3.17x
Add	11.69 nsec/element	53.91 nsec/element	4.61x
IsIn	8.24 nsec/element	39.82 nsec/element	4.83x
AES Tests:			
1 Byte	14.83 nsec/B	36.93 nsec/B	2.49x
16 Byte	7.45 nsec/B	16.95 nsec/B	2.28x
1024 Byte	6.91 nsec/B	15.42 nsec/B	2.23x
4096 Byte	6.96 nsec/B	15.35 nsec/B	2.21x
8192 Byte	6.94 nsec/B	15.47 nsec/B	2.23x
Cell Sized	6.81 nsec/B	14.71 nsec/B	2.16x
Cell Processing:			
Inbound	3378.18 nsec/cell	8418.03 nsec/cell	2.49x
(per Byte)	6.64 nsec/B	16.54 nsec/B	-
Outbound	3384.01 nsec/cell	8127.42 nsec/cell	2.40x
(per Byte)	6.65 nsec/B	15.97 nsec/B	-

Table 6: Performance results on Tor's built-in benchmark program: Native Linux vs. Lind.

tion time.

Results. Table 5 shows the runtime performance for the six real-world applications mentioned above. The Primes application run in Lind has a 6% performance overhead. The small amount of overhead is generated by NaCl's instruction alignment at build time. We expect other CPU bound processes to behave similarly.

The other five applications require repeated calls into SafePOSIX, and this additional computation produced the extra overhead.

A summary of the results for Tor is shown in Table 6. The benchmarks focus on cryptographic operations, which are CPU intensive, but they also make system calls like `getpid` and reads to `/dev/urandom`. The digest operations time the access of a map of message digests. The AES operations time includes encryptions of several sizes and the creation of message digests. Cell processing executes full packet encryption and decryption. In our test, Lind slowed down these operations by 2.5x to 5x. We believe these slowdowns are due to the increased code size produced by NaCl, and the increased overhead from Lind's SafePOSIX system call interface.

Results for the Apache benchmarking tool `ab` are presented in Table 7. In the set of experiments, Lind produced performance slowdowns around 2.7x. Most of the overhead was incurred due to system call operations inside the SafePOSIX re-creation.

Performance overhead in Lind is reasonable, considering that we did not specifically optimize any part of the code to improve speed. It should also be noted that

# of Requests	Native Code	Lind	Impact
10	900 ms	2400 ms	2.67x
20	1700 ms	4700 ms	2.76x
50	4600 ms	13000 ms	2.83x
100	10000 ms	27000 ms	2.70x

Table 7: Performance results on Apache benchmarking tool `ab`: Native Linux vs. Lind.

performance slowdown is common in virtualization systems. For example, Graphene [43] also shows an overhead ranging from 1.4x to 2x when running applications such as the Apache web server and the Unixbench suite [44]. In many cases, Lind shares the same magnitude of slowdown with Graphene. Lind's ability to run a variety of programs demonstrates the practicality of our "popular paths" metric.

6 Limitations

One of our challenges in conducting this study was deciding where to place the limits of its scope. To explore any one strategy in depth, we felt it was necessary to intentionally exclude consideration of a few other valid approaches. These choices may have placed some limitations on our results.

One limitation is that there are some types of bugs that are difficult to evaluate using our metric. For example, bugs caused by a race condition, or that involve defects in internal kernel data structures, or that require complex triggering conditions across multiple kernel paths, may not be immediately identified using our metric. As we continue to refine our metric, we will also look to evolve our evaluation criteria to find and protect against more complex types of bugs.

Another limitation is that our current metric concludes that certain lines of code in the kernel were reached or not. Though this is an important factor in exploiting a bug, it may not be fully sufficient for all bugs. While a stronger conclusion about bug exploitation conditions would be ideal, it would be hard to do so using a quantitative metric. Instead, it would require a more complicated manual process, which was outside the scope of this study.

7 Related Work

This section summarizes a number of earlier initiatives to ensure the safety of privileged code. The literature referenced in this section includes past efforts to design and build virtualized systems, as well as background information on technologies incorporated into Lind.

Lind incorporates a number of existing virtualization techniques, which are described below.

System Call Interposition (SCI) tracks all the system calls of processes such that each call can be modified or denied. Goldberg, et al. developed Janus [20, 48], which adopted a user-level "monitor" to filter system call requests based on user-specified policies. Garfinkel, et al.

proposed a delegating architecture for secure system call interposition called Ostia [18]. Their system introduced emulation libraries in the user space to mediate sensitive system calls issued by the sandboxed process. SCI is similar to the Lind isolation mechanism. However, SCI-based tools can easily be circumvented if the implementation is not careful [41].

Software Fault Isolation (SFI) transforms a given program so that it can be guaranteed to satisfy a security policy. Wahbe, et al. [49] presented a software approach to implementing fault isolation within a single address space. Yee, et al. from Google developed Native Client (NaCl) [51], an SFI system for the Chrome browser that allows native executable code to run directly in a browser. As discussed in Section 5, Lind adopts NaCl as a key component to ensure secure execution of binary code.

Language-based virtualization. Programming languages like Java, JavaScript, Lua [27], and Silverlight [40] can provide safety in virtual systems by "translating" application commands into a native language. Though many sandboxes implement the bulk of standard libraries in memory-safe languages like Java or C#, flaws in this code can still pose a threat [21, 34]. Any bug or failure in a programming language virtual machine is usually fatal. In contrast, the main component of Lind is built using Repy, which is a programming language with a very small TCB, minimizing the chance of contact with kernel flaws.

OS virtualization techniques include bare-metal hardware virtualization, such as VMware ESX Server, Xen [4], and Hyper-V, container systems such as LXC [28], BSD's jail, and Solaris zones, and hosted hypervisor virtualization, such as VMware Workstation, VMware Server, VirtualPC and VirtualBox. Security by isolation [2, 9, 23, 50] uses containment to provide safe executing environments for multiple user-level virtual environments sharing the same hardware. However, this approach is limited due to the large attack surface exposed by most hypervisors.

Library OSes allow applications to efficiently gain the benefits of virtual machines by refactoring a traditional OS kernel into an application library. Porter, et al. developed Drawbridge [36], a library OS that presents a Windows persona for Windows applications. Similar to Lind, it restricts access from usermode to the host OS through operations that pass through the security monitor. Baumann, et al. presented Bascule [5], an architecture for library OS extensions based on Drawbridge that allows application behavior to be customized by extensions loaded at runtime. The same team also developed Haven [6], which uses a library OS to implement shielded execution of unmodified server applications in an untrusted cloud host. Tsai, et al. developed Graphene [43], a library OS that executes both single and multi-process applications with low performance overhead.

The key distinction between Lind and other existing library OSes is that Lind leverages our "popular paths" metric to verify that it only accesses the safer part of the kernel. Existing library OSes trust the underlying host kernel to perform many functions, and filter only certain system calls. Our work and previous library OSes are orthogonal, but we provide useful insights with our "popular paths" metric.

8 Conclusion

In this paper, we proposed a new security metric based on quantitative measures of kernel code execution when running user applications. Our metric evaluates if the lines of kernel code executed have the potential to trigger zero-day bugs. Our key discovery is that popular kernel paths contain significantly fewer bugs than other paths. Based on this insight, we devise a new design for a secure virtual machine called *Lock-in-Pop*. As the name implies, the design scheme locks away access to all kernel code except that found in paths frequently used by popular programs. We test the *Lock-in-Pop* idea by implementing a prototype virtual machine called Lind, which features a minimized TCB and prevents direct access to application calls from less-used, riskier paths. Instead, Lind supports complex system calls by securely re-creating essential OS functionality inside a sandbox. In tests against Docker, LXC, and Graphene, Lind emerged as the most effective system in preventing zero-day Linux kernel bugs.

So that other researchers may replicate our results, we make all of the kernel trace data, benchmark data, and source code for this paper available [24].

Acknowledgements

We thank our shepherd, Dan Williams, and the anonymous reviewers for their valuable comments. We would also like to thank Lois Anne DeLong for her efforts on this paper, as well as Chris Matthews, Shengqian Ji, Qishen Li, Ali Gholami, Wenzheng Xu, and Yanyan Zhuang for their contributions to this project. Our work on *Lock-in-Pop* was supported by U.S. National Science Foundation Award 1223588.

References

[1] Debian Popularity Contest. http://popcon.debian.org/main/index.html. Accessed December 2014.

[2] Qubes OS. http://www.qubes-os.org. Accessed November 2015.

[3] Intel Virtualization Technology Specification for the Intel Itanium Architecture (VT-i), April 2005.

[4] BARHAM, P., DRAGOVICH, B., FRASER, K., HAND, S., HARRIS, T., HO, A., NEUGEBAUER, R., PRATT, I., AND WARFIELD, A. Xen and the art of virtualization. In *Proceedings of the SOSP'03* (2003), pp. 164–177.

[5] BAUMANN, A., LEE, D., FONSECA, P., GLENDENNING, L., LORCH, J. R., BOND, B., OLINSKY, R., AND HUNT, G. C. Composing os extensions safely and efficiently with bascule. In *Proceedings of the Eurosys'13* (2013).

[6] BAUMANN, A., PEINADO, M., AND HUNT, G. Shielding applications from an untrusted cloud with haven. In *Proceedings of the OSDI'14* (2014).

[7] Berkeley software distribution. http://www.bsd.org. Accessed September 2016.

[8] CAPPOS, J., DADGAR, A., RASLEY, J., SAMUEL, J., BESCHASTNIKH, I., BARSAN, C., KRISHNAMURTHY, A., AND ANDERSON, T. Retaining sandbox containment despite bugs in privileged memory-safe code. In *Proceedings of the CCS'10* (2010).

[9] CHEN, X., GARFINKEL, T., LEWIS, E. C., SUBRAHMANYAM, P., WALDSPURGER, C. A., BONEH, D., DWOSKIN, J., AND PORTS, D. R. Overshadow: A virtualization-based approach to retrofitting protection in commodity operating systems. *SIGPLAN Not. 43*, 3 (Mar. 2008), 2–13.

[10] CHOU, A., YANG, J., CHELF, B., HALLEM, S., AND ENGLER, D. *An empirical study of operating systems errors*, vol. 35. ACM, 2001.

[11] Common Vulnerabilities and Exposures. https://cve.mitre.org.

[12] CVE-2016-5195. Dirty COW - (CVE-2016-5195) - Docker Container Escape. `https://web.nvd.nist.gov/view/vuln/detail?vulnId=CVE-2016-5195`, 2016.

[13] CVE Details Datasource. http://www.cvedetails.com/vulnerability-list/vendor_id-33/product_id-47/version_id-163187/Linux-Linux-Kernel-3.14.1.html. Accessed October 2014.

[14] CVE DETAILS. 14 CVE Docker Vulnerabilities Reported. http://www.cvedetails.com/product/28125/Docker-Docker.html?vendor_id=13534, 2017.

[15] Docker. https://www.docker.com. Accessed September 2016.

[16] Exploit Database. https://www.exploit-db.com. Accessed October 2014.

[17] FBI Tweaks Stance On Encryption BackDoors, Admits To Using 0-Day Exploits. http://www.darkreading.com/endpoint/fbi-tweaks-stance-on-encryption-backdoors-admits-to-using-0-day-exploits/d/d-id/1323526.

[18] GARFINKEL, T., PFAFF, B., AND ROSENBLUM, M. Ostia: A delegating architecture for secure system call interposition. In *Proceedings of the NDSS'04* (2004).

[19] gcov(1) - Linux man page. http://linux.die.net/man/1/gcov. Accessed October 2014.

[20] GOLDBERG, I., WAGNER, D., THOMAS, R., AND BREWER, E. A secure environment for untrusted helper applications (confining the wily hacker). In *Proceedings of the USENIX UNIX Security Symposium'96* (1996).

[21] Learn about java technology. http://www.java.com/en/about/.

[22] KELLER, E., SZEFER, J., REXFORD, J., AND LEE, R. B. Nohype: virtualized cloud infrastructure without the virtualization. In *ACM SIGARCH Computer Architecture News* (2010), vol. 38, ACM, pp. 350–361.

[23] LI, C., RAGHUNATHAN, A., AND JHA, N. Secure virtual machine execution under an untrusted management os. In *Cloud Computing (CLOUD), 2010 IEEE 3rd International Conference on* (July 2010), pp. 172–179.

[24] Lind, a new generation of secure virtualization. https://lind.poly.edu/.

[25] Linux kernel zero-day flaw puts 'tens of millions' of PCs, servers and Android devices at risk. http://www.v3.co.uk/v3-uk/news/2442582/linux-kernal-zero-day-flaw-puts-tens-of-millions-of-pcs-servers-and-android-devices-at-risk.

[26] Linux Test Project. https://linux-test-project.github.io/. Accessed February 2015.

[27] The programming language Lua. www.lua.org. Accessed October 2015.

[28] Linux Container (LXC). https://linuxcontainers.org. Accessed September 2016.

[29] MAYER, A., AND SYKES, A. A probability model for analysing complexity metrics data. *Software Engineering Journal 4*, 5 (1989), 254–258.

[30] NSA Discloses 91 Percent Of Vulns It Finds, But How Quickly? http://www.darkreading.com/vulnerabilities—threats/nsa-discloses-91-percent-of-vulns-it-finds-but-how-quickly/d/d-id/1323077.

[31] National Vulnerability Database. https://nvd.nist.gov/. Accessed September 2015.

[32] OZMENT, A., AND SCHECHTER, S. E. Milk or wine: does software security improve with age? In *Usenix Security* (2006).

[33] PALIX, N., THOMAS, G., SAHA, S., CALVÈS, C., LAWALL, J., AND MULLER, G. Faults in linux: ten years later. In *ACM SIGARCH Computer Architecture News* (2011), vol. 39, ACM, pp. 305–318.

[34] PAUL, N., AND EVANS., D. Comparing java and .net security: Lessons learned and missed. In *Computers and Security* (2006), pp. 338–350.

[35] Poisson Distribution. https://en.wikipedia.org/wiki/Poisson_distribution.

[36] PORTER, D. E., BOYD-WICKIZER, S., HOWELL, J., OLINSKY, R., AND HUNT, G. C. Rethinking the library os from the top down. In *Proceedings of the ASPLOS'11* (Newport Beach, California, USA, 2011), pp. 291–304.

[37] Qemu. http://wiki.qemu.org/Main_Page. Accessed September 2016.

[38] Seattle's Repy V2 Library. https://seattle.poly.edu/wiki/RepyV2API. Accessed September 2014.

[39] SHIUE, W.-K., AND BAIN, L. J. Experiment size and power comparisons for two-sample poisson tests. *Applied Statistics* (1982), 130–134.

[40] Microsoft Silverlight. http://www.microsoft.com/silverlight/. Accessed October 2015.

[41] TAL GARFINKEL. Traps and Pitfalls: Practical Problems in System Call Interposition Based Security Tools.

[42] Trinity, a Linux System call fuzz tester. http://codemonkey.org.uk/projects/trinity/. Accessed November 2014.

[43] TSAI, C. C., ARORA, K. S., BANDI, N., JAIN, B., JANNEN, W., JOHN, J., KALODNER, H. A., KULKARNI, V., OLIVEIRA, D., AND PORTER, D. E. Cooperation and security isolation of library oses for multi-process applications. In *Proceedings of the EuroSys'14* (Amsterdam, Netherlands, 2014).

[44] Unixbench. https://github.com/kdlucas/byte-unixbench. Accessed September 2016.

[45] Virtualbox. https://www.virtualbox.org. Accessed September 2016.

[46] Vmware server. https://my.vmware.com/web/vmware/info?slug=infrastructure_operations_management/vmware_server/2_0.

[47] Vmware workstation. https://www.vmware.com/products/workstation. Accessed September 2016.

[48] WAGNER, D. A. Janus: An approach for confinement of untrusted applications. In *Tech. Rep. CSD-99-1056, University of California, Berkeley* (1999).

[49] WAHBE, R., LUCCO, S., ANDERSON, T. E., AND GRAHAM, S. L. Efficient software-based fault isolation. In *SIGOPS Oper. Syst. Rev. 27, 5* (1993), pp. 203–216.

[50] YANG, J., AND SHIN, K. G. 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* (New York, NY, USA, 2008), VEE '08, ACM, pp. 71–80.

[51] YEE, B., SEHR, D., DARDYK, G., CHEN, J. B., MUTH, R., ORMANDY, T., OKASAKA, S., NARULA, N., AND FULLAGAR, N. Native client: A sandbox for portable, untrusted x86 native code. In *Proceedings of the IEEE Symposium on Security and Privacy* (Berkeley, CA, USA, 2009), pp. 79–93.

[52] 0-day exploits more than double as attackers prevail in security arms race. http://arstechnica.com/security/2016/04/0-day-exploits-more-than-double-as-attackers-prevail-in-security-arms-race/. Accessed September 2016.