Verified Programs Can Party: Optimizing Kernel Extensions via Post-Verification Merging

Hsuan-Chi Kuo  
University of Illinois at Urbana-Champaign  
Champaign, Illinois, USA  
hckuo2@illinois.edu

Kai-Hsun Chen  
University of Illinois at Urbana-Champaign  
Champaign, Illinois, USA  
kaihsun2@illinois.edu

Yicheng Lu  
University of Illinois at Urbana-Champaign  
Champaign, Illinois, USA  
yl53@illinois.edu

Dan Williams  
Virginia Tech  
Blacksburg, Virginia, USA  
IBM T.J. Watson Research Center  
Yorktown Heights, New York, USA  
djwillia@vt.edu

Sibin Mohan  
Oregon State University  
Corvallis, Oregon, USA  
sibin.mohan@oregonstate.edu

Tianyin Xu  
University of Illinois at Urbana-Champaign  
Champaign, Illinois, USA  
tyxu@illinois.edu

Abstract

Operating system (OS) extensions are more popular than ever. For example, Linux BPF is marketed as a “superpower” that allows user programs to be downloaded into the kernel, verified to be safe and executed at kernel hook points. So, BPF extensions have high performance and are often placed at performance-critical paths for tracing and filtering.

However, although BPF extension programs execute in a shared kernel environment and are already individually verified, they are often executed independently in chains. We observe that the chain pattern has large performance overhead, due to indirect jumps penalized by security mitigations (e.g., Spectre), loops, and memory accesses.

In this paper, we argue for a separation of concerns. We propose to decouple the execution of BPF extensions from their verification requirements—BPF extension programs can be collectively optimized, after each BPF extension program is individually verified and loaded into the shared kernel.

We present KFuse, a framework that dynamically and automatically merges chains of BPF programs by transforming indirect jumps into direct jumps, unrolling loops, and saving memory accesses, without loss of security or flexibility. KFuse can merge BPF programs that are (1) installed by multiple principals, (2) maintained to be modular and separate, (3) installed at different points of time, and (4) split into smaller, verifiable programs via BPF tail calls. KFuse demonstrates 85% performance improvement of BPF chain execution and 7% of application performance improvement over existing BPF use cases (systemd’s Seccomp BPF filters). It achieves more significant benefits for longer chains.

CCS Concepts: • Software and its engineering → Operating systems.

Keywords: eBPF, BPF, kernel extension, retropine, Spectre, transient attack, indirect jump

ACM Reference Format:

1 Introduction

Operating system (OS) extensions are more popular than ever. Linux BPF is marketed as a “superpower” that allows user programs to be downloaded into the kernel, verified to be safe and executed at kernel hook points. Currently, BPF extensions are used for system call security (e.g., seccomp-BPF [24]), performance tracing (e.g., tracepoints [29] and bcc [2]), and network packet processing (e.g., express data path (XDP) [31, 50]); many other use cases are proposed recently [2, 11, 13, 30, 31, 36–38, 40, 42–45, 51, 59, 63, 69–71, 74, 76–78]. BPF extensions have high performance, because BPF code is highly optimized [20] and BPF programs are executed entirely in the kernel. For example, sandboxing system call with seccomp BPF filters is reported to be 2× faster, compared with ptrace-based system call filtering [53]; using XDP [31, 50] to implement firewall rules improves...
network throughput by 11\times compared to iptable-based imple-mentsations [21].

Currently, BPF extensions are individually verified by an in-kernel BPF verifier and then executed independently. A common execution pattern for BPF extensions is to execute a chain of independently-loaded BPF extensions that are attached at the same hook point. The chain pattern is a natural implementation choice to support BPF extensions that are (1) installed by multiple principals [7, 24, 29]; (2) installed at different points of time (e.g., for temporally-specialized security policies [49, 60]), and (3) maintained to be modular and separate (for maintainability and debuggability) [17].

More fundamentally, due to the limited scalability of BPF program verification, the BPF verifier enforces each BPF extension to be small in size—each BPF program is limited to 4096 instructions with a maximum stack size of 512 bytes; therefore, large BPF features need to be split into small BPF programs, each of which can be verified in time [12].

However, we observe that the chain pattern has large performance overhead, due to indirect jumps penalized by security mitigations (e.g., Spectre), loops, and memory accesses. The typical chain pattern is implemented by a sequence of indirect jumps. The kernel maintains an array of pointers of loaded BPF extensions which are called one by one via indirect jumps to these pointers. While indirect jumps were optimized in modern CPUs via branch prediction, recent security mitigations to speculative vulnerabilities [57] (Spectre, specifically Variant 2) incur significant overhead to indirect jumps, negatively affecting the execution of the chained BPF execution. For example, the de facto software mitigation, Retpoline [23], makes indirect jumps 13.3\times slower [8, 67]. Besides loops, BPF extension chains can also be formed by tail calls; a BPF program can call another BPF program, which takes a memory access and an indirect jump.

To make the matter worse, the overhead of BPF extension chains increases with the increase of the length of the chain. We have seen real-world BPF use cases with long BPF extension chains. For example, systemd [26] installs 19 seccomp BPF filters to its services. Our benchmark on Redis initialized by systemd shows that the chained BPF extensions can cause a 10\% slowdown of Redis. We expect such long chains to be commonplace in future BPF use cases for fine-grained policies and richer features. The overhead of the chains would be untenable.

In this paper, we argue that execution and verification should be a separation of concerns. We propose to decouple the execution of BPF extensions from their verification requirements and limits—kernel extension programs can be collectively optimized, after each extension is individually verified and loaded into the shared kernel. We demonstrate that the decoupling could lead to new opportunities for performance optimizations, while maintaining the safety of the verified extension programs.

We present KFuse, a framework that dynamically and automatically optimizing chains of BPF extensions by transforming indirect jumps into direct jumps, unrolling loops, and saving memory accesses, without loss of security or flexibility. Specifically, KFuse merges a chain of BPF extensions into a fused BPF program by rewriting the return instructions into direct jump instructions and updating the jump offsets. For loop-based chains, KFuse updates the return values based on the original loop semantics (tail calls do not return). KFuse ensures that the fused BPF program maintains all the safety properties required by the BPF verifier. This can be proved by induction-based on the safety properties of the original BPF extension programs in the chain (ensured by the BPF verifier). Lastly, KFuse is fully transparent to user space. It achieves the transparency by maintaining the original verified BPF program and only performing the optimizations for execution. We integrate KFuse into three Linux kernel subsystems that adopt BPF: seccomp, tracepoint, and cgroup. We show that KFuse can be easily integrated in existing BPF use cases—it takes only 20, 28, and 34 lines of code changes for the three subsystems respectively.

We evaluate the effectiveness of KFuse in optimizing execution of BPF extension chains with real-world BPF use cases (the systemd BPF filters) as well as synthetic BPF extensions for scalability analysis. KFuse demonstrates 85\% performance improvement of BPF chain execution and 7\% of application performance improvement over existing BPF use cases (systemd’s seccomp BPF filters). When the chain is long (160 BPF filters), KFuse is able to improve NGINX performance by as much as 2.3 times.

This paper makes the following contributions:

- We propose a new perspective of optimizing a chain of verified BPF extensions collectively;
- We benchmark the overhead of executing BPF extension chains, based on which we identify the root causes and analyze their performance impacts;
- We present KFuse, an in-kernel framework for effectively optimizing BPF extension chains.

We will release all the code and dataset in the paper.

2 BPF Extension and The Chain Pattern

In this section, we introduce BPF extensions for the Linux kernel and identify how admitting extensions into the shared kernel creates a common chain pattern where multiple extensions are attached to the same hook.

2.1 BPF extensions

Linux supports kernel extensions implemented in the BPF (Berkeley Packet Filter) language [64]. BPF was originally used for network packet filters; later, it is adopted by many kernel subsystems, including security [14, 24], tracing [2, 3, 29], and process management [7]. Recent research has proposed a variety of BPF use cases, including storage [81],
Verified Programs Can Party: Optimizing Kernel Extensions via Post-Verification Merging

Figure 1. Common patterns of executing BPF extensions: A chain of BPF programs are store in an array or a list and are executed in loops of indirect jumps.

2.2 Emergence of BPF chain patterns

A chain of BPF extensions is formed when multiple extensions are attached to the same kernel hook point. This is typically implemented in the form of loops. Figure 1 shows three simplified examples from seccomp (Figure 1a), tracepoint (Figure 1b) and cgroup (Figure 1c). Another way to form a chain is through tail calls—a BPF program (the caller) can tail-call another BPF program (the callee), and the callee can tail-call other BPF programs.

We categorize four reasons of the chain pattern: (1) extensions are installed from different principals, (2) developers enforce modularity of kernel extensions, (3) extensions are installed at different points in time, and (4) extensions need to be verified independently.

Extensions installed by different principals. Extensions from different principals can not be combined together before being loaded to the kernel, because they may not trust or even be aware of each other. These extensions are verified independently and stored in an array of extensions, and, therefore, a chain is formed. Take seccomp-BPF as an example. Seccomp BPF filters can also be installed from different principals. In a cluster management system, host-specific seccomp BPF filters are installed [72]; the container runtime (e.g., Docker) can install container-wide BPF filters; and applications can install application-specific BPF filters. All these filters are attached at the same location (the system call entry point). In tracepoint, multiple extensions can be attached to the same tracepoint. Processes from different principals collect performance statistics by installing their own tracepoint extensions. Another example is infrared (IR) decoding—decoding protocols are expressed as BPF extensions that can come from different principals.

Modularity of extensions. Modularity is a reason why developers maintain extensions to be single-purpose and separate. A systemd developer writes in the mailing list [17] regarding the usage of seccomp filters: “Keeping the filters separate made things a lot easier and simpler to debug, and our log output and testing became much less of a black box. We know exactly what worked and what didn’t, and our test validates each filter”. Cgroup firewall rules are also easier
to maintain, debug and reuse when these rules are simple. For instance, a common way to build a firewall is to have a rule that denies all packets and rules to allow specific traffic. Each rule is implemented as a BPF extension, which can be reused and maintained for different policies. One use case of the tail call is to separate big and complicated extensions into smaller and modular ones and chain them together.

Extensions installed at different points in time. Not all extension programs are known a priori so they cannot be combined before being loaded. For example, temporal system call specialization [49, 60] is known for security benefits by installing filters at different execution phases of the application (e.g., the initialization phase and the serving phase). Recent work shows that temporal system call specialization can reduce 51% more of the attack surface, in terms of the system call interface. Seccomp ensures that a BPF filter, once installed, cannot be removed during the process lifecycle. Therefore, temporal specialization leads to multiple seccomp filters to be installed at different points in time, forming a BPF program chain. Besides seccomp, tracepoint and cgroup also allow BPF extensions to be dynamically installed at different points in time, on demand.

Extensions that need to be verified individually. A more fundamental reason of the chain pattern comes from the limited scalability of BPF verification. The verification time depends on the size and complexity of the BPF program. To make sure that a user-supplied BPF extension can be verified in time, the verifier imposes limitations on the size and complexity of the BPF extensions, including the number of instructions, the number of jumps, the stack size, etc. Therefore, to pass the verification, a large, complex BPF extension need to be split into small-sized programs in a chain.

3 The Cost of Chaining Extensions

In this section, we benchmark the overhead of executing a BPF extension chain. The main overhead comes from the indirect jumps that are invoked to switch extensions on the chain and the loops that are used to iterate extensions.

3.1 Methodology

We design the benchmark for measuring the overhead of a chain of BPF executions. A chain can be formed when multiple extensions are attached to the same kernel hook point, or when an extension tail-calls other extensions. When multiple BPF extensions are attached to the same hook point, these extensions are executed in a loop (exemplified in Figure 1). The loop iterates a program array and indirectly jumps to each item on the array. We measured call instruction as an approximation of a jump instruction (Figure 2) so the results include jumping to the target address, setting up the stack and pushing the return address on the stack.

We firstly establish the baseline and measure a chain of $N$ extensions by calling indirect jumps to 3 different functions in a loop via function pointers, as shown in Figure 2a. Then we unroll the loop, as shown in Figure 2b, to obtain the loop cost by comparing the results from Figure 2a and Figure 2b. Finally, we convert indirect jumps to direct jumps, as shown in Figure 2c to obtain the indirect jump cost by comparing the results from Figure 2b and Figure 2c. A extension that uses tail calls needs to load the target program from a program array. This operation incurs memory access cost which is also part of the cost of executing extensions by a loop. Therefore, we reuse the benchmark and present the result in Section 3.2.3.

We conduct all the benchmarking experiments on a single server with a 16 core Intel Xeon Silver 4110 CPU at 2.10GHz and 64 GB of memory. We use the rdtscp instruction to read the CPU cycles (i.e., get_timestamp). We use gcc\(^1\) to compile our benchmark programs with optimization level `-O0`.

3.2 Benchmark results

Table 1 summarizes our main results. In short, executing a function directly with unrolled loops takes 2.6 cycles; executing a function in loops with indirect jumps takes up to 46.3 cycles when retropoline is enabled, a 17.8x increase.

<table>
<thead>
<tr>
<th></th>
<th>Retpoline</th>
<th>No Retpoline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loop + Indirect jumps (Fig. 2a)</td>
<td>46.3</td>
<td>7.7</td>
</tr>
<tr>
<td>Unrolled + Indirect jumps (Fig. 2b)</td>
<td>32.6</td>
<td>2.6</td>
</tr>
<tr>
<td>Unrolled + Direct jumps (Fig. 2c)</td>
<td>2.6</td>
<td>2.6</td>
</tr>
</tbody>
</table>

Table 1. Cycles needed to call a BPF extension: Indirect jumps incur 30 additional cycles with retpoline. Loops incur 13.7 additional cycles with retpoline and 5.1 without.

\(^1\)https://gcc.gnu.org/
We also observe that the branch miss rate is merely 0.01%.

As shown in Table 1, without retpoline, the indirect jump with a function pointer takes 2.6 cycles which is as fast as a direct jump. The is due to the high accuracy (up to 98% [79]) of branch target predictors in modern CPUs. We also observe that the branch miss rate is merely 0.01%. However, with retpoline enabled, an indirect jump takes 30 extra cycles making it 12.5× slower because retpoline forces a branch prediction miss on every indirect jump and stops the CPU from speculatively executing more instructions (i.e., 100% branch miss rate). Also, retpoline itself introduces more instructions including jumping to the retpoline thunk, moving the target function on the stack and returning to the target function. In particular, jumping to the thunk takes 14.79% of the sampled cycles, moving the target function on the stack takes 8.87% and returning to the target function takes 76.33%.

**Alternative mitigations.** Recent work [16, 34, 58] proposes to use control-flow integrity (CFI) to regulate speculation targets, and to use indirect call promotion (ICP) as an optimization to avoid indirect jumps. CFI-based approaches need hardware supports to speculate only known targets for protection against Spectre-like attacks. When misspeculation happens, an expensive serializing instruction (e.g., lfence) is used.

Promoting indirect calls in the loops we target (e.g., through LLVM [28] or BOLT [4]) would produce different code. ICP needs to predict the indirect branch target and convert an indirect jump into a conditional jump i.e., a comparison and a direct jump. Its performance depends on the prediction. An indirect jump with retpoline is still taken for misprediction. On the other hand, KFuse results in one indirect jump (to the merged program) and unconditional jumps (internally stitching together that program). Our narrowed scope (focusing on extensions rather than more general techniques) enables an approach with unconditional jumps.

### 3.2.2 Loop cost. A loop for executing a BPF extension chain includes the following operations. An index counter represents the index in the current iteration. The counter needs to be incremented by 1 to find the next program at the end of each iteration. Within each iteration, the index counter is used to retrieve the corresponding element in the array. Components of a loop result in more instructions and branches compared to the unrolled loop. To measure the cost of the loop, we manually unroll the loop in Figure 2a and call functions via function pointers as shown in Figure 2b. We compare the two cases to obtain the extra cycles per iteration that the loop contributes which are 13.7 (29.6% of the extension call) with retpoline and 5.1 (66% of the

| void func1(); | void func1(); | void func1(); |
| void func2(); | void func2(); | void func2(); |
| void func3(); | void func3(); | void func3(); |
| void (*fp1)() = func1; | void (*fp1)() = func1; | void (*fp1)() = func1; |
| start = get_timestamp(); | start = get_timestamp(); | start = get_timestamp(); |
| for(int i = 0; i < 3; i++) funcs[i](); | (*fp1)(); (*fp2)(); (*fp3)(); | (*fp1)(); (*fp2)(); (*fp3)(); |
| end = get_timestamp(); | end = get_timestamp(); | end = get_timestamp(); |

(a) Indirect jumps in a loop (b) Indirect jumps with an unrolled loop (c) Direct jumps with an unrolled loop

**Listing 1. Retpoline assembly**

```c
void func1();
void func2();
void func3();
void (*fp1)() = func1;
start = get_timestamp();
for(int i = 0; i < 3; i++) funcs[i]();
end = get_timestamp();
```

```c
1. call retpoline_call_target
2. ...
3. lfence /* stop speculation */
4. jmp 0b
5. retpoline_call_target:
6. lea 8(%rsp), %rsp
7. ret
```

**Figure 2. Benchmarks for analyzing the cost of executing a chain of BPF extensions:** (a) measures the cost of loops and indirect jumps. (b) measures the cost of indirect jumps. (c) measures the cost of direct jumps. Comparing (a) and (b) to obtain the cost of loops. Comparing (b) and (c) to obtain the cost of indirect jumps.

1. call retpoline_call_target
2. ...
3. lfence /* stop speculation */
4. jmp 0b
5. retpoline_call_target:
6. lea 8(%rsp), %rsp
7. ret

The loop in the middle (line 2–4) is never executed, and the purpose of it is to fill the instruction pipeline with dummy values to prevent the CPU from speculating the actual target address, effectively nullifying the substantial performance benefits of speculation along with the risks.

As shown in Table 1, without retpoline, the indirect jump with a function pointer takes 2.6 cycles which is as fast as direct calls in Figure 2c. The is due to the high accuracy (up to 98% [79]) of branch target predictors in modern CPUs. We also observe that the branch miss rate is merely 0.01%. However, with retpoline enabled, an indirect jump takes 30 extra cycles making it 12.5× slower because retpoline forces a branch prediction miss on every indirect jump and stops the CPU from speculatively executing more instructions (i.e., 100% branch miss rate). Also, retpoline itself introduces
extension overheads for executing BPF extension chains. Note that retopoline stalls CPU pipelines and magnifies the cost.

We further investigate and break down the cost of a loop. The cost of a loop per iteration includes four parts, (1) checking if the loop condition holds to determine if the loop is over (usually on the index is less than the array size), (2) incrementing the index counter, (3) loading the extension pointer from the array based on the index counter, and (4) jumping to the beginning of the loop. We use perf to profile the loop by CPU sampling. Checking the loop condition takes 28.91% of the sampled cycles. 51.76% of the cycles are spent incrementing the index counter stored on the stack. Loading extensions from the array takes 10.99% of the sampled cycles. Finally, 8.34% of the cycles are spent jumping back to the beginning of the loop. The common approach to mitigate the cost is to unroll the loop. However, this obvious solution of loop unrolling can not be done, if the execution of BPF extensions are tightly coupled to the verification and admission of them.

3.2.3 Memory access cost. To use tail calls, a program needs to firstly look up the target program from a program array and jump to the target. Looking up the target program is essentially the same operation as one of the loop operations – loading the extension from the array based on the index counter (Section 3.2.2). These two operations both read the target program address from an array and therefore, need memory access. According to our measurement of loops, we find that the memory access cost can take up 11% of the loop cost which is 1.5 cycles with retopoline.

3.3 Discussions and Implications
The chain overhead can be significant when the BPF extensions are placed at a performance-critical path. For instance, 54% performance regression is reported for XDP packet filtering when there are only two indirect function calls (jumps) per packet [32]. When the chain is long, the overhead will be accumulated and impact application performance. Some real-world use cases have much longer BPF extension chains than the three-extension chain used in our measurement. For example, systemd [26] installs tens of seccomp BPF filters for different system services, which has significant performance overhead (as shown in Section 5.1.1).

The chain overheads come from the fact that BPF extensions are individually verified and loaded into the kernel. We argue that the execution of these extensions should not be constrained by how they are verified and loaded. Instead, the execution of the chained BPF programs can be collectively optimized in the kernel. Specifically, we can see that replacing indirect jumps with direct jumps, unrolling loops, and saving memory accesses, and each leads to in remarkable performance improvements (Table 1). Therefore, in KFuse, we start from supporting these three optimizations.

4 KFuse Design and Implementation
4.1 Overview
KFuse is a in-kernel framework for collectively optimizing verified BPF extensions that are installed at the same hooking point, in the pattern of chains. In essence, it separates the concerns of verification and execution of BPF extensions. BPF extensions that are installed by different principals or at different points of time are verified independently to ensure safety. KFuse is invoked after the verification to optimize the execution performance.

Currently, KFuse supports three generic optimizations: (1) converting indirect jumps into direct jumps, (2) unrolling loops, and (3) saving memory accesses by rewriting BPF tail calls. It can be easily extended to incorporate other optimizations, such as domain-specific ones.

A BPF extension is loaded from user space, verified by the kernel, and optionally just-in-time (JIT) compiled to native machine code. Then, the target kernel subsystems (e.g., seccomp, tracepoint, and group) can install it by attaching it to the corresponding hook points. KFuse can be invoked dynamically with the input of a chain of BPF extensions; it generates one optimized BPF extension that “fuses” the original BPF extension chain and replaces it with the fused extension at the hook point. Currently, we invoke KFuse after a new BPF extension is loaded and verified. KFuse fuses BPF programs at the BPF-instruction level, rather than native machine code—the fused BPF extension can still be (optionally) JITed.

Listing 2 shows how KFuse is invoked in seccomp. It invokes the currently-attached BPF filter (current_filter) with the newly-admitted BPF filter (prepared). The current filter can be a fused BPF program.

```
1. prepared = seccomp_prepare_user_filter(filter);
2. if (current_filter != NULL)
3.   merge_bpf_procs(current_filter->prog,
4.                     prepared->prog, SECCOMP_RETURN_POLICY);
```

Figure 3 further illustrates the process, where Figure 3a illustrates the original BPF chains at different hook points and Figure 3b illustrates the fused BPF extension when KFuse is used. KFuse first fused b, c, and d at the hook point H2 into an optimized program, bcd, and f and g at H3 into fg. When a new BPF extension E is loaded and hooked at H3, it fuses fg and E into fgE.

The fused program maintains the safety properties enforced by the BPF verifier, which can be proved by induction-based on the safety properties of the original BPF program (ensured by the verifier). On the other hand, the fused program is not limited by the size or complexity constraints required by the BPF verifier to bound the verification time.
KFuse is an in-kernel mechanism that are transparent to user space. It preserves the original extension structure and creates the illusion of a chain of multiple BPF extensions. BPF extensions are loaded and installed in exactly the same way and no user-space change is needed.

Why not fuse at user space? Fusing the BPF extensions at user space is not an option. The fused BPF extension could be difficult for the existing BPF verifier to verify due to the increased complexity. Furthermore, certain BPF use cases (e.g., seccomp) do not allow dynamic updating installed BPF extensions for security reasons.

### 4.2 Optimizing a loop of BPF extensions

As discussed in Section 3, using a loop to iteratively execute loaded BPF extensions is the most common implementation pattern to support BPF extension chains. KFuse optimizes such patterns by merging the BPF extension chain into one large BPF extension by unrolling the loop and converting indirect jumps into direct jumps.

Conceptually, such an optimization is as simple as concatenating two BPF extensions. Two important issues need to be taken care of: return instructions and jump offsets.

#### 4.2.1 Rewriting return instructions into direct jumps

A return instruction terminates the execution of a BPF program. To merge two BPF programs A and B into A+B, KFuse needs to automatically rewrite the return instructions in A into two conceptual instructions, one “jump” instruction destined to the first instruction in B and one “update” instruction for updating an aggregated return value. The “jump” instruction can be implemented using a direct jump. The “update” instruction may need multiple instructions to implement.

KFuse locates every return instruction in the BPF program and replaces it with a direct jump instruction destined to the start of the next BPF program. If the BPF program is the last one in the chain, KFuse replaces the return instructions with direct jumps to a global return instruction (which is the last instruction of the merged program, created by KFuse).

Figure 4 shows an example of merging two BPF programs for seccomp, FILTER_1 and FILTER_2. KFuse replaces each return instruction in FILTER_1 with a direct jump instruction destined to FILTER2_START, and replaces each return instruction in FILTER_2 with a direct jump destined to END.

#### 4.2.2 Updating return values

Updating return values needs to follow the loop semantics to ensure the fused program has the same semantics as the original loop. Different applications aggregate the return value of individual BPF programs in the loop differently, as shown in Figure 1. We refer to the aggregation semantics as a return value policy.

Currently, KFuse supports all the loop semantics of BPF use cases in Linux, including returning the minimal value (seccomp, Figure 1a), returning the bitwise-AND of the values (tracepoint and cgroup, Figures 1b and 1c), and returning the first non-zero values (tc and lsm-bpf). We discuss the return value policies of seccomp, tracepoint, and cgroup:

- **Seccomp.** A seccomp filter returns an action such as allowing the system call or killing the process. The more restrictive the action is, the smaller the return value is. Consequently, we implemented a return value policy that initializes the return value to the maximum integer value and only updates the new return value only if it is smaller than the current value to match the semantics of seccomp (Figure 1a).

- **Tracepoint.** The final return value of tracepoint BPF programs is calculated by performing an AND operation on each return value of the individual programs. The return value policy is implemented by initializing the return value to 1 and updating it via an AND operation on the current and new values.

- **Cgroup.** When a network packet arrives to a process in a cgroup, programs attached to the cgroup will be executed to determine the action for the packet based on the return values of BPF programs. The final value is also calculated by an AND operation like tracepoint. Therefore, the return value policy is the same as the one for tracepoint (initializing the return value to 1 and updating it via an AND operation on the current and new values).

KFuse initializes the global return value stored in temporal storage, and we use BPF_REG_AX in our implementation. It updates the return value at the locations of the original return instructions based on the return value policy (which is an input of KFuse). The global return value will be returned by the fused BPF program. In Figure 4, the return values are updated based on seccomp’s return value policy that returns the smallest value (the most restrictive policy).
A BPF program can tail-call other BPF programs to form a bounded call loop by tail-calling back to the caller (e.g., Prog B, C and D in Figure 5a). BPF verifier imposes a constraint on the number of tail calls (MAX_TAIL_CALL_CNT) to ensure that the program does not execute infinitely by a call loop. KFuse eliminates the need of a memory access that reads the jump target from the extension array by merging the target extension into the caller. Moreover, KFuse converts the indirect jumps to direct jumps if any. Note that this optimization is done independently from JIT, and is before JIT if it is enabled. Therefore, it is available to architectures without a JIT compiler implementation.

BPF programs can use tail calls to form a bounded call loop by tail-calling back to the caller (e.g., Prog B, C and D in Figure 5a). BPF verifier imposes a constraint on the number of tail calls (MAX_TAIL_CALL_CNT) to ensure that the program does not execute infinitely by a call loop. KFuse eliminates the need of a memory access that reads the jump target from the extension array by merging the target extension into the caller. Moreover, KFuse converts the indirect jumps to direct jumps if any. Note that this optimization is done independently from JIT, and is before JIT if it is enabled. Therefore, it is available to architectures without a JIT compiler implementation.

KFuse merges two seccomp BPF filters by rewriting the return instructions and updating jump offsets. Dotted boxes: the actual instructions of the conceptual instruction.

4.2.3 Recalculating jump offsets. Jump offsets need to be recalculated based on return instructions because replacing one single return instruction with multiple instructions changes the extension size and misaligns the original jump offsets. KFuse calculates the new offset by counting the number of return instructions between a jump and its destination. For instance, at line 3 in Figure 4, the jump offset is updated to 5 from 3 because there is a return instruction between the jump and its destination in the original program, assuming that updating the return value takes 2 additional instructions.

4.3 Optimizing tail calls

A BPF program can tail-call other BPF programs to form a chain (e.g., Prog A, B, C, D and E in Figure 5a). To do so, the caller first reads the address of callee from a program array and then performs an indirect jump to the target address. Given the high cost of indirect jumps with retpoline, the JIT compiler, if enabled, will try to optimize the tail call by rewriting the indirect jumps into direct jumps if possible. Note that it is not always possible, e.g., on x86, the target address of the direct jump instruction is stored in a 32-bit signed integer so jumping further than a 32-bit signed integer can only be indirect. Also, this requires to enable JIT, which is not always available or secure [68].
4.4 Transparency to user space

KFuse is fully transparent to user space, by maintaining the original BPF extension structures. It decouples the execution from other operations and only adds a pointer to the fused extensions, as illustrated in Figure 6. Therefore, user space APIs, such as looking up, dumping or removing an extension, are not affected. From the user space view, these BPF extensions are still maintained as a chain but only the fused extension is executed when the chain is invoked. On the other hand, the fused extension needs to be regenerated whenever the chain is updated.

KFuse also supports sharing of BPF extensions, when a BPF extension needs to be attached to different chains (e.g., Prog C in Figure 6). For example, seccomp BPF filters can be shared by different processes. KFuse merges the shared extension to generate different fused extensions. In our experience, BPF extensions are small in size and the additional memory usage is not concerned, as measured in Section 5.6.

4.5 Maintaining Safety Properties

Since kernel extension programs are provided from the less privileged user space, the verifier imposes various constraints on them (Section 2.1). The BPF verifier ensures the safety properties listed in Table 2 hold, including that programs are directed acyclic graphs (DAG), that programs have no loops/unreachable code and that programs have no invalid memory accesses or jumps. These properties ensure the safety of the kernel when loading user-supplied BPF programs. The verifier imposes a number of constraints to satisfy these properties, also shown in Table 2.

### 4.5.1 Satisfying safety properties

The fused BPF program maintains all safety properties enforced by the verifier on every original BPF program. This can be proven by induction. We sketch the ideas below.

- **Directed acyclic graph (DAG).** Instructions of a BPF program can only be a DAG (no backward jumps or unbounded loops). The DAG property guarantees the termination of the program and also ease the detection of unreachable code. KFuse merges two programs by connecting the exits of one program to the start of the other. The fused BPF program is proven to be a DAG—a graph is a DAG iff it can be topologically ordered; the fused program can be topologically ordered by appending the topological order of the second program to the topological order of the first program.

- **No unreachable code** A BPF program cannot contain unreachable code—it needs to be a connected graph.

- **No invalid memory access or malformed jumps** A BPF program cannot access invalid memory region (e.g., memory beyond the stack) or include malformed jumps (e.g., backward or out-of-bound jumps). Every program is individually verified before being merged by KFuse. KFuse does not change its memory access and only adds jumps to well-specified destination.

### 4.5.2 Ignoring constraints for bounding verification time

KFuse is not limited to the constraints imposed by the BPF verifier, if the constraints are solely purposed for bounding the verification time. This is because each individual BPF program is already verified and the fused BPF program does not need to be verified again. The safety properties are maintained by construction.

Table 2 shows those constraints that are used by the verifier. Specifically, BPF_COMPLEXITY_LIMIT_INSNS and BPF_COMPLEXITY_LIMIT_JMP_SEQ are used to bound the program size and complexity; MAX_USED_MAPS is used to limit

<table>
<thead>
<tr>
<th>Property</th>
<th>Purpose</th>
<th>Maintained</th>
</tr>
</thead>
<tbody>
<tr>
<td>Directed acyclic graph (DAG)</td>
<td>Safety</td>
<td>✓</td>
</tr>
<tr>
<td>No unreachable code</td>
<td>Safety</td>
<td>✓</td>
</tr>
<tr>
<td>No invalid memory access</td>
<td>Safety</td>
<td>✓</td>
</tr>
<tr>
<td>No invalid jump</td>
<td>Safety</td>
<td>✓</td>
</tr>
<tr>
<td>MAX_BPF_STACK</td>
<td>Size</td>
<td>✓</td>
</tr>
<tr>
<td>MAX_CALL_FRAMES</td>
<td>Size</td>
<td>✓</td>
</tr>
<tr>
<td>MAX_TAIL_CALL_CNT</td>
<td>Complexity</td>
<td>✓</td>
</tr>
<tr>
<td>BPF_COMPLEXITY_LIMIT_JMP_SEQ</td>
<td>Complexity</td>
<td>✓</td>
</tr>
<tr>
<td>BPF_COMPLEXITY_LIMIT_INSNS</td>
<td>Complexity</td>
<td>✗</td>
</tr>
<tr>
<td>MAX_USED_MAPS</td>
<td>Size</td>
<td>✗</td>
</tr>
<tr>
<td>BPF_MAX_SUBPROGS</td>
<td>Complexity</td>
<td>✗</td>
</tr>
</tbody>
</table>

Table 2. Properties and constraints imposed the verifier. KFuse preserves all the safety properties.
the number of maps, and BPF_MAX_SUBPROGS is used to limit the subprograms in a BPF program.

4.5.3 Domain-specific constraints. Some kernel subsystems also impose domain-specific constraints. For example, cgroups and tracepoints both limit the number of BPF programs that can be attached to the same hook point by BPF_CGROUP_MAX_PROGS and BPF_TRACE_MAX_PROGS. These constraints are agnostic to KFuse and KFuse always respect them. This is because the constraints are placed on the original program structure (Section 4.4). If the constraints are not satisfied, KFuse will not be invoked. We note that these constraints are not for safety but for performance purposes as they bound the total execution time.

4.6 Implementation

KFuse is implemented in around 500 lines of C code, on top of the Linux kernel (v5.9). As discussed in Section 4.1, the main interface is merge_bpf_progs(), which takes three input parameters, the pointers of the original BPF programs, and a policy flag that specifies the return value policy, and returns a pointer of the fused program, as shown below.

1. struct bpf_prog * merge_bpf_progs(
2. struct bpf_prog * old_prog,
3. struct bpf_prog * new_prog,
4. enum return_policy);

KFuse uses BPF_REG_AX (a special register reserved for the kernel) to hold the temporal return value, which is updated according to the return value policy. The merged program has two-instruction prologue that initializes the return value. Each return instruction is replaced with 5 instructions for the Seccomp return policy and 3 instructions for the tracepoint and cgroup return policy. Finally, the merged program is appended with a two-instruction epilogue which loads the return value from BPF_REG_AX to BPF_REG_A and returns.

Integrating KFuse. Kernel subsystems can choose to integrate KFuse. The integration is straightforward with a few lines of code. We integrate KFuse in three kernel subsystems, including seccomp, tracepoint, and cgroup. We apply KFuse to seccomp with 20 lines of code (LoC), tracepoint with 28 LoC and cgroup with 34 LoC. We implement a syscall configuration to allow users to enable/disable KFuse.

5 Evaluation

We measure the effectiveness of KFuse in optimizing performance, as well as its overhead in terms of the time it takes for optimization and additional memory cost. To measure performance, we run performance benchmarks using the three Linux subsystems that integrates KFuse (seccomp, tracepoint and cgroup). All experiments are run on a single server with a 16 core Intel Xeon Silver 4110 CPU at 2.10GHz and 64 GB of memory. We run the experiments in KVM virtual machines. Each VM is configured with 2 VCPUs and 8 GB memory, and the kernel is compiled with the microVM configuration. We use wrk [15] and memtier_benchmark [18] as benchmark clients for NGINX and Redis. We also evaluate the cases where the kernels are compiled without retpoline, which gives a lower bound of the KFuse effectiveness.

Overall, KFuse effectively improves the performance for real-world workloads (7% throughput increase for Redis deployed by systemd). The performance benefits of KFuse grows linearly along with the length of chains.

5.1 Seccomp

We measure the effectiveness of KFuse in improving the syscall call performance under seccomp BPF filters. We conduct three experiments:

- We use KFuse to optimize the seccomp BPF filters launched by systemd which provides application sandboxing using seccomp. We run Redis and NGINX as real-world applications to assess the benefit of using KFuse on existing applications and BPF use cases.
- We conduct a scaling analysis on both system call performance and application performance, with increasing numbers of seccomp BPF filters.
- We present the benefit of KFuse for incremental tightening which installs seccomp filters during application execution to tighten the system call policy. Incremental tightening is meant to be one of the new BPF use cases that lead to a long chain of BPF extensions and motivate the need to reduce the cost.

5.1.1 Systemd BPF filters. Systemd provides a sandboxing feature that restricts the system calls the sandboxed applications can invoke. We benchmark Redis launched by systemd. Redis is shipped on Ubuntu with a systemd configuration that installs 19 seccomp BPF filters sandboxing the service including loading kernel modules, gaining new privileges and modifying control groups. Systemd also installs filters for supporting 32-bit systems (i.e., i386). These filters are sized from 6 instructions to 58 instructions with an average 19.6 instructions. Figure 8a shows the throughput of Redis, with the original 19 filters and with the fused filter by KFuse. The results are normalized by the Redis throughput with no BPF filter. We can see that the 19 filters installed by systemd lead to non-negligible overhead (∼ 10% throughput decrease). KFuse can significantly reduce the overhead (with only 3%) by speeding up the end-to-end BPF execution time from 957 to 148 nanoseconds (reducing 85%) with the same security policy of original filters.

5.1.2 Scaling analysis. We conduct a scaling analysis to measure the performance overhead of different number of seccomp filters and the benefit of using KFuse. We create a simple BPF filter with 4 instructions: checking if the system call number equals 450, if true, kill the process. Otherwise,
Verified Programs Can Party: Optimizing Kernel Extensions via Post-Verification Merging

Verified Programs Can Party: Optimizing Kernel Extensions via Post-Verification Merging

Figure 8. (a) The throughput of Redis with 19 seccomp filters installed by systemd. (b)(c) The latency of getpid and NGINX with different numbers of seccomp filters installed for scaling analysis. The results are normalized to baseline, 0 seccomp filters and retpoline enabled.

Table: Redis throughput and CPU cycles

<table>
<thead>
<tr>
<th>Filters</th>
<th>Redis Throughput</th>
<th>CPU Cycles</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.95</td>
<td>120</td>
</tr>
<tr>
<td>19</td>
<td>0.95</td>
<td>100</td>
</tr>
<tr>
<td>19 merged</td>
<td>0.95</td>
<td>80</td>
</tr>
</tbody>
</table>

Figure 9. Incremental tightening blocks system calls during program execution. This simplified program only uses 4 system calls once, open, read, write, and close. The system call is blocked when it will not be again.

We generate the seccomp BPF filters for different phases of NGINX and Memcached by dynamically tracing their runtime behavior continuously. There are non-deterministic system calls such as those related to timer or signals. We always allow non-deterministic system calls and only incrementally tighten deterministic system calls. We modify the C library to load the profile and install a seccomp filter when a system call will no longer be needed.

We install 18 seccomp filters for NGINX and 22 for Memcached. Figure 11b shows the number of allowed system calls over time for each application. Figure 11a shows that the application performance drops 7%–16% due to the chain of filters. By merging these filters, KFuse eliminates the overheads of calling extensions by 45%–66%.

5.2 Tracepoint

Tracepoint enables live kernel debugging and profiling. Multiple BPF programs can be attached to a single tracepoint and all of them are executed when an event is triggered. To experiment with realistic workloads, we use a tracepoint BPF program that not only performs arithmetic instructions but also write data to a BPF map. The program counts the number of times that a system call, getpid, is called and write the counter to a map. Figure 10a shows the latency of getpid with various numbers of programs attached. With 20 programs, KFuse can speed up latency by 1.71× with retpoline and 1.12× without. With 160 programs, KFuse can speed up latency by 5.19× with retpoline and 1.82× without. In the average case of 80 programs, KFuse can speed up latency by 4.09× with retpoline and 1.7× without.

5.3 Cgroup

Cgroup provides the mechanism for users to attach multiple BPF programs to implement firewall rules. We attach multiple BPF programs with type BPF_CGROUP_INET_INGRESS to experiment with the scenario where multiple firewall rules are applied. The BPF programs simply permit the packet for performance evaluation. We use iPerf3 [10] to measure the

We generate the seccomp BPF filters for different phases of NGINX and Memcached by dynamically tracing their runtime behavior continuously. There are non-deterministic system calls such as those related to timer or signals. We always allow non-deterministic system calls and only incrementally tighten deterministic system calls. We modify the C library to load the profile and install a seccomp filter when a system call will no longer be needed.

We install 18 seccomp filters for NGINX and 22 for Memcached. Figure 11b shows the number of allowed system calls over time for each application. Figure 11a shows that the application performance drops 7%–16% due to the chain of filters. By merging these filters, KFuse eliminates the overheads of calling extensions by 45%–66%.

5.2 Tracepoint

Tracepoint enables live kernel debugging and profiling. Multiple BPF programs can be attached to a single tracepoint and all of them are executed when an event is triggered. To experiment with realistic workloads, we use a tracepoint BPF program that not only performs arithmetic instructions but also write data to a BPF map. The program counts the number of times that a system call, getpid, is called and write the counter to a map. Figure 10a shows the latency of getpid with various numbers of programs attached. With 20 programs, KFuse can speed up latency by 1.71× with retpoline and 1.12× without. With 160 programs, KFuse can speed up latency by 5.19× with retpoline and 1.82× without. In the average case of 80 programs, KFuse can speed up latency by 4.09× with retpoline and 1.7× without.

5.3 Cgroup

Cgroup provides the mechanism for users to attach multiple BPF programs to implement firewall rules. We attach multiple BPF programs with type BPF_CGROUP_INET_INGRESS to experiment with the scenario where multiple firewall rules are applied. The BPF programs simply permit the packet for performance evaluation. We use iPerf3 [10] to measure the
TCP bandwidth with various numbers of programs attached. When attaching less than 60 programs, we do not observe obvious performance enhancements. With 60 programs, KFuse improves the bandwidth by 10% with retpoline and 3% without. With 160 programs, KFuse improves the bandwidth by 31% with retpoline and 11% without.

5.4 Tail call
To measure KFuse’s saving of tail call rewriting, we use the benchmark used by Joly et al. [52] for evaluating tail call overhead, in which a BPF program tail-calls various numbers of other BPF programs and forms chains with different lengths, from 1 to 32. We verify that these tail calls are optimized to direct calls and thus are not slowed down by retpoline. We also evaluate the case when tail calls are not optimized and are indirect jumps. Figure 12 shows the average of 50 iterations in three cases. When tail calls are optimized to direct jumps, KFuse saves memory access cost. The average speed-up (16 tail calls) is 1.4× with the minimum (1 tail call) and maximum (32 tail calls) being 1.1× and 1.7×. When tail calls are indirect jumps, KFuse saves both memory access and indirect jump cost. The average speed-up (16 tail calls) is 16.1× with the minimum (1 tail call) being 2× and maximum (32 tail calls) being 32.88×.

5.5 Cost of updating a chain
We measure the time KFuse takes to fuse a BPF program chain into an optimized BPF program. KFuse works dynamically by fusing the existing program (the merged chain) with the new program. Specifically, KFuse needs to combine the new program with existing one, reallocate memory for the merged program and generate JITed image for the merged program so the time to merge depends on the length of chain (the size of existing program) and the new program. We measure the time by loading up to 160 programs. As shown in Figure 10c, the latency starts from ≈ 27 microseconds when there are only 1 program on the chain and grows to ≈ 138 microseconds when there are 160. When there are 80 programs on the chain, the merging time is ≈ 85 microseconds.

5.6 Memory usage
KFuse needs to keep the original BPF bytecode to support sharing of BPF programs and maintain transparency to the user space (Section 4.4). This could lead to additional memory overhead. To quantify the overhead, we measure the
size of 366 real-world BPF programs from six projects collected by [48] (including ovs, linux, prototype-kernel, suricata, bpf_cilium_test and cilium). The average BPF extension size is merely 4.3 KB, with the median and the 90-th percentile being 0.9 KB and 13.8 KB, respectively. Given the abundant memory of modern computers, the additional memory usage is acceptable.

6 Related work
BPF extensions have achieved phenomenal adoptions in recent years, including security [36, 40, 44, 45], storage [30, 59, 70, 76], profiling [2, 13, 42] and networking [11, 31, 37, 38, 43, 51, 63, 69, 71, 74, 77, 78]. Recent advances on safe BPF infrastructures and implementations [33, 47, 48, 55, 56, 61, 65, 66, 75, 78] further eliminate the barrier of adoptions. KFuse respects the safety properties ensured by BPF verifiers, while optimizing the performance of BPF extension chains.

A main optimization of KFuse is to convert indirect jumps into direct jumps. The costs of indirect jumps come from the mitigation of Spectre [57], a recent class of hardware vulnerabilities, which degrades the efficiency of conditional branches. In addition to retopine [23], recent research proposals for Spectre mitigation includes control-flow hardening [46, 58], page table separation [39] or cache isolation [54] to remove the side channels. Recently, Intel released indirect branch restricted speculation (IBRS) [25] and enhanced IBRS [25], new hardware features to prevent less-privileged processes from controlling the branch target predictor in a more-privileged mode. However, these features either incur huge performance overheads or are only available on new CPUs. We expect the overhead of indirect jumps to remain high in the near future, because free speculation is no longer an option [41, 67].

A few efforts have been made to reduce the overhead of indirect jumps. JumpSwitch [34] optimizes retopine by indirect call promotions (ICP), which depends on an effective target predictor. JumpSwitch also requires dynamic kernel patching. However, JumpSwitch has not been adopted by the Linux kernel, due to its frequent kernel patching that breaks instruction decoding [22]. KFuse does not bear those limitations as it only acts on BPF extensions.

A few Linux patches [1, 27] try to convert unnecessary indirect jumps to direct jumps to improve XDP performance [32]. The BPF JIT compiler optimizes tail calls to direct jumps [19] if the target address is within an offset that can be expressed as a 32-bit integer. BPF trampoline [5, 9] is a mechanism to directly call BPF programs from kernel by a dynamically generated code image which needs to be regenerated when an extension program is added or removed. Optimizations such as trampoline [9] or converting tail calls to direct jumps [19] require BPF programs to be JIT-compiled, which may not be available. Differently, KFuse focuses on eliminating the indirect jumps across the chained BPF extensions. As we shown in the paper, the overhead of chained extension can be significant as the BPF programs are typically small in size. KFuse does not rely on JIT compilation.

7 Conclusion and Discussion
We demonstrate that multiple BPF programs in a chain can be collectively optimized after each of them is verified. Such optimization can lead to remarkable performance gains. Specifically, when the BPF programs are small in size and the chain is long, the overhead of indirect jumps and loops can be as high as 85% of the end-to-end execution time, leading a 7% application performance loss. We have already seen long chains of BPF extensions in real deployments (e.g., 19 Sec-comp BPF filters deployed by systemd) and we expect such long chains to be even more common in future use cases, given the explosion of BPF use cases in recent years and the need of fine-grained, domain-specific policies.

We believe that more advanced cross-optimizations for verified BPF programs (such as application-specific optimizations) can be implemented, beyond the basic optimizations currently supported by KFuse. For example, redundant code (e.g., checks [80]) across the BPF extension programs can be removed, in the scope of the entire BPF chain, to further improve the performance; BPF instructions can be reordered across the BPF extension chain (e.g., by placing hot code early in the fused BPF program to save subsequent computation). Those optimizations can be supported in the current KFuse framework. Further, multiple BPF programs at different hooking points are often orchestrated to collectively implement a feature (e.g., socket filters hooked at cgroup and traffic control [6]). Those BPF programs can potentially be optimized together, while maintaining the safety properties.

Acknowledgements
We thank the anonymous reviewers of EuroSys’22 and our shepherd, Baptiste Lepers, for the valuable feedback. We thank Neil Zhao for the discussion on speculative attacks as well as hardware and software defenses. Mohan’s group is supported in part by Office of Naval Research (ONR) grant N00014-17-1-2889. Xu’s group is supported in part by NSF CNS-1956007, CCF-2029049, CNS-2130560, CNS-2145295, and CCF-1816615.

References


[20] [PATCH net-next 8/9] net: filter: rework/optimize internal BPF inter- 


[34] Luca Deri, Samuele Sabella, and Simone Mainardi. Combining System Support for Containers.


[43] Nadav Amit and Michael Wei. The Design and Implementation of Hypercal...


A Artifact Appendix

A.1 Abstract

This document provides step-by-step instructions to reproduce KFuse’s experiment results including seccomp, tracepoint, cgroup, and tail-call. All experiments are running in QEMU virtual machines. We provide compiled kernel binaries with KFuse enabled and file system images (20GB for each) with all benchmarks installed. We also provide scripts to automate the process.

A.2 Description & Requirements

A.2.1 How to access. The artifact is available at https://doi.org/10.5281/zenodo.6360914.

A.2.2 Hardware dependencies. The experiments require an amd64 machine with virtualization capability because they are run in a KVM virtual machine. We run the experiments on a server with 16 core Intel Xeon Silver 4110 CPU at 2.10GHz and 64 GB of memory. In addition, 60GB disk space is required for file system images and kernel binaries.

A.2.3 Software dependencies. We attach a script that install software dependencies. scripts/install_deps.sh installs necessary software packages for our experiments on Ubuntu 20.10. Furthermore, this artifact needs a native Ubuntu machine that supports KVM. That is, installing Ubuntu on a VM on a Mac will not work.

A.2.4 Benchmarks. None

A.3 Set-up

Run scripts/install_deps.sh in the artifact directory to install software dependencies.

A.4 Evaluation workflow

A.4.1 Major Claims.

- (C1): KFuse can significantly reduce the overhead caused by seccomp BPF filters installed by systemd. This is shown by the experiment (E1) described in Section 5.1.1 whose results are illustrated in Figure 8 (a).
- (C2): KFuse can improve the latency of system calls by merging their seccomp BPF filters. This is shown by the experiment (E2) described in Section 5.1.2 whose results are illustrated in Figure 8 (b)/(c).
- (C3): KFuse can improve the performance of tracepoint by merging its BPF filters. This is shown by the experiment (E3) described in Section 5.2 whose results are illustrated in Figure 10 (a).
- (C4): KFuse can improve the performance of iPerf3 by merging cgroup BPF filters. This is shown by the experiment (E4) described in Section 5.3 whose results are illustrated in Figure 10 (b).
- (C5): KFuse can improve the performance of our tail-call prototype benchmark. This is shown by the experiment (E5) described in Section 5.4 whose results are illustrated in Figure 12.

A.4.2 Experiments. All the experiment instructions and expected results are in the experiment_docs directory in the artifact repository. We describe the details of the benchmarks in the following sections.

- Experiment (E1): systemd [30 human-minutes + 1 compute-hour]: This experiment uses KFuse to optimize the 19 seccomp BPF filters launched by systemd which sandboxes applications with seccomp. We run Redis to assess the benefits of using KFuse on existing applications and their BPF use cases.
- Experiment (E2): seccomp [30 human-minutes + 1 compute-hour]: We create a seccomp BPF filter with 4 instructions: checking if the system call number equals 450, if true, kill the process. Otherwise, allow the call. The system call numbered 450 does not exist, so all the system calls are allowed. This experiment installs different number of filters, from 0 to 160, with four kernel configurations (kfuse-retpoline, kfuse-norettopline, ...

---


nokfuse-retpoline, and nokfuse-noretpoline). Then, we run getpid for each configuration.

- **Experiment (E3): tracepoint [1 human-hours + 1 compute-hours]:** We design a tracepoint BPF program that performs arithmetic instructions and writes data to a BPF map. The program counts the number of times that the system call, getpid, is called and writes the counter to a map. This experiment will install different number of this filter, from 0 to 160, with four kernel configurations (kfuse-retpoline, kfuse-noretpoline, nokfuse-retpoline, and nokfuse-noretpoline). Then, we call getpid and measure the CPU cycle for each configuration.

- **Experiment (E4): cgroup [1 human-hours + 1 compute-hours]:** We creates a BPF program that allows every packet. This experiment attaches this filter with type BPF_CGROUP_INET_INGRESS to define our network traffic rules. To elaborate, the experiment will install different number of this filter, from 0 to 160, with four kernel configurations (kfuse-retpoline, kfuse-noretpoline, nokfuse-retpoline, and nokfuse-noretpoline). Next, we use iPerf3 to measure the TCP bandwidth for each configuration.

- **Experiment (E5): tail-call [30 human-minutes + 2 compute-hours]:** This tail-call benchmark makes BPF programs tail-call one another BPF program and forms chains with different lengths, from 1 to 32. This experiment shows that the overhead of indirect jumps of tail calls are optimized away.

---

2https://iperf.fr/iperf-download.php