Sharma SD, Dagenais M. Enhanced userspace and in-kernel trace filtering for production systems. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 31(6): 1161–1178 Nov. 2016. DOI 10.1007/s11390-016-1690-y

Enhanced Userspace and In-Kernel Trace Filtering for Production Systems

Suchakrapani Datt Sharma, *Student Member, IEEE*, and Michel Dagenais, *Senior Member, IEEE*

Department of Computer and Software Engineering, École Polytechnique de Montréal, Montréal, H3T 1J4, Canada

E-mail: {suchakrapani.sharma, michel.dagenais}@polymtl.ca

Received September 17, 2015; revised May 22, 2016.

Abstract Trace tools like LTTng have a very low impact on the traced software as compared with traditional debuggers. However, for long runs, in resource constrained and high throughput environments, such as embedded network switching nodes and production servers, the collective tracing impact on the target software adds up considerably. The overhead is not just in terms of execution time but also in terms of the huge amount of data to be stored, processed and analyzed offline. This paper presents a novel way of dealing with such huge trace data generation by introducing a Just-In-Time (JIT) filter based tracing system, for sieving through the flood of high frequency events, and recording only those that are relevant, when a specific condition is met. With a tiny filtering cost, the user can filter out most events and focus only on the events of interest. We show that in certain scenarios, the JIT compiled filters prove to be three times more effective than similar interpreted filters. We also show that with the increasing number of filter predicates and context variables, the benefits of JIT compilation increase with some JIT compiled filters being even three times faster than their interpreted counterparts. We further present a new architecture, using our filtering system, which can enable co-operative tracing between kernel and process tracing VMs (virtual machines) that share data efficiently. We demonstrate its use through a tracing scenario where the user can dynamically specify syscall latency through the userspace tracing VM whose effect is reflected in tracing decisions made by the kernel tracing VM. We compare the data access performance on our shared memory system and show an almost 100 times improvement over traditional data sharing for co-operative tracing.

Keywords tracing, debugging, filtering, operating system, interpreter

1 Introduction

With the traditional debugging approach, it becomes quite difficult to gather very low level as well as time accurate details about the systems' behavior in quasi real-time or soft real-time systems. Sampling-based profiling tools are also of limited use in such cases. Therefore, a fast logging mechanism, called tracing, is employed. Tracing can be divided according to the functional aspect (static or dynamic) or its intended use (kernel or userspace tracing — also known as tracing domains).

Tracing usually involves adding special tracepoints in the code. A tracepoint looks like a simple function call, which can be inserted anywhere in the code (in the case of userspace applications) or be provided as part of the standard kernel tracing infrastructure (tracepoint "hooks" in the Linux kernel). Each tracepoint hit is usually associated with an event. For instance, the events in Linux kernel are very low level and occur frequently. Some examples are syscall entry/exit, scheduling calls, etc. For userspace applications, these can be any function call entry in the program. This indeed is a very efficient way to follow a program execution, rather than traditional debugging, especially in scenarios where the effect of pausing, waiting for user interaction and collecting data, can alter the behavior of a normal execution and yield incorrect results. Sometimes, the error cannot be reproduced in normal scenarios, due to the presence of time dependent errors in programs, which do not arise systematically or even frequently (for example, a heisenbug)[1]. For such

Regular Paper

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) under Grant No. CRDPJ424666-11 and the research grants from Ericsson, EfficiOS and PROMPT Québec.

cases, low overhead and low disturbance tracing tools are invaluable.

Tracing involves storing the associated data in a special buffer whenever an event occurs. For a detailed execution trace of a very fast system, with high frequency trace events, this data can be huge and contains precise time-stamps of the tracepoints hit, along with any optional event-specific information (values of variables, registers, etc.). All this information can be stored in a specific format for later retrieval and analysis. In many cases, the trace data contains a lot of uninteresting, redundant information during normal execution and needlessly consumes a lot of storage space. There can be situations where the target system is resource-constrained, such as an embedded network controller, where a huge number of trace events can be generated at very high speed for hundreds of days in a row[2-3]. It would be very inefficient to store all the traced data and try to retrieve it for offline analysis. In such situations, trace filters can be used to discard unwanted tracepoints and record only those specific ones that are of interest. The trace filters are composed of multiple filter predicates which essentially are the conditions to be checked. The predicates are joined together with Boolean operators and form a Boolean expression that returns either TRUE or FALSE. More about this will be discussed in Subsection 3.2 and Section 4.

Most tools employ some form of filtering. We observed that the filtering schemes used in most state-of-the-art tools are the same. This can be seen in Fig.1, where we generally 1) define the filter predicates in a high-level statement form, 2) create a predicate tree and possibly a more efficient bytecode representation, and 3) when the tracepoint is hit, walk the associated predicate tree while evaluating the conditions, or interpret the associated bytecode and evaluate the filter outcome. Another approach, as in Fig.1(c), is to JIT (Just-In-Time) compile the filter bytecode to native code and execute it on the machine. This yields a significant performance improvement as compared with interpreting the bytecode directly.

Some interesting prototyping results were reported by Alexei Starovoitov. An implementation of JIT compiled Berkeley Packet Filter (BPF) bytecode to kernel tracing demonstrated an improvement from 32 ns to 4 ns per call (best-case scenario[4]). These improvements are in contrast to the state-of-the-art trace filtering approaches such as those in LTTng-UST[5-6], which use bytecode generation-interpretation for evaluating filters (Fig.1(b)). We therefore improve tracing performance using JIT compiled filters by enhancing the current tracing architecture as discussed in this paper. It has proven to be more robust than the current filtered tracing techniques and has led to reduced trace storage size, and hence the efficient diagnosis of problems.

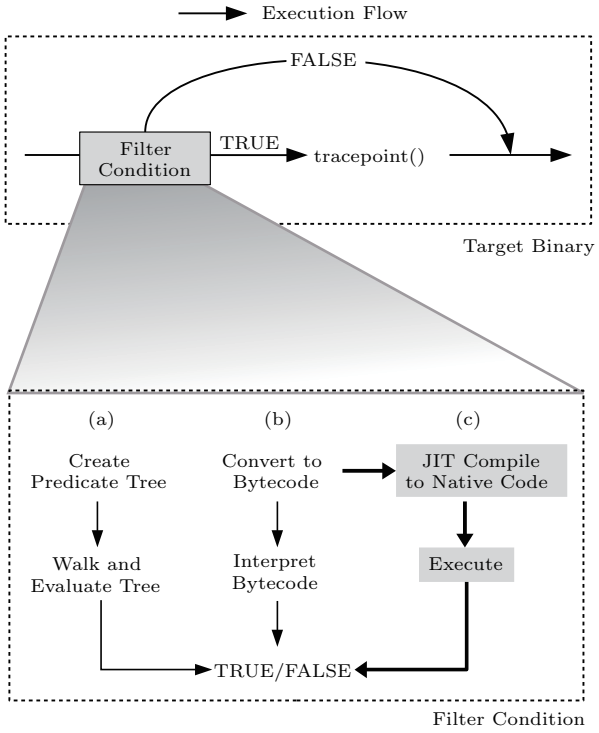

Fig.1. Overview of filtering in trace and debug context. The bold path (c) is the approach which yields minimum overhead.

The benefits of having a very fine control over tracing have always been important from the developers' perspective, but the filter computation overhead has always been a hindrance. In this paper, we present a new tracing scheme which tries to minimize this overhead and hence allows a more flexible use of the JIT technique for conditional tracing.

We also introduce the concept of co-operative tracing where, through an efficient sharing mechanism, kernel tracing can be guided from userspace or vice versa. Important data, such as performance counters or kernel-aggregated values, can be shared between the kernel and the userspace filters, to achieve co-operative tracing.

Our major contributions in this paper are as follows:

• an improved userspace tracing framework which utilizes JIT-based trace filtering, thus reducing the trace footprint and overhead;

• kernel and userspace co-operative tracing through a kernel VM — user VM direct data sharing mechanism

that enables events in userspace to affect kernel tracing and vice versa;

• a userspace library that provides an implementation of the two above contributions as well as a means of incorporating efficient filtering in generic tools.

The remainder of the paper is organized as follows. We start with a discussion on the basic building blocks of tracing. The techniques such as static and dynamic code instrumentation, on which the Linux tracing infrastructure is built, are explained. We also discuss its relevance in our context of trace filtering and our new scheme for kernel-userspace co-operative tracing. We then explain some commonly used filtering techniques used in scenarios like in-kernel network packet filtering and see how different tools like DTrace, LTTng and SystemTap approach filtering of traces. We move on to explain the design of bytecode interpreters relevant for filter design. Subsequently, JIT compilation techniques and their use in tracing are discussed. We introduce our proposed method and architecture for a JIT-based optimized trace filtering framework, its design and its benefits. Our evaluation of its performance against current interpreted filtering techniques is presented. We move on to propose our co-operative tracing system as an extension to the JIT-based trace filtering system, and to achieve high-speed kernel-userspace tracing on resource-constrained soft-realtime systems. We discuss how the current byte code based filtering systems evolve into a generic system, and the efficient data sharing mechanism we propose which yields close to 100x improvements over current data sharing mechanisms. We also expose how this architecture can be used independently, not just for conditional trace filtering, but for taking certain actions (like recording a trace, aggregating data, sharing data with userspace) based on whether conditions are met or not. We see how this architecture differs from approaches taken by tools such as DTrace and SystemTap. Finally, the results from the performance benchmarks, the inferences drawn from the results, and the directions for future work have been presented.

2 Literature Review

Most of the previous relevant work on filtering focused on network packets but not on tracing. McCanne and Jacobson[7] proposed quite early a bytecode-based virtual machine for in-kernel BSD network packet filtering, called as Berkeley Packet Filter (BPF). This interpreted technique delivered a performance of up to 20 times faster than the original tree-based designs such as those of the CMU/Stanford Packet Filter (CSPF)[8].

In Pathfinder, Bailey *et al.*[9] proposed a new way of specifying filters declaratively. This reduces the times a pattern has to be evaluated. Engler and Kaashoek presented DPF[10] where they showed an improvement of about 13x∼26x as compared with Pathfinder's implementation, due to the use of native compilation techniques, while keeping the same declarative language format of filter description.

BPF in its original format was further improved by Begel *et al.* in BPF+[11], where they performed compiler optimizations to eliminate redundant predicates during filter generation. They also eventually implemented an elementary JIT compiler for BPF to improve its performance further — similar to what DPF (Dynamic Packet Filter) had been done with the Pathfinder's implementation.

Wu *et al.* recently proposed Swift[12], a new and complete packet filtering system based on a CISC ISA, and a BPF compatible API to simplify the code specification. They showed up to 3x performance of corresponding in-kernel BPF implementations, mainly due to the aggressive use of SIMD instructions provided by i386 and x86_64.

Very recently, BPF was improved and evolved into an extended BPF (eBPF) implementation in the Linux kernel[4,13] with enhancements to register management, bytecode generation and optimizations using a modern compiler infrastructure. Its architecture was slightly modified to provide a more versatile in-kernel tiny virtual machine. This provided some of the foundation work for the userspace trace filtering and the kernel-userspace tracing improvements we propose in this paper. Earlier versions of BPF syntax have also been used for the packet filtering in Windows.

DTrace[14], originally developed for Solaris, is a purely script-driven tool which consists of a new language (D language) for defining trace scripts. The trace scripts get compiled into an intermediate format (DIF) and are subsequently executed in DTrace's own in-kernel virtual machine. It is now possible to insert probes in userspace applications but this simply generates an interrupt, and the probe handler still executes in kernel space. For sharing data between probe executions, DTrace supports global variables, thread-level variables and aggregations. Aggregations can use per-CPU buckets and can thus be incremented with low overhead, without locking, at high frequency. The actual aggregation, with heavier locking, is only needed

when extracting the aggregated value, typically at the end. Thread-level storage also avoids locking. Reentrancy could be an issue if DTrace allowed the same thread-level variable to be accessed from normal and from interrupt context. Global variables in DTrace are not lock-protected, and concurrent access can lead to corruption. Thus, although being a very elaborate, popular and convenient scripting system for tracing and monitoring, DTrace suffers from several limitations. All scripts execute from kernel space and the only userspace to kernel interaction is achieved using tracepoints in applications generating costly traps. Furthermore, DTrace suffers from scalability problems and offers limited support for sharing global variables.

SystemTap has been developed along similar lines, to gather trace data dynamically. However, for kernel tracing, SystemTap generates C code to be compiled as a kernel module and loaded dynamically. This differs from the BPF and DTrace approach of executing bytecode within the kernel[15]. While this approach, in theory, offers the best performance with native code, it suffers from the requirement of needing a full compilation environment for the target kernel at runtime. SystemTap scripts can define and use global variables. They are automatically read- or write-locked when accessed from the scripts, in case the scripts could be executing concurrently in probe handlers. This severely limits the scalability in scenarios requiring data sharing. Furthermore, while probes can now be hooked to userspace code, they generate an interrupt and the corresponding scripts execute in kernel space, just like DTrace. There is thus no provision for scripts executing in userspace and sharing data with the kernel. We discuss this further in Subsection 3.2 and Section 5, where comparisons with the newer eBPF approach are made.

3 Background

Most tracing tools are built on underlying mechanisms which deliver different performance under various scenarios. In terms of performance, the most important factor across all tools is the reduced overhead. As each tracepoint execution incurs some time, this added time can potentially slow down the normal execution of the software and yield different results. The goal is therefore to have negligible overhead, ensuring that the behavior is the same, with or without tracing. We now discuss some basic concepts and relevant techniques that many state-of-the-art tracing tools employ.

Static Instrumentation. Instrumentation in computing is the process of adding a certain code in any given application, with the inserted code snippet performing tasks related to diagnosing errors, profiling activities or gathering traces. The piece of code is intended to run fast with very little overhead. In many cases, this code can be added statically, where it is added before compilation — for example, as a small function call at the trace target function entry and exit. When compiled with this instrumentation, each call to the trace target function entry and exit will lead to the instrumentation being run. This static instrumentation can also be done at compile time, where the code can be inserted by the compiler backend. The Linux kernel provides manually inserted static trace points using the `TRACE_EVENT` macro[16]. It exposes trace hooks on which other kernel tracing systems can be built. In addition to static code instrumentation, compiled code can also be inserted into binaries on disk, without any source code being available.

Dynamic Instrumentation. The other type of instrumentation is dynamic instrumentation, sometimes also called dynamic binary instrumentation (DBI). Traditional static techniques insert code at compile time, and this inserted code is persistent. Whenever the specific function is called, the instrumentation code also runs and incurs some overhead — even when the developer does not necessarily want the instrumentation code to run. It also limits the instrumentation only to software for which the code is available for recompilation. Instrumentation can either be performed on the binary residing on disk or by attaching to running processes (for example, attaching a debugger to a running process). Dynamic instrumentation tools and frameworks can be built using either 1) TRAP-based approach such as in older Kprobes[17-18] and GDB's normal tracepoints[19], or 2) a trampoline-based approach such as in Dyninst[20] and PIN[21], or 3) a more elaborate JIT technique as used in tools like Valgrind① and PIN.

3.1 Tracing

Static tracing for LTTng, in kernel and userspace, is implemented using the static instrumentation techniques where a `tracepoint()` call may be placed anywhere in a function, and with supporting macros can generate very fast and accurate tracing data[5]. During compilation, this call gets expanded to an actual tracing

①http://valgrind.org/docs/manual/manual.html, July 2016.

function, according to the tracing context. This is the most optimum tracing mode. The Linux kernel's own tracing infrastructure, `ftrace`, provides static as well as dynamic tracing, depending on how it is used. Other tracing tools like SystemTap provide dynamic tracing through the use of Kprobes, Jprobes and Uprobes[22]. SystemTap also uses Dyninst for userspace tracing to gain some performance as well. The Kprobe approach has been used extensively to insert instrumentation code in the non-blacklisted kernel functions. These have traditionally been TRAP-based, but trampoline-based probes have also been made available recently. Dynamic tracing with LTTng is based on the kernel's Kprobe technique.

Irrespective of what technology they are built upon, activated tracepoints may generate a lot of data. This motivates the work on filters and how they can be used to filter out a large fraction of uninteresting trace data.

3.2 Filters

Filtering is widely used in computing — from filter queries supplied to SQL databases to providing sandboxed secure execution environments by filtering out syscalls[23]. The basic idea of a filter F is to find a small subset S from a large input set L. The criterion of selecting S is that the application of filter F to each element i of L returns TRUE.

$$S = \{i : i \in L, F(i)\}.$$

Here $F(i)$ can be defined as a Boolean function whose outcome depends on the filter predicates $P_1, P_2, ..., P_n$. These predicates are the heart of the filter itself and are joined with Boolean operands. In our tracing context, a filter function F with an expression E and operator $(\star)$ can be defined as follows.
For every $i \in L$, let

$$F(i) = \begin{cases} \text{TRUE}, & \text{if } E = \{P_1 \star P_2 \star ... \star P_n\} \\ & \text{is TRUE for } i, \\ \text{FALSE}, & \text{otherwise.} \end{cases}$$

In operating systems and software applications, the need for filtering is the most prominent in network packets. A lot of network traffic on the system causes packets of various protocols, sizes, having different sources and destinations, to pass through the networking device and the kernel. A user would specify its needs in the form of a Boolean expression. There are multiple approaches for evaluating the filter expressions. The concept of building trees and evaluating them for Boolean outcomes has been used before in filters like the NIT[7] in SunOS and Linux kernel's internal network packet filter. In their earlier stages, these filters had a predicate tree walker which walked the nodes, evaluated them, and eventually reached a final binary decision. From the seemingly infinite number of packets being transferred from the device, the predicate tree formation and walking algorithm requires a considerable amount of computation to evaluate each packet. To overcome this, an initial version of the Berkeley Packet Filter (BPF) introduced a bytecode interpretation based filtering[7,24]. The predicates in an expression are expressed as nodes of a control flow graph (CFG). The nodes were converted to bytecode and interpreted by a small in-kernel register based BPF interpreter. At the time of its introduction in BSD, BPF gave an improvement of 20 times over earlier techniques. This was also evident in recent patches to the Linux kernel where, in certain scenarios, BPF based filtering brought down the filtering costs from 139 ns to 32 ns[4]. We now discuss the ways to improve this further by techniques such as JIT compiling the bytecode.

Filter Performance Optimizations. The maximum time consumed in VM execution is actually the cost of instruction dispatch[25-26]. The computation can be equivalent to a few machine instructions but the dispatch mechanism usually takes a maximum of 10∼12 machine instructions and involves a time consuming indirect branch. The dispatch mechanisms are typically of either switch or threaded type. To give a short overview, a switch dispatch may contain a large `switch-case` statement where, for each opcode of the VM, there would be one case statement to fetch and evaluate the opcode — which is part of the interpretation phase shown in Fig.1(b). Then, the next bytecode is fetched and evaluated until the bytecode program is finished.

The upgraded BPF+ implementation[11] incorporated many tiny data-flow optimizations such as removing redundant predicates from CFG during the BPF bytecode generation phase, the identification of potential lookup tables, the optimization of register usage and so on. The authors of [11] also did an early JIT implementation and converted the bytecode to native code with a simple register assignment scheme. They obtained a speedup of up to 6.6x between unoptimized BPF code and JIT compiled native code in certain scenarios with a varying number of predicates.

For every bytecode instruction passed to the `switch`, instead of interpreting and dispatching the

equivalent operation, this minimal JIT compiler emits the x86 opcodes and stores them into a code cache upon running for the first time. For the subsequent filter runs, the code is executed natively from the code cache and bypasses the instruction dispatch mechanism (Fig.1(c)). This considerably reduces the overhead, as stated before. The main performance gain by JIT compiling filter bytecode is achieved when the events occur at a high frequency, and run long enough, such as in "always on" systems. We have used a similar principle for our filtering architecture. Along with micro-optimizations to the BPF system and the usage of the fastest tracing approach, we have proposed a very fast trace filtering system, as described in Section 4.

3.3 Trace Filtering

The need for filtering in tracing tools has been addressed before in tools such as DTrace and LTTng. For high frequency events, to sieve out uninteresting events becomes a high priority task. Filters are applied in the execution path of each tracing event. To reduce the overhead, many systems defer the trace filtering to analysis time. Trace viewing and analysis frameworks such as TraceCompass are optimized for performing complex analysis[27]. Cantrill *et al.* discussed the importance of runtime filtering earlier[14]. With a better filtering infrastructure, it is possible to filter out traces at runtime as well. We now discuss some trace filtering approaches that have been used before, and then move on to explain our filtering design in Section 4.

Speculative Tracing. DTrace provides a filtered tracing mechanism called speculative tracing. The basic idea is to record the trace data tentatively in a separate speculation buffer, and then decide whether to commit data to the main tracing buffers or discard it based on checking the data with `speculate()` function. An example is shown in [14] and [28] where the authors described how a filtered trace of all functions entries is only committed if a particular syscall such as `ioctl()` returns a failure. While it is seen as a runtime filtering approach, the speculation involves writing the data to the buffer and possibly copying it to the principal buffer. The DTrace filter execution architecture itself consists of custom bytecode generation and interpretation using a small in-kernel DTrace virtual machine. This predicate condition interpretation, coupled with the data copies, makes the overhead of this approach comparable to that of tools using bytecode interpretation.

LTTng Trace Filtering. LTTng works similarly. The expressions are converted to bytecode and then interpreted. We take an example of LTTng User Space Tracer (UST) where a filter is set on an event. As shown in Fig.2, when the client encounters a filter expression for a specific userspace event to be enabled, the client first parses the expression using a custom lexer-parser and then converts it into a syntax tree.

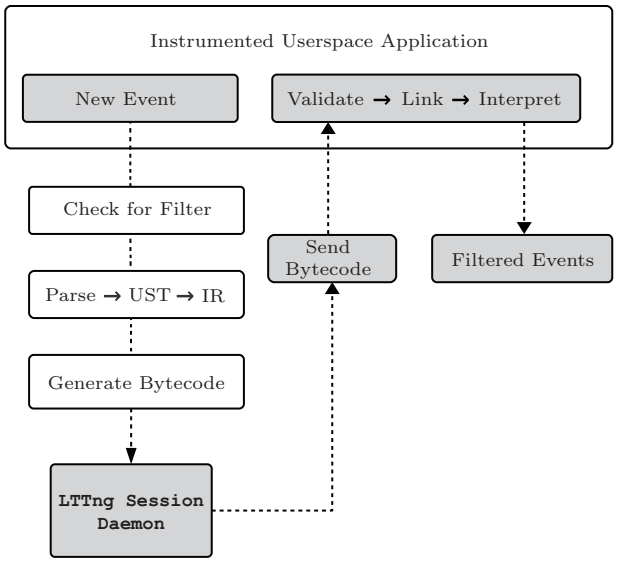

Fig.2. The client is responsible for the conversion of a filter expression to the bytecode, which is sent through `lttng-sessiond` to the instrumented userspace application for validation, linking and an eventual interpretation per event.

The nodes of the syntax tree are visited and classified. Then, the intermediate representation (IR) is generated and a small verification is done on IR. Currently, there is no support for binary arithmetic operations, as the trace filtering needs were very limited. Only logic and comparisons operations are provided. Also, except for logical operators, the nesting of other operators is not allowed. IR is checked to ensure that no wildcard is used in-between string literals and that only valid operators are used. Then, the bytecode is generated by traversing the tree in post-order. The generated bytecode and data is saved to the context and transmitted to the session daemon, `lttng-sessiond`, which sends the bytecode to the userspace application targeted for event filtering. There, the bytecode execution process starts. First, the bytecode is linked to the target event to create a bytecode runtime. Second, a range overflow check for different instruction classes is done and the bytecode is validated for illegal instructions. Finally, the bytecode is sent to LTTng's own filtering virtual machine.

LTTng's interpreter is a hybrid stack/register based virtual machine. As seen in Fig.3, it is a stack-based VM consisting of two registers, `ax` and `bx`, aliased to the top of stack. This makes operations easy, just like on register-based machines, as the push and pops are reduced. At the same time, this gives more flexibility just as for stack machines. The interpreter is a threaded and instruction dispatch based interpreter[29-30], but can be used as a normal dispatch in scenarios where compiler support is not available.

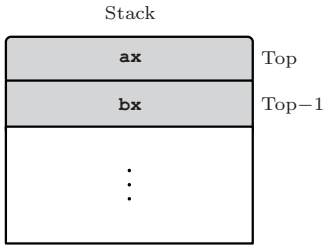

Fig.3. LTTng interpreter is stack-based with two registers (`ax` and `bx`) aliased to the top of stack.

Fig.4 shows how the "signed not-equal-to" operator is interpreted. Here, the operation is performed directly using the macros `estack_bx_v` and `estack_bx_a` which point to the two values to be tested on the execution stack. The LTTng interpreter is quite efficient in relation to the limited scope it has (simple filter execution). However, as we have observed, both by the analysis of the source code and through performance numbers discussed in Subsection 6.2, further optimization is possible with the use of JIT compilation, better optimizations in the bytecode compiler and adding more features (arithmetic operations) to make it more flexible. The overhead is within the range of those tools using bytecode interpretation.

```
OP(FILTER_OP_NE_S64):
{
    int res;
    res = (estack_bx_v != estack_ax_v);
    estack_pop(stack, top, ax, bx);
    estack_ax_v = res;
    next_pc += sizeof(struct binary_op);
    PO;
*
```

Fig.4. LTTng's machine interpreting a bytecode.

4 Filtered Tracing Architecture

We propose a novel userspace trace filtering architecture, with an improved overall tracing performance, as compared with available tracing tools. We choose LTTng and eBPF as the main drivers for this tracing architecture. We now describe the underlying framework on which our filtered trace architecture is based, and present a justification behind that choice.

4.1 Base Framework

eBPF. The idea to convert BPF bytecode to native code, as discussed in Subsection 3.2, has been exploited recently again by Starovoitov for an improved BPF implementation in Linux kernel[4,31]. The earlier implementations, also called classic BPF in the Linux kernel, consisted of two 32-bit registers — A and K. The conditional branch had two jump targets JT (jump if true) and JF (jump if false). There were 32-bit memory slots for filter data. As the main goal of BPF was packet filtering, there are dedicated "extensions" where the developer can load and store data from packets directly. Keeping in mind the good performance and the simplicity of BPF, efforts have been ongoing to make it more generic and modern. The newer version called, extended BPF (or eBPF)[4,13] has many improvements. The instruction set has been changed, and was designed with emphasis on the importance of JIT and underlying architectures on which it is run. eBPF now has 10 internal registers and one frame pointer. The calling convention is similar to current architectures, like ARM64 and x86_64, avoiding extra copies in calls[31]. With this calling convention, the eBPF registers also map one-to-one to the x86_64 and other hardware registers. This simplifies the JIT compiler implementation as well. The main target of eBPF is a generic kernel interpretation framework. It sports a robust verifier and has a concept of "BPF maps", an abstract data type to share data between the kernel and the userspace. There are various helper functions as well, and a dedicated `bpf()` syscall has been proposed to update and access the maps that the BPF programs keep on updating. However, for tracing purposes in userspace, eBPF needs to be optimized for filtering, so that filtering operations can directly occur in userspace. Our adaptation aims to achieve that. Apart from filtering, our extensions can provide co-operative conditional tracing from userspace.

LTTng. The Linux Trace Toolkit next generation (LTTng)[6] is a very fast and extremely low-overhead tracing tool developed at DORSAL②. With a non-activated tracepoint inserted in the code, it gives near-

②http://dorsal.polymtl.ca/en, Oct. 2016.

zero impact on the overall execution of the target application. This distinguishes LTTng from the other tools, making it an excellent choice for real-time applications. Its tracing technique implements a fast wait-free read-copy-update (RCU) buffer for storing data from tracepoint execution[32]. Its efficiency and scalability has been demonstrated in various performance comparisons[6,33-34]. LTTng-UST is the userspace tracing counterpart of LTTng. The major factor for such an increase in performance is the use of a lock-less ring buffer in LTTng-UST, as it efficiently manages multiple readers trying to access the same resource simultaneously[5]. However, LTTng-UST still lacks in areas such as providing an improved dynamic tracing mechanism and an efficient filtering mechanism for userspace tracing. Our contributions also lead to a JIT compiler based bytecode for LTTng, in addition to its interpreted filter bytecode, provided by default in userspace. It also provides a base to add an initial support for JIT-based kernel filtering as well.

Coupled with eBPF's efficient JIT-based filtering technique, LTTng-UST's fast tracing performance can lead to an improved overall performance as compared with interpreted approaches used by DTrace and LTTng's default interpreter. The design of our new eBPF-based JIT compiler and interpreter framework, for userspace and kernel tracing, is influenced by the network filtering approach for which eBPF was originally designed. Our filtering scheme, however, deviates from this network-centric approach. It aims to provide improved performance specifically for userspace tracepoint filtering, and for combined kernel and userspace tracing. The reach of eBPF usage has also been extended by allowing LLVM/GCC-based backends to generate very efficient BPF bytecode from a restricted C interface, while maintaining similar performance.

The system architecture is shown in Fig.5. The filter arguments either are declared by the user manually, in eBPF bytecode, or can be generated by the LLVM-based backend which converts those simple "C"-like expressions into eBPF bytecode. The filter also needs the information about the trace payload and the tracepoint context, which can be obtained from the target binary in which the filter is run. It can then be fed to our userspace implementation of the eBPF library③. The library either checks for the JIT support on the architecture on which it is run, or can be configured to always JIT compile the bytecode. The JIT-compiled code is saved to a code cache and the filter is run around the `tracepoint()` call. As a fallback, the bytecode could be interpreted if the JIT compilation fails. Even though our library implementation is directed towards trace filters, it can also be easily used as a basis to build generic filtering tools such as syscall filtering, or database filters in userspace. We now explain in detail the various steps taken during filtering in this proposed architecture.

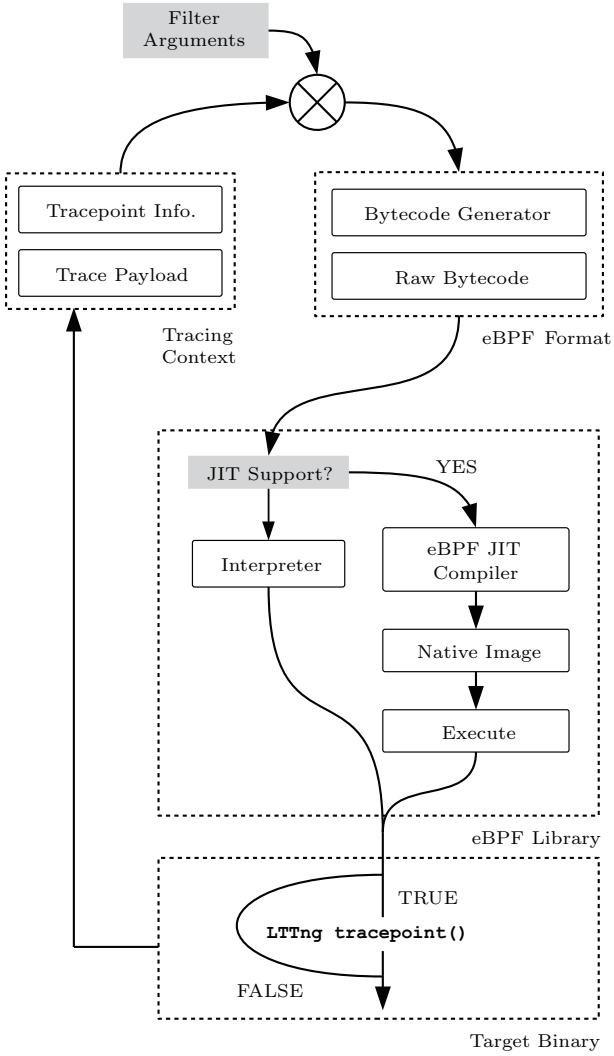

Fig.5. Architecture of our proposed eBPF-based trace filtering system.

Bytecode Preparation. As stated before, there are two ways to provide bytecode to the interpreter (and for later JIT compilation). In the first mode, the user specifies the filter by hard-coding the eBPF opcode macros such as `BPF_LD_IMM64(BPF_REG_0, 1)`, `BPF_EXIT_INSN()` in the target program, or manually assembling bytecodes for an eBPF program and loading it as shown in Fig.6. This is useful only when the

③http://step.polymtl.ca/~suchakra/libebpf.tar.gz, July 2016.

filter is small or the developer is proficient enough to write BPF assembly and manually. The other option is to specify the filter in C and let the recently developed LLVM's eBPF backend generate the eBPF bytecode binary. The compiler converts the eBPF filter, specified in a restrictive C format, to a binary with a `.text` section containing the executable filter eBPF bytecodes. We implemented a small method to extract the opcodes from the section and pass opcodes on to the interpreter or the JIT compiler library. This approach is beneficial because there is an opportunity for the developer to use the optimization routines from the LLVM tools.

```
    idd r1, (0)r1
    mov r2, 42
    jeq r1, r2 goto TRUE
    mov r0,
    ret
TRUE:
    mov r0, 1
    ret
```

Fig.6. eBPF program for a sample filter.

We now discuss some characteristics of the bytecode itself. As mentioned earlier in Subsection 4.1, the newer bytecode of eBPF is closer to native architectures like x86. Using a similar format leads to a more uniform and portable design. The register layout is shown in Table 1, derived from the filter documentation in the Linux kernel[31].

Table 1. Register Mapping for eBPF-x86

eBPF	x86	Purpose
R0	`rax`	Return value from function/ exit value from eBPF
R1	`rdi`	First argument
R2	`rsi`	Second argument
R3	`rdx`	Third argument
R4	`rcx`	Fourth argument
R5	`r8`	Fifth argument
R6	`rbx`	Callee saved
R7	`r13`	Callee saved
R8	`r14`	Callee saved
R9	`r15`	Callee saved
R10	`rbp`	Frame pointer

The R0 register in eBPF is the place where the exit value from eBPF programs is stored. Upon eBPF program return, R0 is set to 1 for TRUE and 0 for FALSE, just as shown in Fig.6.

Register R1 is the place where the filter context is loaded. For example, the context is often made available to the target program through a structure, filled with arguments on which filtering is to be performed. These arguments can be the payload fields from the LTTng tracepoint or, for more complex scenarios, can be obtained at runtime (LTTng's context such as PID/TID). A pointer to this structure can be passed in register R1, which is then accessed as filter context by the eBPF program. In addition, tracing filters regularly need to compare strings, since several tracepoint payload fields are formatted as strings (e.g., the filename in the `open()` syscall). We therefore implemented a `bpf_strcmp()` function which can be called from within the eBPF code. Such helper functions make eBPF filter programs more flexible. As discussed later in Subsection 5.3, we used these helper functions to further extend the filtering system.

Native Code Compilation. The main feature of the system, and the leading reason for improved performance is the JIT compilation of the bytecode. The JIT compilation process for this library is a simple one-to-one JIT and for each instruction (or group of eBPF instructions) there is a direct translation to native code instructions. The compiler backend is non-optimizing. Indeed, the LLVM `clang` compiler frontend performs most of the interesting optimizations, before sending the intermediate representation to the bytecode generation backend. The native code then follows closely the generated bytecode. For illustrative purposes, in Fig.7, we explain the machine code compilation for Fig.6 on an x86-64 system.

```
0   push %rbp                          }
1   mov %rsp, %rbp                     }  (1)
4   sub $0x228, %rsp                   }
b   mov %rbx, -0x228(%rbp)             }
12  mov %r13, -0x220(%rbp)             }  (2)
19  mov %r14, -0x218(%rbp)             }
20  mov %r15, -0x210(%rbp)             }
27  xor %rax, %rax                     }  (3)
29  xor %r13, %r13                     }
2c  mov (%rdi), %rdi                   }
30  mov $x2a, %rsi                     }  (4)
3a  cmp %rsi, %rdi                     }
3d  jz  0x4b                           }
3f  mov $0x0, %rax                     }  (5)
49  jmp 0x55                           }
4b  mov $0x1, %rax                     }
55  mov -0x228(%rbp), %rbx             }
5c  mov -0x220(%rbp), %r13             }
63  mov -0x218(%rbp), %r14             }  (6)
6a  mov -0x210(%rbp), %r15             }
71  leave                              }
72  ret                                }
```

Fig.7. JIT-compiled eBPF program for a sample filter.

The compiler first emits some standard instructions to build the function preamble ((1)). Some variables are allocated on the stack as well for later use, and the values of callee saved registers are saved ((2)). This is a standard preparation for a JIT-compiled filter binary. eBPF's R0 and R7 registers, used as the old A and K registers, are cleared ((3)). The filter context value supplied in R1 (`rdi`) is loaded and compared with a predefined value ((4)). Based on this comparison, 0 or 1 is loaded in R0 (`rax`) and a jump to the exit routine is taken ((5)). A standard set of bytecodes is also emitted for the exit, the callee-saved registers are restored and the filter function is exited ((6)).

Deviating from the Linux kernel's eBPF approach, our eBPF library is lighter, and the JIT compiler faster, by excluding the support for special instructions that perform direct computations in the kernel on network packet data structures. Since we do not need the BPF map data structures, the compiler and the interpreter now being in the userspace, these have been removed as well. Instead, to extend the filtering library to a generic assisted-tracing library, we propose our own shared-memory based communication system between the kernel and the userspace eBPFs, as detailed in Subsection 5.3. For filtering, the native code also supports calls to new helper routines for tracing specific string comparison functions, such as `bpf_strcmp`. The architecture is kept flexible so that other helper functions can be added as desired.

We tested the performance of the filter, and the filtered tracing architecture, in relation to various factors such as filter execution speed, and compared it with the performance of LTTng's userspace trace filtering system based on bytecodes. For practical reasons, because of the limitations of the C pre-processor for defining variable length argument lists, LTTng's bytecode filter currently limits the number of filter predicates to 10, which limited us for our test cases. However, the design of eBPF-based filters has no such restrictions for similar tests. For now, the number of instructions that can be executed with eBPF is kept at 4 096, with the support for tail-calls so that multiple filters can be chained as desired. In our tests, for a similar filter predicate type, we could filter on 50 predicates with our design, as compared with 10 with LTTng's current interpreted filter. The design of the experiments and our findings are elaborated in Section 6.

5 Improved Tracing Infrastructure

In Section 4, we discussed how eBPF and LTTng can be used to develop a new and efficient filtered tracing architecture. We now explore the use of eBPF to provide a new way of performing dynamic tracing in the kernel. We eventually propose and present a new cooperative kernel-userspace tracing system, which supports dynamically defining conditional tracing, and a more efficient data sharing mechanism. We now briefly describe similar approaches taken by other tools.

5.1 Dynamic Tracing

Some of the most interesting developments in the tracing infrastructure have been the ability to dynamically insert tracing probes and take actions when these dynamic probes are hit. The dynamic tracing tools at kernel level are available with different granularities. One of the approaches that has proven to be very flexible is defining a scripted tracing language, which is dynamically compiled at runtime to some IR or bytecode, and then intended to be interpreted in-kernel. Based on the instructions, certain "functions" or "actions" can be executed to gather data into buffers, to be read later from the userspace. Some famous examples are ProbeVue[35] and DTrace[28]. These tools provided scripting languages like D and Vue which would be compiled to an intermediate format. For example, in the case of DTrace, the D program input through the `dtrace` command or a userspace application, would go through the same process of lex-parse to generate the parse tree, and then be compiled to a D intermediate format (DIF). A visual survey of the DTrace code reveals that the DTrace compiler offers very limited optimizations (integer constant folding and peephole optimization) as compared with the enhanced optimizations performed in LLVM for eBPF bytecode. This DIF would be compiled by the assembler to the DIF object (DIFO). This is then coupled with data tables (strings and variables) to form the DTrace object format (DOF), which is the actual bytecode interpreted by the in-kernel DTrace VM. VM is a RISC machine with a fixed register set. The instruction length is fixed to four bytes. To retrieve values from the kernel, DTrace provides a driver that communicates with a userspace library (which can be used with other DTrace consumers like `lockstat` and `intrstat`). This is one of the most comprehensive dynamic tracing infrastructures available. However, it requires a custom VM in-kernel, and the interpretation cost can be high for long running

or badly written scripts. Another approach was that of SystemTap, where the SystemTap scripts would be translated into pure C language and then compiled as kernel modules. These could then be loaded at runtime in the kernel to provide tracing support. It eliminates the need for an in-kernel VM, but the cost of tracepoint executions and data accesses has been high as compared with other dynamic tools[36].

5.2 Data Sharing

Apart from the cost of the tracepoint execution, the cost of collecting and aggregating data is an important consideration as well. Most tools employ a producer-consumer design where the trace events can send data (producer) to a buffer (either in kernel or userspace), and the filled buffers become available (in userspace) for analysis, storage or display (consumer). Neira-Ayuso *et al.* discussed various kernel-userspace data sharing mechanisms before[37]. For very large data bandwidth, the best strategy is to minimize the number of context switches or syscalls.

In DTrace, the `libdtrace` library is responsible for consuming data retrieved from the buffers at probe execution. DTrace provides per-CPU buffers in the kernel that are filled with relevant data. Based on the `ioctl()` arguments in the library, an action is taken on the buffer, such as copying data to the relevant userspace buffer. LTTng provides very efficient shared memory per-CPU ring buffers for one-way sharing of data from userspace to kernel. With the support for Kprobes as well, it is an efficient dynamic tracing system for the kernel. However, there is no specific tracing script support in LTTng, for more advanced analysis or aggregation, as can be done with tools like DTrace. In its current form, eBPF allows the two-way sharing of data (in BPF maps) from userspace to kernel, based on the `bpf` syscall (refer to Fig. 8).

Aggregated or filtered values, stored in hash-tables or array-maps, can also be accessed and updated directly from within an eBPF program bytecode, using the `BPF_CALL` instruction and helper functions, since the program has already been in kernel context. Even though eBPF is efficient and flexible, as it can be dynamically compiled and be used to aggregate data, it would benefit from a more efficient way to transfer data. eBPF itself is not a complete tracer but an infrastructure upon which tracers can be built. The main benefit of eBPF is that it generates dynamically compiled JIT code for tracing. With a more efficient data sharing system used cooperatively with the LTTng tracing system, it can provide an overall benefit, in terms of speed as well as flexibility, to scripted tracing. The resulting system provides a better tracing infrastructure than that offered by currently available tools. We now discuss our co-operative tracing approach based on eBPF and LTTng.

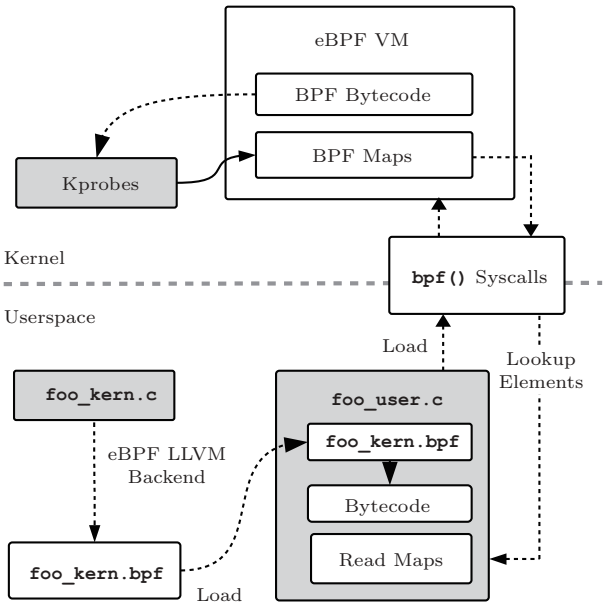

Fig.8. eBPF program in its current form, with the kernel part (`foo_kern.c`) and a userspace part (`foo_user.c`). The userspace part uses the `bpf()` syscall to load the bytecode in the eBPF kernel VM, as well as reading and updating data in BPF maps.

5.3 KeBPF and UeBPF Interactions

Now that we have a system to dynamically execute JIT-compiled code in our programs, we can think beyond just filtering, and make decisions and take "actions" based on aggregated values, in kernel as well as userspace. On the kernel side, this effort is presently ongoing in the form of small eBPF scripts that can aggregate data and share it with userspace[38] (refer to Fig. 8). The kernel eBPF (KeBPF) machine provides access to the shared values in the form of array-maps or hash tables using a syscall. The user can decide to perform aggregations on the values in hash tables in kernel and concurrently read them from userspace. With some effort, eBPF programs can also be used to take decisions based on remembered state (e.g., aggregated values). Our implementation of userspace eBPF (UeBPF) as a library opens new possibilities to collaboratively trace and share data from userspace to kernel and vice versa.

5.3.1 Illustrative Use Case

We show the importance of the interaction between KeBPF and UeBPF programs using an example. For diagnosing system performance, it can be beneficial for the user to track the latency of syscalls issued by a particular userspace process. For that, we developed a custom module where the userspace process registers itself using some `ioctl()` and then probes the `sys_enter` and `sys_exit` trace events along with the time-stamps for each syscall. We can thus compute, for each syscall, how much time the syscall was taking, and thus track the particular syscall latency. We can keep track of all syscalls and set a threshold to decide when to record an event or not, based on the syscall latency threshold. If the elapsed time for a syscall is more than the threshold, the event can be recorded, or otherwise discarded. However, the latency threshold should not be the same for each syscall. It can vary from syscall to syscall and can vary based on the complexity of the request and the underlying hardware speed. We can therefore add specific hooks in the userspace application which specify expected thresholds to the kernel. These hooks then can be set from within eBPF programs so that the user is able to dynamically change the threshold values even at function granularity.

On the kernel side, the kernel can share data with the userspace application to assist it in tracing, based on conditions such as checking if CPUs have been switched, if we are in a blocking state while waiting for a device, etc. All such process states can be shared and the process can then conditionally decide to trace or not. This requires a fast data sharing mechanism between KeBPF and UeBPF programs for minimum overhead. We therefore implemented a `mmap` based shared memory between kernel and userspace so that KeBPF and UeBPF programs could share data directly. Other approaches, such as Perf-based events and LTTng's data sharing, use fast shared memory as well, but only in the context of tracing data, flowing from the producer to the consumer. Also, as discussed before, SystemTap and DTrace are limited in how variables can be shared between different probes executed in kernel mode, and offer no way of executing code in userspace, thus communicating with such code. DTrace's buffers are accessible from the userspace `libdtrace` library, but this involves copies from kernel buffers.

Our sharing is between two VMs (KeBPF and UeBPF). Therefore, there is a direct access to take decisions on tracing from both sides, right at the bytecode level. In that context, a shared memory access enables very efficient communications, and useful usage scenarios. Coming back to the example, as shown in Fig.9, we have a process with PID 42 that registers with our syscall latency tracker module.

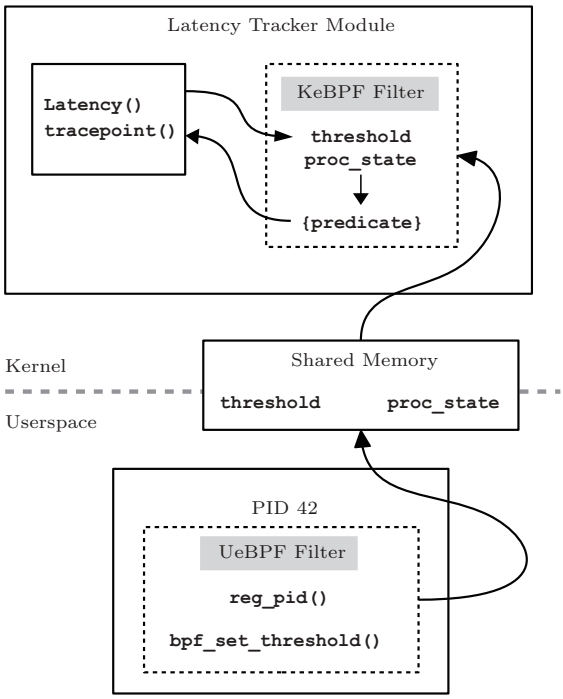

Fig.9. KeBPF-UeBPF shared memory implementation showing syscall latency thresholds being set dynamically from within a UeBPF filter program.

The process contains a UeBPF filter attached to certain function level hooks in the application (as discussed before) which calls our implemented eBPF helper function `bpf_set_threshold()`. This helper function, when called from within the UeBPF filter, writes the updated threshold for the given process/syscall in a shared memory mapped location, shared with KeBPF. The sharing itself is based on a `mmap()` based implementation where KeBPF allocated the required memory during initialization and then mapped it to the userspace address space for direct access. The user can write data using the helper function, and thus can have direct bidirectional access and a zero-copy overhead. Therefore, the thresholds can be dynamically adjusted, and the kernel tracing output can be controlled. In addition, the userspace can continually fill the `proc_state` structure with current process state so as to control other parts of kernel/userspace tracing. In contrast, as of now, the state of the art is to use the `bpf()` syscall and create/update BPF maps from userspace. Each `bpf()` syscall, however, incurs more cost for the same operation, as compared with a direct read or write in our

shared memory. We implemented the VM-VM sharing by allocating the page in the kernel and then mapping the memory to grant access from userspace. In addition, in the default eBPF maps, each value requires an explicit copy in the kernel from userspace, which is avoided in our shared memory approach.

In our tracing approach, kernel and userspace scripts are each executed in their respective context. This enables very fast userspace tracing, avoiding context switches or traps at each userspace tracepoint. This is the reason behind the unrivaled performance of the LTTng userspace library. The excellent communication performance between userspace and kernel space, enabled by the shared memory implementation, then opens up a lot of possibilities, as shown later in the experiments, because of this high-performance architecture. There are, however, precautions required for using this shared memory channel. From the security point of view, the kernel scripts should treat appropriately these userspace supplied values. Furthermore, appropriate synchronization mechanisms must be used, depending on the access protocol.

For single-threaded synchronous access, no synchronization is required. For instance, a userspace script, executed from a single-threaded process, may specify a threshold just before the application issues a system call. Upon finishing the system call, while the application is still blocked, a kernel script would check if the threshold was exceeded, in which case it could write a stack dump to the trace. In this scenario, no synchronization is needed.

There are several cases where thread-level storage can also avoid synchronization issues. Thread-level storage can easily be built using arrays indexed by the thread ID, or using similar mechanisms. One common scenario is aggregating counts (e.g., the number of bytes read, the number of packets received). This could lead to severe scalability problems if a single global variable protected by a lock was used. Instead, one variable per thread (or per core) is typically used and no synchronization is required. The variables can then be read and aggregated at the end, once the scripts are deactivated, not being concurrently updated any more. Alternatively, the variables can be read, even while they are being incremented, as accesses to aligned word-size variables are atomic.

For shared, concurrently accessed global variables, the situation is more problematic. For instance, tracers in probe handlers either avoid any locking, like LTTng with atomic lockless operations, or only allow probes in specific contexts where locking is possible. For example, SystemTap limits the context where probes can be inserted, avoiding NMI interrupts for instance, and automatically protects accesses to global variables with locks. Our implementation currently does not impose any particular access scheme or locking protocol. As an example, the userspace RCU algorithms have proven to provide near-zero read-side overhead for concurrency control applications. It accomplishes this by allowing reads to occur concurrently with updates. RCU maintains multiple versions of objects and makes sure they are not discarded until all pre-existing read side critical sections are finished[32,39]. These algorithms would be applicable to a mixed kernel and userspace environment, but the current URCU library implementation would need to be extended to communicate with a kernel counterpart.

Our proposed approach allows a direct link between two VMs, one in userspace and the other in kernel, to aggregate data, share data with zero copy overhead, and set filtered tracing and conditional actions for each other. Results and inferences from our performance tests on our shared memory implementation, for cooperative KeBPF-UeBPF tracing, are presented in the next section.

6 Experiments and Results

In order to demonstrate the effectiveness of the proposed architecture and algorithms, we divide the experiments into two sets. The first set focuses on the pure performance of native code filters, and their performance when tracing is enabled with varying parameters. The second set evaluates how our shared memory implementation performs as compared with the `bpf()` syscall based approach, used by the default in-kernel eBPF implementation.

6.1 Test Environment

All tests were done on a machine running Fedora 20 with the default 64-bit kernel 3.15 and the eBPF patched kernel 3.17-rc7 for kernel eBPF tests. We used LTTng v2.6 on our workstation running an Intel i7-3770 featuring 4 cores, with hyper-threading disabled, and 16 GB of memory, for tracing and observing its interpreter performance.

6.2 Filter Experiment Set

There are multiple factors on which the trace filters performance can be measured. The most important

is overhead, which can be defined as the extra time or effort required to complete a task when an external factor acts upon a control experimental setup. In terms of tracing/filtering, the time taken due to the addition of tracing and filtering can be compared with a baseline value (the normal execution time of the target process). This extra time is the overhead and is the primary measure of the impact caused by any proposed addition to the tracing system. To evaluate the performance, we designed a synthetic benchmark④ with operator chaining. As shown in Fig.10, the filter predicates $(P_1, P_2, ..., P_N)$ are simple string comparisons connected with a Boolean operator $(\star)$, which is usually an AND/OR. The important time measurements for us here are the time required to build and set up the filter (t_K), the time to evaluate the filter (t_e), and upon evaluation, the time taken to execute the tracepoint code (t_t). Thus, the total time relevant for our observations is,

$$T = t_K + t_e + t_t.$$

To evaluate t_e, we took AND/OR operator chained predicates, doing string comparisons, and observed them for a varying number of events, under a biased condition (the filter always returned TRUE) (refer to Fig.11). This measured the performance of eBPF-JIT vs interpreted and hardcoded filters, so that we could understand better how the native compiled filters in userspace were functioning. We then devised another comprehensive test to compare AND/OR chained predicates, doing string comparisons with varying depth-of-evaluation (DoE), for a run consisting of 100 million events. Varying the DoE meant that the filter was evaluated to be FALSE, and skipped the remaining predicates, after P_X predicates. This is the same as having a filter length equal to the position of the P_X predicate, and the filter evaluated to be TRUE. We varied the DoE from $P = 5$ to $P = 40$ (refer to Fig.12). In the following test, we included the tracepoint time factor t_t as well. For this, we used similar tests and varied the events and the number of predicates, but the filter is kept biased as TRUE, so that the tracepoint was called and we could measure t_t. This gave us the total $(t_e + t_t)$, needed to fully characterize our system. We neglected t_K as the preparation time for filters is amortized over a large number of executions, and is negligible under such conditions. The intended use case is high performance trace filters, with high frequency events observed over long durations.

In this same experiment, we compared the time taken by similar LTTng-UST's interpreted filters with that by our eBPF interpreted and JIT approach. However, we limited the number of predicates to only nine variables, due to LTTng currently allowing no more than nine distinct string variables as trace payload (refer to Fig.13 for results).

Observations. In the first test case for filter optimizations, also shown in Fig.11, we observed that for 100 million events and a 50 predicates filter, the interpreted eBPF filter was 4.3x slower than the hard-coded filter (considered as a lower-bound reference). The native compiled eBPF filter, however, was only 1.4x slower than the hard-coded reference. Even though the JIT compiled filter performance is expected to be similar to that of the actual hard-coded filter, since both were executing native machine code, we could see that this small overhead was due to extra instructions being executed for each filter, as seen in Fig.7. The overall filter performance was consistent for 1M and 10M events, indicating that there would be a consistent filter overhead reduction in the 3x range, when using native compiled eBPF filters, as compared with similar interpreted filters, for tracing scenarios with long predicates and high event frequency.

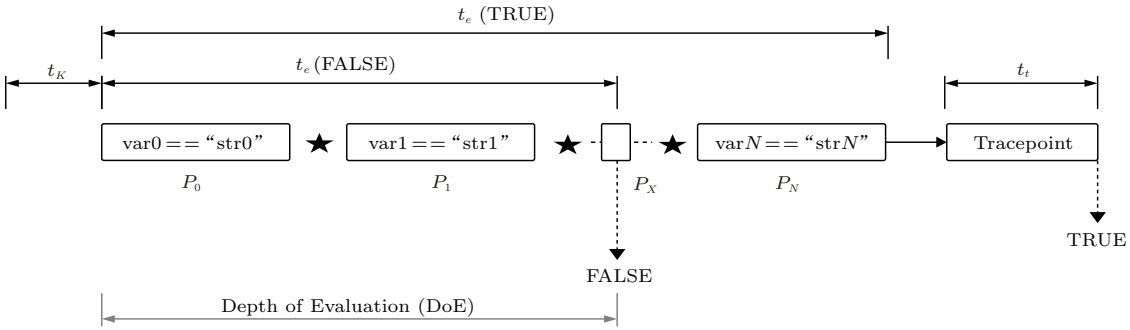

Fig.10. Design of the trace filter benchmark. Evaluation time t_e depends on DoE.

④http://step.polymtl.ca/∼suchakra/libebpf-benchmarks.tar.gz, July 2016.

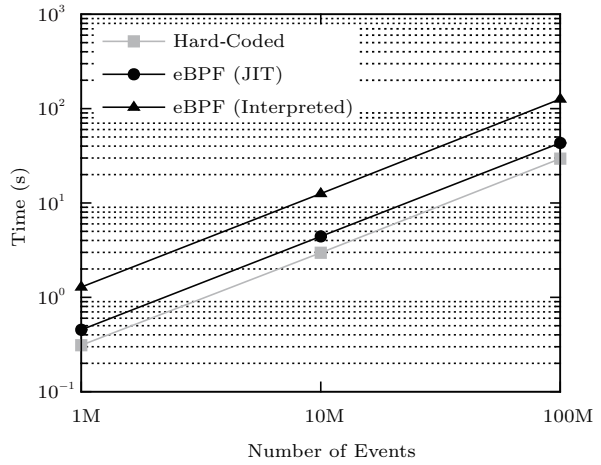

Fig.11. Pure eBPF filter performance with a 50 predicate TRUE biased AND chain.

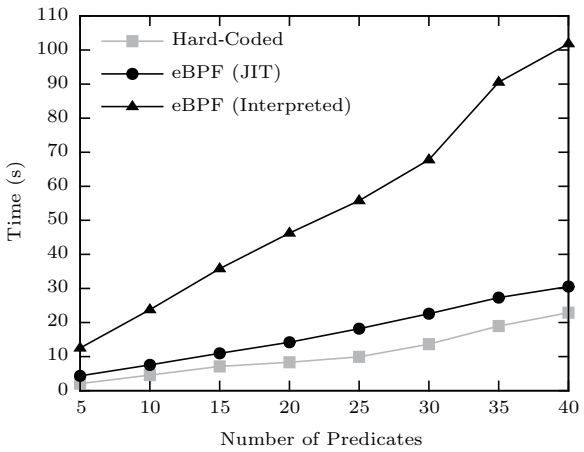

Fig.12. Pure eBPF filter performance with 100M events and a TRUE biased AND chain.

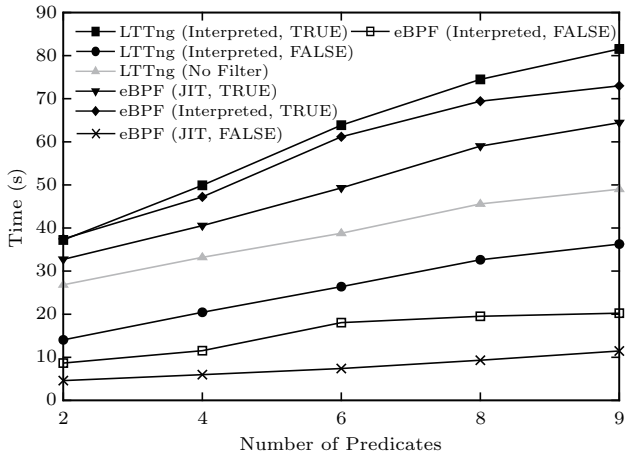

Fig.13. eBPF vs LTTng's filter performance with an increasing number of TRUE/FALSE biased AND chain predicates.

In the second test case, as shown in Fig.12, we observed that with constant 100 million events and an increasing number of predicates, the benefit of natively compiled eBPF trace filters increased marginally. The performance of a 10-predicate JIT compiled eBPF filter was 3.1x better than a similar interpreted filter. This increased to 3.2x for 20 predicates, and a little over 3.3x for 40 predicates. This shows that even for filters with unusually long predicate chains, the performance remains consistent with that of natively compiled filters.

In the third test scenario, we compared LTTng's interpreted filter performance with eBPF's JIT compiled filter performance by observing the biased FALSE cases in Fig.13.

For nine predicates, eBPF's JIT filter was 3.1x faster than a similar LTTng's interpreted filter. We further observed that eBPF's interpreted filter itself was 1.8x faster than LTTng's interpreted filter, pointing to a better register-based eBPF interpreter. We now see how these filters fared with LTTng tracing enabled. In that case, if the filter was evaluated to TRUE, the tracepoint was recorded and the observed time included the tracepoint time. We compared our observations with the LTTng (no filter) mode as the reference line, where no filter is set and all tracepoints are recorded. For a filter of nine AND chained string comparisons, biased to TRUE, the interpreted LTTng had an overhead of 325 ns/event, as compared with eBPF-JIT filter's overhead of 154 ns/event, when LTTng (no filter) was taken as reference. Our JIT-based approach was therefore 2.1x faster. This demonstrates the fact that, with the small cost of JIT-based filtering with our library (154 ns/event in this case), the user can implement filtering at little cost to potentially save a lot of resources by cutting down on unnecessary events that can easily be filtered.

6.2.1 Memory Overhead

Tracing itself may consume a significant amount of memory, some added instructions and data in the program executable, but more importantly, per-CPU structures to buffer events until they are streamed to disk or the network. For instance, in a typical high throughput tracing setup, LTTng can even be configured to use 64 sub-buffers of 8 MB each. For each tracepoint for which a condition is added with our proposed scheme, a small data structure is needed for the interpreted case to represent the trace actions and predicates. When JIT compilation is used for the condition bytecode, additional memory is required to store the compiled code. Either way, the additional memory con-

sumed is insignificant and much lower at 32 KB for a complex 50 predicates eBPF filter, as compared with the memory requirements of the trace buffers themselves.

6.3 Shared Memory Experiment Set

For this experiment set, we created a synthetic benchmark to evaluate the performance of our KeBPF-UeBPF shared memory implementation and compared it with the default syscall based sharing system used in KeBPF. We started off by creating an eBPF program which populates an eBPF array-map with 1 000 integers with random values. We then looked up these values from userspace using the `bpf()` syscall with arguments `BPF_MAP_LOOKUP_ELEM` and measured the time for multiple runs. We compared this with the time taken to read the same values updated in Kernel eBPF using our shared memory, and then read from userspace using a simple `read()`, or a helper function in UeBPF which calls `read()`.

Observations. In the case of KeBPF-UeBPF shared memory implementation, we got an overall improvement of 99x over the default implementation. The time taken by 1 000 reads of an integer array-map is shown in Table 2. The default `bpf()` implementation took 218 ns/read whereas our shared memory implementation took 2.2 ns/read, which can be explained by the fact that the shared memory can be directly accessed, without any system call, as detailed in Subsection 5.3.

Table 2. Time Taken for 1 000 Reads of an Integer Array Map

	Time (ns)	Std. Dev.
Baseline	2 120	210
eBPF-shm	2 247	984
eBPF-syscall	218 203	801

Our eBPF-shm shared memory was close to the baseline values taken using simple read calls. eBPF-syscall, in Table 2, shows the time taken to read using the `bpf()` syscall. Going through the eBPF code in the kernel, we observed that a similar process, of using this syscall to update a map value in the kernel, would incur a syscall time as well as the time involved in copying the value to the kernel space. In our shared memory system, however, there was no extra copy involved, and KeBPF and UeBPF could share data directly at a high speed, as observed in our test.

7 Conclusions

In this paper we presented two contributions to tracing techniques in userspace. First, we improved the trace filtering mechanism by using Just-In-Time (JIT) compilation to convert trace filter bytecode into native machine code. We used the Linux kernel's eBPF-based bytecode technique, and improved it for tracing in userspace context. We targeted LTTng-UST as the tracer, due to its low overhead, and observed that our native filtering approach surpasses the filtering performance of similar high performance state-of-the-art tools. We showed that with our technique, we could filter traces in record time to have smaller traces and provide more efficient tracing, in long running high frequency in production tracing scenarios, such as embedded soft-realtime systems and networked nodes. We devised a rigorous benchmark and found that the performance of the new JIT-based filtered tracing architecture is 3x that of the current interpreted approaches. We provided the implementation in the form of a generic userspace eBPF filtering library that can be used to improve other trace filtering scenarios such as network packets or syscall-based sand-boxing in userspace.

As the second contribution, we developed a shared memory system between the default Kernel eBPF and our Userspace eBPF, by extending the eBPF system at both levels. This enables sharing data at greater speeds and using it to do co-operative tracing from kernel to userspace and vice versa. We demonstrated this using a basic syscall latency tracing example, where the thresholds could be dynamically adjusted at function-level granularity using hooks in the userspace application, right from within UeBPF. KeBPF could then access it to make decisions on recording or discarding syscall events. The interaction between a kernel VM and a userspace VM is significant as it allows a direct interaction between decision making sections. Along with the benefit of zero-copy overhead, it provides the flexibility for performing conditional actions on kernel-userspace shared data — such as performance counter values, process states (off-CPU state, wait threshold, syscall latency threshold, resources thresholds, etc.).

There are, however, some limitations in our current approach, which will motivate our future work. We have observed that very specific and long filter predicate usecases in trace filtering can have a negative effect. The overall trace becomes small, and important events which give a context to the tracing scenario get

missed out. To overcome this, we can use profile-guided tracing where the generated bytecode can perform some non-intrusive profiling on the tracing, get some feedback and also record traces which are relevant to the filter scenario — even if it does not satisfy the filter condition. Triggers for system-wide (kernel+userspace) tracing can be defined and, when enabled, in addition to the intended filtered tracepoint, would also record a system-wide trace for some predefined or dynamically defined duration. We can also utilize LLVM's compiler infrastructure to support some high-level meta language to define tracing specific scripts, and move towards traditional script-based filtering when required, while keeping all the benefits of low overhead and speed provided by LTTng.

The UeBPF library could also benefit from more explicit support for data sharing through multiple threads. We can also port it to non-Linux platforms as a generic library for trace, network packet or syscall filtering. This would expand it from its current Linux specific implementation. In some specific usecases, it may also be worthwhile to investigate hardware-based trace filtering, where an eBPF machine would be implemented not just as a JIT compiler but as specialized hardware. Our current userspace eBPF implementation could also be extended to provide support for program flow tracing, such as with Intel Processor Trace (PT)[40], where eBPF programs from userspace could conditionally trigger a PT snapshot.

Acknowledgment We would like to thank Alexei Starovoitov for his recent eBPF related work in the Linux Kernel, authors of LTTng project for their technical guidance and Francis Giraldeau for his comments.

References

[1] Ball T, Burckhardt S, de Halleux J, Musuvathi M, Qadeer S. Deconstructing concurrency heisenbugs. In *Proc. the 31st International Conference on Software Engineering*, May 2009, pp.403-404.

[2] Bligh M, Desnoyers M, Schultz R. Linux kernel debugging on Google-sized clusters. In *Proc. the Linux Symposium*, June 2007, pp.29-40.

[3] Ezzati Jivan N. Multi-level trace abstraction, linking and display [Ph.D. Thesis]. École Polytechnique de Montréal, 2014.

[4] Starovoitov, A. Tracing: Accelerate tracing filters with BPF [LWN.net]. http://lwn.net/Articles/598545/, May 2016.

[5] Goulet D. Unified kernel/user-space efficient Linux tracing architecture [Master Thesis]. École Polytechnique de Montréal, 2012.

[6] Desnoyers M. Low-impact operating system tracing [Ph.D. Thesis]. École Polytechnique de Montréal, 2009.

[7] McCanne S, Jacobson V. The BSD packet filter: A new architecture for user-level packet capture. In *Proc. the USENIX Winter Conference*, January 1993, pp.259-269.

[8] Mogul J, Rashid R, Accetta M. The packet filter: An efficient mechanism for user-level network code. *ACM SIGOPS Operating Systems Review*, 1999, 21(5): 39-51.

[9] Bailey M L, Gopal B, Pagels M A, Peterson L L, Sarkar P. Pathfinder: A pattern-based packet classifier. In *Proc. the 1st Symposium on Operating Systems Design and Implementation*, Nov. 1994, pp.115-123.

[10] Engler D R, Kaashoek M F. DPF: Fast, flexible message demultiplexing using dynamic code generation. *ACM SIGCOMM Computer Communication Review*, 1996, 26(4): 53-59.

[11] Begel A, McCanne S, Graham S L. BPF+: Exploiting global data-flow optimization in a generalized packet filter architecture. *ACM SIGCOMM Computer Communication Review*, 1999, 29(4): 123-134.

[12] Wu Z, Xie M, Wang H. Design and implementation of a fast dynamic packet filter. *IEEE/ACM Transactions on Networking*, 2011, 19(5): 1405-1419.

[13] Sobel L. Secure input overlays: Increasing security for sensitive data on Android [Master Thesis]. Massachusetts Institute of Technology, 2015.

[14] Cantrill B M, Shapiro M W, Leventhal A H. Dynamic instrumentation of production systems. In *Proc. the USENIX Annual Technical Conference*, June 27-July 2, 2004, pp.15-28.

[15] Jacob B, Larson P, Leitao B H, Silva S A M M. SystemTap: Instrumenting the Linux kernel for analyzing performance and functional problems. IBM Redpaper, 2009. http://www.redbooks.ibm.com/abstracts/redp4469.html, Oct. 2016.

[16] Rostedt S. Using the TRACE_EVENT() macro (Part 1) [LWN.net]. http://lwn.net/Articles/379903/, May 2016.

[17] Keniston J, Panchamukhi P S, Hiramatsu M. Kernel probes (Kprobes). https://www.kernel.org/doc/Documentation/kprobes.txt, May 2016.

[18] Hiramatsu M. The Enhancement of kernel probing — Kprobes jump optimization. http://tracingsummit.org/w/images/f/fa/HiramatsuLinuxCon2010.pdf, Oct. 2016.

[19] Brown A, Wilson G. The Architecture of Open Source Applications, Volume II. CreativeCommons, 2012.

[20] Buck B, Hollingsworth J K. An API for runtime code patching. *International Journal of High Performance Computing Applications*, 2000, 14(4): 317-329.

[21] Reddi V J, Settle A, Connors D A, Cohn R S. PIN: A binary instrumentation tool for computer architecture research and education. In *Proc. the Workshop on Computer Architecture Education*, June 2004.

[22] Prasad V, Cohen W, Eigler F C, Hunt M, Keniston J, Chen J. Locating system problems using dynamic instrumentation. In *Proc. the Linux Symposium*, June 2005, pp.49-64.

[23] Kim T, Zeldovich N. Practical and effective sandboxing for non-root users. In *Proc. USENIX Conference on Annual Technical Conference*, June 2013, pp.139-144.

[24] Corbet J. BPF: The universal in-kernel virtual machine [LWN.net]. http://lwn.net/Articles/599755/, May 2016.

[25] Shi Y, Casey K, Ertl M A, Gregg D. Virtual machine showdown: Stack versus registers. *ACM Transactions on Architecture and Code Optimization*, 2008, 4(4): 2:1–2:36.

[26] Davis B, Beatty A, Casey K, Gregg D, Waldron J. The case for virtual register machines. In *Proc. the Workshop on Interpreters, Virtual Machines and Emulators*, June 2003, pp.41-49.

[27] Gebai M, Giraldeau F, Dagenais M. Fine-grained preemption analysis for latency investigation across virtual machines. *Journal of Cloud Computing*, 2014, 3(1): Article No. 23.

[28] McDougall R, Mauro J, Gregg B. Solaris Performance and Tools(c) DTrace and MDB Techniques for Solaris 10 and OpenSolaris. Prentice Hall, 2006.

[29] Ertl M A, Gregg D. The behavior of efficient virtual machine interpreters on modern architectures. In *Proc. the 7th International Euro-Par Conference on Parallel Processing*, Aug. 2001, pp.403-412.

[30] Gagnon E M, Hendren L J. SableVM: A research framework for the efficient execution of java bytecode. In *Proc. the 1st USENIX Java Virtual Machine Research and Technology Symposium*, April 2001, pp.27-40.

[31] Schulist J, Borkmann D, Starovoitov A. Linux socket filtering aka Berkeley packet filter (BPF). https://www.kernel.org/doc/Documentation/networking/filter.txt, May 2016.

[32] Desnoyers M, McKenney P E, Stern A S, Dagenais M R, Walpole J. User-level implementations of read-copy update. *IEEE Transactions on Parallel and Distributed Systems*, 2012, 23(2): 375-382.

[33] Chakraborty S. Efficiency of LTTng as a Kernel and Userspace Tracer on Multicore. VDM Verlag Dr. Müller, Saarbrücken, Germany, 2011.

[34] Desfossez J. LTTng-UST vs SystemTap userspace tracing benchmarks. https://sourceware.org/ml/systemtap/2011-q1/msg00244.html, May 2016.

[35] Lascu O, Bodily S, Harvala M, Singh A K, Song D, Berg F V D. IBM AIX Continuous Availability Features. IBM Redpaper, 2008. http://www.redbooks.ibm.com/redpapers/pdfs/redp4367.pdf, Oct. 2016.

[36] Beamonte R, Dagenais M R. Linux low-latency tracing for multicore hard real-time systems. *Advances in Computer Engineering*, 2015, 2015: Article ID 261094.

[37] Neira-Ayuso P, Gasca R M, Lefevre L. Communicating between the kernel and user-space in Linux using netlink sockets. *Software — Practice and Experience*, 2010, 40(9): 797-810.

[38] Gregg B. eBPF: One small step. http://www.brendangregg.com/blog/2015-05-15/ebpf-one-small-step.html, May 2016.

[39] McKenney P E (ed.). Is parallel programming hard, and, if so, what can you do about it? https://www.kernel.org/pub/linux/kernel/people/paulmck/perfbook/perfbook.html, May 2016.

[40] Reinders J. Processor tracing. https://software.intel.com/en-us/blogs/2013/09/18/processor-tracing, May 2016.

Suchakrapani Datt Sharma is a Ph.D. candidate at École Polytechnique de Montréal, Montréal. He received his M.Tech. degree in instrumentation and control (biomedical instrumentation) from the College of Engineering, Pune, India, in 2011. He worked as an automotive infotainment systems developer for Tata Consultancy Services in 2012. His research interests are bytecode interpreters, compilers, tracers and associated performance analysis tools.

Michel Dagenais is a professor in the Department of Computer and Software Engineering at École Polytechnique de Montréal, Montréal. He authored or co-authored over 100 scientific publications, as well as numerous free documents and free software packages in the fields of operating systems, distributed systems and multicore systems, in particular in the area of tracing and monitoring Linux systems for performance analysis. Most of his research projects are in collaboration with industry and generate free software tools among the outcomes. The Linux Trace Toolkit next generation, developed under his supervision, is now used throughout the world and is part of several specialized and general purpose Linux distributions.