BeeBox: Hardening BPF against Transient Execution Attacks
Di Jin, Alexander J. Gaidis, and Vasileios P. Kemerlis, Brown University
https://www.usenix.org/conference/usenixsecurity24/presentation/jin-di

This paper is included in the Proceedings of the 33rd USENIX Security Symposium.
August 14–16, 2024 • Philadelphia, PA, USA
978-1-939133-44-1

Open access to the Proceedings of the 33rd USENIX Security Symposium is sponsored by USENIX.
Abstract
The Berkeley Packet Filter (BPF) has emerged as the de-facto standard for carrying out safe and performant, user-specified computation(s) in kernel space. However, BPF also increases the attack surface of the OS kernel disproportionately, especially under the presence of transient execution vulnerabilities. In this work, we present BeeBox: a new security architecture that hardens BPF against transient execution attacks, allowing the OS kernel to expose eBPF functionality to unprivileged users and applications. At a high level, BeeBox sandboxes the BPF runtime against speculative code execution in an SFI-like manner. Moreover, by using a combination of static analyses and domain-specific properties, BeeBox selectively elides enforcement checks, improving performance without sacrificing security. We implemented a prototype of BeeBox for the Linux kernel that supports popular features of eBPF (e.g., BPF maps and helper functions), and evaluated it both in terms of effectiveness and performance, demonstrating resilience against prevalent transient execution attacks (i.e., Spectre-PHT and Spectre-STL) with low overhead. On average, BeeBox incurs 20% overhead in the Katran benchmark, while the current mitigations of Linux incur 112% overhead. Lastly, BeeBox exhibits less than 1% throughput degradation in end-to-end, real-world settings that include seccomp-BPF and packet filtering.

1 Introduction
The Berkeley Packet Filter (BPF) [9, 74] is increasingly gaining traction as an in-kernel, “universal” virtual machine (VM): that is, a generic mechanism for extending or customizing OS kernel functionality by “pushing” (i.e., performing) user-specified computation(s) in kernel space, in a safe and performant manner. BPF greatly improves the efficiency of certain tasks by moving application logic closer to where data is generated, processed, or received, thereby reducing kernel-to-user I/O and avoiding expensive context switches. It enables a plethora of mature and emerging applications, including security monitoring and policy enforcement [23, 69, 96], high-performance networking [36, 95], observability and tracing [46], storage [104], as well as scheduling [39]. More importantly, many BPF applications require, or can benefit from, allowing the use of the BPF infrastructure in an unprivileged manner [19]. Unfortunately, however, BPF increases the OS attack surface [50], and it has been extensively used as the underlying apparatus for mounting transient execution attacks [11] against the kernel (from userland).

Transient execution attacks abuse the speculative-code execution capabilities built into modern CPUs, effectively leaking data through (forced) mis-speculation and side channels [33, 71]. BPF facilitates such attacks [55, 56] as BPF code can be highly customizable and is executed in the same context as the OS kernel itself, hence bypassing deployed defenses [32, 41]. Modern OSes place a wide range of security-sensitive data in kernel space, such as cryptographic material [63], the memory contents of user processes (i.e., physmap/direct map) [53], as well as (secret) metadata that impact the effectiveness of system-wide security protection mechanisms, like (K)ASLR [25]. Hence, it is of paramount importance to mitigate such BPF-based information leakage, especially in the presence of unprivileged BPF.

The Linux kernel currently deploys a combination of mitigations [58] against BPF-assisted, transient execution attacks, including: (1) selectively blocking speculative code execution; (2) limiting pointer arithmetic operations; and (3) verifying the safety of all branch-instruction combinations, even when some of them are impossible. As we discuss in Section 3, these mitigations primarily extend the BPF code verification and interpretation/JITing process via a collection of ad-hoc checks. This approach of progressively retrofitting security to the BPF infrastructure has caused a range of problems, in terms of usability and performance, without providing clear protection guarantees [8, 91, 92]—and, as expected, it has also led to various bypasses [7, 29, 57]. For these reasons, Linux distributions restrict eBPF to privileged users only [19]. In this work, we propose BeeBox: a new BPF security architecture that protects against transient execution attacks through a principled, security-first design approach.
The main idea behind BeeBox is to sandbox the BPF data in an SFI-like (software fault isolation) [99] manner. Moreover, by using static analysis and various domain-specific properties, BeeBox elides enforcement checks to improve performance, without sacrificing security. BeeBox improves the security posture of BPF, thereby allowing unprivileged users to safely reap the benefits of the BPF (sub)system and utilize it to its full potential.

In summary, we make the following contributions:

• We present the design of BeeBox, a security architecture against challenging and prevalent transient execution attacks in BPF—namely, Spectre-PHT and Spectre-STL [56]—, which isolates sensitive data from BPF’s speculative execution reach. Our design also covers the security of BPF helper functions and cBPF.

• We provide a prototype implementation of BeeBox, which supports popular BPF applications, as well as a collection of tools and methods that simplify the future expansion of BeeBox’s supported features.

• We evaluate our prototype on synthetic and real-world BPF programs and workloads, demonstrating that BeeBox provides increased security, compatibility, and efficiency than Linux’s current mitigations. BeeBox achieves 0–23% overhead in kernel micro-benchmarks, and < 1% overhead in end-to-end, real-world settings.

2 Background

2.1 Kernel Security

OS kernels are usually the most privileged piece of software on a system, making them a lucrative target for attackers. Unfortunately, OS kernels are typically written in memory- and/or type-unsafe languages, which are prone to memory errors [97]. An attacker armed with a memory corruption vulnerability in the kernel can tamper with control data (e.g., code pointers [59]) to hijack the kernel’s control flow and eventually gain arbitrary code execution [54, 81]. The current, predominant technique for achieving arbitrary code execution is code reuse [10], where an attacker chains together snippets of existing (i.e., benign) code “out-of-context.”

Several defense strategies have been proposed to mitigate code-reuse attacks, with the most popular relating to information hiding or integrity checking. Schemes of the former attempt to conceal sensitive information by either randomizing [25] or diversifying [26, 60] the layout of memory, or by employing new memory protection policies, such as execute-only memory (XOM) [83, 89]. Schemes of the latter attempt to ensure that computed control-flow transfers are valid according to a given policy [5, 28, 59, 61]. Another technique for defending against code-reuse attacks in the kernel relies on isolating kernel memory from the attack payload, since real-world exploits may require complex payloads that do not fit into a limited corrupted region. Defenses in this direction prevent the kernel from using attacker-controlled payloads in user space [20, 54, 103] or in kernel memory regions that are effectively user-controlled [50, 53].

Privilege escalation can also be achieved by data-only attacks [38, 47], where, instead of control-flow data, the attacker overwrites targeted data and escalates privilege without triggering any control-flow integrity violations. Defenses against generic data-only attacks, including DFI (data-flow integrity) [16] and full memory safety [76, 77], remain impractical for performance reasons. Selective isolation schemes such as xMP [84] and ISLAB [75] were proposed to protect critical data such as page tables, process credentials, and memory management metadata.

Apart from privilege escalation, another goal of kernel attacker is information disclosure, which can be leveraged to obtain sensitive information managed by the kernel, such as cryptographic keys or secrets of other processes [53]. However, exploiting memory vulnerabilities in software is not the only way to achieve information disclosure; hardware vulnerabilities, like Meltdown [56] or cache-based side channels [33, 71], are an effective, alternate route.

2.2 Berkeley Packet Filter

Overview. BPF [74] was originally designed to accelerate network monitoring by performing user-defined packet filtering in-kernel rather than in user space. Since its inception, BPF has grown into a generic facility in Linux that acts as a universal, in-kernel VM for user-defined computation [18].

While BPF continues to support efficient networking applications, such as packet filtering [31] and load balancing [80], a variety of other applications capitalize on it to increase security and support new functionality. Several popular applications, such as Docker, Firefox, Chromium, and OpenSSH, employ seccomp-BPF [69] to filter system calls (syscalls) according to user-defined policies, with recent research exploring automatic generation [23] and secure enforcement [27] of syscall policies. Additionally, BPF underpins many tracing tools [9], since users can attach BPF programs to Kprobes (kernel probes) [93], granting increased visibility into the workings of a system. Other use cases for BPF include reducing context switching overhead in FUSE (Filesystem in Userspace) [6] and improving throughput and latency of the kernel’s storage stack [104].

Design. When a BPF program is pushed into the kernel, it is first verified for safety and then attached to a kernel hook point. Later, when a requisite event fires (e.g., a syscall is made or a packet is received), the installed program is executed. Linux currently supports two types of BPF: classic BPF (cBPF) and its more feature-rich successor, extended BPF (eBPF). The typical use cases for cBPF involve syscall [69] and packet filtering [31]. The cBPF machine abstraction consists of two 32-bit registers and 64 bytes of addressable scratch memory.
cBPF code is translated internally to eBPF, allowing the kernel to maintain a single BPF runtime. eBPF greatly expands upon cBPF’s capabilities and use-cases, offering: 10 general-purpose, 64-bit registers; a read-only frame-pointer register; and up to 512 bytes of addressable scratch memory (i.e., stack space). The instruction set of both cBPF and eBPF allows for memory loads and stores, moving values between registers, ALU operations, and branching. However, eBPF additionally provides the ability to call kernel-native functions, dubbed helpers [1], from within eBPF programs. Helpers provide various utilities and help maintain BPF maps—a variety of data structures implemented atop generic, kernel-resident key/value stores that allow eBPF programs to persist state across invocations and expose data to userland. All BPF programs also have types, which inform the kernel where the programs can be attached, what helpers and maps can be accessed, and what context (i.e., arguments and data) should be provided with upon invocation [68]. For example, an eBPF program with type xdp (BPF_PROG_TYPE_XDP) will run on a network device and the context passed to it will consist of pointers to the raw packet data, which can be used to perform packet rewriting and forwarding.

Safety and Security. BPF is designed to be safe to run within kernel context. To achieve that, when BPF programs are loaded into the kernel they pass through a static verifier, which is guaranteed to be sound but not complete—i.e., a safe program may be rejected, but an unsafe program will not pass. The BPF verifier ensures similar properties for both cBPF and eBPF, but verifies them separately, performing a much simpler analysis for cBPF programs. For eBPF, the verifier mainly verifies three properties: (1) program termination; (2) memory safety; and (3) type safety when calling helpers. For (1), the verifier simply ensures that all branching instructions only “jump forward.” For (2) and (3), the verifier uses static analysis techniques to determine the register types, whether they are pointers or scalars, etc. For a scalar, the analysis tries to figure out the range of its possible values; for a pointer, the analysis needs to determine what kind of memory it points to and whether it points within bounds. Further, the BPF verifier is also in charge of rewriting BPF program code to adjust helper-call targets and inline various helpers as BPF instructions when possible. Unfortunately, while performing static analysis to avoid runtime checks is crucial for the performance of BPF, the verifier is known to be error prone [62] and susceptible to speculative execution attacks [83].

Support. Both LLVM and GCC maintain eBPF backends, which compile programs written in C-like syntax to eBPF ELF files. Additionally, both of these BPF compilers—which are different from the in-kernel (BPF) JIT compiler—provide various utilities for working with eBPF ELFs. In a similar vein, libbpf [2] is a C library maintained by the Linux kernel which provides a low-level API for loading eBPF ELFs and managing eBPF programs, maps, and events from userland.

2.3 Transient Execution Attacks

Modern CPUs maximize their performance by employing various techniques to avoid CPU idling, such as out-of-order execution and speculative execution [35]. Instructions that are executed out-of-order, or speculatively, but never committed to the CPU’s architectural state are considered to be transiently executed. Although transiently executed instructions are never architecturally visible, they can leave observable side effects in the CPU’s micro-architectural state (e.g., the cache [56], load ports [90], line-fill buffers [98], store buffers [14]). This has enabled transient execution attacks, where an adversary coerces the CPU into transiently accessing unintended information before using micro-architectural side channels [33, 71, 98, 100, 102] to leak information.

The Spectre family of attacks is one class of transient execution attacks that abuse CPU predictors, via (mis)training or tampering, to trigger speculative execution of instructions that access sensitive data, which can later be inferred through side channels. There are a handful of Spectre variants, classified by the CPU predictor they target. Spectre-PHT (Spectre-v1) [11] targets the pattern history table (PHT)—used to predict the outcome of conditional branches—to trigger the speculative execution of instructions preceded by a conditional branch. Spectre-BTB (Spectre-v2) [56] targets the indirect branch predictor, i.e., the branch target buffer (BTB), to trigger the speculative execution of forward-edge indirect branches. Spectre-STL (Spectre-v4) [48] abuses the memory disambiguator to load data before a prior, dependent store completes.

A variety of defenses have been proposed to address transient execution attacks, spanning both software [17] and hardware [37]. In the Linux kernel, the Spectre-PHT mitigation relies on developer annotations to identify when an array index is not trusted and add bit-wise operations to bound array indices during mis-speculation [65]. Spectre-BTB is mitigated via retpolines [43] and IBRS [41].

There are other root causes of transient execution, but predictor-based speculation and unexpected exceptions are the most prevalent and well-researched [85]. For exception-based transient execution attacks such as Meltdown [56] and Fallout [14], the problems are typically fixed through hardware changes [45], since it is generally clear what should happen in such situations: transient execution should respect potential exceptions. For predictor-based transient execution, isolation can be enforced such that the predictors cannot be influenced across contexts [41, 44] (e.g., influencing the kernel from userland), which can mitigate attacks that rely on abusing shared micro-architectural states [11, 72]. But for intra-context attacks, the safety of speculation is inherently software-defined, and solutions that prevent speculation, in general, are expensive [56]. Hence, selective application of various mitigations (such as lfence [40] or SSBD [13]) in targeted places is recommended by vendors [42] and has begun to receive attention by researchers [24].
3 Motivation

The combination of BPF and Spectre attacks allows an adversary to mount transient execution attacks to leak sensitive data from the kernel with little noise and without reliance on software vulnerabilities [55,56]. As a result, eBPF functionality is disabled by default for unprivileged users [19,87,94]. Features that use eBPF, such as seccomp-BPF, have seen proposals to support eBPF, but were rejected due to the complexity and difficulties that stem from maintaining certain mitigations [51]. Yet, the full set of BPF functionality is desirable for a lot of applications. Apart from seccomp-BPF, there are a plethora of proposed applications that try to use BPF in an unprivileged manner, including high-speed storage systems [104] and scheduling [39,52]. XDP-based applications [36], such as Cilium [95] and Katran [80], would also benefit from a reduction in privilege required by the BPF runtime.

Currently, if unprivileged BPF is enabled, the BPF verifier and JIT engine are responsible for applying BPF-specific Spectre mitigations at load time [58], which we collectively refer to as Linux provisional mitigations (LPM). For Spectre-PHT, the verifier rewrites pointer arithmetic to bound results within a given object, similar to the array_index_nospec macro [65]. To protect type safety under Spectre-PHT, the verifier's static analysis enforces type consistency across all combinations of branch choices, even impossible ones (e.g., a path where two branches with contradicting conditions are both taken). For Spectre-BTB, the JIT engine adds instrumentation for indirect branches in accordance with the kernel-wide defenses (such as retpolines [43]). For Spectre-STL, the verifier emits speculation barriers (e.g., lfence [40]) after stores to the stack, whenever pointers are involved.

A summary of defenses the Linux kernel deploys against transient execution attacks is provided in Table 1. Spectre-BTB and Meltdown are mitigated by generic, kernel-wide defenses, such as retpolines [43], IBRS [41], and KPTI [66], while generic Spectre-PHT and Spectre-STL defenses, such as SLH [15,82] and SSBD [13], are expensive to apply to the whole kernel. Consequently, the LPM defend against these attacks specifically in the context of BPF; however, the LPM have various issues regarding compatibility, complexity, scope, and performance, which we summarize below.

Compatibility. The LPM introduce additional analysis complexity and safety requirements to the verifier, causing it to reject (previously) safe code patterns in BPF programs. For example, Cilium [95] breaks because it performs variable-offset stack accesses, and Katran [80] breaks because it requires conditional pointer arithmetic. In addition, valid BPF programs experience verifier state explosion due to the verification regarding impossible branch combinations (§7.1.3). This places a burden on application developers to either address the verifier's requirements by rewriting code patterns targeted by the mitigations, or decide to forgo protection and load the program with elevated privilege.

Complexity and Scope. The LPM were formed over time, in an ad-hoc manner, as a response to repeated bypasses of Spectre-PHT and Spectre-STL mitigations caused by incompleteness and/or corner cases [7,29,57]. Progressively retrofitting security in this manner adds complexity to the BPF (sub)system that makes it difficult for kernel developers to reason about the safety of interactions between proposed verifier changes and the LPM [8,91,92]. Further, despite all this added complexity, the mitigations still provide incomplete protection, not covering eBPF or native BPF helper functions.

Performance. The mitigation against Spectre-STL inserts speculation barriers (e.g., lfence) after any pointer spill, and subsequent load, targeting the same stack slot. As a result, the LPM incur significant runtime overhead, especially in non-trivial BPF programs (§7.2.2). In general, it is difficult to stop mis-speculation in a way that does not hinder performance [12], which is the root cause of complexity in LPM.

4 Threat Model

Adversarial Capabilities. We assume an attacker that has access to BPF functionality and seeks to disclose sensitive information from the kernel. The attacker also has the appropriate rights to use the desired types of BPF programs (e.g., cap_net_admin for XDP), but is otherwise unprivileged. We do not assume the attacker can install BPF programs with cap_perfmon, since this capability can disclose a wide range of sensitive values already (such as register values at any point) [67]. As a result, performance-related BPF types are not available to the attacker, like the BPF programs that attach to Kprobes [93] and tracepoints [9].

As an unprivileged user of the system, the attacker can create any BPF program in the kernel that passes verification, as well as control the contents of BPF maps within allowed semantics. The attacker is also assumed to have the ability to arbitrarily trigger speculation at any point between the entry and exit of their BPF program, including when it invokes helper functions—these are only limited by potential speculation barriers. Under such speculative execution scenarios, the attacker can also arbitrarily flip conditional jumps and skip or forward any store to any load targeting the same address.

Hardening Assumptions. Regarding the kernel, we assume that the concrete execution of any kernel or BPF-runtime code is safe, and that there are no errors in the BPF verifier or JIT engine. Any execution of kernel code, concrete or speculative, before and after execution of a BPF program, is also assumed to be safe—i.e., the kernel does not leak any sensitive information into BPF programs’ speculation, nor does it incorrectly use unsafe values after BPF programs’ speculation. The only defense against Spectre-PHT and Spectre-STL for BPF is BeeBox. Further, we assume that the kernel has appropriate defenses, in either software or hardware, against other cross-privilege transient execution attacks, such as Meltdown [56].
Table 1: Existing Linux kernel defenses and BeeBox’s coverage over transient execution attacks and compatibility features.

<table>
<thead>
<tr>
<th>Category</th>
<th>Feature</th>
<th>LPM [58]</th>
<th>retpoline [43]</th>
<th>IBRS [41]</th>
<th>KPTI [66]</th>
<th>BeeBox</th>
</tr>
</thead>
<tbody>
<tr>
<td>Security</td>
<td>Block Spectre-PIT in BPF code</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Block Spectre-PIT in BPF helpers</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Block Spectre-STL in BPF code</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Block Spectre-STL in BPF helpers</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Block Spectre-BTB</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Block Meltdown</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Compatibility</td>
<td>Allow conditional ptr. arithmetic in unpriv. BPF</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Avoid verifier state explosion in unpriv. BPF</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Figure 1: The memory layout of BeeBox. The left part represents the CPU register states when a JITed BPF program is running; the middle part is a magnified view of the memory layout of a beebox; and the right part represents the virtual address space, where beeboxes are placed after the VMALLOC region, with a 4GB poison region in-between adjacent pairs.

Defenses against memory errors, such as control-flow integrity [5], page-table protection [22], pointer integrity [59], and data isolation schemes [75, 84] are orthogonal to BeeBox.

5 Design

The goal of BeeBox is to prevent BPF programs from accessing sensitive data, even under speculative execution. Fortunately, the majority of data that BPF programs access (e.g., the BPF stack and BPF maps) are designed to be visible to users. This key observation allows BeeBox to identify what data should be isolated, which does so by leveraging SFI techniques [99], tailored to the speculative execution domain. Specifically, BeeBox places all data a given BPF program will access inside a dedicated memory region, called beebox, and applies SFI-style isolation to the program, confining all of its memory accesses to this region. BeeBox thus ensures memory safety, as the BPF program can only access data inside the beebox—concretely or speculatively—, which does not contain any sensitive data.

5.1 BeeBox Sub-address Space

BeeBox isolates data directly accessed by BPF programs in beeboxes, as shown in Figure 1. A beebox is created for each user, such that no information can be leaked to, or from, other users (including user root). For example, as shown on the right of Figure 1, beebox_1 and beebox_2 are two separate beeboxes that were created for two different users. In addition, 4GB poison zones are placed between each beebox to support optimizations (§5.4). The kinds of memory objects placed inside each beebox sub-address space are magnified in the middle portion of Figure 1 and described below.

BPF Stack. The BPF stack is a region where BPF programs can temporarily store intermediate computation results that are not semantically shared with other components (e.g., the kernel itself or user applications). Since BPF programs are executed with (CPU, core)-migrations disabled—i.e., the BPF program invocation will stay on the same CPU—, we allocate a BPF stack region for each CPU in a beebox.

BPF Maps. BPF maps hold user-owned data, which can be accessed by both BPF programs and their corresponding user process (via the bpf system call). To support maps securely, their defining data structures are split: map data and some metadata essential to map operations (e.g., lookup and update) are safely placed in the beebox, since their contents are either controlled or can be inferred by the user; the remaining metadata used by the kernel itself (e.g., associated function pointers) remain outside of the beebox. Some metadata, such as the map size, are duplicated for the kernel code to access, because their copy in beebox is untrusted under speculation.

BPF Context. Handling contexts poses unique challenges since context data structures can interleave private kernel data and BPF-accessible data. For example, a BPF program of type BPF_PROG_TYPE_SOCKET_FILTER will receive a struct sk_buff in its context that has a large number of fields, out of which only a handful are made accessible to the BPF program itself (via struct __sk_buff). For these BPF program types, BeeBox copies the relevant parts of the context data structure that are visible to the BPF program into the BPF stack region. In cases where private kernel data is not interleaved with BPF-accessible data, the context data structure is usually generated and filled on the kernel’s stack, just before the BPF program’s invocation (as in seccomp-BPF).
Here, the non-interleaved context can be generated on the BPF stack region and no other changes are required. Network-related BPF programs usually have access to a packet buffer, and a pointer to a packet, which, in this case, is part of the context object. This is a common pattern in BPF, and the packet can be seen as an extension to the context. Through accurate static analysis done by the BPF verifier, accesses to these packets are identified and protected against overflowing. We handle packets the same way we handle context: we copy them into beebox. Note that copying context data structures (and packets) can be costly; we designed certain optimizations that can be applied to different types of BPF programs (§5.4).

**Space Considerations.** The size of the beebox region is 4GB. We chose this size mainly to allow for efficient SFI instrumentation (§5.2). Additionally, the stack region of each CPU reserved for BPF programs should be \((c + p) \cdot l + s\), where: \(c\) is the size of the largest context, \(p\) is the size of the largest packet we allow BPF to handle, \(l\) is the level of nested interrupts which can have BPF running, and \(s\) is the size of the kernel stack. This equation stems from the fact that all BPF stack frames are allocated on the kernel stack in the vanilla Linux kernel, and each nested interrupt can have at most \(c + p\) bytes added on top of that. We have chosen the total reserved area per-CPU to be 280KB to cover \(p = 64\)KB, \(c = 4\)KB, \(l = 4\), and \(s = 8\)KB. (These sub-regions are allocated and mapped during bootstrap, so they cannot be arbitrarily sized.)

### 5.2 Securing BPF Programs

**Instrumentation.** BeeBox uses SFI-style isolation to ensure pointers used by a BPF program are confined to a 4GB beebox region that contains only user-accessible data. Pointers embedded in a BPF program during JITing, and pointers passed to the program by the calling context, are transformed into boxptrs: 32-bit offsets within the beebox region. When a memory access needs to be performed, a usable 64-bit pointer is created from a given boxptr by clearing the upper 32 bits of the boxptr and adding the result to a register—the boxbase register—, which contains the starting address of the beebox region (1 in Figure 1). To initialize the boxbase register, the JIT engine emits a native instruction at the beginning of a BPF program which loads the base pointer, encoded as an immediate value, into the boxbase register. The boxbase register is never spilled or used in any other way, ensuring it cannot be controlled speculatively (or leaked); BeeBox guarantees the target of memory accesses from within BPF programs are always confined inside the beebox region. (Figure 3 shows an example of this transformation in x86-64.)

**BPF Stack.** There is a per-CPU pointer keeping track of the current top of the BPF stack allocated by BeeBox. The pointer is updated by locally atomic instructions to be safe against interrupts, since BPF programs can run with interrupts enabled and can also be invoked within interrupt context.

Figure 2: Memory layout and CPU states when helper functions are executed. Blue indicates the value is speculatively trusted; yellow indicates otherwise. The left part represents the CPU state when running an SU-hardened helper; the right part corresponds to running an ST-hardened helper.

BeeBox modifies the JIT compiler to emit inlined prologues and epilogues to move the BPF stack pointer and set the frame pointer register accordingly (2 in Figure 1).

**BPF Maps.** When BPF maps are accessed by the kernel on behalf of the user’s query via system calls, the kernel uses normal pointers to dereference the map data inside a beebox (instead of boxptrs). However, values loaded from the beebox speculatively should not be trusted and should be handled safely (e.g., masking the value before using it), which is consistent with the kernel’s current security model regarding user input [70]. When BPF maps are accessed directly from BPF programs (e.g., 3 in Figure 1), BeeBox uses 32-bit boxptrs, and the metadata is either embedded in the JITed program or looked up from the copy inside beebox. When BPF maps are accessed by helpers, the BPF program passes boxptrs to (modified) helper functions which use them safely.

**BPF Context.** Regardless of whether private kernel data is interleaved with BPF accessible data, BeeBox transforms the context pointer (4 in Figure 1) into a boxptr at the call site of the kernel that invokes a given BPF program. If the context includes a packet, the packet is also copied into the beebox. All pointers in the context data structure that point into the packet are also converted to boxptrs.

### 5.3 Securing Helper Functions

There are two hardening strategies for securing BPF helper functions. **SU-hardening** (Speculatively Untrusted, §5.3.2) protects helper functions such that they can be safely run during (untrusted) speculation. **ST-hardening** (Speculatively Trusted, §5.3.3) ensures that helper functions are run under the precondition that there is no influence from untrusted speculation. Figure 2 illustrates the memory layout and CPU states when helper functions are executed under both schemes.
Listing 1: A compiled helper function, bpf_get_current_uid_gid, where all memory loads are via clean pointers.

```
1 bpf_get_current_uid_gid:
2    mov rax, QWORD PTR gs:0x1ad00 # struct task_struct
3    test rax, rax
4    je 1f
5    mov rax, QWORD PTR [rax+0x638] # struct cred
6    mov rax, QWORD PTR [rax+0x4] # uid, gid
7    ret
8 1:
9    mov rax, 0xffffffffffffffea # -EINVAL
10   ret
```

5.3.1 Clean Pointers

We first introduce the definition of clean pointers, an important concept for our discussion later. All memory accesses using boxptrs can be confined with BeeBox’s SFI scheme (§5.2); however, certain BPF helpers use native pointers that point outside of a beebox, and hence cannot be transformed into boxptrs. These native pointers can potentially be speculatively-hijacked by the BPF program, escaping BeeBox’s defense. Since we assume the correctness of the helper functions’ concrete execution, we can avoid speculatively targeting unsafe locations by placing speculation barriers (e.g., _fence instructions) before native pointer dereferences. This naïve solution can be greatly improved upon with a key insight: native pointers in helper functions that are uninfluenced by any untrusted input (e.g., arguments passed from BPF programs) do not require a preceding speculation barrier—they can be safely dereferenced without leaking sensitive kernel data. We consider such pointers as clean.

Clean pointers are identified by static taint analysis, where the sources of taint are untrusted input (e.g., helper function arguments, values loaded from the beebox region, values loaded from the stack). If a native pointer is not tainted when dereferenced, then the dereference is considered to be done via a clean pointer. (boxbase is handled like a clean pointer.) Importantly, clean pointers are native pointers, so their values must be prevented from leaking into attacker visible realms (e.g., beebox memory and the branch predictor). Also, clean pointers are read-only, because any write to a kernel address can contaminate future reads to that address (as specified by our threat model), which may compromise the “cleanliness” of other (clean) pointers derived from the address.

Listing 1 shows a BPF helper function, bpf_get_current_uid_gid, which is used for getting the current task’s user ID (UID) and group ID (GID). This function uses three native pointers to: get the current struct task_struct (line 2); extract its security context, struct cred (line 5); and read the task’s UID and GID together, as a single quadword (line 6). Since all memory accesses in the function refer to private kernel data, and since all pointers are uninfluenced by BPF program data, all three native pointers are considered clean pointers and do not require any hardening from BeeBox.

5.3.2 Speculatively Untrusted Hardening

Speculatively untrusted hardening (SU-hardening) assumes that the execution of BPF helper functions can be run under untrusted speculation. Given this, BeeBox ensures that: (1) all memory accesses into the beebox by the helper are SFI-hardened (6 in Figure 2); (2) no native pointers are leaked into the beebox; and (3) all memory accesses to external (i.e., kernel) data are either prefixed with a speculative barrier or accessed via a clean pointer (5 in Figure 2). As far as (2) goes, BeeBox modifies the helpers to store the local variables using the BPF stack, instead of the kernel stack, while keeping the return addresses outside of beebox. This is to prevent the attacker from mounting a Spectre-STL against stale kernel stack values through the helper.

Since SU-hardening requires speculation barriers prior to external memory accesses that are not through clean pointers, this technique is better suited to BPF helper functions that do not access lots of external data. For example, certain utility helpers (e.g., bpf_strtol) and the majority of BPF map helpers only operate on information already in the beebox. However, some helpers that do access external data mainly through clean pointers (e.g., bpf_get_smp_processor_id) do benefit from this approach as well.

SU-hardening is achieved through three main steps. First, the interface between BPF programs and helper functions is modified to replace native pointers with boxptrs, which maintain their 32-bit representation throughout the helper’s execution. Since boxptrs cannot be accessed directly, they go through a special C macro, unbox, which applies SFI instrumentation and makes the boxptrs usable. (The unbox macro uses the volatile keyword to nail down the exact point of dereference in the compiled code and avoid spilling any native pointers before they are used.) Second, BeeBox ensures that helper functions maintain the “cleanliness” guarantee of the boxbase register by ensuring it is only used in the unbox macro and never spilled to the stack or overwritten. Finally, BeeBox ensures that external memory accesses are exclusively done via clean pointers, else they are prefixed with a speculation barrier. (This is currently done manually; a binary analysis tool based on our clean pointer identification algorithm is needed if one wishes to automatically ensure function is safe. We leave this for future work.)

5.3.3 Speculatively Trusted Hardening

Speculatively trusted hardening (ST-hardening) secures a helper by providing the guarantee that all speculation in a given BPF helper function will be safe, and thus no additional considerations are needed for accessing kernel data (7 in Figure 2). This hardening technique is fitted for helper functions that access kernel data much more than data produced by BPF programs. For example, bpf_skb_clone_redirect will clone the sk_buff passed in, which involves accessing a lot of kernel data, including the heap allocator’s metadata.
To achieve this guarantee, BeeBox sanitizes any data flow originating from the calling context or the beebox. Thus, at a given helper’s entry point, all boxptrs are converted into native pointers and a speculation barrier is inserted to avoid speculative execution on unsafe arguments. Additionally, accessing beebox memory requires a speculation barrier (in Figure 2), which is provided by another utility macro BeeBox defines. By default the speculation barriers are inserted; however, if a developer desires, they can (carefully) optimize away the barriers. Finally, in the JITed BPF program, another speculation barrier is used to avoid unintended values from leaking back into the BPF’s CPU state.

5.4 Performance Optimizations

We refer to the scheme combining all the techniques up until this point as BeeBox-Basic. It comprehensively stops any Spectre-PHT or Spectre-STL attack originating from unprivileged BPF programs; however, it can be costly, due to the copying of context and dense SFI-instrumentation. We address these problems in the following subsections.

5.4.1 Generic Optimizations

Reduced Copy. The first optimization variant, called BeeBox-RC (Reduced Copy), improves performance by reducing the amount of data that is copied into a beebox. Notably, this optimization strategy is applicable to every type of BPF program that does not have a stack-based context. At a high-level, by using the static analysis done by the verifier, a BPF program’s accesses to contexts and packets are tracked and only the used portion of them is copied into the beebox.

BPF does not support variable-offset context accesses. Leveraging this fact, BeeBox-RC records up to 16 distinct context access offsets during the verifier’s static analysis, and stores them along with the BPF program. If the number of accessed fields is within the limit, BeeBox-RC only copies each accessed field of the context data structure into the beebox.

For packets, because BPF supports variable-sized accesses, BeeBox-RC reuses the range analysis results of the max (packet) access offset, only copying-in the packet up to the size of the max offset that the BPF program will access. Because networking-related BPF programs typically operate on only the header of packets, the maximum size of packet copying is usually bounded to a reasonable number.

5.4.2 Case-specific Optimizations

Clean Context and Packet Pointers. The second optimization variant, called BeeBox-CP (Clean Pointer), makes the context pointer a clean pointer (i.e., a native pointer instead of a boxptr; in Figure 2) to avoid copying contexts. Consequently, the context pointer is handled with the guarantees that clean pointers require: it is never spilled into memory and does not leak into other data flows—aside from being used as a base for dereferencing. As a result, the context pointer is treated similar to boxbase, where calls to helper functions avoid clobbering or spilling it. Other guarantees of the clean context pointer are checked in the verifier, and the optimization only applies when the context pointer is indeed clean. However, to ensure that BPF programs produced by a BPF compiler adhere to such requirements, and can hence utilize the optimization, BPF compiler changes are necessary. (For the purposes of this work, we manually produced compliant BPF programs in order to make use of the optimization.)

Importantly, since the context is a clean pointer pointing to the original context data structure outside of the beebox, the packet pointer enclosed inside the context is also a clean pointer. To safely allow variable-size accesses into the packet using clean pointers, BeeBox-CP (re)uses the static analysis done by the verifier to accurately identify and rewrite packet accesses into calls to custom helper functions that utilize the array_index_nospec macro in the kernel.

Ring-buffer No-copy. The third optimization variant applies when the type of the BPF program already requires the user to have network-device access privileges, e.g., xdp. In this case, we avoid copying packets by directly placing the device’s ring buffer into non-DMA memory, if the driver does not have support for XDP natively) in the beebox. Note that this optimization cannot work with higher-level, network-related BPF programs (such as socket-level BPF), because when such a packet is allocated, either at the driver or NAPI level, the destination socket is unknown to the kernel. We name this optimization variant BeeBox-RB.

Optimizing memcpy and Friends. In certain BPF helpers, functions such as memcpy and memcmp are used. These functions are simple but have extensive pointer dereferences, which seriously impact performance if SFI instrumentation is applied to all of them. We optimize this type of function by “unboxing” the arguments to native pointers when calling them, and manually ensuring that these native pointers are not changed or spilt in the function. Because the size of copying, or comparing, is a 32-bit integer value, we place a 4GB poison zone after the beebox region to confine all potential memory accesses by these functions.

6 Implementation

We implemented a prototype of BeeBox on x86-64 Linux v6.1. Due to the sheer number of features available to BPF, our prototype of BeeBox does not cover everything: we only BeeBox-hardened a subset of all the available features (of vanilla BPF) to support popular BPF applications. Our implementation consists of three inter-connected components, which we describe below.
6.1 BeeBox-aware JIT Compiler

We reserve beebox regions in the unused part of the kernel’s address space, after the vmalloc region, and we implemented a new allocation function, beebox_alloc, which wraps __vmalloc_node_range to provide custom ranges for allocating inside each beebox. In the JIT compiler, we added an instruction to load the base of the appropriate beebox into r12, which is not used by the current JIT compiler and acts as the boxbase pointer (e.g., line 1 in Figure 3b). We also modified the verifier to insert a new BPF (pseudo) instruction, BPF_BOXMEM, directly before load or store instructions in the BPF bytecode. BPF_BOXMEM marks the following instruction, signalling to the JIT process that the access should be SFI-hardened. During JITting, BPF_BOXMEM-ed memory accesses are transformed into two instructions that make up the core SFI instrumentation. First, a mov with a 32-bit register operand (e.g., eax for rax) is inserted to clear the upper 32 bits of the register (line 5 in Figure 3b). Then, second mov instruction is inserted with a base + index addressing mode, where r12 (i.e., the boxbase pointer) is used as the base and another 64-bit register, which stores the offset into the beebox, is used as the index (line 6 in Figure 3b). These two moves bound memory accesses inside the 4GB beebox region.

In each beebox region, we allocate a 280KB stack for every CPU. Each BPF stack has a corresponding variable representing the top of the stack as a boxptr. We modified the JIT compiler’s handling of function prologues and epilogues in order to “open” and “close” stack frames in the beebox stack region, instead of the normal stack, and to move the frame pointer into ebp, which is a 32-bit boxptr. All BPF stack accesses are frame-pointer-based so we avoid having to move rsp and handle additional stack swapping. Lastly, all stack accesses identified by the verifier are prepended with the BPF_BOXMEM BPF instruction, which are later correctly JIT-compiled to SFI-hardened native code.

6.2 Maps, Helpers, and Language Tricks

To simplify development and debugging, when writing C code that interacts with pointers to a beebox, we added a special attribute, i.e., ‘__beebox’, to annotate boxptrs in the kernel—the __beebox attribute is defined as

Listing 2: Example of data structures placed inside beebox, which only contains information that can be exposed to the user. All pointer fields are annotated with ‘__beebox’.

| struct bb_pcpu_freelist_head { |
| struct bb_pcpu_freelist_node __beebox ‘first; |
| raw_spinlock_t lock; |
| } |
| struct bb_pcpu_freelist { |
| ...
| struct bb_pcpu_freelist_head extralist; |
| } |
| struct bpf_htab_inner { |
| ...
| struct bucket __beebox ‘buckets; |
| struct bb_pcpu_freelist freelist; |
| u32 n_buckets; |
| ...
| } |

__attribute__((noderef)) when the source code is inspected by the static analysis tool sparse [64]. We chose to implement boxptrs as an attribute instead of defining a new type since the attribute can transparently keep the pointer’s original type information. This makes incorrect propagation and accidental dereference noticeable at compile time, which streamlined the development of our prototype (otherwise, it would be difficult to apply the unbox macro correctly in expressions like: \&array_map->map->elem[i].lock).

We currently support five types of maps commonly used by BPF programs: array map, per-CPU array map, array map of maps, hash map, and LRU hash map. When data structures such as linked lists—which are needed for hash map—are moved into a beebox, their internal pointers are also converted into boxptrs. In Listing 2, we show how certain data structures placed in beebox are defined. All the embedded pointers are annotated with the __beebox attribute, and their value represents the offset within the beebox. We also duplicated and adapted the accessor macros, and inline functions, for these data structures to work with boxptrs.

We reserved r12 when compiling helper functions to ensure it is never spilled or contaminated. In addition, we changed the function signature of the helpers we ported to BeeBox to use boxptrs; thus, together with the changes in the data structures, the compiled helper functions will not contain any accidental un-instrumented pointer dereferences. The counter-part to these helper functions is the call sites in the JITed BPF programs; they are transformed by the verifier to use boxptrs to point at the in-beebox part of the BPF maps.

Array map, per-CPU array map, and array map of maps each have three helper functions that need to be adjusted for the BPF runtime: we support two in array map, three in per-CPU array map, and one in array map of maps. For other helpers, we support everything used by the BPF programs in our evaluation, including 2/42 common helpers, 6/14 socket-specific helpers, and 1/29 xdp-specific helpers.
6.3 Supported BPF Program Types

**cBPF.** We support cBPF in seccomp-BPF and packet filtering. Our analyses and BPF-level transformations are implemented in the cBPF verifier. The stack accesses in cBPF use dedicated cBPF instructions, hence we insert \texttt{BPF\_BOXMEM} before the translated eBPF instruction(s). seccomp-BPF’s context is stack-based so it is just moved to the BPF stack. For packet filtering cBPF, we applied BeeBox-CP (§5.4.2), while packet accesses go through instrumented helpers.

**Socket Filter eBPF.** For eBPF-based socket filters, we implemented both BeeBox-RC (§5.4.1) and BeeBox-CP (§5.4.2). To implement BeeBox-RC, all stack accesses and map accesses remain the same; for context access, we insert \texttt{BPF\_BOXMEM} before each context-memory access. We also changed the invocation of BPF programs to copy selected parts of the context and packets into the \texttt{beebox}, which we transformed accordingly. In the case of BeeBox-CP, we performed additional static analysis on the BPF programs to verify that the context pointer is not leaked into other registers or spilled to memory. Due to calling convention considerations, we chose the callee-saved register \texttt{rbx} to hold the context pointer. We also duplicated the helper functions to avoid using or clobbering \texttt{rbx}, similar to \texttt{r12}.

**XDP.** Our support of XDP BPF programs includes two optimizations: BeeBox-RC (§5.4.1) and BeeBox-RB (§5.4.2). For XDP, the context is generated dynamically on the (kernel) stack; thus, we simply generate it on the BPF stack instead. In BeeBox-RC, part of the packet is copied into the \texttt{beebox} region. Although the copying only requires part of the packet, we allocate enough memory to fit potential XDP-supported modifications to the packet (e.g., extending the header). The range of the packet copied out is also adjusted in the event that packet pointers are modified in the XDP context. When enhanced with BeeBox-RB, depending on whether the network driver supports XDP natively, the allocation of the packet buffer at the driver level or the NAPI level is adjusted to appropriately handle allocating directly in the \texttt{beebox} region.

6.4 Engineering Effort

The core implementation of BeeBox required \approx 900 additional lines of code (LOC) added to Linux v6.1. Specifically, we added \approx 700 LOC to the BPF JIT engine and verifier to implement the main functionality and optimization schemes of BeeBox, and \approx 200 LOC to adapt the hook points that invoke BPF programs. Atop this, we also migrated \approx 1000 lines of helper code into separate compilation units and manually added the instrumentation at the source code as necessary. Regarding the helpers that we instrumented (§6.2): support for the three array map variants took a graduate student (with knowledge of Linux kernel internals) one day to develop; support for the two hash map variants took three days to develop; and the program-type-specific helpers took four hours.

7 Evaluation

We evaluated BeeBox in terms of effectiveness and performance. For the former, we analyzed the impact of BeeBox on BPF’s functionality compared against no-mitigations and existing mitigations in Linux v6.1. For the latter, we evaluated BeeBox’s performance on synthetic eBPF benchmarks and real-world BPF applications, including Katran, seccomp-BPF, and cBPF packet filtering. Our experiments were run on a machine equipped with a 16-core, 3.7GHz Intel Xeon W-2145 CPU and 64GB RAM, running 64-bit Ubuntu 20.04 LTS.

7.1 Effectiveness of BeeBox

7.1.1 Security Analysis

When hardened with BeeBox, four different kinds of pointers are considered for the security of the system. The first kind is pointers into untrusted regions, like the \texttt{beebox} and the kernel stack, when used by SU-hardened helpers. They are speculatively-safe because they cannot target outside the untrusted region(s) and there is no sensitive data in those regions. The second kind is clean pointers; they are speculatively-safe because they can only target the same locations as concrete execution, which is assumed to be trusted. The third kind is pointers used in ST-hardened functions. These pointers are secured since ST-hardened functions use speculation barriers to ensure no attacker-controlled speculation occurs during their execution. The forth kind is developer exempted pointers, including ones handled with \texttt{array\_index\_nospec}, whose value can be influenced by speculation, but are explicitly permitted by developers. In summary, BeeBox comprehensively protects all memory accesses during BPF execution.

7.1.2 Security Evaluation

We created three synthetic exploits that simulate speculative execution attacks launched from BPF programs, which attempt to (speculatively) access a target location to leak sensitive data. We consider BeeBox able to stop a given attack if we can no longer detect a signal in the cache corresponding to accessing the sensitive data. To probe the cache for a signal, we created a kernel module that performs Flush+Reload [102].

**Exploit #1: Spectre-PHT.** For the first exploit, we crafted a BPF program that performs a dynamically indexed access into a BPF map element. Here, the mis-speculation starts from a conditional branch, similar to the original Spectre attack [56]. The exploit works without mitigations and fails against BeeBox (and LPM).

**Exploit #2: Spectre-STL.** For the second exploit, we targeted Spectre-STL. We crafted a BPF program that triggers erroneous store-to-load forwarding by saturating store ports [58]. Similar to the first exploit, this exploit works without mitigations and fails against BeeBox (and LPM).
The LPM prevent certain BPF program patterns from being accepted by the verifier. In contrast, BeeBox is designed with compatibility in mind, allowing more programs to pass verification, while also receiving superior protection. We describe two such examples, which we also summarize in Table 1.

**Verifier State Explosion.** The BPF verifier keeps an internal count of analyzed instructions to decide whether a program is safe after processing 611 instructions. Since BeeBox comprehensively stops access of sensitive information from BPF programs, it avoids this analysis and achieves verification scalability as program size and complexity increase.

### 7.1.3 Compatibility Evaluation

The LPM prevent certain BPF program patterns from being accepted by the verifier. In contrast, BeeBox is designed with compatibility in mind, allowing more programs to pass verification, while also receiving superior protection. We describe two such examples, which we also summarize in Table 1.

**Conditional Pointer Arithmetics.** Under the current Spectre-PHT mitigation(s), the BPF verifier rejects BPF programs where a pointer can have different value sets due to different branches (with the exception of map value pointers that point to different elements). This is a common code pattern in BPF programs; for example, one of Katran’s BPF programs cannot be loaded by an unprivileged user since it has pointers that target different packet fields based on the IP version (e.g., IPv4 vs. IPv6). BeeBox can avoid this restriction; even though boxptrs can speculatively perform arbitrary arithmetic, they cannot point to sensitive data (outside beebox).

### 7.2 Performance of BeeBox

**7.2.1 Basic Operations**

We crafted a synthetic micro-benchmark that consists of small BPF programs (which we make available along with our prototype of BeeBox) that stress different kinds of operations, in order to better understand BeeBox’s behavior in comparison with LPM. For each case, the running time of the given BPF program was measured using the `bpf` system call with the option `BPF_PROG_TEST_RUN`, which is modified to use `rdtscp` for accurate timing. We ran each BPF program in a tight loop for one million iterations with interrupts disabled. The BPF program type we use is `socket`, and the synthetic packets are 64 bytes, except in the `packet` benchmark, where the packet size is 150 bytes. Finally, all benchmark programs explicitly avoid using the stack unnecessarily.

Results are shown in Figure 4. We compared four different settings: (1) no mitigation, (2) LPM, (3) BeeBox-RC (§5.4.1), and (4) BeeBox-CP (§5.4.2). Setting (1) executes BPF programs as root, while (2) – (4) execute BPF programs in an unprivileged manner. LPM incur a 270% runtime over-

---

**Listing 3: The packet-loading helper.**

```c
u64 bpf_skb_load_helper_32(const struct sk_buff *skb,
const void *data,
int headlen,
int offset) {
    __be32 tmp, *ptr;

    /* be32 tmp, 'ptr; */
    const int len = __sizeof(tmp);
    if (likely(offset >= 0)) {
        if (headlen - offset >= len)
            return get_unaligned_be32(data + offset);
    }

    /* return */
    return -EFAULT;
}
```

**Figure 4: Synthetic micro-benchmark that involves various socket filter eBPF programs and different hardening schemes (no protection vs. LPM vs. BeeBox).**
head compared to no protection (No Mitigation) on the stack benchmark, but do not exhibit any significant slowdown in any other case. Both BeeBox schemes incur a 7%–23% overhead on hash map operations due to the instrumentation of the helpers. BeeBox-RC has additional overhead due to copying, which peaks at 110% for the shortest BPF programs (tiny and context). This experiment also demonstrates that since LPM only add instrumentation around stack accesses, BeeBox’s runtime performance gain over the LPM comes from better handling of stack load/store operations and Spectre-STL.

### 7.2.2 Real-world BPF Programs

**Katran Benchmarks.** Real-world eBPF programs can be more complex than the aforementioned, synthetic microbenchmark programs, unavoidably using the stack for register spills. The performance of the LPM severely degrades in such situations, whereas BeeBox scales better with the complexity of the BPF program. To test this, we evaluated BeeBox on Katran’s load balancer XDP program. In this setting, we benchmarked two BeeBox optimizations, applied separately to XDP: BeeBox-RC (§5.4.1) and BeeBox-RB (§5.4.2). The benchmarked BPF program is the standalone Katran XDP program, and the driving workload comes from Katran’s -perf_testing option in its test program. To run Katran’s XDP program as a non-root user we provided the corresponding process with CAP_NET_ADMIN and CAP_BPF.

The results of the benchmark are shown in Figure 5. The overhead of LPM is consistently above 70%, with an average overhead of 112%, which confirms our hypothesis that most of the overhead is caused by register spilling. The overhead of BeeBox-RC ranges between 28%–50% in most benchmarks, but spikes up to 200% for packets that are passed-through or dropped early; therefore, on average, it has an overhead of 71%. For BeeBox-RB, the average overhead is 20%, with a maximum overhead of 39% when the workload consists of IPv6 ICMP packets (workload 9). When we compare BeeBox-RB to BeeBox-RC, we observe that the copying of context contributes 42% overhead on average. We also constructed a modified version of BeeBox-RB that applies instrumentation only to helper functions to determine their contribution to the overall overhead. This modified scheme has an average overhead of ≈17%, indicating that most of the overhead in BeeBox-RC is due to the instrumentation of the helpers.

**Packet Filter Benchmark.** We tested the performance of raw socket filtering using cBPF. To efficiently support cBPF packet filtering, we applied the BeeBox-CP scheme (§5.4.2). On the benchmark machine, we setup a UDP server listening on localhost and a UDP client that continuously sends packets with 32-byte payloads. We then attached a cBPF filter to a raw socket of the loopback (lo) device; the filter on the raw socket gets executed in the softIRQ raised by the client sending the packets, and is thus on the same CPU as the client. We pinned the client process to a CPU, and made sure that

<table>
<thead>
<tr>
<th>Filter</th>
<th>No Mitigation</th>
<th>BeeBox-CP</th>
<th>%-Chg</th>
</tr>
</thead>
<tbody>
<tr>
<td>bpf1</td>
<td>325927 (± 3611)</td>
<td>327778 (± 3006)</td>
<td>+0.57%</td>
</tr>
<tr>
<td>bpf2</td>
<td>324615 (± 3960)</td>
<td>323374 (± 5375)</td>
<td>−0.38%</td>
</tr>
<tr>
<td>bpf3</td>
<td>324114 (± 3977)</td>
<td>323834 (± 5088)</td>
<td>−0.09%</td>
</tr>
<tr>
<td>bpf4</td>
<td>328610 (± 4827)</td>
<td>325568 (± 7818)</td>
<td>−0.93%</td>
</tr>
<tr>
<td>bpf5</td>
<td>328072 (± 3883)</td>
<td>325395 (± 7352)</td>
<td>−0.82%</td>
</tr>
<tr>
<td>bpf6</td>
<td>314810 (± 2025)</td>
<td>313618 (± 2650)</td>
<td>−0.38%</td>
</tr>
</tbody>
</table>

(a) Packet filtering performance in pkts/s with 95% CIs.

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>No Mitigation</th>
<th>BeeBox</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nginx</td>
<td>0.81% (± 1.09%)</td>
<td>0.32% (± 1.47%)</td>
</tr>
<tr>
<td>Redis</td>
<td>0.98% (± 0.44%)</td>
<td>0.84% (± 0.74%)</td>
</tr>
</tbody>
</table>

(b) Tput degradation of seccomp-BPF with 95% CIs.

Table 2: cBPF performance results.

The CPU’s utilization was close to 100%. Finally, to collect the results, we measured this scenario’s throughput (pkts/s) as packets are filtered using rules that are similar to prior works [50,101]. Results are shown in Figure 2a; the overhead introduced by BeeBox is < 1%.

**seccomp-BPF Benchmark.** Because seccomp-BPF’s context is generated on the stack, BeeBox supports it by replacing the (kernel) stack allocation with a BPF stack allocation. To measure BeeBox’s impact on seccomp-BPF performance, we applied sysfilter [23], a framework that generates and enforces seccomp-BPF policies, to Nginx [3] and Redis [4], and measured the performance of the hardened binaries. Nginx was configured to have two working processes. The test traffic was generated by wrk [30] with two running threads, each opening 128 connections, requesting 1KB payloads in every request. Redis was benchmarked using memtier [86] with two worker threads, each having 128 clients. The GET:set request ratio was 10:1, and the data object size was 32 bytes. Each experiment ran for one minute. We ensured the CPU running the Nginx server and the Redis server was saturated with the testing configuration. Given that we tuned the packet sizes to be small, both tasks are syscall-intensive. The Nginx experiment handles over 43.5K req/s, and the Redis experiment handles more than 155K req/s. The results are shown in Figure 2b. The JITed cBPF programs already have limited impact on the services, and hence our defenses do not incur any observable overhead.

**Memory Usage.** We ran the real-world benchmarks and monitored the system’s total memory usage. We measured this in a KVM VM with 4 vCPUs and 8GB of RAM, hosted on the benchmarking machine. The results of this experiment are shown in Table 3. The memory usage of the Linux kernel with no mitigations is reported in the second column (Vanilla Usage). The memory usage for BeeBox is given in the third
Figure 5: Katran’s load balancer XDP program/benchmark. The most efficient scheme, BeeBox-RB, also includes a breakdown into baseline, helper instrumentation, and BPF instrumentation (colored dark to light).

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Vanilla Usage</th>
<th>BeeBox Usage</th>
<th>Overhead</th>
</tr>
</thead>
<tbody>
<tr>
<td>At rest</td>
<td>176MB (178MB)</td>
<td>180MB (180MB)</td>
<td>2.4%</td>
</tr>
<tr>
<td>Packet filter</td>
<td>182MB (183MB)</td>
<td>186MB (188MB)</td>
<td>2.3%</td>
</tr>
<tr>
<td>Katran</td>
<td>580MB (582MB)</td>
<td>592MB (592MB)</td>
<td>2.0%</td>
</tr>
<tr>
<td>Nginx (seccomp)</td>
<td>189MB (190MB)</td>
<td>196MB (197MB)</td>
<td>3.5%</td>
</tr>
<tr>
<td>Redis (seccomp)</td>
<td>212MB (213MB)</td>
<td>218MB (221MB)</td>
<td>3.0%</td>
</tr>
</tbody>
</table>

Table 3: Memory usage of BeeBox compared to vanilla Linux. At rest means no workload is running. The reported numbers are formatted as 'avg (max)'.

## 8 Discussion

**Compatibility.** Regarding ISA-compatibility, although BeeBox is implemented for x86-64, the instrumentation techniques are generic and can be applied to other architectures. The main assumption required to enable our SFI scheme is that register-only masking and addition operations will enforce the pointer range during speculation. For existing BPF application compatibility, BeeBox does not require any changes from legacy BPF applications to benefit from its protection, unlike the current mitigations in the Linux kernel.

**Future Extensions.** To extend BeeBox to cover more BPF functionality, developer effort should be concentrated on migrating and instrumenting more helper functions and BPF data structures. The division between kernel data and BPF data in BPF maps is ultimately developer-defined, therefore, there is no easy way to automate the entire process. However, sparse [64] has made the manual instrumentation process much easier and less error-prone for us. An extension to sparse’s capabilities to allow multiple pointer annotations working together (e.g., __percpu and __beebox) can streamline the process even more.

## 9 Related Work

**Sandboxing BPF.** SandBPF [62] also uses SFI [99] to sandbox BPF. However, it is designed to defend against errors in the verifier and the JIT compiler, and does not defend against transient execution attacks. It also assumes the input to the helpers is completely trusted. Lastly, its overhead is much higher than BeeBox: ≈7% in end-to-end experiments.

**BPF Hardening.** The BPF infrastructure has been used to assist kernel exploitation. BPF JIT-spraying [73, 88] and EPF [50] use BPF to bypass ret2usr defenses [20, 54]. These attacks and proposed defenses revolve around memory safety, and assume the presence of memory-safety vulnerabilities. In contrast, BeeBox deals with transient execution attacks, which do not strictly require the presence of memory errors.

**Correctness of BPF.** Given the complexity of the BPF infrastructure, techniques have been proposed to enhance its security. Nelson et al. [79] proposed to use formal verification to secure the JIT compiler, and Jia et al. [49] proposed to use Rust to improve the runtime security of helpers. Similar to the work regarding BPF hardening above, these works also focus on memory safety problems, and hence differ from BeeBox’s focus on transient execution attacks.

**Wasm Hardening.** WebAssembly (Wasm) [34] is a portable, low-level bytecode designed, in part, to safely and efficiently run untrusted code in web browsers. Unfortunately, Wasm is susceptible to Spectre attacks, which allow the attacker to escape the sandbox and access sensitive data outside of it. To combat Spectre attacks in Wasm, Swivel [78] presents two hardening schemes: (1) a software-only SFI approach, similar to BeeBox; and (2) a hardware-assisted approach that makes use of Intel’s CET and MPK hardware extensions [21]. While the SFI scheme of Swivel is similar in concept to BeeBox, the two schemes have separate, domain-specific requirements and features that mold their respective designs.
10 Conclusion

We presented the design, implementation, and evaluation of BeeBox: a new security architecture that hardens BPF against transient execution attacks, allowing the OS kernel to expose its functionality to unprivileged users and applications. BeeBox sandboxes the BPF runtime against speculative code execution in an SFI-like manner, and uses a combination of static analyses and domain-specific properties to selectively remove enforcement checks to improve performance. Our BeeBox prototype for the Linux kernel supports popular features (e.g., BPF maps and BPF helper functions) and incurs low runtime overhead against prevalent transient execution attacks, such as Spectre-PHT and Spectre-STL. On average, BeeBox incurs 20% overhead in Katran’s macro-benchmarks and < 1% throughput degradation in end-to-end, real-world settings that involve seccomp-BPF and packet filtering.

Availability

Our prototype implementation of BeeBox is available at: https://gitlab.com/brown-ssl/beebox

Acknowledgments

We thank the anonymous reviewers for their valuable feedback. This work was supported by the National Science Foundation (NSF), through award CNS-2238467. Any opinions, findings, and conclusions or recommendations expressed herein are those of the authors and do not necessarily reflect the views of the US government or NSF.

References


[57] Piotr Krysiuk and Benedict Schlueter. bpf: Fix leakage due to insufficient speculative store bypass mitigation. https://git.kernel.org/pub/scm/linux/kernel/git/torvalds/linux.git/commit/?id=2039f26f3aca5b0e419b98f65d36481337b8e6e.


[83] Marios Pomonis, Theofilos Petsios, Angelos D. Keromytis, Michalis Polychronakis, and Vasileios P. Kemerlis. kR’ X: Comprehensive Kernel Protection


