

# Securing Monolithic Kernels using Compartmentalization

Soo Yee Lim  
University of British Columbia  
Canada

Sidhartha Agrawal  
University of British Columbia  
Canada

Xueyuan Han  
Wake Forest University  
United States of America

David Eysers  
University of Otago  
New Zealand

Dan O’Keeffe  
Royal Holloway University of  
London  
United Kingdom

Thomas Pasquier  
University of British Columbia  
Canada

## ABSTRACT

Monolithic operating systems, where all kernel functionality resides in a single, shared address space, are the foundation of most mainstream computer systems. However, a single flaw, even in a non-essential part of the kernel (e.g., device drivers), can cause the entire operating system to fall under an attacker’s control. Kernel hardening techniques might prevent certain types of vulnerabilities, but they fail to address a fundamental weakness: the lack of intra-kernel security that safely isolates different parts of the kernel. We survey kernel compartmentalization techniques that define and enforce intra-kernel boundaries and propose a taxonomy that allows the community to compare and discuss future work. We also identify factors that complicate comparisons among compartmentalized systems, suggest new ways to compare future approaches with existing work meaningfully, and discuss emerging research directions.

## CCS CONCEPTS

• **Security and privacy** → **Operating systems security**; Virtualization and security; **Software security engineering**.

## 1 INTRODUCTION

Monolithic operating systems (OSs), such as Linux, Microsoft Windows, and BSD Unix variants, underpin the majority of today’s desktop, cloud, and mobile computing environments. While these OSs support a wide variety of software and hardware, their size and complexity leads to frequent, severe security flaws [64]. It cannot be assumed that these OSs’ kernels can be correctly implemented to contain security threats from untrusted user-space applications.

Improving the security of monolithic kernels is challenging, because any potential solution must incur low performance overhead, retain compatibility with existing applications, and ensure maintainability, to be practically adopted. For example, formally-verified microkernels [103] and OSs

written in memory-safe languages [30, 63] offer desirable security properties, but they fail to satisfy some, if not all, of these requirements. Consequently, their adoption is limited to more niche markets (e.g., L4 running on Apple’s Secure Enclave processors). On the other hand, incremental approaches to hardening monolithic kernels [26, 163] are easy to deploy but offer more limited security benefits. For example, the Kernel Self Protection Project [26] developed a variety of defenses for the upstream Linux kernel, but each eliminates only a specific class of bugs or a specific method of exploitation.

We discuss *kernel compartmentalization* as a way forward to secure existing monolithic kernel architectures with minimal changes to their existing code bases (and therefore relatively low engineering effort). Kernel compartmentalization techniques divide a monolithic kernel into multiple compartments, each within their own security domains, and provide strong isolation between these domains: Interactions between domains are restricted to prevent an attacker from taking control of the entire kernel by exploiting a vulnerability in just one domain.

After decades of research into kernel compartmentalization, the time is ripe for a systematic look at past techniques that compartmentalize kernel subsystems to achieve desirable security properties, such as confidentiality, integrity, and availability. Two additional reasons make this survey particularly timely: First, emerging hardware to help support software compartmentalization (e.g., Arm Morello [6]) can substantially reduce performance overhead, bringing into production kernel compartmentalization approaches that were previously considered too expensive. Second, recent interest [41, 51, 94, 164] in using the extended Berkeley Packet Filter (eBPF) to import user-supplied code into production kernels at runtime motivate the development of kernel compartmentalization to reduce the security risks of running in-kernel eBPF programs. While eBPF employs a static verifier to guarantee the safety of its programs, soundness remains a challenge due to many vulnerabilities in the

verifier [19–22, 92]. Kernel compartmentalization could enforce runtime isolation of eBPF programs and therefore allow users to safely extend the kernel [116].

## Contributions

We synthesize kernel compartmentalization research spanning over two decades and make the following contributions:

- **Taxonomy of kernel compartmentalization techniques:** We provide a unifying framework to categorize kernel compartmentalization techniques into two broad architectures, *sandbox* and *safebox*, based on their distinct threat models. These two architectures differ in ways they address two important challenges in kernel compartmentalization: (1) identifying suitable compartment boundaries and (2) enforcing isolation between the identified boundaries at runtime.
- **Categorization of existing work:** We use our taxonomy to discuss the existing body of work and identify potential future research directions.
- **Evaluation of kernel compartmentalization:** We examine current evaluation strategies used in the literature to understand why adopting a general benchmarking strategy to compare compartmentalization systems is difficult. We propose ways to simplify future comparisons with existing work.

## Structure

In §2, we discuss security issues that plague monolithic kernels to motivate the need for kernel compartmentalization. We categorize prior compartmentalization approaches into two architectures, *sandbox* and *safebox*, and discuss their differences in §3. In §4, we analyze common compartmentalization units and how their boundaries are identified. In §5, we classify isolation mechanisms that enforce these boundaries. Then, we examine emerging hardware technologies that will likely influence future compartmentalization landscape in §6. Finally, in §7, we discuss the limitations of the evaluation strategies currently used in the literature and propose ways to better understand and compare different compartmentalization solutions.

## 2 BACKGROUND AND MOTIVATION

A monolithic OS design (Fig. 1a) is problematic with respect to security mainly for two reasons. First, the size of a monolithic kernel’s code base and its complexity make it easy to introduce bugs and vulnerabilities, which could allow untrusted user-space applications to compromise the kernel and thus the entire system. For example, the SLOccount

tool [170], which we use to measure the size of the Linux v6.4 kernel code base, reported ~24 million lines of code, 98% of which is written in memory-unsafe C and 1.2% in assembly. While individual OS instances typically use only a subset of this large code base, their sizes are still substantial compared to those of the vast majority of software. With a steady development rate of over 75,000 commits per year [1], more than 1,000 vulnerabilities were discovered in the last five years [128]. Applying formal verification [102] and static program analysis [76] to soundly prove the absence of bugs and vulnerabilities in the code base is difficult, because existing approaches do not scale to the size of the kernel. For instance, the seL4 project took 11 person-years to formally verify a microkernel that has about 9,000 lines of code [103]. Similarly, recent kernel fuzzing approaches [91, 99, 149, 155] can detect previously unknown vulnerabilities, but they cannot *guarantee* the absence of vulnerabilities in the kernel. Finally, rewriting existing kernel code in a safer language (e.g., Rust for Linux [28]) can reduce the prevalence of some (e.g., out-of-bounds array access) but not all types of vulnerabilities. For instance, Rust cannot prevent memory leaks that lead to resource exhaustion [101].

Second, all OS services, including the core functionality (scheduling, memory management, etc.), modules that extend the kernel (e.g., audit systems), and device drivers, operate in a single security domain. This lack of isolation can give an attacker full privilege to access, modify, and control the entire system’s data and resources when they exploit a vulnerability *anywhere* in the kernel. Even a bug in a peripheral kernel module can constitute a single point of failure, causing essential OS functionality to fall fully under the attacker’s control. For example, a heap-based buffer overflow vulnerability (CVE-2020-12654) in the Linux kernel’s Marvell Wi-Fi driver can cause arbitrary code execution and denial-of-service (DoS) [18].

### 2.1 Kernel Hardening Techniques

One strategy to address the fundamental brittleness of a monolithic kernel architecture is to increase the barrier to an attack. For example, address space layout randomization (ASLR) [78] randomizes the locations of libraries, the heap, and the stack in a process’ address space so that an attacker cannot reliably find useful instructions to create a return-oriented programming attack [78]. Other techniques to make bugs harder to exploit include data execution prevention (DEP) [33], stack canaries [165], and shadow stacks [55, 66, 178]. All of these techniques, commonly referred to as *kernel hardening*, are effective at reducing the possibility of a vulnerability being successfully exploited, but they can also be bypassed by attackers [45, 71, 79, 88, 109, 153, 157]. For example, attackers can exploit information leakage from side

channels to deduce [39] and even derandomize [72] address space layout. Kernel compartmentalization is complementary to existing kernel hardening techniques and can be deployed alongside them.

## 2.2 Alternative Kernel Designs

Microkernels and unikernels are the two main alternatives to monolithic kernels, see Fig. 1b and Fig. 1c. We briefly discuss them here and refer the interested reader to existing literature [68, 122] for more detailed discussions and different interpretations of these designs. Other kernel designs (e.g., exokernel [69] and multikernel [40]) exist, but since we focus on monolithic kernels, an in-depth discussion of these designs is beyond the scope of this survey.

**Microkernel:** Unlike a monolithic kernel, a microkernel [68] implements only a subset of kernel functionality (e.g., scheduling and IPC), moving other functionality, such as file systems and device drivers, into user space. This *compartmentalized* design (where individual subsystems run in their own, mostly isolated, domains) significantly reduces the system’s trusted computing base (TCB) to the point where the kernel code can be formally verified [103]. However, microkernels can incur large performance overhead, because switching between the kernel and user-mode functionality is costly, especially for high-throughput devices such as network cards. As we see in §5, some kernel compartmentalization systems use similar mechanisms to separate functionality in a monolithic kernel, inheriting similar performance issues.

**Unikernel:** A unikernel (and a library OS in general) builds a single address space OS stack that only contains the OS subsystems required by a specific application. Therefore, compared to a microkernel, a unikernel does not incur the expensive cost of context switches; compared to a classic monolithic kernel, it reduces the TCB by excluding unnecessary kernel functionality. However, since all kernel code essential to a target application shares the same address space, a vulnerability in one kernel component (e.g., a buggy device driver) can still affect the entire kernel stack. Consequently, some unikernels [111, 148] resort to compartmentalization techniques. Moreover, while recent unikernels (e.g., Lupine OS [106] and Unikraft [105]) can run legacy programs, unlike a monolithic kernel, their highly specialized, application-specific organization is unsuitable for general-purpose computing.

**The future of monolithic kernels:** Alternatives to monolithic kernel designs are gradually conquering a wider portion of the global OS market (e.g., MINIX [82] on Intel chips, and embedded L4 [68] on Apple devices). However, a significantly large proportion of devices still depend on monolithic OSs and will continue to do so in the foreseeable future. We believe that compartmentalization of a monolithic OS

is essential to an *incremental* path towards a safer OS [113]. Anecdotaly, while not strictly kernel compartmentalization, the Linux community has shown interest in address space isolation techniques to combat the Spectre/Meltdown family of vulnerabilities [57]. Some compartmentalization systems, such as VirtuOS [134], have gone as far as effectively transforming monolithic OSs into microkernels, blurring the line between these two architectures.

## 2.3 Software Compartmentalization

Kernel compartmentalization is a type of *software compartmentalization*, which is more generally defined as the process of decomposing a software artifact into isolated but collaborative components to mitigate the exploitation of vulnerabilities [81]. Many software compartmentalization techniques have been proposed in the user-space context, which generally fall into three categories: (1) protecting the OS from untrusted *user-space* applications [67, 90, 100, 175], (2) protecting security-sensitive user-space applications or pieces of application logic (PAL) from a compromised OS [37, 74, 84, 96, 124, 125, 156], and (3) protecting *both* the OS from misbehaving user-space applications and user-space applications from a malicious OS [114]. The interested reader can refer to the survey by Shu et al. [151].

While we see similar use cases emerge in kernel space (§3), we focus on two important aspects of software compartmentalization in the kernel context: (1) identifying compartment boundaries (§4), and (2) enforcing these boundaries while maintaining good performance (§5). Differences between the techniques employed in user space and within the kernel are significant enough that a direct application from one domain to the other is challenging (see §4.1).

## 3 OVERVIEW OF KERNEL COMPARTMENTALIZATION APPROACHES

Generally speaking, all kernel compartmentalization systems create boundaries between *untrusted* system components, which might be buggy or malicious, and *trusted* ones, which are assumed to be benign. Kernel compartmentalization is enforced by (1) *resource access control* to restrict the code and data accessible to a kernel compartment, and (2) *control transfer restriction* when execution crosses a compartment boundary. However, existing compartmentalization systems differ in the *direction* in which these controls are performed at the compartment boundary due to different assumptions of the threat model they adopt (Fig. 2). We categorize existing kernel compartmentalization work into the *sandbox* architecture and the *safebox* architecture (Table 1).

**Sandbox:** Some kernel compartmentalization systems [52, 73, 145] target kernel subsystems, such as device drivers, that traditionally run in kernel space but are typically considered

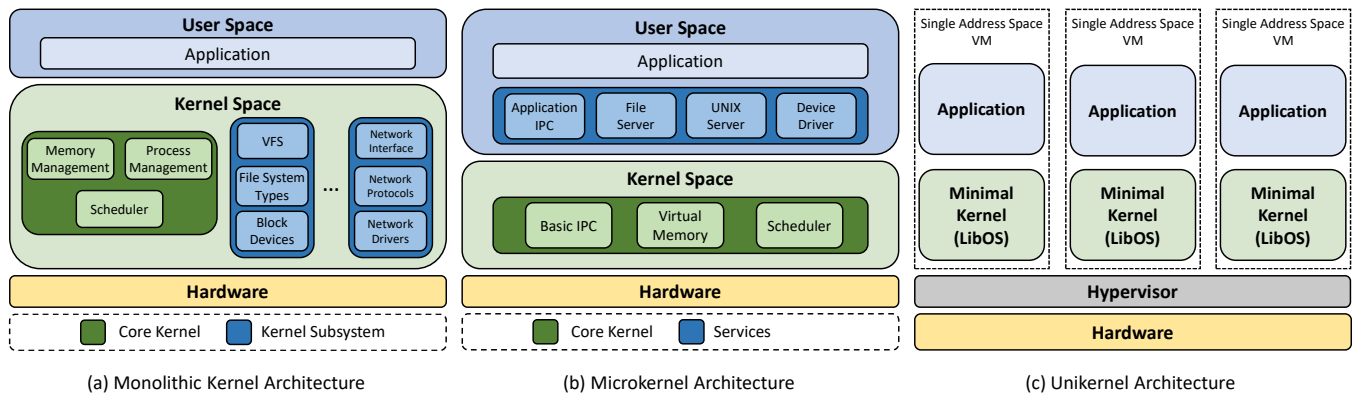


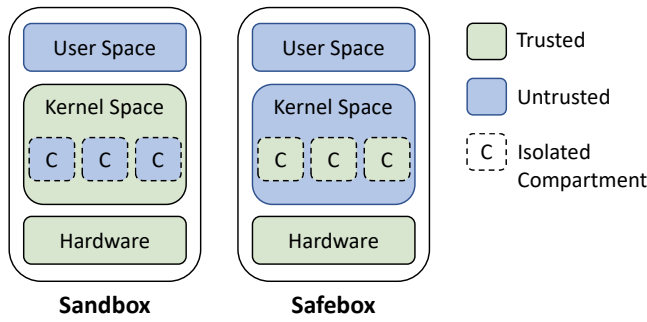
Figure 1: Alternative kernel designs

Table 1: An overview of existing kernel compartmentalization systems (M = Manual, A = Automated, N/R = Not Required)

|                           |                        | Use Case                 | Paper             | Security Properties |           |             |                | Boundary      |                      | Isolation Mechanism |                   |             |                   |               |   |
|---------------------------|------------------------|--------------------------|-------------------|---------------------|-----------|-------------|----------------|---------------|----------------------|---------------------|-------------------|-------------|-------------------|---------------|---|
|                           |                        |                          |                   | Confidentiality     | Integrity | Reliability | Recoverability | Decomposition | Glue Code Generation | Intra-kernel        |                   | User/Kernel | Kernel/Hypervisor |               |   |
|                           |                        |                          |                   |                     |           |             |                |               |                      | Domain Based        |                   |             |                   | Address Based |   |
|                           |                        |                          |                   |                     |           |             |                |               |                      | Software Based      | Hardware Assisted |             |                   |               |   |
| Architecture              | Sandbox                | Fault Isolation          | Mondrix [173]     |                     | •         | •           |                | M             | M                    |                     | •                 |             |                   |               |   |
|                           |                        |                          | Microdrivers [73] |                     | •         | •           |                | A             | A                    |                     |                   |             | •                 |               |   |
|                           |                        |                          | Decaf [145]       |                     | •         | •           |                | A             | A                    |                     |                   |             | •                 |               |   |
|                           |                        | Fault Resistance         | Nooks [160]       |                     | •         | •           | •              | M             | M                    |                     |                   | •           |                   |               |   |
|                           |                        |                          | BGI [52]          |                     | •         | •           | •              | M             | N/R                  | •                   |                   |             |                   |               |   |
|                           |                        |                          | SafeDrive [177]   |                     | •         | •           | •              | M             | N/R                  | •                   |                   |             |                   |               |   |
|                           |                        |                          | VirtuOS [134]     |                     | •         | •           | •              | M             | M                    |                     |                   |             |                   |               | • |
|                           |                        | Vulnerability Isolation  | XFI [70]          |                     | •         |             |                | A             | N/R                  | •                   |                   |             |                   |               |   |
|                           |                        |                          | LXFI [123]        |                     | •         |             |                | M             | N/R                  | •                   |                   |             |                   |               |   |
|                           | LXD <sub>s</sub> [131] |                          | •                 | •                   |           |             | M              | A             |                      |                     |                   |             |                   | •             |   |
|                           | LVD [132]              |                          | •                 | •                   |           |             | M              | A             |                      |                     |                   |             |                   | •             |   |
|                           | HAKC [126]             |                          | •                 | •                   |           |             | M              | M             |                      | •                   |                   |             |                   |               |   |
|                           | Safebox                | Security State Isolation | KCoFI [59]        |                     | •         |             |                | M             | N/R                  |                     | •                 |             |                   |               |   |
|                           |                        |                          | KENALI [154]      |                     | •         |             |                | A             | N/R                  |                     | •                 |             |                   |               |   |
| xMP [142]                 |                        |                          | •                 | •                   |           |             | M              | N/R           |                      |                     |                   |             |                   | •             |   |
| IskiOS [80]               |                        |                          | •                 | •                   |           |             | M              | N/R           |                      | •                   |                   |             |                   |               |   |
| Secure Policy Enforcement |                        | PerspikuOS [65]          |                   | •                   |           |             | N/R            | N/R           |                      |                     | •                 |             |                   |               |   |
|                           |                        | SKEE [35]                |                   | •                   |           |             | N/R            | N/R           |                      |                     | •                 |             |                   |               |   |

to be less trustworthy than the rest of the kernel. These compartmentalization systems enforce security rules on data

and control flow going *out of* the compartment boundary, so that any bugs or vulnerabilities within the compartment



**Figure 2: Threat models of kernel compartmentalization architectures.**

cannot adversely affect the kernel. The sandbox architecture reduces *the kernel’s attack surface* by confining an untrustworthy component within its own isolated compartments.

**Safebox:** Safebox systems [35, 65] follow the opposite threat model. They assume that the kernel is untrustworthy and therefore isolate some critical kernel functionalities, such as mandatory access control, to protect them from potential threats originating from the rest of the kernel. These systems enforce security rules on data and control flow going *into* the compartment boundary. From the point of view of an isolated component, the safebox architecture reduces *the component’s attack surface* by excluding the rest of the kernel from its TCB.

We discuss both architectures in more detail in the remainder of this section, identifying the security properties that each architecture serves, for various use cases.

### 3.1 Sandbox Architecture

Kernel extensions, such as loadable kernel modules and device drivers, have long been a major source of errors [56] and vulnerabilities [53, 128, 138] in the Linux kernel. The sandbox architecture protects the kernel from their errors or exploits by isolating each of these components in a separate compartment while restricting resource accesses and control transfers *out of* the compartment’s boundary.

**Resource Access Control.** Resource access control preserves the integrity of the kernel’s *non-control data* (e.g., configuration data) by restricting write access to the data outside of an isolated compartment. This protects the rest of the kernel from any modification that could corrupt kernel memory and change the kernel’s behavior unexpectedly.

**Control Transfer Restriction.** The corruption of *control data* (e.g., function pointers, jump targets, and return addresses) diverts a program’s normal control flow. To prevent an isolated component from executing arbitrary kernel code in the event of an error or kernel exploit, sandbox systems interpose on all control transfers *out of* the compartment’s

boundary. Specifically, they restrict the component’s control flow to an immutable set of valid entry points, which can be defined at compile time e.g., as an allowlist of control flow transfer targets [52, 70, 160].

**Use Cases and Security Properties.** The literature covers three main types of threat models regarding kernel extensions against which the sandbox architecture protects the core kernel: (1) *buggy* extensions containing bugs that could affect system up-time; (2) *exploitable* extensions containing bugs that can be leveraged by a user-space program to corrupt kernel state (e.g., privilege escalation); (3) *malicious* extensions actively trying to compromise the system (e.g., kernel rootkits). These models represent different real-life attack scenarios. For example, the *buggy* threat model is suitable for an enterprise workstation where employees have no root access, and kernel extensions are centrally managed. A user under the *exploitable* threat model might unintentionally install a piece of untrusted software that exploits bugs in kernel extensions. Instead, under the *malicious* threat model, an insider or supply-chain attack might be used to install a kernel rootkit to infiltrate the victim machine. Sandbox systems confine *buggy* extensions through *fault isolation* and *fault resistance*, and contain *exploitable* and *malicious* extensions via *vulnerability isolation*. Overall, they achieve *integrity*, *confidentiality*, *reliability*, and *recoverability*.

**3.1.1 Fault Isolation.** Faults in the kernel could corrupt and crash the system. Unfortunately, they often arise due to bugs in the large and complex kernel code base. Fault isolation [73, 126, 145, 173] confines the impact of an error in the buggy component to improve a system’s *integrity* and *reliability*.

**Integrity.** Commodity monolithic kernels are implemented in memory unsafe languages and thus are vulnerable to memory corruption. A bug in a kernel component can e.g., corrupt critical kernel state that leads to system failure or otherwise unexpected behavioral changes in the system [44]. Fault-isolating systems restrict write access to kernel memory to ensure that a sandboxed kernel component can modify only its own memory or the memory it has legitimate access to. This preserves *data integrity*, which prevents an error from modifying kernel memory outside the sandbox. Additionally, a sandboxed component cannot execute arbitrary kernel functions without having legitimate access to them. This restriction enforces *control-flow integrity*, which, combined with resource access control (i.e., data integrity), prevents the buggy kernel component from affecting the behavior of the rest of the system.

**Reliability.** Reliability measures a system’s ability to remain operational despite the presence of misbehaving components. By compartmentalizing unreliable kernel components and restricting their interactions with the rest of the system, existing kernel compartmentalization techniques [52, 73, 134,

145, 160, 173, 177] allow faulty components to fail without crashing the system, improving its reliability. The end goal is to increase overall system up-time.

**3.1.2 Fault Resistance.** While a fault-isolating system contains the impact of an error, a faulty component remains unavailable until it is reloaded or when the system reboots. To restore (some of) the component's functionality, a fault-resistant system [52, 134, 159, 177] not only isolates but also automatically recovers from most (but not necessarily all) faults within a compartment. Therefore, in addition to a fault-isolation mechanism, a fault-resistant system also implements a recovery mechanism. For example, a virtual machine (VM) based compartmentalization system [134] can leverage the hypervisor's ability to terminate and restart VMs to recover from a failed compartment. Thus, such a system also guarantees *recoverability*.

**Recoverability.** A kernel component's failure affects the availability of its dependent components. To facilitate automatic recovery, a fault-resistant system tracks the component's updates to kernel objects. When a failure is detected, the recovery mechanism unloads the faulty component, restores kernel state, releases kernel resources held by the faulty component, and restarts the kernel component in its compartment. This restores the faulty component to a functional state from which it can continue; as a result, other affected components can return to normal operations.

**3.1.3 Vulnerability Isolation.** Both fault-isolating and fault-resistant systems assume that faults in a kernel component are unintended errors. In contrast, vulnerability-isolating systems [70, 123, 131, 132] assume a stronger threat model where the component is either exploitable or malicious. Since they target kernel attacks rather than faults, they provide no reliability or recoverability guarantees; instead, they guarantee *confidentiality* and *integrity*.

**Confidentiality.** An adversary with an arbitrary read primitive can access any sensitive information in kernel memory. For example, a vulnerability [13] in the Xen network back-end driver allows for out-of-bounds access, which leaks kernel information. A sandbox system can restrict an untrustworthy component's read access to kernel memory outside of its compartment to protect sensitive kernel data. However, software-based vulnerability-isolating systems generally do not restrict read access for performance concerns; consequently, confidentiality attacks are out of their scope [70, 123]. Fortunately, with the availability of efficient hardware isolation primitives, hardware-based vulnerability-isolating systems can guarantee the confidentiality of sensitive kernel data without incurring substantial performance overhead [126, 131, 132].

**Integrity.** An adversary can abuse write primitives of a compromised kernel component to corrupt critical kernel state

(e.g., a page table) or tamper with run-time checks. Therefore, vulnerability-isolating systems provide additional integrity guarantees, on top of data and control-flow integrity discussed in §3.1.1, to protect security-critical code and data. For example, both BGI [52] (a fault-isolating system) and XFI [70] (a vulnerability-isolating system) instrument run-time checks in untrustworthy kernel components to enforce isolation. However, XFI additionally verifies the integrity of its checks at load time to prevent an adversary with an arbitrary write primitive from modifying these checks.

## 3.2 Safebox Architecture

Attackers can exploit kernel vulnerabilities to bypass security mechanisms (e.g., ASLR [10, 11, 17] and SELinux [12]). The safebox architecture isolates security-critical kernel components, such as security-related kernel state and in-kernel policy enforcement mechanisms, from the rest of the monolithic system to prevent security bypass. To do so, it restricts resource accesses and control transfers *into* an isolated compartment.

**Resource Access Control.** The safebox architecture compartmentalizes security-related kernel resources and restricts their access to only the trusted code within the compartment. Since the rest of the kernel has no access to the isolated kernel data, its integrity is preserved even in the presence of a compromised kernel.

**Control Transfer Restriction.** A safebox system allows protected kernel operations to execute only within the safebox. The rest of the kernel transfers control into the safebox whenever a protected kernel operation is requested.

**Use Cases and Security Properties.** Kernel security mechanisms can be protected under two threat models: (1) the kernel is benign but contains *exploitable* vulnerabilities; (2) the kernel is compromised and potentially *malicious*. In the *exploitable* threat model, an unprivileged attacker can exploit kernel vulnerabilities to achieve privilege escalation. The *malicious* threat model assumes that a kernel has fallen under the control of an attacker, i.e., kernel data and kernel code can be modified to exhibit malicious behavior. Safebox systems defend against the *exploitable* kernel through *isolation of security-critical state*, and protect against a *malicious* kernel by providing an isolated execution environment for *secure in-kernel policy enforcement*. We discuss the differences in their *integrity* and *confidentiality* guarantees below.

**3.2.1 Security State Isolation.** Attackers can exploit kernel vulnerabilities to obtain arbitrary read/write primitives [9], which enable them to corrupt security-critical state to bypass security mechanisms. For example, KCoFI leverages hardware trap vector tables and page mapping information to enforce control-flow integrity [59]. However, attackers can bypass this ad-hoc security mechanism by corrupting this

security-critical state. Therefore, to protect these security mechanisms, safebox systems isolate such state, enforcing *confidentiality* and/or *integrity*.

**Confidentiality.** The leakage of kernel state could lead to security bypass, so confidentiality is important to safeguard defense mechanisms. For example, prior attacks [14–16] exploit memory safety bugs, which reveal internal kernel state, to bypass kernel address space layout randomization (KASLR). Safebox systems preserve confidentiality by restricting the kernel’s read access to data stored within a safebox. However, similar to sandbox systems, software-based safebox systems [59, 154] often do not guarantee confidentiality due to performance concerns while hardware-based ones [80, 142] do.

**Integrity.** To prevent tampering of security-critical state, safebox systems restrict the kernel’s write access to the safebox. For safebox systems that rely on the memory management unit (MMU) for correct configuration of access permissions or page table mappings, a corrupted MMU can violate safebox isolation. Therefore, MMU integrity must also be preserved in such systems [59, 80, 142, 154].

**3.2.2 Secure Policy Enforcement.** When a kernel is compromised, security mechanisms are at risk because attackers can arbitrarily modify kernel state and code. Although the state (data) of a security mechanism can be isolated with security-state-isolating safeboxes, the policy (code) itself can still be corrupted. Existing safebox systems [35, 65] leverage compartmentalization to provide a safe execution environment for policy enforcement, in which the TCB no longer includes the entire kernel. In other words, the *integrity* of the isolated security mechanism is not affected even in the presence of a compromised kernel.

**Integrity.** Similar to security state isolation, the kernel is restricted from writing to the safebox, and MMU isolation is required if it is involved in compartmentalization. However, the MMU isolation strategy differs. In security state isolation (§3.2.1), the MMU management code, though sometimes instrumented with run-time checks [59], executes outside the safebox so the kernel retains control over the kernel memory. This MMU isolation strategy works because the kernel is assumed benign. However, secure-policy-enforcement systems assume a malicious kernel, so leaving the MMU in the hands of a malicious kernel could break MMU isolation and potentially the safebox itself. Therefore, these systems deprive the kernel of critical MMU functionalities so it cannot modify the MMU without switching to the safebox. This allows the safebox to gain exclusive control over the entire kernel memory, preventing the malicious kernel from bypassing security mechanisms by e.g., changing access permissions or injecting malicious code.

### 3.3 Summary

Our categorization of sandbox and safebox architectures is broad, abstracting away nuanced differences among a wide variety of threat models adopted in the literature to emphasize their high-level commonalities. However, *the devil is in the details*: With almost no published work making the exact same threat assumptions, a direct comparison between two systems is unequivocally difficult – an important but unfortunate aspect of kernel compartmentalization work that we will revisit in §7. It is therefore crucial to examine *explicit* and *implicit* threat assumptions made by authors when identifying a suitable mechanism for a given use case.

## 4 COMPARTMENT BOUNDARIES

Kernel compartmentalization involves 1) creating compartments and 2) enforcing isolation. In this section, we describe the process of compartment creation, which requires *kernel decomposition* (to separate the code and data of a kernel component to be compartmentalized) and *glue code generation* (to marshal and synchronize data between an isolated component and the rest of the kernel). We discuss the challenges in both steps and solutions proposed by existing kernel compartmentalization systems to address them.

### 4.1 Kernel Decomposition

Prior work on software compartmentalization (§2.3) splits a user-space application into a privileged and an unprivileged component. This process, known as *privilege separation*, can be done either manually [98, 124, 161] or automatically [46, 48, 117, 119, 120]. Manually compartmentalizing any non-trivial legacy application is impractical, due to e.g., extensive uses of function pointers (such as in OpenSSL). Work such as Privtrans [48], Wedge [46], and SeCage [119] instead leverages program analysis techniques to automatically partition user-space applications. However, the techniques they use cannot be applied directly to kernel code for the following reasons:

**Kernel Size:** Current program analysis techniques, regardless of whether they can be automated or not, do not scale well to the large code base of a monolithic kernel. This is because program analysis rapidly becomes prohibitively expensive as the number of program states exponentially increases [83, 162, 179]. Although scalability can be improved at the expense of soundness, which is a reasonable trade-off for bug-finding tools [121], an unsound analysis in the context of partitioning could lead to correctness issues if some critical data are left unprotected. Thus, this trade-off must be considered carefully in kernel space.

**Complex Kernel Data Dependencies:** Kernel compartmentalization must enforce restrictions on all resources required by an isolated kernel component, including resources



shared between the compartment and the rest of the kernel. For correctness and performance, decomposition frameworks must infer the component’s data dependencies soundly and precisely. However, pointers in data structures make precisely identifying decomposition boundaries difficult. To understand pointer aliasing, a common form of data dependency in C/C++, these frameworks must perform complex global pointer analysis, which does not scale to large programs [118]. As such, many software compartmentalization systems [46, 48, 119] lack support for dependency analysis involving pointers; others [73, 145] often over-estimate shared resources to ensure soundness, thus impacting performance (e.g., as a result of unnecessary synchronization) [87].

**Multi-threading:** Most user-space decomposition mechanisms [48, 81, 117, 119] do not support multi-threading, but it is a key feature in commodity kernels. Moreover, multi-threading introduces concurrency issues that must be addressed with synchronization mechanisms during glue code generation (§4.2).

Depending on the architecture (§3), decomposition focuses on different parts of the kernel. In the *sandbox* architecture, a kernel compartmentalization system typically performs *device driver decomposition*, while *security state decomposition* and *policy enforcement decomposition* are usually the focus of a *safebox* system.

**4.1.1 Device Driver Decomposition.** Device driver decomposition techniques can be divided into *manual* and *automated* approaches. The majority of published research [52, 123, 126, 131, 132, 160, 173, 177] manually decomposes device drivers. For example, LXFI [123] supports fine-grained privilege separation by allowing users to manually annotate kernel interfaces and specify security policies for each instance of a shared kernel module (e.g., a block device). If a module instance is compromised, the attacker can only exploit the privileges associated to the instance (e.g., write only to the affected block device). While LXFI trusts any annotations provided by the user, manual decomposition is unsustainable and prone to human error, especially as the size and complexity of driver code increases.

To automate decomposition, Microdriver [73] and Decaf [145] use static analysis to identify code that can be delegated to user space, effectively decomposing a device driver into a kernel-level driver (containing high-priority code such as interrupt handling) and a user-level driver (containing infrequently invoked, low-priority code), which allows the user-level driver to be written in safer languages than C (e.g., Java [145] in Decaf). This approach resembles the typical mechanism/policy split in microkernel architectures (§2.2). However, only a subset of driver functionality is moved to user space, so faults remaining in the kernel part of the

decomposed driver can still pose a threat to the system’s reliability. To scale the decomposition framework to the entire device driver, program analysis approaches need to overcome the aforementioned challenges (e.g., alias analysis). For example, KSplit [87] addresses the scalability issue faced by many user-space privilege separation techniques [110, 158] via the parameter tree approach [118], which is a modular way of computing dependence graphs from programs containing complex uses of pointers.

**4.1.2 Security State Decomposition.** To isolate security-critical kernel state (e.g., page tables and process credentials), some work [59, 80, 142] *manually* identifies the critical data of a security mechanism and stores them in a protected memory region. For example, KCoFI [59] isolates a region of memory reserved for storing security-sensitive data like page mapping information.

While manual decomposition is suitable for well-scoped kernel data, automated static analysis is necessary when the data to be isolated is non-trivial to infer (e.g., when the sensitive data is mixed with other kernel data). For example, KENALI [154] protects access control checks scattered throughout the kernel by isolating all data that is relevant to the checks. It leverages the implicit semantics of security checks, which specify that every check must return a security-related error code when it fails. The semantics help KENALI reduce its search space to only the functions that could return a “permission denied” error code. A sound dependency analysis to ensure that no sensitive data is left unprotected then becomes feasible, as long as the kernel code strictly adheres to the semantics.

**4.1.3 Policy Enforcement Decomposition.** Unlike device drivers and security-critical state, a host’s policy enforcement mechanism is often *directly* loaded into a secure compartment during secure boot, *without* the need for any decomposition analysis. For instance, PerspicuOS [65] and SKEE [35] provide a secure environment where the policy enforcement mechanism can safely execute even in a malicious kernel. Since a secure compartment already isolates the policy enforcement mechanism from the kernel, the mechanism itself does not need to be decomposed to protect its critical data.

## 4.2 Glue Code Generation

The main purpose of “glue code” is to *marshal* and *synchronize* data across compartment boundaries. Marshaling is necessary, because some isolation techniques (§5) place a kernel compartment into a separate virtual address space, so a cross-compartment procedure call is needed to execute a function in another address space. Synchronization is necessary, because all Unix kernels are reentrant [47], i.e., multiple kernel processes can safely execute concurrently without



introducing any consistency issues among kernel resources. Shared resources and synchronization primitives (e.g., spinlocks and mutexes) must be updated upon every function invocation that crosses the compartment boundary. We discuss glue code generation for each type of decomposition we presented in §4.1.

**4.2.1 Device Driver Decomposition.** Most software-based kernel compartmentalization systems [52, 70, 123, 177] isolate device drivers in a *logically separate* portion of the kernel address space, so they remain in the same virtual address space as the kernel. These systems typically enforce isolation via run-time checks on read, write, and jump instructions, restricting access to arbitrary kernel memory while sharing with the rest of the kernel a single copy of kernel data. Unless these checks require additional information from the kernel that must be updated upon cross-compartment transitions (e.g., in BGI [52], an access control list is used for permission checking; see also §5.4), data marshaling and synchronization is generally not necessary. On the other hand, Nooks [160] requires glue code, but only for data synchronization, because it keeps a shadow copy of shared kernel resources for isolated device drivers to directly modify; marshaling is not needed, because kernel pages are mapped into a compartment’s page table for read access.

Glue code can be generated *manually* [126, 134, 160, 173] or *automatically* [73, 131, 132, 145]. Nooks [160] does so manually; developers implement synchronization routines to copy shared resources between an isolated device driver and the kernel. Microdriver [73] and Decaf [145] leverage field access analysis, a static analysis technique that identifies the fields of a complex data structure crossing the compartment boundary, to automatically identify data fields to be marshaled and synchronized. However, as discussed in §4.1, static analysis is complicated by the proliferation of pointers in the kernel. For example, C/C++ pointers do not carry bounds information, making data marshaling difficult [118]. To ensure completeness, Microdriver and Decaf still rely on user annotations to understand the semantics of pointers in kernel data structures (e.g., to determine the size of an array). *Automated marshaling* of pointer data has subsequently been made possible by PtrSplit [118]. PtrSplit instruments the compartmentalized program to track the bounds of pointers, so that all pointers that cross the compartment boundary have the bounds information necessary for automatically marshaling pointer data.

LXDs [131] and LVD [132] do not use static analysis but instead define an Interface Definition Language (IDL), a special-purpose language for describing interfaces between software components, to automatically generate glue code. However, they still require developers to manually define the IDL specification for device drivers. To reduce the effort involved,

KSplit [87] uses static analysis to generate IDL specifications for marshaling and synchronizing data across compartment boundaries; user annotations are needed only when the analysis cannot resolve ambiguities (e.g., unions or void pointers). Once a working IDL specification is defined (either manually in LXDs and LVD, or semi-automatically in KSplit), the IDL compiler can automatically generate the glue code.

**4.2.2 Security State and Policy Enforcement Decomposition.** Sharing protected data between a compartment and the kernel is disallowed *by design* in mechanisms that isolate security state or support secure policy enforcement, because the goal of a safebox compartmentalization system [35, 59, 65, 80, 142, 154] is to ensure that the protected memory can *only* be accessed by the trusted code in the isolated compartment. Therefore, such a system does not require any glue code for data marshaling or synchronization, irrespective of the isolation technique used for compartmentalization.

### 4.3 Summary

A monolithic kernel is typically built from a large and complex code base. Dependencies between sub-components are difficult to identify, frequently hidden behind programmatic abstractions. Relying completely on manual decomposition effort is impractical, especially given the sheer amount of code to analyze and the large number of modifications to the code base required by compartmentalization (e.g., inserting synchronization primitives). Meanwhile, fully automated approaches also fall short of expectations. *Finding a middle ground with acceptable trade-offs seems inevitable:* We must continue to improve automated program analysis techniques to identify compartment boundaries and insert necessary code for across-boundary interactions, and involve humans in the loop only when code complexity exceeds the capacity of current analysis techniques. While advances in automated techniques continue to reduce the amount of manual intervention required, it is unclear from present work how often humans are involved in practice. We further discuss this issue in §7.

## 5 COMPARTMENT ISOLATION MECHANISMS

A kernel compartmentalization system must enforce an isolation policy on a decomposed kernel component to control resource accesses and restrict control transfers. However, unlike user-space isolation, where policy enforcement typically runs at a higher privilege level than the isolated user-space code [95, 120, 130, 137], kernel isolation faces the challenge that the isolated component executes in the same privileged kernel mode as the rest of the kernel. Since kernel privilege grants an attacker access to critical kernel state and code in memory, enforcing kernel isolation requires some form

of privilege separation from the decomposed kernel component to prevent security bypass. Existing work leverages the *user/kernel* (§5.1) or *kernel/hypervisor* (§5.2) privilege separation to enforce isolation at a higher privilege level than the decomposed kernel component. Isolation can also be enforced at the same privilege level (i.e., *intra-kernel* isolation) using either *address-based* (§5.3) or *domain-based* (§5.4) techniques. We categorize isolation mechanisms based on these privilege separation models and discuss how they enforce sandbox- and safebox-based kernel compartmentalization.

## 5.1 User/Kernel

Commodity OSs leverage processor modes (e.g., user and kernel mode) to enforce privilege separation. All user code runs in a separate virtual address space and must rely on the kernel via system calls to execute privileged instructions.

**Sandbox-based Isolation.** The hardware-enforced user/kernel boundary can be used to sandbox untrusted kernel code (e.g., device drivers) in various ways. At one extreme, early work effectively converts a monolithic kernel into a microkernel architecture (Fig. 1). For example, Sawmill [75] partitions the monolithic kernel into a core kernel and a number of services, and moves these services to user space. However, switching between the user- and kernel-mode code is costly, especially for high-throughput devices such as network cards. To address this issue, Microdriver [73] sandboxes only a subset of device driver functionality (Fig. 3a), relegating rarely used code (e.g., device startup and shutdown) to user space while leaving performance-critical code (e.g., data transmission) in the kernel. The performance gain comes at the cost of security, since the kernel-mode code, which might contain faults to corrupt the kernel, continues to execute in the same address space as the rest of the kernel. However, unused or less-frequently-used code – which is isolated by Microdriver – is the subset of the code base that is known to contain more bugs [38, 77, 107] and thus more likely to be leveraged as an attack vector. Renzelmann et al. [145] take this approach one step further, implementing user-space code in a safer language (Java in their prototype, see Fig. 3b) that is less prone to memory vulnerabilities common in C.

**Safebox-based Isolation.** Many user-space compartmentalization systems [95, 137] leverage Trusted Execution Environments (TEEs) to protect applications from a malicious kernel. For example, Intel’s Software Guard Extensions (SGX) [58], a widely used TEE, uses hardware-based cryptography to protect the confidentiality and integrity of user-mode code from the rest of the system. However, kernel functionality cannot execute within an SGX enclave; moreover, as with sandbox-based isolation, any communication between an enclave and trusted kernel functionality introduces large overhead. Therefore, we expect that user-space TEEs such

as SGX will remain of limited practical use for compartmentalizing kernel functionality. However, as we discuss in §5.2, ARM’s TrustZone technology [85] has been used for kernel compartmentalization.

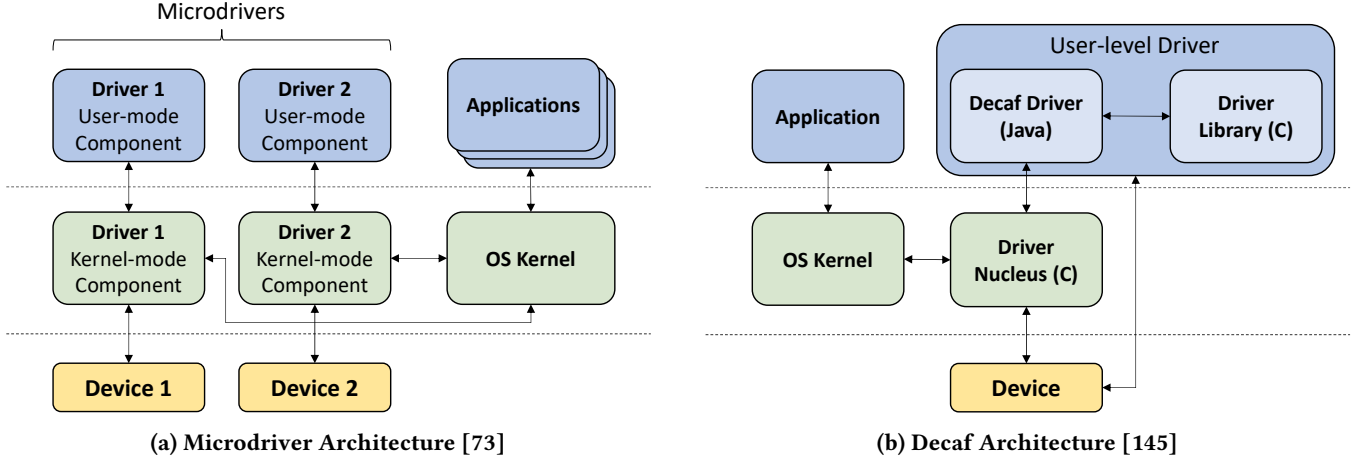
## 5.2 Kernel/Hypervisor

Hypervisors run at a higher privilege level than the kernel, managing virtual machines (guest OSs) that share physical computing resources. Kernel compartmentalization isolates a kernel component within a VM, which executes in a separate virtual address space, while enforcing security policies from the higher-privileged hypervisor. Unlike user/kernel isolation (§5.1), compartments in kernel/hypervisor isolation execute in privileged kernel mode and therefore have direct access to privileged instructions. However, since attackers can abuse privileged instructions to break the isolation, a compartment is subject to some restrictions on its use of those instructions, whereby protected sensitive instructions (e.g., programming of Interrupt Controllers [132]) require a hypercall into the hypervisor. We discuss techniques [131, 132] that mask and thus reduce the cost of transitioning across the kernel/hypervisor boundary in §5.5.

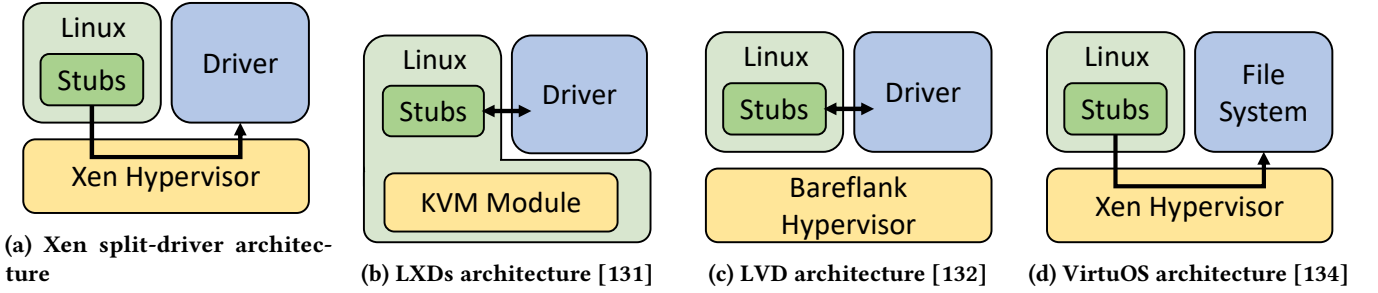
**Sandbox-based Isolation.** The Xen hypervisor’s *split-driver* model (Fig. 4a) is an early example of using virtualization for kernel compartmentalization [141]. In this model, a front-end driver running in a DomU VM forwards all driver requests to a back-end driver that can interact with the physical device. By default, the back-end driver runs in the Dom0 VM, a privileged VM started by the hypervisor to manage DomU VMs. However, a faulty back-end driver in Dom0 can affect the hypervisor; moreover, it cannot easily *recover* from an error, since the Dom0 VM cannot be restarted. To improve *reliability* and *recoverability*, the split-driver model allows the back-end driver to run in a separate DomU VM, called the *driver domain* [32], so that a faulty driver can be killed or restarted independent of other VMs (including the one running the trusted kernel).

LXDs [131] (Fig. 4b) use the KVM hypervisor [27], instead of Xen, to compartmentalize device drivers. Like in Xen’s split-driver model, LXDs run an untrusted device driver in a separate VM. However, rather than running in a separate VM, LXDs’ trusted kernel acts as the hypervisor to manage VMs running untrusted compartments.

Both the split-driver model and LXDs incur costs when transitioning between the kernel and hypervisor privilege levels, which is more expensive than transitioning between user and kernel levels through system calls [133]. To reduce these costs, LVD [132] (Fig. 4c) uses VMFUNC [89] to enable direct communications between VMs. Like in the split-driver model, the trusted kernel and untrusted device drivers run



**Figure 3: Different user/kernel compartmentalization approaches.** In Microdriver, the user-mode driver is written in the same C language as the kernel-mode driver; in Decaf, the user-mode driver can be written in languages other than C.



**Figure 4: Different hypervisor-based compartmentalization approaches.**

in separate VMs, but communications between them do not require going through the hypervisor (see also §5.5).

Finally, we note that some work has explored ways to compartmentalize beyond device drivers. For example, VirtuOS [134] (Fig. 4d) runs a complete kernel subsystem (e.g., a network stack or a file system) in a separate VM.

**Safebox-based Isolation.** The hypervisor can enforce fine-grained memory permissions on VMs running trusted compartments by managing multiple *views* of memory. Each view defines a set of access permissions for a VM to restrict code and data accessible from outside the VM. xMP [142] uses this approach to ensure that only trusted compartments can access security-critical state. Conceptually, this approach shares some similarities with the ones we see in §5.3.

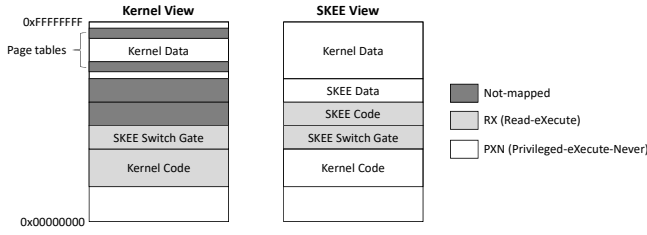
Similar techniques have been applied using ARM’s TrustZone [85, 140]. For example, Azab et al. [36] inserts hooks in the kernel to transfer control to a kernel running in the TrustZone. The safebox is fully protected from the kernel being monitored and has full access to its memory. Others [112] go a step further and propose extensions to provide secure

access to devices through the safebox, which can be used to implement I/O reference monitors [97, 167].

**Optimizing Domain Switches.** As seen in §5.1 and §5.2, moving kernel code to user space or a different hypervisor domain is an attractive way to compartmentalize a monolithic kernel, as doing so leverages pre-existing and well-known hardware-enforced privilege separation. However, calling kernel functionality outside of a compartment requires an expensive domain switch. We can minimize this overhead in three ways [131, 132]: 1) carefully identifying compartment boundaries to reduce the number of calls (see §4); 2) using hardware support (see §5.5); and 3) adopting an asynchronous model and a more efficient IPC mechanism (see §5.5).

### 5.3 Address-based Intra-kernel

*Intra-kernel* isolation means that an isolated component runs at the same privilege level as the rest of the kernel. As a result, the compartment has supervisor privilege to execute privileged instructions or access privileged registers. We classify



**Figure 5: An example of address space separation on ARMv7 in SKEE [35]**

intra-kernel approaches into two types: *address-based* (this section) and *domain-based* (see §5.4).

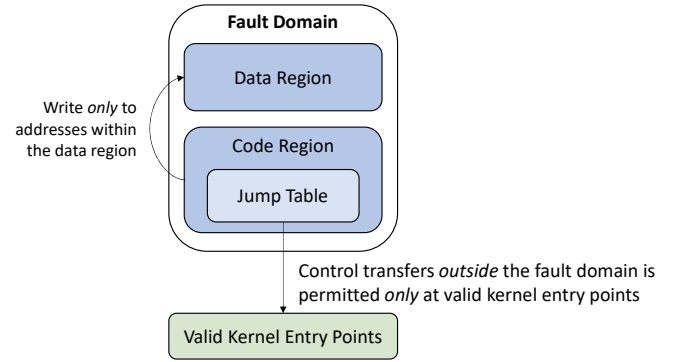
Address-based compartmentalization uses the Memory Management Unit (MMU) to run a compartment in a separate address space for isolation. Since address translation is not shared between compartments, an isolated compartment cannot access data from other compartments. However, modification to MMU control registers could allow an untrusted compartment to access the data of a trusted compartment. Therefore, depending on the threat model, additional security measures are required.

#### Sandbox-based Isolation.

Nooks [160] leverages the MMU to isolate kernel extensions. Each compartment has its own set of page tables with appropriate page permissions. Specifically, isolated kernel extensions have read-only kernel access and read-write access to their own domains (while the core kernel has read-write access to the entire address space). Configuring the page table for a domain to remove write access to the rest of the kernel address space thus suffices to prevent inadvertent memory corruption. Note that Nooks is a fault-resistant system designed to prevent a buggy extension from inadvertently corrupting important kernel data structures; a malicious extension can bypass its isolation by reloading the hardware page-table base register.

Since compartments are not permitted to modify kernel data structures directly, Nooks supports cross-compartment interaction through interposition and object tracking mechanisms that copy kernel data to compartments and copy them back atomically after changes have been applied, sanitizing the changes where possible. If a check (e.g., parameter validation) fails, or a hardware fault occurs during domain execution, Nook’s object tracking and recovery mechanisms release kernel resources held by the domain and reset the system to a safe state.

**Safebox-based Isolation.** Safebox systems such as PerspicuOS [65] and SKEE [35] use the MMU to enforce intra-kernel isolation on commodity hardware even when the kernel is malicious. In these systems, a *trusted* compartment is responsible for mediating all page table modifications. Its page



**Figure 6: Software fault isolation [166]**

table pages are either read-only by the rest of the (potentially malicious) kernel or completely inaccessible from outside the compartment.

For example, as illustrated in Fig. 5, SKEE instruments the kernel’s memory translation tables to 1) unmap entries related to SKEE’s isolated environment; and 2) restrict the kernel’s write access to the kernel code and SKEE’s switch gate, which mediates *all* jumps from the kernel into the isolated environment. As such, it safeguards the kernel’s only entry point to the isolated environment.

However, as with Nooks, a malicious compartment with access to the MMU control registers can change its address translations to map the trusted compartment’s pages. To prevent this, both PerspicuOS and SKEE scan the kernel code to remove any instruction that might be used to modify MMU control registers. The registers to be protected are architecture specific. For example, PerspicuOS protects CR0’s write-protection (WP) bit on x86/x86-64 and several other control registers, while SKEE protects ARM’s translation table base registers (TTBR0/1) and the translation table base control register (TTBCR). We note that a similar approach could be applied to sandbox systems (like Nooks) to provide stronger isolation between compartment boundaries.

## 5.4 Domain-based Intra-kernel

Relying on privilege levels to enforce compartment boundaries comes at the cost of expensive privilege-level transitions. To avoid this overhead, researchers have explored ways to enforce restrictions on data accesses and control transfers within a single address space at the same privilege level. Domain-based intra-kernel compartmentalization divides an address space into multiple *domains* that share the same address translation with each other and the rest of the kernel. To create a domain, existing work proposes both software- and hardware-based mechanisms to control access to specific regions of the virtual memory. The *granularity* of a domain varies depending on the mechanism.

**5.4.1 Software-based Isolation.** Software-based techniques typically instrument untrusted code with run-time checks to ensure that data accesses and control flow transfers are within specified bounds as illustrated in Fig. 6.

**Sandbox-based Isolation.** Early work built on ideas from *software fault isolation* (SFI) [166] to enforce isolation. In comparison to address-based techniques, SFI’s run-time checks introduce additional overhead for intra-compartment execution. Conversely, SFI avoids costly page table updates and data synchronization on compartment transitions [166]. Moreover, SFI enables fine-grained isolation at the object or even byte level with minimal or no changes to the existing code and data layout.

For example, the SafeDrive compiler [177] inserts run-time checks on pointer variables in untrusted extensions to enforce type and memory safety within the extension, which prevents spatial bounds violations in kernel API calls and direct accesses to shared kernel objects. BGI [52] creates an Access Control List (ACL) for each byte of memory to specify access permissions for untrusted extensions. ACLs are stored in a protected area in the main memory to prevent any modification from within an extension. The BGI compiler instruments write instructions with run-time checks to restrict an extension’s access to kernel memory. Since each run-time check costs additional CPU cycles, to reduce performance overhead, BGI does not instrument read instructions; as a result, BGI can only guarantee *integrity* but not *confidentiality*. Trading off confidentiality for performance is in fact common among SFI-based systems [70, 123].

**Safebox-based Isolation.** Using software-based isolation techniques to protect security-critical state in an untrusted kernel is challenging. The reason is that the kernel has the privileges to perform a range of low-level operations that can violate language-level safety assumptions; these operations can be abused to corrupt *arbitrary* kernel memory. Criswell et al. [61] identified MMU configurations, DMA, and page swapping as the three kernel functionalities subject to such manipulation. They also identified four types of kernel memory that must be protected to prevent security bypass: 1) the processor state when held in registers or memory, 2) stack values for kernel threads, 3) memory-mapped I/O locations, and 4) code pages in memory.

To address some of these challenges, they proposed secure virtual architecture (SVA) [61]. SVA is a compiler framework that compiles the kernel to a virtual instruction set based on LLVM’s intermediate representation and statically checks that the intermediate bytecode meets several safety properties (e.g., control-flow integrity, type safety for a subset of objects). For safety properties that cannot be guaranteed statically (e.g., dynamic bounds and indirect calls), SVA inserts run-time checks. In addition, SVA disallows explicit assembly code in the kernel (e.g., code that handles low-level

hardware interactions to manipulate the MMU); instead, it requires the assembly code to be replaced by special SVA instructions [62], implemented as high-level, well-defined intrinsic functions, to enforce memory safety guarantees. For example, the `sva.save.integer(void* buffer)` function is used to save the processor’s integer *control state* (i.e., control and general-purpose registers) into the memory pointed to by `buffer` (e.g., on a context switch).

KCoFI [59] uses SVA to enforce *control-flow integrity* (CFI) on kernel code. To protect the security-critical state of the CFI mechanism (e.g., page mapping information), KCoFI stores them in an isolated memory region. KCoFI instruments all kernel store instructions to prevent errant writes from corrupting them. To ensure the integrity of the kernel code and the address space layout, KCoFI uses SVA to restrict how the kernel configures the MMU. Further, it relies on SVA to protect against DMA attacks that manipulate the IOMMU and several other low-level, security-sensitive kernel operations and states as identified by Criswell et al. [61].

**5.4.2 Hardware-based Isolation.** Hardware features from different CPU vendors vary, but existing domain-based work typically leverages them to provide page-sized or even finer-grained (e.g., byte-level) isolation. Some hardware features can be used in both sandbox and safebox architectures to optimize the techniques mentioned in software-based isolation (§5.4.1). Therefore, we organize the remainder of this section based on the granularity of the protection, instead of on the sandbox/safebox architecture.

**Page-sized Isolation.** *Memory Protection Keys* (MPK) were introduced in Intel’s Skylake processor [89] to separate the virtual address space into 16 domains using four designated bits in every page table entry. Each CPU core has a special `pkru` register that stores the core’s permissions for every domain. However, while only the kernel can modify a page’s domain (since it is stored in the page table), the `pkru` register can be modified in user mode. If MPK is used for kernel compartmentalization, then any user process can invalidate compartmentalization by changing the value of the `pkru` register. To circumvent this limitation, IskiOS [80] reserves eight domains for kernel compartmentalization and allows user-space applications to use only the remaining domains. When a user process attempts to write to the `pkru` register, IskiOS traps to the kernel, which intercepts the instruction and checks that only user-space domains are modified. However, user applications running on IskiOS incur performance overhead on every write to `pkru`. It is important to note that IskiOS assumes that the kernel might have buffer overflow bugs, but is not compromised by the attacker.

Recently, Intel announced *Supervisory Keys* (PKS) [29] which adds a new `pkrs` register that is accessible only from

the kernel mode. PKS can be used to implement kernel compartmentalization, just as described by IskiOS, but without trapping on writes to the `pkru` register and instead using the `pkrs` register. At the time of this writing, however, no system has used PKS for kernel compartmentalization.

Similar to MPK, *Memory Domains* (MD) on ARM-v7 CPUs enables an address space to be split into compartments. The kernel assigns every page to one of 16 domains using a 4-bit flag in the Page Directory Entry (PDE). The Domain Access Control Register (DACR), similar to the `pkru` register, dictates each core's permission for every domain. However, unlike the `pkru` register, both PDE and DACR can be accessed only in kernel mode. Therefore, MD-enabled kernel compartmentalization seems plausible, but prior work [54] has used MD to create user-space compartments only. Similar to IskiOS, such an approach trusts parts of the kernel to set up these domains.

**Finer-grained Isolation.** Mondrix [173] proposes custom hardware extensions to support fine-grained compartmentalization of the Linux kernel. Mondrix builds on Mondrian [172], which supports flexible, controlled memory-sharing between isolated domains in user space. Mondrian associates a domain to a permission table, which describes word-level access permissions in a process' address space. Each core has a register that points to the permission table associated with the current domain and a dedicated lookaside buffer for performance. Permission tables can only be edited by a privileged memory supervisor running in domain 0. Mondrix adapts Mondrian's memory protection mechanisms to the Linux kernel. As with Mondrian, only the memory supervisor can access all of the memory without the mediation of a permissions table. There is one memory supervisor in the kernel, and it is trusted. To improve performance of domain switches, Mondrix stores context information (e.g., stacks) in user memory that is writeable by hardware but read-only to software other than the memory supervisor. This reduces the overhead from context switching between kernel threads on cross-domain calls. Mondrix also adds hardware to protect stacks of different threads resident in the same protection domain. Apart from the need for custom hardware, a disadvantage of Mondrix is the need for a separate protection table for each domain, which may limit domain scalability. In addition, although it provides finer granularity than that of MMU approaches (§5.3), Mondrix still only provides limited protection for sub-word granularity allocations (e.g., array entries or stack frames).

To support byte-level memory protection, Intel introduced *Memory Protection Extensions* (MPX) [89] in 2015. With MPX, programmers can create and enforce bounds by specifying two 64-bit address denoting the start and end of a memory range. Two new instructions, `bndcl` and `bndcu`, perform

bounds checking on an address. For compartmentalization, the address space is partitioned using MPX bounds. By defining a single bound and adding a single bound check before every memory access, we ensure that every access is within the allowed partition on the address space. However, MPX's bounds registers (`bnd0` to `bnd3`) can store at most four bounds; additional bounds must then be stored in memory, which affects performance [104]. Therefore, MPX-based approaches can only efficiently support four domains, and the performance degrades with more domains. Koning et al. [104] show that MPX outperforms SFI in most cases, but since this work is done only for user-space compartmentalization, the applicability of their results to kernel space remains to be confirmed. Note that MPX was discontinued in 2019 due to performance overhead [135] and side-channel vulnerabilities [50].

ARM's Memory Tagging Extensions (MTE) [34] and Pointer Authentication Codes (PAC) [5] are new security features on ARM processors. With respect to kernel compartmentalization, MTE is similar to both Intel MPX in that it supports fine-grained memory protection, and to Intel MPK in that it associates a domain with a virtual memory region. Unlike MPK, its granularity is much smaller (16 bytes instead of 4KB). In MTE, each protected memory region is tagged with one of up to 16 *colors*. A straightforward application of MTE to kernel compartmentalisation is to tag each domain with a unique color. However, as with MPK, this approach suffers from limited domain scalability.

HAKC [126] combines PAC and MTE to overcome the scalability limitations of MTE for kernel compartmentalisation. PAC introduces new instructions for code and data *pointer authentication* to protect the integrity of code and/or data pointers. On pointer creation, PAC computes a cryptographic message authentication code (MAC), referred to as a *pointer authentication code* (PAC), over the pointer and additional non-secret context data. The PAC is then stored in unused higher-order bits of the pointer. Subsequent attempts to load and/or dereference the pointer can be instrumented with instructions to check whether the pointer still matches its PAC and has not been corrupted by an adversary. PAC does not provide full memory safety and is potentially vulnerable to malicious PAC generation and reuse attacks [115]. Nonetheless, precise code pointer integrity [108] prevents all known control-flow hijacks, and data pointer integrity protects against many powerful data-oriented attacks (e.g., [86]).

To combine PAC with MTE, HAKC proposes a two-level design where compartments are subdivided into up to 16 different *cliques*. Cliques can access data in other cliques in the same compartment, subject to a *clique access policy*, but are prevented from accessing other compartments. Control transfers between compartments are restricted according

to a *compartment transition policy*, and can only jump to well-defined entry points in other compartments.

To enforce the access and transition policies, HAKC assigns different MTE colors to the cliques within a compartment. Importantly, MTE colors are *not* stored directly in pointers nor are MTE instructions used to check data accesses or control transfers. Instead, HAKC signs each pointer using PAC, but concatenates as part of the PAC context a unique compartment identifier and the colors of accessible cliques within the compartment. This prevents a clique from accessing cliques in other compartments even though colors may be reused. However, control transfers to a new compartment must also transfer any data needed from the old compartment, introducing additional overhead. Since there is no limit to the number of compartments, HAKC’s two-level design thus introduces a trade-off whereby adding more compartments improves security through finer-grained compartmentalisation, at the cost of increased overhead caused by data transfers between compartments.

## 5.5 Optimizations in Cross-compartment Communication

Apart from a few SFI approaches (e.g., [52]), *cross-compartment communication* constitutes the main source of overhead after compartments are created [130–132, 134, 142]. Many of the kernel compartmentalization mechanisms in this section use a range of techniques to speed up or mask these costs. While many of the optimizations were proposed for specific mechanisms, their applicability can be considered more widely.

**Asynchronous Communication.** VirtuOS [134], LXDs [131], and LVD [132] employ asynchronous communication between compartments to mask the cost of domain switches. This transforms the programming model from a blocking to a non-blocking one, and thus some re-engineering of isolated compartments might be required to perform useful work while waiting for an asynchronous call to complete. To avoid the engineering effort, an alternative approach is to retain the blocking programming model but use a cooperative threading library to multiplex multiple lightweight threads onto a single hardware thread. When a lightweight thread makes a cross-domain call, it transparently yields to another lightweight thread in the same domain. This approach also facilitates batching and pipelining of cross-domain calls made by different lightweight threads.

**Leveraging Cache Coherence.** Removing the kernel from the critical path in IPC was first proposed by URPC [43]. LXDs [131] take this idea one step further: Instead of using memory as the point of coherence for cross-domain invocations, they use cache coherence to speed-up communication

(in a fashion similar to Barrelfish [40]). Switching across domains to transfer information becomes unnecessary, but it requires switching to an asynchronous programming model (as described above) and pinning domains that need to communicate to different cores.

**Sparing the Hypervisor.** Switching between address spaces of different virtualization domains requires a VM exit, which is significantly more expensive than a traditional system call [171]. Intel introduced VMFUNC that allows *intra-domain* switching between a list of predefined extended page tables (EPTs) with overhead comparable to a system call. LVD [132] uses VMFUNC to speed up cross-domain invocations. However, there are security implications to using VMFUNC instead of trapping to the hypervisor. For instance, unlike the case of VM enter/exit, the entry/exit into the domain is not happening at predefined points, as the VMFUNC changes the current address view of the domain and continues to execute the next instruction. This means that an attacker inside an isolated domain can potentially find executable bytes that make up the VMFUNC instruction, and trigger an illegal VMFUNC call. LVD uses two additional techniques to safeguard against this attack vector: 1) Scanning (and rewriting) the executable code to find arbitrary bytes that can form a VMFUNC instruction, 2) Ensuring that the list of allowed EPTs for the VMFUNC is restricted to only two domains i.e., the kernel and the isolated domain. These techniques are generally applicable (or rather needed) in any implementation that uses VMFUNC to compartmentalize the kernel. xMP [142] uses Virtualization Exceptions (#VE) to handle memory access violations instead of trapping to the hypervisor.

Both of these techniques give more capabilities to an isolated compartment to reduce the number of calls to the hypervisor; conceptually similar approaches could be employed by other compartmentalization mechanisms too. For example, Dune [42] elevates the capabilities of a user process to manage its own page table so that it does not have to trap to the kernel on page faults. This technique could be applied to user-space drivers to reduce the number of boundary crossings between user space and the kernel.

## 6 EMERGING HARDWARE FEATURES

In this section, we discuss emerging hardware features that could be used to improve the performance of different aspects of kernel compartmentalization mechanisms described in §5.

### 6.1 Encryption-based Techniques

AMD [3] and Intel [24] have proposed new hardware extensions to encrypt pages belonging to a guest VM. This encryption prevents the hypervisor from reading the data pages of the VM. Just as virtualization has been used to implement



sandboxes in the kernel [131, 132, 134], we speculate that these hardware extensions could be adapted to implement safeboxes to protect data in VMs from the hypervisor.

## 6.2 Hardware Capabilities

CHERI [168] is an extension to RISC [174] and ARM-v8 [6] instruction set architectures (ISAs). It replaces a memory address (pointer) with a 128-bit *capability* and a 1-bit out-of-band *tag*. As shown in Fig. 7, along with the value of an address, a capability holds additional information about the memory region being accessed, e.g., its bounds and associated permissions. The capability is considered *valid* only when the tag bit is set. When a new capability needs to be created, it must be derived from a valid source capability. A legal derivation requires the bounds and permissions of the derived capability to be lesser than those of the source capability. Otherwise, the hardware clears the new capability’s tag bit, thus making it invalid. An exception is raised at run time when an invalid capability is used in a *load*, *store*, or *jump* instruction. The tag bit cannot be explicitly set for a capability; therefore, it is impossible for software to forge capabilities.

When a program accesses a memory location outside the range specified by the capability (e.g., a buffer overrun), the hardware raises an exception, providing *spatial memory safety*. If the user directly manipulates the address of a jump target, the corresponding tag bit will be cleared, and the capability becomes invalid. Executing an instruction that uses the capability in a jump will raise an exception, and thus, CHERI also provides *control flow integrity*. Prior work [147] uses CHERI to build user-space compartments to achieve both properties.

CheriBSD [168] is a port of FreeBSD to CHERI, in which every pointer in the kernel is a CHERI capability. At the time of this writing, CheriBSD does not create compartments inside the kernel using its capabilities. However, just like Mondrix used Mondrian for kernel compartmentalization, CHERI could be used to compartmentalize the kernel in CheriBSD. There also exists an ongoing effort to port Linux to the CHERI hardware [7], but to the best of our knowledge, similar to CheriBSD, it does not create compartments inside the kernel using CHERI.

## 6.3 Control Flow Integrity

Intel Control-Flow Enforcement Technology (CET) [31] provides two new ISA capabilities to defend against control-flow attacks [150]: Shadow Stack [49] (SS) and Indirect Branch Tracking [139] (IBT). Shadow Stack (SS) is an additional stack for return addresses only. The stack has a Shadow Stack Pointer (SSP) pointing to the last address a program is expected to return to. If the program returns to a different

address, an exception is triggered. Indirect Branch Tracking (IBT) is a feature that defends against jump/call-oriented programming (JOP and COP) attacks. By keeping track of the expected targets of indirect branches, it raises an exception if the target is not one of the expected targets. CET could be used to optimize software-based approaches for CFI mentioned in §5.4.1 such as KCoFI [59, 61].

## 7 EVALUATING KERNEL COMPARTMENTALIZATION

In this section, we describe three categories of evaluation criteria used in current kernel compartmentalization literature, i.e., *performance*, *security*, and *practicality*. We discuss the difficulties in comparing among different compartmentalization techniques, motivating the need for our community to agree on *standardized benchmarks* for meaningful comparison. Table 2 summarizes the experimental setups and three main factors (i.e., availability, documentation, and reliance on specific hardware features) that could affect comparability and reproducibility of the systems we survey.

### 7.1 Performance Evaluation

Using only a single performance measure is insufficient to fully evaluate a compartmentalized kernel, because different techniques can affect the performance of different aspects of the system. As such, existing work uses various *micro-* and *macro-benchmarks* to measure and understand performance overhead.

**Microbenchmarks.** The most commonly used microbenchmark to evaluate core kernel functionality is LMBench [127], which measures the performance impact on individual system calls [59, 65, 154]. To measure the overhead on network throughput, systems that isolate network drivers [123, 131, 132, 160] typically use netperf [93] and iperf [25] to evaluate the overall send/receive bandwidth. KENALI [154] and SKEE [35] use the ARM cycle count register (CCNT) to measure the latency of a round-trip context switch as the cost of crossing isolated compartments. SFI-based systems [70, 123] use the SFI microbenchmarks [152] to measure the overhead on tasks performed by a wide range of kernel extensions, such as page eviction, MD5 fingerprinting, and logical disk operations.

**Macrobenchmarks.** In contrast to microbenchmarks, macrobenchmarks can provide good indicators of a system’s performance under realistic workloads. Network applications like Apache web servers make heavy use of kernel functionality, so existing systems [59, 60, 65, 160] use ApacheBench [2] and httpperf [129] to measure the overhead on a web server’s bandwidth. Besides network servers, some systems also perform macrobenchmarking using client-side applications. For example, SVA [60] measures file conversion time (from WAV

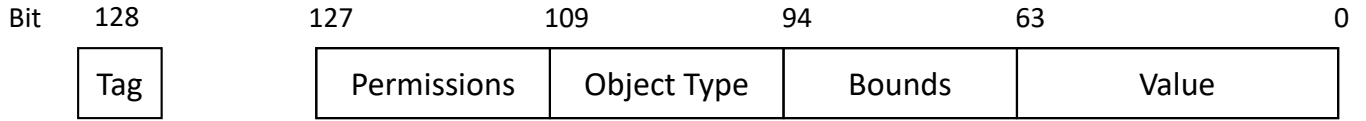


Figure 7: 128-bit CHERI capability representation (labelled bit 0–127) with an out-of-band tag (bit 128) [169]

to MP3) using the LAME MP3 encoding benchmark [136] and file compression time of bzip2 to evaluate its performance overhead on end-user applications. KENALI [154] and SKEE [35] use Android benchmarks, such as AnTuTu [4], Geekbench [23], and Vellamo [143], to simulate user activities like web browsing and gaming on their Android compartmentalization prototypes.

## 7.2 Security Evaluation

Security evaluation is used to verify the effectiveness of a kernel compartmentalization mechanism against faults and vulnerabilities. To evaluate fault-isolating and fault-resistant systems (see §3.1), *fault injection* is used to validate that they isolate the kernel from faults that would otherwise lead to undefined behavior or system failure. For example, Nooks [160] adapted a synthetic fault injector to emulate programming errors (e.g., uninitialized local variables) in kernel extensions.

To evaluate a system’s ability to mitigate vulnerabilities, authors either perform experiments with real-world exploits from *Common Vulnerabilities and Exposures* (CVEs) (a database of publicly disclosed cyber-threats) [52, 60, 73, 123, 134, 142, 154, 160, 173], or provide case studies of a vulnerability or a class of vulnerabilities [80, 126]. The former replicates CVEs that match the threat model assumptions, evaluating a kernel compartmentalization system against the selected real-world exploits to test its efficacy. For example, LXFI [123] evaluates its ability to thwart privilege escalation vulnerabilities by transmitting crafted messages over reliable datagram sockets, triggering an exploit that allows an attacker to execute arbitrary code in the kernel [8]. The latter focuses on analyzing how an attack *would* be mitigated by a kernel compartmentalization system, assuming that the system correctly implements its design specifications.

Quantitative metrics also exist to measure the extent to which a compartmentalization mechanism reduces the attack surface. For example, KCoFI [59] uses the *average indirect target reduction* (AIR) metric [176] to quantify the reduction of possible targets of indirect control-flow transfer, as well as the open-source ROPgadget tool [146] to measure the reduction of possible ROP gadgets. KCoFI and PerspicuOS [65] both use the SLOCCount tool [170] to measure a kernel’s TCB size (in lines of code) as an indirect measure of its attack surface.

## 7.3 Practicality Evaluation

Evaluating the practicality of a kernel compartmentalization mechanism through the amount of *engineering effort* helps identify potential barriers to its adoption. One measure of such an effort is the number of lines of code that are modified or added. For example, PerspicuOS [65] counts code modifications in Git change logs to quantify the effort of porting a commodity kernel to the nested kernel architecture. LXFI [123] and Microdriver [73] use the number of annotations to measure the engineering overhead of modifying kernel modules.

## 7.4 Limitations of current evaluations

It is difficult to understand how a given kernel compartmentalization approach compares to another, because different techniques have been implemented over a multitude of OSs. For example, BGI [52] and LXFI [123], while similar in many aspects, are implemented in Windows and Linux, respectively. Different hardware used in evaluation also makes direct comparison difficult. For example, Mondrix [173] requires hardware that supports Mondrian Memory Protection, which is not widely available. Even when two approaches use the same OS, their performance can still vary significantly across OS *versions* or as a result of misconfigurations [144]. In addition, two solutions may assume slightly different threat models, which lead to different sets of trades-offs (see §3).

**Performance.** We suggest that future work evaluate the overhead introduced by common compartmentalization operations, such as the transitioning cost, against other systems (or at a minimum, against some well-known “cost” such as function calls or system calls) to facilitate order-of-magnitude comparison. Unfortunately, only a few evaluations provide such order-of-magnitude comparisons [131, 132]. We acknowledge the difficulty of comparing different solutions. However, some baselines, such as the SFI microbenchmark [152] frequently used to evaluate SFI-based kernel compartmentalization systems, should be adopted to provide a common frame of reference. For example, LXFI compares its SFI microbenchmark results to those of XFI without the need to re-run the same experiments on XFI.

**Security Effectiveness.** Vulnerabilities are hard to replicate as they get patched out over time. To evaluate whether two competing systems mitigate the same vulnerability, forward- or back-porting of these approaches to the same affected OS

**Table 2: An overview of the possible factors that may affect the comparability of systems. For each selected system, we identify its experimental setup, the availability of its source code and documentation, and its reliance on hardware-specific features. The last column indicates whether a system was compared with similar work.**

| System                 |            | Experimental Setup                        |                                           | Reproducibility |               |                                        | Comparison w/ Other Systems |
|------------------------|------------|-------------------------------------------|-------------------------------------------|-----------------|---------------|----------------------------------------|-----------------------------|
| Name                   | Venue      | Processor                                 | OS                                        | Open Source     | Documentation | Reliance on Hardware-Specific Features |                             |
| Nooks [160]            | SOSP '03   | Intel Pentium IV                          | Linux 2.4.18                              | ✓               | ✓             |                                        |                             |
| Mondrix [173]          | SOSP '05   | Simulator (SimICS/Bochs x86)              | Linux 2.4.19                              |                 |               | ✓                                      |                             |
| SafeDrive [177]        | OSDI '06   | Dual Xeon                                 | Linux 2.6.15.5                            |                 |               |                                        |                             |
| XFI [70]               | OSDI '06   | Intel Pentium M                           | Windows (version unknown)                 |                 |               |                                        | ✓                           |
| Microdrivers [73]      | ASPLOS '08 | Intel Pentium D / AMD Opteron             | Linux 2.6.18.1                            |                 |               |                                        |                             |
| Decaf [145]            | ATC '09    | Intel Pentium D / Intel Core2 Quad        | Linux 2.6.18.1                            |                 |               |                                        | ✓                           |
| BGI [52]               | SOSP '09   | Intel Core2 Duo                           | Windows Vista Enterprise SP1              |                 |               |                                        |                             |
| LXFI [123]             | SOSP '11   | Intel i3-550                              | Linux 2.6.36                              |                 |               |                                        |                             |
| VirtuOS [134]          | SOSP '13   | Intel Xeon E5520                          | Linux 3.2.30                              | ✓               | ✓             |                                        |                             |
| KCoFI [59]             | S&P '14    | Intel i7-3770                             | FreeBSD 9.0                               | ✓               | ✓             |                                        |                             |
| PerspicuOS [65]        | ASPLOS '15 | Intel i7-3770                             | FreeBSD 9.0                               | ✓               | ✓             |                                        |                             |
| KENALI [154]           | NDSS '16   | Nvidia Tegra K1                           | Android 5.0-5.1.1 (exact version unclear) | ✓               | ✓             |                                        | ✓                           |
| SKEE [35]              | NDSS '16   | Qualcomm Snapdragon APQ8084 / Exynos 7420 | Linux 3.10.61 (Android 5.1.1)             |                 |               |                                        |                             |
| LXD <sub>s</sub> [131] | ATC '19    | Intel E5-4620                             | Linux 4.8.4                               | ✓               | ✓             | ✓                                      |                             |
| LVD [132]              | VEE '20    | Intel E5-2660                             | Linux 4.8.4                               | ✓               | ✓             | ✓                                      |                             |
| xMP [142]              | S&P '20    | Intel i7-7700                             | Linux 4.18.x                              | ✓               | ✓             | ✓                                      | ✓                           |
| IskiOS [80]            | RAID '21   | Intel Xeon Silver 4114 (Skylake)          | Linux 5.10.x                              | ✓               |               | ✓                                      |                             |
| HAKC [126]             | NDSS '22   | Simulator (ARMv8.5a FAST model)           | Linux 5.10.x                              | ✓               | ✓             | ✓                                      |                             |

(and the same OS version) would require significant engineering effort, arguably beyond reasonable evaluation expectations. As a consequence, the set of CVEs used to perform evaluation is inconsistent, making it difficult to generalize from a specific CVE (or a class of CVEs) the effectiveness

of an approach. Many approaches also have different threat models, which further complicates comparison (e.g., some types of vulnerabilities are out of scope for one system but in scope for others). We suggest that future work adopt standardized approaches to quantitatively measure the reduction

of the attack surface [146, 176] and the TCB [170]. While not a panacea, it provides a common frame of reference without an unreasonable amount of labor.

**Engineering Effort.** Only a few authors have attempted to *quantitatively* evaluate the effort required to isolate a kernel component e.g., by counting the number of manual annotations [65, 73, 123, 145] (see §7.3). These measurements can be comparable between two competing systems, if they share the *same* isolated kernel component. For example, LXFI [123] and Decaf [145] both measured the number of annotations on the E1000 network device driver. Qualitative evaluation is more common, however. For example, SKEE [35] only described the effort needed to modify the kernel, without any data provided to quantify that effort. While a qualitative description is useful, it provides only a *subjective* comparison across systems. We suggest that future work at a minimum quantitatively measure the amount of code that needs to be written and the number of files/subsystems modified.

It is equally important to understand the difficulty of *identifying* compartment boundaries, which is different from the complexity of the implementation code that enforces those boundaries. We observe that measuring the intellectual load required to identify appropriate compartment boundaries or to provide hints (and annotations) to an automated system remains relatively unexplored. However, this is of particular importance when deciding if a given approach is practical to adopt (§4).

## 8 CONCLUSION

We survey operating system kernel compartmentalization work spanning more than two decades. We identify two main categories of compartmentalization: *sandbox* systems that protect the kernel from untrusted kernel extensions and *safebox* systems that isolate security-sensitive mechanisms from the rest of the kernel. For each type of compartmentalization, we systematize the techniques used to identify boundaries between security domains and enforce isolation across those boundaries. We describe various hardware features that can be adopted to improve the performance of kernel compartmentalization and thus facilitate its practical adoption. Finally, we identify barriers that lead to inconsistent evaluation strategies among different solutions and suggest ways to address them.

## ACKNOWLEDGMENT

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC). Nous remercions le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG) de son soutien.

## REFERENCES

- [1] 2020 Linux Kernel History Report. (Accessed 16th April 2024). [https://project.linuxfoundation.org/hubfs/Reports/2020\\_kernel\\_history\\_report\\_082720.pdf?hsLang=en](https://project.linuxfoundation.org/hubfs/Reports/2020_kernel_history_report_082720.pdf?hsLang=en)
- [2] ab - Apache HTTP server benchmarking tool. online. (Accessed 16th April 2024). <http://httpd.apache.org/docs/2.4/programs/ab.html>.
- [3] AMD Secure Encrypted Virtualization. (Accessed 16th April 2024). <https://developer.amd.com/sev/>
- [4] AnTuTu Benchmark. online. (Accessed 16th April 2024). <https://www.antutu.com>.
- [5] ARMv8.5-A Pointer Authentication. (Accessed 16th April 2024). [shorturl.at/GHM69](https://shorturl.at/GHM69)
- [6] Arm® Architecture Reference Manual Supplement Morello for A-profile Architecture. online (accessed 16th April 2024). (Accessed 16th April 2024). <https://developer.arm.com/documentation/ddi0606/aj/?lang=en>.
- [7] CHERI Linux. online. (Accessed 16th April 2024). <https://github.com/cheri-linux>.
- [8] CVE-2010-3904. online. (Accessed 16th April 2024). <https://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2010-3904>.
- [9] CVE-2013-6282. (Accessed 16th April 2024). <https://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2013-6282>
- [10] CVE-2014-9585. (Accessed 16th April 2024). <https://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2014-9585>
- [11] CVE-2015-1593. (Accessed 16th April 2024). <https://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2015-1593>
- [12] CVE-2016-10044. (Accessed 16th April 2024). <https://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2016-10044>
- [13] CVE-2018-15471. online. (Accessed 16th April 2024). <https://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2018-15471>.
- [14] CVE-2018-7273. online. (Accessed 16th April 2024). <https://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2018-7273>.
- [15] CVE-2018-7755. online. (Accessed 16th April 2024). <https://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2018-7755>.
- [16] CVE-2019-10639. online. (Accessed 16th April 2024). <https://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2019-10639>.
- [17] CVE-2019-11190. (Accessed 16th April 2024). <https://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2019-11190>
- [18] CVE-2020-12654. online. (Accessed 16th April 2024). <https://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2020-12654>.
- [19] CVE-2020-8835. (Accessed 16th April 2024). <https://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2020-8835>
- [20] CVE-2021-31440. (Accessed 16th April 2024). <https://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2021-31440>
- [21] CVE-2021-33200. (Accessed 16th April 2024). <https://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2021-33200>
- [22] CVE-2021-3490. (Accessed 16th April 2024). <https://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2021-3490>
- [23] Geekbench. online. (Accessed 16th April 2024). <https://www.geekbench.com/>.
- [24] Intel Multi-Key Total Memory Encryption (MK TME). (Accessed 16th April 2024). <https://software.intel.com/sites/default/files/managed/a5/16/Total-Memory-Encryption-Multi-Key-Spec.pdf>
- [25] iPerf - The ultimate speed test tool for TCP, UDP and SCTP. online. (Accessed 16th April 2024). <https://iperf.fr/>.
- [26] Kernel Self Protection Project. (Accessed 16th April 2024). [https://kernelsec.org/wiki/index.php/Kernel\\_Self\\_Protection\\_Project](https://kernelsec.org/wiki/index.php/Kernel_Self_Protection_Project).
- [27] Linux KVM. (Accessed 16th April 2024). <https://www.linux-kvm.org/>
- [28] Next steps for Rust in the kernel. (Accessed 16th April 2024). <https://lwn.net/Articles/908347/>.

- [29] PKS/PMEM: Add Stray Write Protection [LWN.Net]. (Accessed 16th April 2024). <https://lwn.net/Articles/887532/>
- [30] Redox. online. (Accessed 16th April 2024). <https://www.redux-os.org/>.
- [31] A Technical Look at Intel's Control-flow Enforcement Technology (CET). (Accessed 16th April 2024). [shorturl.at/EKMN6](https://shorturl.at/EKMN6)
- [32] Xen Driver Domain. (Accessed 16th April 2024). [https://wiki.xenproject.org/wiki/Driver\\_Domain](https://wiki.xenproject.org/wiki/Driver_Domain)
- [33] Starr Andersen and Vincent Abella. 2004. Data Execution Prevention. Changes to Functionality in Microsoft Windows XP Service Pack 2, Part 3: Memory Protection Technologies. (2004).
- [34] ARM. ARMv8.5-A Memory Tagging Extension. (Accessed 16th April 2024). [https://developer.arm.com/-/media/Arm%20Developer%20Community/PDF/Arm\\_Memory\\_Tagging\\_Extension\\_Whitepaper.pdf](https://developer.arm.com/-/media/Arm%20Developer%20Community/PDF/Arm_Memory_Tagging_Extension_Whitepaper.pdf)
- [35] Ahmed Azab, Kirk Swidowski, Rohan Bhutkar, Jia Ma, Wenbo Shen, Ruowen Wang, and Peng Ning. 2016. SKEE: A Lightweight Secure Kernel-level Execution Environment for ARM. In *Network and Distributed System Security Symposium (NDSS'16)*. Internet Society.
- [36] Ahmed M Azab, Peng Ning, Jitesh Shah, Quan Chen, Rohan Bhutkar, Guruprasad Ganesh, Jia Ma, and Wenbo Shen. 2014. Hypervision across worlds: Real-time kernel protection from the ARM TrustZone secure world. In *ACM Conference on Computer and Communications Security (CCS'14)*. 90–102.
- [37] Ahmed M Azab, Peng Ning, and Xiaolan Zhang. 2011. Sice: a hardware-level strongly isolated computing environment for x86 multi-core platforms. In *Proceedings of the 18th ACM Conference on Computer and Communications Security*. 375–388.
- [38] Babak Amin Azad, Pierre Laperdrix, and Nick Nikiforakis. 2019. Less is more: quantifying the security benefits of debloating web applications. In *Security Symposium (USENIX Sec'19)*. USENIX, 1697–1714.
- [39] Antonio Barresi, Kaveh Razavi, Mathias Payer, and Thomas R. Gross. 2015. CAIN: Silently Breaking ASLR in the Cloud. In *Workshop on Offensive Technologies (WOOT'15)*. USENIX.
- [40] Andrew Baumann, Paul Barham, Pierre-Evariste Dagand, Tim Harris, Rebecca Isaacs, Simon Peter, Timothy Roscoe, Adrian Schüpbach, and Akhilesh Singhanian. 2009. The multikernel: a new OS architecture for scalable multicore systems. In *Symposium on Operating Systems Principles (SOSP'09)*. ACM, 29–44.
- [41] Maxime Bélaïr, Sylvie Lanepce, and Jean-Marc Menaud. 2019. Leveraging kernel security mechanisms to improve container security: a survey. In *International Conference on Availability, Reliability and Security (ARES'19)*. ACM.
- [42] Adam Belay, Andrea Bittau, Ali Mashtizadeh, David Terei, David Mazières, and Christos Kozyrakis. 2012. Dune: Safe user-level access to privileged CPU features. In *Symposium on Operating Systems Design and Implementation (OSDI 12)*. USENIX, 335–348.
- [43] Brian N Bershad, Thomas E Anderson, Edward D Lazowska, and Henry M Levy. 1991. User-level interprocess communication for shared memory multiprocessors. *ACM Transactions on Computer Systems (TOCS)* 9, 2 (1991).
- [44] J. K. Biba. 1977. Integrity Considerations for Secure Computer Systems. (1977).
- [45] Bruno Bierbaumer, Julian Kirsch, Thomas Kittel, Aurélien Francillon, and Apostolis Zarras. 2018. Smashing the Stack Protector for Fun and Profit. In *International Conference on ICT Systems Security and Privacy Protection*. Springer/IFIP.
- [46] Andrea Bittau, Petr Marchenko, Mark Handley, and Brad Karp. 2008. Wedge: Splitting applications into reduced-privilege compartments. In *Symposium on Networked Systems Design and Implementation (NSDI'08)*. USENIX.
- [47] Daniel P Bovet and Marco Cesati. 2005. *Understanding the Linux Kernel: from I/O ports to process management*. O'Reilly Media, Inc.
- [48] David Brumley and Dawn Song. 2004. Privtrans: Automatically partitioning programs for privilege separation. In *Security Symposium*. USENIX.
- [49] Nathan Burow, Xiping Zhang, and Mathias Payer. 2019. SoK: Shining light on shadow stacks. In *Symposium on Security and Privacy (S&P'19)*. IEEE, 985–999.
- [50] Claudio Canella, Jo Van Bulck, Michael Schwarz, Moritz Lipp, Benjamin Von Berg, Philipp Ortner, Frank Piessens, Dmitry Evtyushkin, and Daniel Gruss. 2019. A systematic evaluation of transient execution attacks and defenses. In *Security Symposium (USENIX Sec'19)*. USENIX, 249–266.
- [51] Cyril Cassagnes, Lucian Trestioreanu, Clement Joly, and Radu State. 2020. The rise of eBPF for non-intrusive performance monitoring. In *Network Operations and Management Symposium (NOMS'20)*. IEEE/I-FIP.
- [52] Miguel Castro, Manuel Costa, Jean-Philippe Martin, Marcus Peinado, Periklis Akritidis, Austin Donnelly, Paul Barham, and Richard Black. 2009. Fast byte-granularity software fault isolation. In *Symposium on Operating Systems Principles (SOSP'09)*. ACM.
- [53] Haogang Chen, Yandong Mao, Xi Wang, Dong Zhou, Nickolai Zeldovich, and M. Frans Kaashoek. 2011. Linux Kernel Vulnerabilities: State-of-the-Art Defenses and Open Problems. In *Asia-Pacific Workshop on Systems*. ACM.
- [54] Yaohui Chen, Sebassujeen Reymondjohnson, Zhichuang Sun, and Long Lu. 2016. Shreds: Fine-grained execution units with private memory. In *Symposium on Security and Privacy (S&P'16)*. IEEE, 56–71.
- [55] Tzi-cker Chiueh and Fu-Hau Hsu. 2001. RAD: A compile-time solution to buffer overflow attacks. In *International Conference on Distributed Computing Systems (ICDCS'01)*. IEEE.
- [56] Andy Chou, Junfeng Yang, Benjamin Chelf, Seth Hallem, and Dawson Engler. 2001. An empirical study of operating systems errors. In *Proceedings of the eighteenth ACM Symposium on Operating Systems Principles*. 73–88.
- [57] Jonathan Corbet. 2022. Generalized address-space isolation. online. (2022). <https://lwn.net/Articles/886494/>.
- [58] Victor Costan and Srinivas Devadas. 2016. Intel SGX explained. *Cryptology ePrint Archive* (2016).
- [59] John Criswell, Nathan Dautenhahn, and Vikram Adve. 2014. KCoFI: Complete control-flow integrity for commodity operating system kernels. In *Symposium on Security and Privacy (S&P'14)*. IEEE, 292–307.
- [60] John Criswell, Nicolas Geoffray, and Vikram Adve. 2009. Memory Safety for Low-Level Software/Hardware Interactions. In *Security Symposium (USENIX Sec'09)*. USENIX.
- [61] John Criswell, Andrew Lenharth, Dinakar Dhurjati, and Vikram Adve. 2007. Secure Virtual Architecture: A Safe Execution Environment for Commodity Operating Systems. In *Symposium on Operating Systems Principles (SOSP'07)*. ACM, Association for Computing Machinery.
- [62] John Criswell, Brent Monroe, and Vikram Adve. 2006. A virtual instruction set interface for operating system kernels. In *Workshop on the Interaction between Operating Systems and Computer Architecture*.
- [63] Cody Cutler, M Frans Kaashoek, and Robert T Morris. 2018. The benefits and costs of writing a POSIX kernel in a high-level language. In *Symposium on Operating Systems Design and Implementation (OSDI'18)*. USENIX.
- [64] CVE Details. Vulnerabilities in the Linux kernel 2019. (Accessed 16th April 2024). Accessed January 2020.

- [65] Nathan Dautenhahn, Theodoros Kasampalis, Will Dietz, John Criswell, and Vikram Adve. 2015. Nested kernel: An operating system architecture for intra-kernel privilege separation. In *Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS'15)*. ACM.
- [66] Lucas Davi, Ahmad-Reza Sadeghi, and Marcel Winandy. 2011. ROPdefender: A detection tool to defend against return-oriented programming attacks. In *Symposium on Information, Computer and Communications Security (ASIACCS'11)*. ACM.
- [67] John R Douceur, Jeremy Elson, Jon Howell, and Jacob R Lorch. 2008. Leveraging legacy code to deploy desktop applications on the web.. In *OSDI*, Vol. 8. 339–354.
- [68] Kevin Elphinstone and Gernot Heiser. 2013. From L3 to SeL4 What Have We Learnt in 20 Years of L4 Microkernels?. In *Symposium on Operating Systems Principles (SOSP'13)*. ACM, 133–150.
- [69] Dawson R Engler, M Frans Kaashoek, and James O'Toole Jr. 1995. Exokernel: An operating system architecture for application-level resource management. *ACM SIGOPS Operating Systems Review* 29, 5 (1995), 251–266.
- [70] Ulfar Erlingsson, Martín Abadi, Michael Vrbale, Mihai Budiu, and George C Necula. 2006. XFI: Software guards for system address spaces. In *Symposium on Operating Systems Design and Implementation (OSDI'06)*. USENIX.
- [71] Dmitry Evtvyushkin, Dmitry Ponomarev, and Nael Abu-Ghazaleh. 2016. Jump over ASLR: Attacking branch predictors to bypass ASLR. In *International Symposium on Microarchitecture (MICRO)*. IEEE/ACM.
- [72] Austin Gadiant, Baltazar Ortiz, Ricardo Barrato, Eli Davis, Jeff Perkins, and Martin Rinard. 2019. *Automatic Exploitation of Fully Randomized Executables*. Technical Report. MIT CSAIL.
- [73] Vinod Ganapathy, Matthew Renzelmann, Arini Balakrishnan, Michael Swift, and Somesh Jha. 2008. The design and implementation of Microdrivers. *ACM SIGARCH Computer Architecture News* (2008).
- [74] Tal Garfinkel, Ben Pfaff, Jim Chow, Mendel Rosenblum, and Dan Boneh. 2003. Terra: A virtual machine-based platform for trusted computing. In *Proceedings of the nineteenth ACM Symposium on Operating Systems Principles*. 193–206.
- [75] Alain Gefflaut, Trent Jaeger, Yoonho Park, Jochen Liedtke, Kevin J Elphinstone, Volkmar Uhlig, Jonathon E Tidswell, Luke Deller, and Lars Reuther. 2000. The SawMill multiserver approach. In *European workshop: beyond the PC: new challenges for the operating system*. ACM.
- [76] David Gens, Simon Schmitt, Lucas Davi, and Ahmad-Reza Sadeghi. 2018. K-Miner: Uncovering Memory Corruption in Linux.. In *Network and Distributed System Security Symposium (NDSS'18)*. Internet Society.
- [77] Masoud Ghaffarinia and Kevin W Hamlen. 2019. Binary control-flow trimming. In *Conference on Computer and Communications Security (CCS'19)*. ACM, 1009–1022.
- [78] Cristiano Giuffrida, Anton Kuijsten, and Andrew S Tanenbaum. 2012. Enhanced operating system security through efficient and fine-grained address space randomization. In *Security Symposium (USENIX Sec'12)*. USENIX.
- [79] Ben Gras, Kaveh Razavi, Erik Bosman, Herbert Bos, and Cristiano Giuffrida. 2017. ASLR on the Line: Practical Cache Attacks on the MMU. In *Network and Distributed System Security Symposium (NDSS'17)*, Vol. 17. Internet Society.
- [80] Spyridoula Gravani, Mohammad Hedayati, John Criswell, and Michael L. Scott. 2019. IskiOS: Lightweight Defense Against Kernel-Level Code-Reuse Attacks. (2019). <http://arxiv.org/abs/1903.04654>
- [81] Khilan Gudka, Robert NM Watson, Jonathan Anderson, David Chisnall, Brooks Davis, Ben Laurie, Ilias Marinos, Peter G Neumann, and Alex Richardson. 2015. Clean application compartmentalization with SOAAP. In *Conference on Computer and Communications Security (CCS'15)*. ACM.
- [82] Jorrit N Herder, Herbert Bos, Ben Gras, Philip Homburg, and Andrew S Tanenbaum. 2006. MINIX 3: A highly reliable, self-repairing operating system. *ACM SIGOPS Operating Systems Review* 40, 3 (2006).
- [83] Michael Hind. 2001. Pointer analysis: Haven't we solved this problem yet?. In *Workshop on Program Analysis for Software Tools and Engineering*. ACM.
- [84] Owen S Hofmann, Sangman Kim, Alan M Dunn, Michael Z Lee, and Emmett Witchel. 2013. Inktag: Secure applications on an untrusted operating system. In *Proceedings of the eighteenth International Conference on Architectural Support for Programming Languages and Operating Systems*. 265–278.
- [85] ARM Holdings. 2009. ARM Security Technology: Building a Secure System using TrustZone Technology. <http://infocenter.arm.com/help/index.jsp?topic=/com.arm.doc.prd29-genc-009492c/>. (2009).
- [86] Hong Hu, Shweta Shinde, Sendriou Adrian, Zheng Leong Chua, Praatek Saxena, and Zhenkai Liang. 2016. Data-oriented programming: On the expressiveness of non-control data attacks. In *2016 IEEE Symposium on Security and Privacy (SP)*. IEEE, 969–986.
- [87] Yongzhe Huang, Vikram Narayanan, David Detweiler, Kaiming Huang, Gang Tan, Trent Jaeger, and Anton Burtsev. 2022. KSplit: Automating Device Driver Isolation. In *Symposium on Operating Systems Design and Implementation (OSDI'22)*. USENIX, 613–631.
- [88] Ralf Hund, Carsten Willems, and Thorsten Holz. 2013. Practical timing side channel attacks against kernel space ASLR. In *Symposium on Security and Privacy (S&P'13)*. IEEE.
- [89] Intel. Intel® 64 and IA-32 Architectures Software Developer Manuals. (Accessed 16th April 2024). <https://software.intel.com/content/www/us/en/develop/articles/intel-sdm.html>
- [90] Suman Jana, Donald E Porter, and Vitaly Shmatikov. 2011. TxBBox: Building secure, efficient sandboxes with system transactions. In *2011 IEEE Symposium on Security and Privacy*. IEEE, 329–344.
- [91] Dae R Jeong, Kyungtae Kim, Basavesh Shivakumar, Byoungyoung Lee, and Insik Shin. 2019. Razzer: Finding kernel race bugs through fuzzing. In *Symposium on Security and Privacy (S&P'19)*. IEEE, 754–768.
- [92] Jinghao Jia, Raj Sahu, Adam Oswald, Dan Williams, Michael V Le, and Tianyin Xu. 2023. Kernel Extension Verification is Untenable. In *Workshop on Hot Topics in Operating Systems (HotOS'23)*. ACM, 150–157.
- [93] Rick Jones. Netperf. online. (Accessed 16th April 2024). <https://github.com/HewlettPackard/netperf>.
- [94] Kostis Kaffes, Jack Tigar Humphries, David Mazières, and Christos Kozyrakis. 2021. Syrup: User-defined scheduling across the stack. In *Symposium on Operating Systems Principles (SOSP'21)*. ACM.
- [95] Vishal Karande, Erick Bauman, Zhiqiang Lin, and Latifur Khan. 2017. SGX-Log: Securing System Logs With SGX. In *ASIA Conference on Computer and Communications Security (ASIACCS'17)*. ACM.
- [96] Eric Keller, Jakub Szefer, Jennifer Rexford, and Ruby B Lee. 2010. No-hype: virtualized cloud infrastructure without the virtualization. In *Proceedings of the 37th Annual International Symposium on Computer Architecture*. 350–361.
- [97] Arslan Khan, Hyungsub Kim, Byoungyoung Lee, Dongyan Xu, Antonio Bianchi, and Dave Jing Tian. 2021. M2MON: Building an MMIO-based Security Reference Monitor for Unmanned Vehicles. In *Security Symposium (USENIX Sec'21)*. USENIX, 285–302.
- [98] Douglas Kilpatrick. 2003. Privman: A Library for Partitioning Applications.. In *USENIX Annual Technical Conference, FREENIX Track*. 273–284.

- [99] Kyungtae Kim, Dae R Jeong, Chung Hwan Kim, Yeongjin Jang, Insik Shin, and Byoungyoung Lee. 2020. HFL: Hybrid Fuzzing on the Linux Kernel. In *Network and Distributed System Security Symposium (NDSS'20)*. Internet Society.
- [100] Taesoo Kim and Nickolai Zeldovich. 2013. Practical and effective sandboxing for non-root users. In *USENIX Annual Technical Conference (USENIX ATC'13)*.
- [101] Steve Klabnik and Carol Nichols. 2023. *The Rust programming language*. No Starch Press.
- [102] Gerwin Klein. 2009. Operating system verification—an overview. *Sadhana* 34, 1 (2009).
- [103] Gerwin Klein, Kevin Elphinstone, Gernot Heiser, June Andronick, David Cock, Philip Derrin, Dhammika Elkaduwe, Kai Engelhardt, Rafal Kolanski, Michael Norrish, et al. 2009. seL4: Formal verification of an OS kernel. In *Symposium on Operating Systems Principles (SOSP'09)*. ACM.
- [104] Koen Koning, Xi Chen, Herbert Bos, Cristiano Giuffrida, and Elias Athanasopoulos. 2017. No need to hide: Protecting safe regions on commodity hardware. In *European Conference on Computer Systems (EuroSys'17)*. ACM, 437–452.
- [105] Simon Kuenzer, Vlad-Andrei Bădoi, Hugo Lefeuvre, Sharan Santhanam, Alexander Jung, Gauthier Gain, Cyril Soldani, Costin Lupu, Ștefan Teodorescu, Costi Răducanu, Cristian Banu, Laurent Mathy, Răzvan Deaconescu, Costin Raiciu, and Felipe Huici. 2021. Unikraft: Fast, Specialized Unikernels the Easy Way. In *European Conference on Computer Systems (EuroSys'21)*. ACM, 376–394.
- [106] Hsuan-Chi Kuo, Dan Williams, Ricardo Koller, and Sibin Mohan. 2020. A Linux in unikernel clothing. In *Proceedings of the Fifteenth European Conference on Computer Systems*.
- [107] Anil Kurmus, Alessandro Sorniotti, and Rüdiger Kapitza. 2011. Attack surface reduction for commodity OS kernels: trimmed garden plants may attract less bugs. In *European Workshop on System Security*. 1–6.
- [108] Volodymyr Kuznetsov, László Szekeres, Mathias Payer, George Candea, R. Sekar, and Dawn Song. 2018. *Code-Pointer Integrity*. Association for Computing Machinery and Morgan & Claypool, 81–116. <https://doi.org/10.1145/3129743.3129748>
- [109] Bingchen Lan, Yan Li, Hao Sun, Chao Su, Yao Liu, and Qingkai Zeng. 2015. Loop-oriented programming: a new code reuse attack to bypass modern defenses. In *2015 IEEE Trustcom/BigDataSE/ISPA*, Vol. 1. IEEE.
- [110] Chris Lattner, Andrew Lenharth, and Vikram Adve. 2007. Making context-sensitive points-to analysis with heap cloning practical for the real world. *ACM SIGPLAN Notices* 42, 6 (2007), 278–289.
- [111] Hugo Lefeuvre, Vlad-Andrei Bădoi, Ștefan Teodorescu, Pierre Olivier, Tiberiu Mosnoi, Răzvan Deaconescu, Felipe Huici, and Costin Raiciu. 2021. FlexOS: making OS isolation flexible. In *Workshop on Hot Topics in Operating Systems (HotOS)*.
- [112] Matthew Lentz, Rijurekha Sen, Peter Druschel, and Bobby Bhattacharjee. 2018. Secloak: Arm trustzone-based mobile peripheral control. In *Annual International Conference on Mobile Systems, Applications, and Services (MobiSys'18)*. ACM, 1–13.
- [113] Jialin Li, Samantha Miller, Danyang Zhuo, Ang Chen, Jon Howell, and Thomas Anderson. 2021. An incremental path towards a safer OS kernel. In *Workshop on Hot Topics in Operating Systems (HotOS'21)*. ACM.
- [114] Yanlin Li, Jonathan McCune, James Newsome, Adrian Perrig, Brandon Baker, and Will Drewry. 2014. MiniBox: A Two-Way Sandbox for x86 Native Code. In *2014 USENIX annual technical conference (USENIX ATC 14)*. 409–420.
- [115] Hans Liljestrand, Thomas Nyman, Kui Wang, Carlos Chinea Perez, Jan-Erik Ekberg, and N Asokan. 2019. PAC it up: Towards pointer integrity using ARM pointer authentication. In *28th USENIX Security Symposium (USENIX Security 19)*. 177–194.
- [116] Soo Yee Lim, Xueyuan Han, and Thomas Pasquier. 2023. Unleashing Unprivileged eBPF Potential with Dynamic Sandboxing. In *SIGCOMM Workshop on eBPF and Kernel Extensions*. ACM.
- [117] Joshua Lind, Christian Priebe, Divya Muthukumaran, Dan O’Keeffe, Pierre-Louis Aublin, Florian Kelbert, Tobias Reiher, David Goltzsche, David Eyers, Rüdiger Kapitza, et al. 2017. Glamdring: Automatic application partitioning for Intel SGX. In *Annual Technical Conference (ATC'17)*. USENIX, 285–298.
- [118] Shen Liu, Gang Tan, and Trent Jaeger. 2017. PtrSplit: Supporting general pointers in automatic program partitioning. In *Conference on Computer and Communications Security (CCS'17)*. ACM, 2359–2371.
- [119] Shen Liu, Dongrui Zeng, Yongzhe Huang, Frank Capobianco, Stephen McCamant, Trent Jaeger, and Gang Tan. 2019. Program-Mandering: Quantitative Privilege Separation. In *Conference on Computer and Communications Security (CCS'19)*. ACM.
- [120] Yutao Liu, Tianyu Zhou, Kexin Chen, Haibo Chen, and Yubin Xia. 2015. Thwarting memory disclosure with efficient hypervisor-enforced intra-domain isolation. In *Conference on Computer and Communications Security (CCS'15)*. ACM.
- [121] Aravind Machiry, Chad Spensky, Jake Corina, Nick Stephens, Christopher Kruegel, and Giovanni Vigna. 2017. DR. CHECKER: A soundy analysis for Linux kernel drivers. In *Security Symposium (USENIX Sec'17)*. USENIX.
- [122] Anil Madhavapeddy and David J Scott. 2014. Unikernels: the rise of the virtual library operating system. *Commun. ACM* 57, 1 (2014).
- [123] Yandong Mao, Haogang Chen, Dong Zhou, Xi Wang, Nickolai Zeldovich, and M. Frans Kaashoek. 2011. Software fault isolation with API integrity and multi-principal modules. In *Symposium on Operating Systems Principles (SOSP'11)*. ACM.
- [124] Jonathan M McCune, Yanlin Li, Ning Qu, Zongwei Zhou, Anupam Datta, Virgil Gligor, and Adrian Perrig. 2010. TrustVisor: Efficient TCB reduction and attestation. In *2010 IEEE Symposium on Security and Privacy*. IEEE, 143–158.
- [125] Jonathan M McCune, Bryan J Parno, Adrian Perrig, Michael K Reiter, and Hiroshi Isozaki. 2008. Flicker: An execution infrastructure for TCB minimization. In *Proceedings of the 3rd ACM SIGOPS/EuroSys European Conference on Computer Systems 2008*. 315–328.
- [126] Derrick McKee, Yianni Giannaris, Carolina Ortega Perez, Howard Shrobe, Mathias Payer, Hamed Okhravi, and Nathan Burow. 2022. Preventing Kernel Hacks with HAKC. In *Network and Distributed System Security Symposium (NDSS'22)*. Internet Society.
- [127] Larry W McVoy, Carl Staelin, et al. 1996. Imbench: Portable Tools for Performance Analysis. In *Annual Technical Conference (ATC'96)*. USENIX, 279–294.
- [128] Mitre. Linux Kernel Vulnerability Trends Over Time. online (accessed 16th April 2024). (Accessed 16th April 2024). [https://www.cvedetails.com/product/47/Linux-Linux-Kernel.html?vendor\\_id=33](https://www.cvedetails.com/product/47/Linux-Linux-Kernel.html?vendor_id=33).
- [129] David Mosberger and Tai Jin. 1998. httpperf—a tool for measuring web server performance. *ACM SIGMETRICS Performance Evaluation Review* 26, 3 (1998), 31–37.
- [130] Shravan Narayan, Craig Disselkoen, Daniel Moghimi, Sunjay Cauligi, Evan Johnson, Zhao Gang, Anjo Vahldiek-Oberwagner, Ravi Sahita, Hovav Shacham, Dean Tullsen, et al. 2021. Swivel: Hardening WebAssembly against Spectre. In *Security Symposium (USENIX Sec'21)*. USENIX.
- [131] Vikram Narayanan, Abhiram Balasubramanian, Charlie Jacobsen, Sarah Spall, Scott Bauer, Michael Quigley, Aftab Hussain, Abdullah Younis, Junjie Shen, and Moinak Bhattacharyya. 2019. LXDs: Towards isolation of kernel subsystems. In *Annual Technical Conference (ATC'19)*. USENIX.
- [132] Vikram Narayanan, Yongzhe Huang, Gang Tan, Trent Jaeger, and Anton Burtsev. 2020. Lightweight kernel isolation with virtualization



- and VM functions. In *International Conference on Virtual Execution Environments*. ACM.
- [133] Luke Nelson, Helgi Sigurbjarnarson, Kaiyuan Zhang, Dylan Johnson, James Bornholt, Emina Torlak, and Xi Wang. 2017. Hyperkernel: Push-button verification of an OS kernel. In *Symposium on Operating Systems Principles (SOSP'17)*. ACM, 252–269.
- [134] Ruslan Nikolaev and Godmar Back. 2013. VirtuOS: an operating system with kernel virtualization. In *Symposium on Operating Systems Principles (SOSP'13)*. ACM.
- [135] Oleksii Oleksenko, Dmitrii Kuvaiskii, Pramod Bhatotia, Pascal Felber, and Christof Fetzer. 2018. Intel MPX Explained: A Cross-Layer Analysis of the Intel MPX System Stack. *ACM Measurement and Analysis of Computing Systems*. (2018).
- [136] OpenBenchmarking. LAME MP3 Encoding. online. (Accessed 16th April 2024). <https://openbenchmarking.org/test/pts/encode-mp3>.
- [137] Riccardo Paccagnella, Pubali Datta, Wajih Ul Hassan, Adam Bates, Christopher W Fletcher, Andrew Miller, and Dave Tian. 2020. Custos: Practical tamper-evident auditing of operating systems using trusted execution. In *Symposium on Network and Distributed System Security (NDSS'02)*. Internet Society.
- [138] Nicolas Palix, Gaël Thomas, Suman Saha, Christophe Calvès, Julia Lawall, and Gilles Muller. 2011. Faults in Linux: Ten years later. In *International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS'11)*. ACM.
- [139] Vasilis Pappas, Michalis Polychronakis, and Angelos D Keromytis. 2013. Transparent {ROP} exploit mitigation using indirect branch tracing. In *Security Symposium (USENIX Sec'13)*. USENIX, 447–462.
- [140] Sandro Pinto and Nuno Santos. 2019. Demystifying ARM TrustZone: A comprehensive survey. *ACM Computing Surveys (CSUR)* 51, 6 (2019), 1–36.
- [141] Ian Pratt, Keir Fraser, Steven Hand, Christian Limpach, Andrew Warfield, Dan Magenheimer, Jun Nakajima, and Asit Mallick. 2005. Xen 3.0 and the art of virtualization. In *Linux symposium*, Vol. 2.
- [142] Sergej Proskurin, Marius Momeu, Seyedhamed Ghavamnia, Vasileios P Kemerlis, and Michalis Polychronakis. 2020. xMP: selective memory protection for kernel and user space. In *Symposium on Security and Privacy (S&P'20)*. IEEE, 563–577.
- [143] Qualcomm. Vellamo Mobile Benchmark Suite. online. (Accessed 16th April 2024). <https://www.qualcomm.com/news/onq/2012/09/20/vellamo-mobile-benchmark-suite>.
- [144] Xiang (Jenny) Ren, Kirk Rodrigues, Luyuan Chen, Camilo Vega, Michael Stumm, and Ding Yuan. 2019. An Analysis of Performance Evolution of Linux's Core Operations. In *Symposium on Operating Systems Principles (SOSP'19)*. ACM.
- [145] Matthew J Renzelmann and Michael M Swift. 2009. Decaf: Moving Device Drivers to a Modern Language. In *Annual Technical Conference (ATC'09)*. USENIX.
- [146] Jonathan Salwan. ROPgadget Tool. online. (Accessed 16th April 2024). <https://github.com/JonathanSalwan/ROPgadget>.
- [147] Vasily A. Sartakov, Lluís Vilanova, David Eysers, Takahiro Shinagawa, and Peter Pietzuch. 2022. CAP-VMs: Capability-Based Isolation and Sharing in the Cloud. In *Symposium on Operating Systems Design & Implementation (OSDI 22)*. USENIX, 597–612.
- [148] Vasily A. Sartakov, Lluís Vilanova, and Peter Pietzuch. 2021. CubicleOS: A Library OS with Software Componentisation for Practical Isolation. In *International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS'21)*. ACM.
- [149] Sergej Schumilo, Cornelius Aschermann, Robert Gawlik, Sebastian Schinzel, and Thorsten Holz. 2017. kAFL: Hardware-assisted feedback fuzzing for OS kernels. In *Security Symposium (USENIX Sec'17)*. USENIX, 167–182.
- [150] Vedvyas Shanbhogue, Deepak Gupta, and Ravi Sahita. 2019. Security analysis of processor instruction set architecture for enforcing control-flow integrity. In *International Workshop on Hardware and Architectural Support for Security and Privacy (HASP'19)*. ACM, 1–11.
- [151] Rui Shu, Peipei Wang, Sigmund A Gorski III, Benjamin Andow, Adwait Nadkarni, Luke Deshotels, Jason Gionta, William Enck, and Xiaohui Gu. 2016. A study of security isolation techniques. *ACM Computing Surveys (CSUR)* 49, 3 (2016), 1–37.
- [152] Christopher Small and Margo Seltzer. 1998. MiSFIT: Constructing safe extensible systems. *IEEE concurrency* 6, 3 (1998), 34–41.
- [153] Kevin Z Snow, Fabian Monrose, Lucas Davi, Alexandra Dmitrienko, Christopher Liebchen, and Ahmad-Reza Sadeghi. 2013. Just-in-time code reuse: On the effectiveness of fine-grained address space layout randomization. In *Symposium on Security and Privacy (S&P'13)*. IEEE.
- [154] Chengyu Song, Byoungyoung Lee, Kangjie Lu, William Harris, Taesoo Kim, and Wenke Lee. 2016. Enforcing Kernel Security Invariants with Data Flow Integrity. In *Network and Distributed System Security Symposium (NDSS'16)*. Internet Society.
- [155] Dokyung Song, Felicitas Hetzelt, Jonghwan Kim, Brent Byunghoon Kang, Jean-Pierre Seifert, and Michael Franz. 2020. Agamotto: Accelerating kernel driver fuzzing with lightweight virtual machine checkpoints. In *Security Symposium (USENIX Sec'20)*. USENIX, 2541–2557.
- [156] Udo Steinberg and Bernhard Kauer. 2010. NOVA: A microhypervisor-based secure virtualization architecture. In *Proceedings of the 5th European conference on Computer systems*. 209–222.
- [157] Nenad Stojanovski, Marjan Gusev, Danilo Gligoroski, and Svein J Knapskog. 2007. Bypassing data execution prevention on Microsoft Windows XP SP2. In *International Conference on Availability, Reliability and Security (ARES'07)*. IEEE.
- [158] Yulei Sui and Jingling Xue. 2016. SVF: interprocedural static value-flow analysis in LLVM. In *International Conference on Compiler Construction (CC'16)*. ACM, 265–266.
- [159] Michael M. Swift, Muthukaruppan Annamalai, Brian N. Bershad, and Henry M. Levy. 2004. Recovering Device Drivers. In *Proceedings of the 6th Conference on Symposium on Operating Systems Design & Implementation - Volume 6 (OSDI'04)*. USENIX Association, USA.
- [160] Michael M Swift, Brian N Bershad, and Henry M Levy. 2003. Improving the reliability of commodity operating systems. In *Symposium on Operating Systems Principles (SOSP'03)*. ACM.
- [161] Richard Ta-Min, Lionel Litty, and David Lie. 2006. Splitting interfaces: Making trust between applications and operating systems configurable. In *Proceedings of the 7th symposium on Operating systems design and implementation*. 279–292.
- [162] Teck Bok Tok, Samuel Z Guyer, and Calvin Lin. 2006. Efficient flow-sensitive interprocedural data-flow analysis in the presence of pointers. In *International Conference on Compiler Construction*. Springer.
- [163] Jeff Vander Stoep. 2016. Protecting Android with more Linux kernel defenses. (2016). <https://android-developers.googleblog.com/2016/07/protecting-android-with-more-linux.html>.
- [164] Marcos AM Vieira, Matheus S Castanho, Racyus DG Pacifico, Elerson RS Santos, Eduardo PM Câmara Júnior, and Luiz FM Vieira. 2020. Fast packet processing with eBPF and XDP: Concepts, code, challenges, and applications. *Computing Surveys (CSUR)* 53, 1 (2020).
- [165] Perry Wagle, Crispin Cowan, et al. 2003. Stackguard: Simple stack smash protection for gcc. In *GCC Developers Summit*.
- [166] Robert Wahbe, Steven Lucco, Thomas E. Anderson, and Susan L. Graham. 1993. Efficient Software-Based Fault Isolation. In *Symposium on Operating Systems Principles (SOSP'93)*. ACM.
- [167] Jinwen Wang, Ao Li, Haoran Li, Chenyang Lu, and Ning Zhang. 2022. Rt-tee: Real-time system availability for cyber-physical systems using

- arm trustzone. In *Symposium on Security and Privacy (S&P'22)*. IEEE, 352–369.
- [168] Robert N.M. Watson, Jonathan Woodruff, Peter G. Neumann, Simon W. Moore, Jonathan Anderson, David Chisnall, Nirav Dave, Brooks Davis, Khilan Gudka, Ben Laurie, Steven J. Murdoch, Robert Norton, Michael Roe, Stacey Son, and Munraj Vadera. 2015. CHERI: A Hybrid Capability-System Architecture for Scalable Software Compartmentalization. In *Symposium on Security and Privacy (S&P'15)*. IEEE.
  - [169] Robert N M Watson, Simon W Moore, Peter Sewell, and Peter G Neumann. An Introduction to CHERI. (Accessed 16th April 2024).
  - [170] David Wheeler. SLOCCount. online. (Accessed 16th April 2024). <https://dwheeler.com/sloccount/>.
  - [171] Dan Williams, Ricardo Koller, Martin Lucina, and Nikhil Prakash. 2018. Unikernels as processes. In *Symposium on Cloud Computing (SoCC'18)*. ACM.
  - [172] Emmett Witchel, Josh Cates, and Krste Asanović. 2002. Mondrian Memory Protection. In *International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS'02)*. ACM.
  - [173] Emmett Witchel, Junghwan Rhee, and Krste Asanović. 2005. Mondrix: Memory isolation for Linux using Mondriaan memory protection. In *Symposium on Operating Systems Principles (SOSP'05)*. ACM.
  - [174] Jonathan Woodruff, Robert NM Watson, David Chisnall, Simon W Moore, Jonathan Anderson, Brooks Davis, Ben Laurie, Peter G Neumann, Robert Norton, and Michael Roe. 2014. The CHERI capability model: Revisiting RISC in an age of risk. In *International Symposium on Computer Architecture (ISCA'14)*. ACM/IEEE.
  - [175] Bennet Yee, David Sehr, Gregory Dardyk, J Bradley Chen, Robert Muth, Tavis Ormandy, Shiki Okasaka, Neha Narula, and Nicholas Fullagar. 2010. Native client: A sandbox for portable, untrusted x86 native code. *Commun. ACM* 53, 1 (2010), 91–99.
  - [176] Mingwei Zhang and R. Sekar. 2013. Control Flow Integrity for COTS Binaries. In *Security Symposium (USENIX Sec'13)*. USENIX.
  - [177] Feng Zhou, Jeremy Condit, Zachary Anderson, Ilya Bagrak, Rob Ennals, Matthew Harren, George Necula, and Eric Brewer. 2006. SafeDrive: Safe and Recoverable Extensions Using Language-Based Techniques. In *Symposium on Operating Systems Design and Implementation (OSDI'06)*. USENIX.
  - [178] Jie Zhou, Yufei Du, Zhuojia Shen, Lele Ma, John Criswell, and Robert J. Walls. 2020. Silhouette: Efficient Protected Shadow Stacks for Embedded Systems. In *Security Symposium (USENIX Sec'20)*. USENIX.
  - [179] Zhiqiang Zuo, Kai Wang, Aftab Hussain, Ardalan Amiri Sani, Yiyu Zhang, Shenming Lu, Wensheng Dou, Linzhang Wang, Xuandong Li, Chenxi Wang, et al. 2021. Systemizing Interprocedural Static Analysis of Large-scale Systems Code with Graspan. *ACM Transactions on Computer Systems (TOCS)* 38, 1-2 (2021), 1–39.