

Fine Grained Kernel Logging with KLogger: Experience and Insights

Yoav Etsion[†] Dan Tsafirir^{†Ψ} Scott Kirkpatrick[†] Dror G. Feitelson[†]

[†]School of Computer Science and Engineering
The Hebrew University of Jerusalem, Israel

^ΨIBM T. J. Watson Research Center
Yorktown Heights, NY

{etsman,dants,kirk,feit}@cs.huji.ac.il

ABSTRACT

Understanding the detailed behavior of an operating system is crucial for making informed design decisions. But such an understanding is very hard to achieve, due to the increasing complexity of such systems and the fact that they are implemented and maintained by large and diverse groups of developers. Tools like KLogger — presented in this paper — can help by enabling fine-grained logging of system events and the sharing of a logging infrastructure between multiple developers and researchers, facilitating a methodology where design evaluation can be an integral part of kernel development. We demonstrate the need for such methodology by a host of case studies, using KLogger to better understand various subsystems in the Linux kernel, and pinpointing overheads and problems therein.

Categories and Subject Descriptors

D.4.8 [Performance]: Measurements; C.4 [Performance of Systems]: Design studies

General Terms

Measurement, Performance, Experimentation

Keywords

operating systems, Linux, kernel logging, KLogger, performance evaluation, scheduling, locking, overheads

1. INTRODUCTION

In the late 1970s, UNIX version 6 consisted of ~60,000 lines of code [17]. Today, version 2.6 of the Linux kernel consists of over 5,500,000 lines of code, and almost 15,000 source files. This is a great testimony to the complexity of modern operating systems.

Modern, general purpose operating systems need to manage a plethora of hardware devices: storage devices, networking, human interface devices, and the CPU itself. This is done using software layers such as device drivers, file-systems, and communications protocols. The software is designed and implemented by hundreds of programmers writing co-dependent code. This is especially true

for community-developed operating systems such as Linux and the BSD family. While such open-source approaches benefit from the talents and scrutiny of multiple avid developers, they may also lead to situations where different pieces of code clash, and do not inter-operate correctly [2].

Adding to this problem is the power of modern CPUs and the increasing parallelism introduced by symmetric multi-processing and multi-core CPUs. While increasing CPU power might mask performance problems, the increasing parallelism introduces a myriad of issues system designers need to deal with — most of which stem from the need to synchronize parallel events.

The resulting software is too complex for a human programmer to contain, and might even display counter-intuitive behavior [15]. Analyzing system behavior based on measurements is often thwarted by measurement overheads that overshadow the effects being investigated. This sometimes described as the *Heisenberg effect* for software [28]. All this has detrimental effects on the engineering of critical system components. For example, it is not uncommon that code is submitted to the Linux kernel, and sometimes even accepted, based on a subjective “*feels better*” argument [19].

This situation raises the need for better system analysis tools, that will aid developers and researchers in obtaining a better understanding of system behavior. Given systems’ complexity, one cannot expect an all encompassing system analyzer, because that would require a full understanding of the operating system’s code. A more promising approach is a framework allowing developers to build event loggers specific to the subsystem at hand. This framework should be integrated into the kernel development methodology, by designing a subsystem’s event logger along with the subsystem itself. In fact, an event logger based on the subsystem’s logic can also complement the subsystem’s documentation. Ultimately, such a framework may facilitate the creation of a collection of system loggers based on the experience of developers writing the code in the first place.

In this paper we introduce *KLogger*, a fine-grained, scalable, and highly flexible kernel logger. *KLogger* is designed for post-mortem analysis, logging the configured kernel events with very low overhead. It is reliable in the sense that event loss due to buffer overflow is rare, and can be detected by the user by tracking gaps in the event serial numbers (indicating a serial number was allocated but the corresponding event not logged for lack of buffer space). Furthermore, events can be logged from any point in the running kernel. Logging is done into per-CPU buffers, making *KLogger* scalable, a required feature for the increasingly parallel modern processors.

KLogger can be specialized for specific subsystems using an event configuration file, which leads to the generation of event-specific code at kernel compilation time. This structured specialization mechanism, called *KLogger schemata*, allows kernel developers to share their expertise and insights, thus allowing other researchers to analyze code without having to fully understand its intricacies. The idea behind this design is based on the notion that a high level understanding of a subsystem should be enough to evaluate it, rather than having to know the gory details of the implementation.

To better demonstrate the ease of *using* a schema vs. the difficulties in *creating* one, we present the process scheduler as an example. At the base of every multiprogramming operating system there is a point in time where one process is scheduled out (either preempted or blocked), and another is scheduled in. This operation is known as a context switch. However, the exact code snippet performing the context switch — saving one process's state and restoring another's — is implementation dependent; in Linux, it is a preprocessor macro with different versions for each supported architecture (and most versions are implemented using hand crafted assembly code). This is called from a variety of locations, including but not limited to the *context_switch* function.

Pinpointing the base code performing a context switch requires kernel expertise, but the real problem is verifying that this is the *only* place where a process may be resumed. In Linux, a new process is created using the fork system call. However, a process actually starts running only from inside the context switch code, in which one process's state is saved and another's state is restored. Obviously, the state saved is the state of a process as seen by the context switch function. After a process's state is restored, the function performing the context switch completes some accounting and returns, thereby resuming the execution of the newly scheduled process. Had the fork system call duplicated the parent's running state to the child process as seen by the fork system call itself, the child process would have continued to run from inside the fork system call in the first time it is scheduled. But this would skip the post-context-switch accounting, thus threatening the consistency of the process scheduler's internal data structures. Therefore the fork code actually constructs a stack frame for a new process that uses a specialized version of the context switch function for newly created processes — in which the new process will start its first quantum after it is scheduled in. These are extremely intricate implementation details that anyone placing logging calls must be aware of, but are not relevant for higher level tracking of scheduling events, either for evaluation of a process scheduler's performance or for investigating the interaction of scheduling with other subsystems such as the file system or memory manager.

We believe that the KLogger framework offers a solution to this problem in its formalization of KLogger schemata. Specifically to our example, the scheduler developer — who knows all the implementation intricacies — will create a schema for scheduler-related events, including a context switch logging call in the correct code snippets. This will enable the community to use that event for keeping track of which process is running at each instant, or for measuring scheduler performance, without having to overcome the hurdle of fully understanding the fine implementation details described above. Moreover, once a KLogger schema is available, it can be used to shed light on the implementation details of the relevant kernel subsystem, acting as code annotations complementing the subsystem's documentation.

KLogger currently supports the Linux kernel — both the 2.4.x and the 2.6.x kernel versions. Although a newer series exists, we cannot dismiss the 2.4.x kernel series as it is still favored by many administrators, especially since the 2.6.x series has been considered unstable for a long time, even by some kernel developers [5, 33].

To demonstrate the power and flexibility of KLogger, we dedicate over half of this paper to describing several case studies in which KLogger uncovered bottlenecks or mis-features — including examples of what we have learned about the behavior of the Linux kernel using KLogger.

The rest of this paper is organized as follows. Section 2 reviews related work. Sections 3, 4, and 5 describe the design principles, programmer/user interface, and implementation of the KLogger infrastructure. Section 6 describes our testbed, after which we describe the case studies in sections 7 through 10.

2. RELATED WORK

KLogger is a software tool used to log events from the operating system's kernel, with the developers defining the events at compilation time. This is not a novel approach, and there exist several tools which operate on the same principle. Unfortunately, these tools have various limitations, chief among which is high overhead that limits the granularity of events that can be investigated.

The simplest logging tool in Linux is *printk*, the kernel's console printing utility [4, 18], whose semantics are identical to those of C's standard *printf*. This tool incurs a substantial overhead for formatting, and is not reliable — it uses a cyclic buffer which that is read by an external unsynchronized daemon. The buffer can therefore be easily overrun, causing event loss.

The most effective Linux tool we have found is the *Linux Trace Toolkit* (LTT) [35]. LTT logs a set of some 45 predefined events, including interrupts, system calls, network packet arrivals, etc. The tool's effectiveness is enhanced by its relatively low overhead and a visualization tool that helps analyzing the logged data. However, it is not flexible nor easily extensible to allow for specific instrumentation.

IBM's K42 operating system has an integrated logging tool that shares some features with KLogger [34]. These features include fine-grained logging of events from every point in the kernel, variable-length timestamped events and logging binary data that is decoded post-mortem, among others. There is currently an attempt to integrate some of this system's features into LTT, such as efficient transfer of logged data from the kernel to user-level [36].

A more flexible approach is taken by *Kerninst* [30], and what seem to be its successors — *DTrace* [8] on Sun's Solaris 10 operating system, and *Kprobes* [23] from IBM in Linux. These tools dynamically modify kernel code in order to instrument it: either by changing the opcode at the requested address to a jump instruction or by asserting the processor's debug registers, thus transferring control to the instrumentation code. After the data is logged, control returns to the original code. The ability to add events at runtime makes these tools more flexible than KLogger.

None of the above tools provide data about the overhead they incur per logging a single event (with the exception of *Kerninst*, which is the principal metric in evaluating a tool's granularity. We therefore measured them using the KLogger infrastructure and found

that their overhead is typically much higher than that of KLogger. This measurement is described below (in the section dealing with KLogger’s *stopwatch* capabilities), and is summarized in Table 1.

TIPME [10] is a specialized tool aimed at studying system latencies, which logs system state into a memory resident buffer whenever the system’s latencies were perceived as problematic. This tool partly inspired the design of KLogger, which also logs events into a special buffer. It is no longer supported, though.

The Windows family also has a kernel mechanism enabling logging some events, called *Windows Performance Monitors* [29], but very little is known about its implementation.

An alternative to logging all events is to use sampling [1]. This approach is used in *OProfile*, which is the underlying infrastructure for HP’s *Prospect* tool. OProfile uses Intel’s hardware performance counters [14] to generate traps every N occurrences of some hardware event — be it clock cycles, cache misses, etc. The overhead includes a hardware trap and function call, so logging 10,000 events/second can lead to 3-10% overall overhead (depending on which hardware counter is being used). Also, this tool is periodic, and thus bound to miss events whose granularity is finer than the sampling rate.

Yet another approach for investigating operating system events is to simulate the hardware. For example, *SimOS* [25] was effective in uncovering couplings between the operating system and its underlying CPU [26], but is less effective when it comes to understanding the effects of specific workloads on the operating system per-se.

Finally, architectures with programmable microcode have the option to modify the microcode itself to instrument and analyze the operating system, as has been done on the VAX [21]. In principle, this approach is also viable for Intel’s Pentium IV processors, which internally map op-codes to μ ops using some firmware. The problem is that this firmware is one of Intel’s best guarded secrets, and is not available for developers.

3. KLOGGER DESIGN PRINCIPLES

KLogger is a framework for logging important events to be analyzed offline. Events are logged into a memory buffer, which is dumped to disk by a special kernel thread whenever its free space drops below some low-water mark.

The design of KLogger originated from the need for a tool that would enable kernel researchers and developers direct, unabridged, access to the “darkest” corners of the operating system kernel. None of the tools surveyed above provides the combination of qualities we required from a fine grained kernel logging tool. Thus, KLogger was designed with the following goals in mind:

A Tool for Researchers and Developers. KLogger is targeted at researchers and developers, and not at fine tuning production systems. This goal forces us to maintain strict event ordering, so events are logged in the same order as executed by the hardware. Also, events must not get lost so logging must be reliable. These two features also make KLogger a very handy debug tool. On the other hand, this goal also allows for event logging code to incur some minimal overhead even when logging is disabled. An additional requirement was support for logging the hardware’s performance counters. While such counters are now available on most

platforms, we currently only support the Intel Pentium IV performance monitoring counters [14].

Low overhead. When monitoring the behavior of any system, our goal is “to be but a mere fly on the wall”. Thus overhead must be extremely low, so as not to perturb the system behavior. The overhead can be categorized into two orthogonal parts: *direct overhead* — the time needed to take the measurement, and *indirect overhead* — caused by cache lines and TLB entries evicted as a result of the logging. These issues are discussed below in the section dealing with KLogger’s *stopwatch* capabilities.

Flexibility. KLogger must be flexible, in that it can be used in any part of the kernel, log any event the researcher/developer can think of, and allow simplicity in adding new types of events. Also, it must allow for researchers to share methodologies: if one researcher comes up with a set of events that measure some subsystem, she should be able to easily share her test platform with other researchers, who are not familiar with the gritty implementation details of that particular subsystem. This goal is important since it allows for KLogger users to easily incorporate the ideas and insights of others. KLogger’s flexibility is further discussed in the section titled “KLogger Schemata” and demonstrated later on in several case studies.

Ease of Use. Using KLogger should be intuitive. For this reason we have decided to use semantics similar to printing kernel data to the system log, leaving the analysis of the results for later. These semantics, along with the strictly ordered, reliable logging, make KLogger a very handy debugging tool. Another aspect of this goal is that configuration parameters should be settable when the system is up, avoiding unnecessary reboots or recompilations. KLogger’s programmer/user interface is further discussed below.

The design goals are specified with no particular order. Even though we have found them to be conflicting at times, we believe we have managed to combine them with minimal tradeoffs.

4. THE PROGRAMMER/USER INTERFACE

This section will discuss the business end of KLogger — how to operate and configure this tool.

KLogger’s operation philosophy is quite simple: when designing a measurement we first need to define what we want to log. In KLogger’s lingo, this means defining an event and the data it holds. Second, we need to declare when we want this event logged. Third, we have to configure runtime parameters, the most important of which is the toggle switch — start and stop the measurement. The last step is analyzing the data, the only part in which the user is on her own. Since analyzing the data is task specific to the data gathered, the user needs to write a specific analyzing program to extract whatever information she chooses, be it averaging some value, or replaying a set of events to evaluate an alternate algorithm. To simplify analysis, KLogger’s log is text based, and formatted as a Perl array of events, each being a Perl hash (actually, the log is dumped in binary format for efficiency, and later converted into its textual form using a special filter).

To simplify the description of the interface, we will go over the different components with a step by step example: defining an

event that logs which process is scheduled to run. The event should be logged each time the process scheduler chooses a process, and should hold the *pid* of the selected process and the number of L2 cache misses processes experienced since the measurement started (granting a glimpse into the processes' cache behavior).

4.1 Event Configuration File

The event configuration file is located at the root of the kernel source tree. A kernel can have multiple configuration files — to allow for modular event schemata — all of which must be named with the same prefix, *klogger.conf* (unlisted dot-files, following the UNIX convention for configuration files). The configuration file contains both the event definitions and the hardware performance counters definitions (if needed).

Performance counter definitions are a binding between a virtual counter number and an event type. The number of counters is limited only by the underlying hardware, which has a limited number of registers. Sometimes certain events can only be counted using a specific subset of those registers, further limiting the performance counters variety. The KLogger infrastructure defines a set of well known event names as abstractions, and allows the user to bind virtual counters to these event types. When reading the configuration files, the KLogger code generator uses a set of architecture-specific modules to generate the correct (mostly assembly) code for the underlying hardware. In our example we set virtual hardware counter 0 to count L2 cache misses:

```
arch PentiumIV {
    counter0 l2_cache_misses
}
```

Accessing a predefined hardware counter is described below.

Event definitions are C-like structure entities, declaring the event's name and the data fields it contains. The data types are similar to primitive C types, and the names can be any legal C identifier. The event used in our example is

```
event SCHEDIN {
    int pid
    ulonglong L2_cache_misses
}
```

This defines an event called *SCHEDIN* that has three fields — the two specified, and a generic header which contains the event type, its serial number in the log, and a timestamp indicating when the event occurred. The timestamp is taken from the underlying hardware's cycle counter, which produces the best possible timing resolution. This event will appear in the log file as the following Perl hash:

```
{
    header => {
        "type"      => "SCHEDIN",
        "serial"    => "119",
        "timestamp" => "103207175760",
    },
    "pid"          => "1073",
    "L2_cache_misses" => "35678014",
},
```

A more detailed description of the configuration file is beyond the scope of this paper.

4.2 Event Logging

Logging events inside the kernel code is similar to using the kernel's *printk* function. KLogger calls are made using a special C macro called *klogger*, which is mapped at preprocessing time to an *inlined* logging function specific to the event. This optimization saves the function call overhead, as the *klogger* logging code simply stores the logged data on the log buffer.

The syntax of the logging call is:

```
klogger(EVENT, field1, field2, ...);
```

where the arguments are listed in the same order as they are declared in the event definition. KLogger uses C's standard type checks. In our scheduler example, the logging command would be:

```
klogger(SCHEDIN, task->pid,
        klogger_get_l2_cache_misses());
```

with the last argument being a specially auto-generated inline function that reads the appropriate hardware counter.

Note that when KLogger is disabled in the kernel configuration (e.g. not compiled in the kernel), the logging calls are eliminated using C's preprocessor, so as not to burden the kernel with any overhead.

4.3 Runtime-Configurable Parameters

KLogger has a number of parameters that are tunable at runtime, rather than compile time. These parameters are accessible using the Linux *sysctl* interface, or its equivalent */proc* filesystem counterpart — namely by writing values into files in the */proc/sys/klogger/* directory. Accessing these parameters using the general filesystem abstraction greatly simplifies KLogger usage, as it enables users to write shell scripts executing specific scenarios to be logged. It also allows a running program to turn on logging when a certain phase of the computation is reached.

The most important parameter is KLogger's general on/off switch. Logging is enabled by simply writing "1" into the */proc/sys/klogger/enable* file. Writing "0" into that file turns logging off. This file can also be read to determine whether the system is currently logging.

Even though the kernel is capable of logging a variety of events, at times we want to disable some so only a subset of the events actually get logged. Each event is associated with a file named after the event in the */proc/sys/klogger/* directory. Like the main toggle switch, writing a value of 0 or 1 to this file disables or enables the logging of that event, respectively.

Another important configuration parameter is the buffer size, set by default to 4MB. However, as the periodic flushing of the buffer to disk obviously perturbs the system, a bigger buffer is needed in scenarios where a measurement might take longer to run and the user does not want it disturbed. The */proc/sys/klogger/buffer_size* file shows the size of each CPU's buffer (in MBs). Writing a new number into that file reallocates each CPU's buffer to the requested number of MBs (if enough memory is not available an error is logged in the system log).

The last parameter we review is the low-water mark. This parameter determines when the buffer will be flushed to disk, and its units are percents of the full buffer. KLogger's logging buffer acts as an

asymmetric double buffer, where the part above the low-water mark is the main buffer, and the part below the low-water mark is the reserve buffer that is only used when during flushing. This is further explained in Section 5.2. By default, the buffer is flushed when its free space drops below 10%. In some scenarios the flushing action itself may generate events, therefore the threshold should be increased to avoid overflowing the buffer. If an overflow does occur the kernel simply starts skipping event serial numbers until space is available, allowing verification of the log's integrity. Changing the parameter's value is done by simply writing the new low-level mark (in percents) to the `/proc/sys/klogger/lowwater` file.

4.4 Internal Benchmarking Mechanism

The final part of KLogger's interface is its internal benchmarking mechanism. When designing a benchmark, one needs to pay attention to the overhead incurred by the logging itself, in order to evaluate the quality of the data collected. When KLogger generates its logging code, it also generates benchmarking code for each event that iteratively logs this event using dummy data, and measures the aggregate time using the hardware's cycle counter. The number of iterations defaults to 1000, and is settable using the `sysctl/proc` interface at runtime. The average overhead (as an integral division) for each event is reported using per-event files in the `/proc/sys/klogger/benchmarks/` directory. This estimate can then be used by the developer/researcher to evaluate whether an event incurs an intolerable overhead, in which case it can be simply disabled at runtime with no need to recompile the kernel.

The calculated average overhead gives a good estimate for the overhead incurred by KLogger, but it is important to remember that this overhead only accounts for the time spent *logging* the information, but not the time spent *obtaining* the real information to be logged from kernel data structures, or even directly from the hardware. Estimating the overhead of the latter is not feasible within the KLogger framework, and is left for the user to cautiously evaluate.

Finishing with our *SCHEDIN* example, the event's logging overhead takes less than 200 cycles on our 2.8GHz Pentium IV machine — or **70 nanoseconds**. In fact, we have found this value to be typical for logging events containing up to 32 bytes ($8 \times 32\text{bit}$ integers). The overhead incurred *obtaining* the information in this case cannot be neglected, and is mainly attributed to reading the number of cache misses from the hardware's performance monitoring counters — measured at another 180 cycles. Nevertheless, this measurement is a demonstrates KLogger's low logging overhead.

5. KLOGGER IMPLEMENTATION

In this section we discuss the details of KLogger's implementation and how its design principles — mainly the low overhead and flexibility — were achieved.

5.1 Per-CPU Buffers

As noted in previously, KLogger's buffer operates as an asymmetric double buffer, with the low-water mark separating the main buffer from the reserve, flush time, buffer.

KLogger employs per-CPU, logically contiguous, memory locked buffers. In this manner allocating buffer space need not involve any inter-CPU locks, but only care for local CPU synchronization (as opposed to the physical memory buffer used in [9]). On a single CPU, race conditions can only occur between system context and interrupt context, so blocking interrupts is the only synchronization

construct required. In fact, since the buffer is only written linearly, maintaining a *current position* pointer to the first free byte in the buffer is all the accounting needed, and safely allocating *event_size* bytes on the buffer only requires the following operations:

1. block local interrupts
2. `event_ptr = next_free_byte_ptr`
3. `next_free_byte_ptr += event_size`
4. unblock local interrupts

Interrupt blocking is required to prevent the same space allocated to several events, since the *next_free_byte_ptr* pointer is incremented on every event allocation. Furthermore, we want to prevent the possibility that the buffer will be flushed between the event allocation and the actual event logging. As flushing requires the kernel to context switch into the kernel thread in charge of flushing the specific per-CPU buffer (Section 5.2), disabling kernel preemption during the logging operation assures reliability. This simple synchronization imposes minimal interference with the kernel's normal operation, as it only involves intra-CPU operations — allowing KLogger to be efficiently used in SMP environments¹.

Logging buffers are written-to sequentially, and only read-from at flush time. With these memory semantics caching does not improve performance, but quite the contrary: it can only pollute the memory caches. We therefore set the buffers' cache policy to *Write-Combining* (WC). This semantics, originally introduced by Intel with its *Pentium-Pro* processor [14], is intended for memory that is sequentially written-to and is rarely read-from, such as frame buffers. WC does not cache data on reads, and accumulates adjacent writes in a CPU internal buffer, issuing them in one bus burst.

5.2 Per-CPU Threads

During the boot process, KLogger spawns per-CPU kernel threads, that are in charge of flushing the buffers when the low-water mark is reached. Although the logging operation should not disturb the logged system, flushing the buffer to disk obviously does. To minimize the disturbance KLogger threads run at the highest priority under the real time *SCHED_FIFO* scheduler class. This class, mandated by Posix, has precedence over all other scheduling classes, preventing other processes from interfering with KLogger's threads.

Each thread dumps the per-CPU buffer to a per-CPU file. The separate files can be interleaved using timestamps in the events' headers, as Linux synchronizes the per-CPU cycle counters on SMP machines [4, 18].

Flushing the buffer might cause additional events to be logged, so the buffer should be flushed before it is totally full. As described above, KLogger's full buffer is split into an asymmetric double buffer by the low-water parameter. This split enables the flushing thread to safely dump the two parts of the buffer. The full four step flushing process is described in Figure 1.

To prevent the logged data from being tainted by KLogger-induced events, the log is annotated when the flush begins and finishes with two special events: *DUMP_BEGIN* and *DUMP_FINISH*. The presence of these two events in the log allows for cleaning the data from artifacts introduced by the logging function itself, further diminishing the Heisenberg effect.

¹In fact, locking interrupts is only needed because there is no atomic "fetch and add" operation in the x86 ISA. Such an op-code could have further reduced the overhead



Figure 1: The four steps of the flush operation: (1) The log buffer reaches the low-water mark and wakes up the dump thread. (2) Thread writes the data between the beginning of the buffer and the current position, possibly causing new events to be logged to the reserve part. (3) Atomically resetting the buffer’s current position (with interrupts disabled). (4) Events from the reserve part are flushed to disk, possibly causing new events to be logged at the beginning of the buffer.

When KLogger is being disabled, the KLogger threads are awakened in order to empty all CPU buffers, and only then is KLogger ready for another logging session.

5.3 Low Overhead Through Code Generation

KLogger generates specific encoding and decoding code for each user defined event, as two complementing inlined C functions.

The decision to generate specific code for each event, rather than use generic code, is motivated by the desire to reduce the overhead as much as possible. An important part of KLogger is its simple, yet powerful code generator. The generator produces specially crafted code for each event that simply allocates space on the CPU’s buffer and copies the logged data field by field. The code avoids using any taken branches or extra memory which might cause cache misses, in order to reduce the uncertainty induced by the logging action as much as possible. It is optimized for the common code path: successfully allocating space and logging the data without any branches. The resulting overhead is indeed minimal, as reported in Section 7.1. Moreover, if the event is disabled the code incurs an overhead of only a few ALU operation and one forward branch, resulting in minimal runtime interference.

Neither the code generation nor executing the event specific code requires intervention from the user — generating the code is an implicit part of the compilation process, and event logging is done using the generic *klogger* C macro which is replaced with the event-specific inlined function by the C preprocessor.

5.4 Extent of Changes to the Linux Kernel

Knowing the complexity of the Linux kernel, and the rate of code evolution therein, we have tried to make KLogger’s code self contained in its own files, non-intrusive to the kernel sources.

The full KLogger patch consists of about 4600 lines of code, of which, under 40 lines modify kernel sources, and 13 modify kernel Makefiles. The rest of the patch consists of KLogger’s own files. This fact makes KLogger highly portable between kernel versions — the same patch can apply to several minor kernel revisions.

Moreover, KLogger only uses a minimal set of kernel constructs: kernel thread creation, memory allocation, atomic bit operations, and just a few others. As such, porting it to other operating systems should be a feasible task.

5.5 KLogger Schemata

KLogger’s schemata are its most powerful mode of operation. A schema is simply a set of complementary events, that provide comprehensive coverage of a certain subsystem or issue. For example, if KLogger is set up to log all kernel interrupts, we say it is using the Interrupt Logging Schema (our own Interrupt Logging Schema is described later on). Such schemata turn KLogger into a flexible framework enabling easy instrumentation of kernel subsystems and provide a platform with which the research community can discuss and standardize the evaluation of these subsystems. This modular design enables the evaluation of separate subsystems individually, but also as a whole.

In practice, a KLogger schema is composed of one or more configuration files, and a kernel patch incorporating the necessary KLogger calls. Such a kernel patch is considered a light patch, as it just places KLogger calls in strategic locations. This combination gives KLogger schemata the power of simplicity: first, it is very easy to create new schemata, assuming one knows his way around the kernel well enough to place the KLogger calls. Second, using a schema only involves copying its configuration files and applying its kernel patch.

Even though KLogger simplifies the process of evaluating kernel subsystems, creating a new schema requires a good understanding of the subsystem at hand, as demonstrated by the process context switch example described earlier. Similar circumstances apply to almost all the kernel subsystems. For example, the network subsystem is based on a stack of protocols. A network researcher may want to study network latencies, in which case she must know when a packet was received at the Ethernet layer, submitted to the IP layer and so on, until finally the user process is scheduled and reads the payload. While this high level understanding is enough for most studies, having to find the exact places in the network subsystem code when the described events occur is an arduous task. But this task can be avoided once the proper Klogger schema exists — hopefully even created by the developer. Note that this example involves two subsystem — the network and the process scheduler — each with its own intricacies and the resulting code learning curve.

Our vision is to collect a host of schemata, created by kernel researchers and developers, incorporating their knowledge and insights. In particular, developers of new kernel facilities just need to write a schema able to log and evaluate their work. We believe such a collection can be a valuable asset for the operating system research community.

The following sections will describe some case studies utilizing a few of the basic schemata we designed, and show some interesting findings and insights we have gathered when using KLogger.

6. TESTBED

Our case studies demonstrating KLogger’s abilities were conducted on klogger-enhanced 2.6.9 and 2.4.29 Linux kernels, representing the 2.6 and 2.4 kernel series, respectively. KLogger was set to use a 128MB memory buffer, to avoid buffer flushing during measurements.

Our default hardware was a 2.8GHz Pentium IV machine, equipped with 512KB L2 cache, 16KB L1 data cache, 12Kμops L1 instruction cache, and 512MB RAM. Other hardware used is specified when relevant.

| Tool | Direct Overhead | L1 Cache Misses |
|-----------------|-----------------|-----------------|
| <i>KLogger</i> | 321±35.66 | 6.55±0.56 |
| <i>LTT</i> | 1844±1090.24 | 69.03±25.94 |
| <i>printk</i> | 4250±40.80 | 227.40±2.73 |
| <i>H/W Trap</i> | 392±1.95 | N/A |

Table 1: The mean overhead \pm standard deviation incurred by different logging facilities, measured using the *Stopwatch* schema. Direct overheads are shown in cycles, after sanitizing the worst 0.1% of the results for each measurement and subtracting the *Stopwatch* events’ overhead.

7. CASE STUDY: STOPWATCH SCHEMA

The *Stopwatch* schema defines two event types: *START* and *STOP*. As the name suggests, it is used to measure the time it takes to perform an action, simply by locating the two events before and after the action takes place. In fact, when used in conjunction with the hardware performance counters it can measure almost any type of system metric: cache misses, branch mis-prediction, and instructions per cycle (IPC), just to name a few.

7.1 Measuring Kernel Loggers

A good demonstration of *KLogger*’s flexibility is its ability to measure the overhead incurred by other logging tools. We have used three interference metrics: *direct overhead*, the number of computing cycles consumed by the logging action, and *L1* and *L2* cache misses estimating the *indirect overhead* caused by cache pollution — a well known cause of uncertainty in fine grained computation, and in operating systems in general [31].

The logging tools measured are Linux’s system log printing *printk* [18] (whose intuitive semantics *KLogger* uses), the *Linux Trace Toolkit (LTT)* [35], a well known logging tool in the Linux community, and *KLogger* itself. In order to create a meaningful measurement, we needed the logging mechanisms to log the same information, so we implemented a subset of *LTT*’s events as a *KLogger* schema. Another promising tool is Sun’s *DTrace* [8], which is an integral part of the new Solaris 10 operating system. At the time of writing, however, we did not have access to its source code. Instead, we estimated its direct overhead by measuring the number of cycles consumed by a hardware trap (which is the logging method used in the *x86* version of Solaris). A hardware trap is also at the core of Linux’s *Kprobes* tool.

Table 1 shows the results of one of the most common and simple events — checking if there is any delayed work pending in the kernel (*softirq*). This event is logged at a rate of 1000Hz in the 2.6.x Linux kernel series, each time saving just two integers to the log. To eliminate suspicious outliers we have removed the worst 0.1% ($\frac{1}{1000}$ th) of the results for each measurement. This greatly reduced the standard deviation for all measurements, as the removed samples contained extremely high values reflecting system interference. For example, the removed samples of the H/W trap measurements — which mostly contained nested interrupts — reduced the standard deviation from 265 to 1.9.

The table shows that *KLogger* incurs much less overhead than the other tools: by a factor of 5 less than *LTT*, and more than an order of magnitude for *printk*. The difference between indirect overheads is even greater (we only show *L1* misses, as *L2* misses were negligible for all tools). As for *DTrace*, while *KLogger* incurs less overhead than a single hardware trap — *DTrace*’s basic building block on

the *x86* architecture — we only see a small difference in the direct overhead. *DTrace* however, is based on a virtualized environment, so its direct overhead is assumed to be considerably greater.

8. CASE STUDY: LOCKING SCHEMA

Modern operating systems employ fine grained mutual exclusion mechanisms in order to avoid inter-CPU race conditions on SMPs [3, 27]. *KLogger*’s *locking schema* is intended to explore the overheads of using inter-CPU locks.

Fine grained mutual exclusion in Linux is done through two basic busy-wait locks: *spinlock* and *rwlock* [4, 18]. The first is the simplest form of busy-wait mutual exclusion, where only one CPU is allowed inside the critical section at any given time. The second lock separates code that does not modify the critical resource — a *reader* — from code that modifies that resource — a *writer*, allowing multiple readers to access the resource simultaneously, while writers are granted exclusive access.

The goal of the locking schema is to measure lock contention, and identify bottlenecks and scalability issues in the kernel. The schema tracks the locks by the locking variable’s memory address, and is composed of 5 events. The first two are initialization events (*RWINIT/SPININIT*) which are logged whenever *KLogger* first encounters a lock — these events log the lock’s address and name (through *C* macro expansion). The three other events — *READ*, *WRITE*, and *SPIN* — are logged whenever a lock is acquired. Each log entry logs the lock’s address and the number of cycles spent spinning on the lock. The lock’s memory address is logged to uniquely identify the lock, and to allow correlation with the kernel’s symbol table. This schema is the most intrusive as it wraps the kernel’s inlined lock functions with macros to allow for accounting. Still, its overhead is only ~10% of the cycles required to acquire a free lock (let alone a busy one).

8.1 Overhead of Locking

How many cycles are spent by the kernel spinning on locks? Very little data is known on the matter: Bryant and Hawkes [7] wrote a specialized tool to measure lock contention in the Linux kernel which they used to analyze filesystem performance [6]. Kravetz and Franke [16] focused on contention in the 2.4.x kernel CPU scheduler, which has since been completely rewritten. A more general approach was taken by Mellor-Crummey and Scott [20]. Their goal however was to measure the time for acquiring a lock in a Non-Uniform Memory Architecture (NUMA). Unrau et al. [32] extended this work for the experimental *Hurricane* operating system. Both papers did not address the overall overhead of locking on common workloads, hardware, and operating systems. Such an evaluation is becoming important with the increasing popularity of SMP (and the emerging multi-core) architectures both in servers and on the desktop.

Locking is most pronounced with applications that access shared resources, such as the virtual filesystem (VFS) and network, and applications that spawn many processes. In order to identify contended locks, we chose a few applications that stress these subsystems, using varying degrees of parallelization.

- **Make**, running a parallel compilation of the Linux kernel. This application is intended to uncover bottlenecks in the VFS subsystem. In order to isolate the core VFS subsystem from the hardware, compilations were performed both

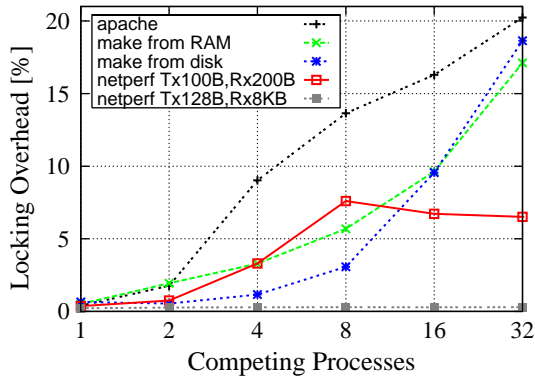


Figure 2: Percentage of cycles spent spinning on locks for each of the test applications.

on memory resident and disk based filesystems.

- Netperf**, a network performance evaluation tool. We measured the server side, with multiple clients sending communications using the message sizes in Netperf’s standard round-robin TCP benchmark — 1:1, 64:64, 100:200, and 128:8192, where the first number is the size of the message sent by the client, and the second is the size of the reply. Each connecting client causes the creation of a corresponding Netperf process on the server machine.
- Apache**, the popular web server was used to stress both the network and the filesystem. Apache was using the default configuration, serving Linux kernel source files from a RAM based filesystem. To simulate dynamic content generation (a common web server configuration), the files are filtered through a Perl CGI script that annotates the source files with line numbers. Stressing was done using the Apache project’s own *flood* tool. Its performance peaked at 117Req/s

In this case study we used the largest SMP available to us: a 4-way Pentium III Xeon processors (512KB L2 cache), equipped with 2GB of RAM. Its network interface card (NIC) is 100Mb/s Ethernet card. The stressing clients are a cluster of 2-way Pentium IV 3.06GHz machines (512KB L2 cache, 4GB RAM), equipped with 1Gb/s Ethernet cards. KLogger was set with a 128MB buffer for each of the server’s CPUs. To verify the results obtained on this somewhat aging hardware, we repeated all measurements running the box with only 2 CPUs, and compared the results with those of the modern 2-way SMP. The similarity of these results indicate that although the processors are older, the SMP behavior of the systems has not changed. For lack of space, we only show the results for the 4-way SMP hardware.

Tests consisted of running each application with different levels of parallelism — 1, 2, 4, 8, 16, and 32 concurrent processes: when N was the degree of parallelism, Make was run with the $-jN$ flag spawning N parallel jobs, while Apache and Netperf simply served N clients. During test execution KLogger logged all locking events within a period of 30 seconds. The reason for this methodology is that the kernel uses locks very frequently, generating a huge amounts of data. The 30 seconds period was set so KLogger could maximize its buffer utilization, while avoiding flushing it and interfering with the measurement.

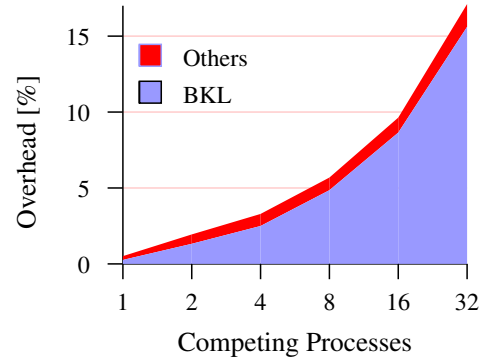


Figure 3: Portion of the cycles spend on BKL and other locks, for ramdisk-based Make.

Using the logged data, we aggregated the total number of cycles spent spinning on locks in the kernel, as percents of the overall number of cycles used. The results are shown in Figure 2.

At the highest level of parallelism, running Apache has the CPUs spend over 20% of their cycles waiting for locks, and both measurements of Make exceed 15% overhead. Netperf however, suffers from only a little more than 6% overhead — simply because the 100Mb/s network link gets saturated.

If we focus on the point of full utilization, which is at 4 competing processes for our 4-way SMP, we see that Apache loses ~9% to spinning. This is a substantial amount of cycles that the CPUs spend waiting.

The case of the Make benchmarks is especially interesting. When using a memory based filesystem vs. a disk based one, we would expect better performance from the memory based filesystem, as it does not involve accessing the slower hard disk media. But when using 4 processes, the results for both mediums were roughly the same. The answer lies in the locking overhead: while the ramdisk based Make loses just over 3% to spinning, the disk based one loses just over 1%. It appears time spent by processes waiting for disk data actually eases the load on the filesystem locks, thus compensating for the longer latencies.

The next step was to identify the bottlenecks: which locks are most contended? It seems the cause of this behavior in all but the Netperf example is just one lock — Linux’s *Big Kernel Lock* (BKL).

The BKL is a relic from the early days of Linux’s SMP support. When SMP support was first introduced to the kernel, only one processor was allowed to run kernel code at any given time. The BKL was introduced somewhere between the 2.0.x and 2.2.x kernel versions as a hybrid solution that will ease the transition from this earlier monolithic SMP support, to the modern, fine grained support. Its purpose was to serve as a wildcard lock for subsystems not yet modified for fine-grained locking. The BKL has been deemed a deprecated feature for quite some time, and developers are instructed not to use it in new code. It is still extensively used, however, in filesystem code, and in quite a few device drivers.

Figure 3 shows the portion of the BKL in the overall lock overhead for the ramdisk based Make benchmark. Results for the disk-

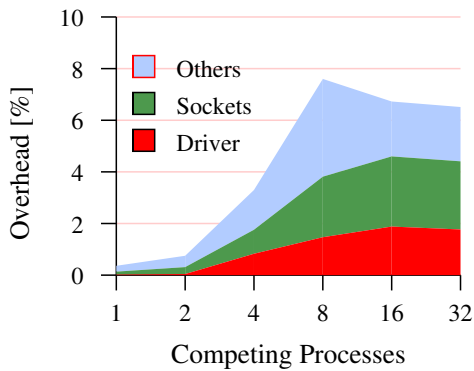


Figure 4: Overheads of different lock types for the Netperf benchmark, using the 100:200 message sizes.

based version and Apache are similar. Obviously BKL accounts for the lion’s share of the overhead, with all other locks taking no more than 2% of the overall cycles, and only roughly 0.5% in the common case. In addition, we found that the memory-based Make accesses BKL twice as often as the disk-based one.

The picture is completely different for the Netperf benchmark (Figure 4). BKL is completely missing from this picture, as both the networking and scheduling subsystems were completely rewritten since the introduction of BKL, and have taken it out of use. Instead, locking overhead is shared by the device driver lock, socket locks, and all other locks. The device driver lock protects the driver’s private settings and is locked whenever a packet is transmitted or received and when driver settings change — even when the device’s LED blinks. basically, this lock is held almost every time the device driver code is executed. In fact, it is locked more times than any other lock in the system by a factor of almost 3. The socket locks refer to *all* the sockets in the system, meaning at least the number of running Netperf processes: each Netperf process owns one socket. This figure is a rough estimate of the aggregate locking overhead caused by the networking subsystem. Both the device driver lock and the socket locks indicate the saturation of the networking link when running somewhere between 8-16 competing processes. All other locks in the system are responsible for ~33% of all the locking activity, peaking when running 8 competing processes. The majority of those cycles are spent on various process wait queues, probably related to the networking subsystem. We did not, however, find any group of locks causing the 8 process peak.

In conclusion, our measurements demonstrate the continuing liabilities caused by BKL even in the recent 2.6.9 kernel, and the harmful effects of device drivers with a questionable design. This is just a simple analysis of critical fine-grained locking mechanisms in the Linux kernel, made possible by KLogger’s low overhead. The fact that we immediately came by such bottlenecks only strengthens the assumption that many more of these performance problems are found in the kernel, but we simply lack the tools and the methodology to identify them.

9. CASE STUDY: SCHEDULER SCHEMA

The *scheduler schema* consists of 8 basic events which allow for an accurate replay of process CPU consumption. Essential information about each event is also logged. The events are:

1. *TRY_TO_WAKEUP* — some process has been awakened.
2. *REMOVE_FROM_RUNQ* — a process has been removed from the run queue.
3. *ADD_TO_RUNQ* — a process has been added to the run queue.
4. *SCHEDOUT* — the running process has been scheduled off a CPU.
5. *SCHEDIN* — a process has been scheduled to run.
6. *FORK* — a new process has been forked.
7. *EXEC* — the *exec* system call was called
8. *EXIT* — process termination.

Using KLogger, creating these events is technically very easy. However, designing this schema and the data it logs requires in-depth knowledge about the design and behavior of the Linux CPU scheduler, as described above.

9.1 Evaluating the Scheduler’s Maximal Time Quantum

KLogger’s scheduling schema can be used to empirically evaluate aspects of the Linux scheduler’s design. The maximal CPU timeslice is an example of a kernel parameter that has changed several times in the past few years. It was 200ms by default in the 2.2.x kernels. The 2.4.x kernels set it to a default 60ms, but it could be changed in the 10–110ms range based on the process’s dynamic priority. Today, the 2.6.x kernels set it to a value in the range of 5–800ms based on *nice* (the static priority), with a 100ms default when *nice* is 0, which it nearly always is. When searching the Linux kernel mailing list we failed to find any real reasoning behind these decisions.

An interesting question is whether these settings matter at all. We refer to an *effective quantum* as the time that passed from the moment the process was chosen to run on a processor, until the moment the processor was given to another process (either voluntarily or as a result of preemption). In this case study, we wish to determine whether the effective quanta of various common workloads correspond to the maximum quantum length.

Using KLogger’s scheduling schema, determining the adequacy of the maximum quanta is very simple. The applications were chosen as representatives of a few common workloads:

- **Multimedia** — Playing a 45 second MPEG2 clip using the multithreaded *Xine* and the single threaded *MPlayer* movie players. Both players are popular in the Linux community, with *xine*’s library being the display engine behind many other movie players.
- **Network** — Downloading a ~30MB kernel image using the *wget* network downloader.
- **Disk Utilities** — Copying a 100MB file, and using *find* to search for a filename pattern on a subtree of the */usr* filesystem.
- **Computation+Disk** — Kernel compilation.
- **Pure Computation** — A synthetic CPU-bound program, continuously adding integers and never yielding the CPU voluntarily, running for 60 seconds.

Measurements were run with the default *nice* value, meaning a 100ms maximum time quantum on the 2.6.9 Linux kernel. We ran

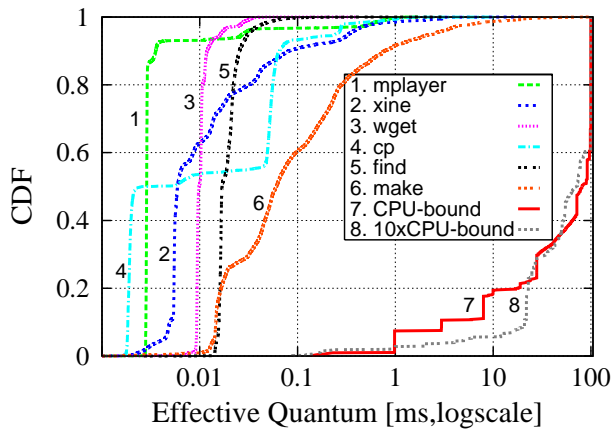


Figure 5: Cumulative distribution function (CDF) of the effective quanta for the different workloads. Quantum length is 100ms. X axis is logarithmic.

each application on a dedicated machine, competing only with the default system daemons. The results are shown in Figure 5.

Let us first discuss the synthetic CPU-bound application: even this synthetic workload only reaches the maximum quantum in ~40% of its runs, with a similar percentage not even reaching half the maximal quantum. Thus ~60 % of the quanta were shortened by some system daemon waking up — the only background load in the system. These interruptions hardly consumed any CPU: only 0.00675% of the total run. They occurred at an average rate of 5 times per second. With the maximal quantum set at 100ms, at most 50% of the quanta should have been affected, contrary to the results displayed in figure 5 which show that 60% were affected. The explanation is simple: an interrupted quantum is split into at least two effective quanta (a quantum can be broken down more than once), so the effect of the noise generated by system applications is actually amplified.

As for the other workloads, it is clear that the maximum quantum is almost a theoretical bound that is never reached: ~90% of the effective quanta of all applications but Make and the CPU-bound are shorter than 100 μ s — a thousandth of the maximum quantum. The kernel Make is an exception, with its 90th percentile lying at 1ms (this is still a negligible 1% of the maximum quantum). In fact, if not for the logarithmic scaling of the X axis we would not have been able to show any differences.

Our conclusion is that although required to prevent starvation, the actual length of the time quantum has little importance in modern systems. The only workload affected by it is CPU-bound. It would need to be shortened by more than 100 to affect other application types (regardless of what the effect would actually be), but as Linux currently uses a 1000Hz clock (on the x86 architecture) it cannot support a sub-millisecond quantum. Lengthening the maximum time quantum on CPU servers in an attempt to reduce the context switch overhead (measured using KLogger to be 3608 \pm 1630 cycles and 140 \pm 38 L1 misses) is also futile in light of the scheduling noise generated by system daemons. This is an example of how a consistent use of logging tools such as KLogger by kernel developers can help make more informed decisions about parameter settings, adapting them to common workloads.

9.2 How Adding Load Can Actually Improve Performance

The advent of chip multiprocessors (CMP) and symmetric multi-threading (SMT) has raised the question whether modern general purpose operating systems are capable of adequately handling the resulting increase in software parallelism [12]. To evaluate the adequacy of the scheduler to such workloads we used the Xine multithreaded movie player and the synthetic CPU-bound stressors described in the previous section.

In our experiment, we ran Xine alongside an increasing number of stressors on a 4-way, 550MHz Pentium III machine. The machine was chosen so that together, Xine and the X server (the two applications involved in displaying the movie) consume less than 100% of a single CPU's cycles. However, since they are both prioritized by the scheduler as interactive applications, assigning them to the same CPU causes them to compete with each other every time a frame is to be displayed — potentially resulting in failure to display the frame on time. In contradistinction, when assigned to a CPU alongside a stressor, their identification as interactive yields a priority boost over a competing stressor.

The results of this experiment are shown in Figure 6. This figure shows that using the default Linux scheduler Xine's performance degrades as the load increases. Surprisingly though, this degradation is not monotonic in the number of CPU stressors — we observe temporary performance *loss* of Xine when running 3 stressors, whereas running 10 stressors results in temporary performance *gain* (this is consistent in repeated runs). Using KLogger's scheduler schema we analysed the migration patterns on all 4 CPUs, and how each CPU's cycles were allocated to the different processes.

When running less than 3 stressors, both Xine and the X server get a dedicated CPU each. When running 3 stressors however, the scheduler attempts to balance the load by pushing both Xine and X onto the same processor — as each consumes less than 100% of a CPU's cycles — making them compete against each other, and leading to degraded performance. When running 4-7 stressors the scheduler is unable to balance the load and again separates Xine and X, letting each compete with a different stressor; but since it considers both Xine and the X server to be interactive, they both receive an interactive priority gain, giving them priority over the co-located stressors. Note that while intuitively running 7 stressors may seem similar to running only 3, that is not the case from the scheduler's point of view: the scheduler sees all CPUs engaged in computation and that some imbalance is necessary. The result is that Xine and the X server are not pushed to the same CPU, but rather compute on different CPUs with one or two stressors — over which they have priority. Conversely, when running 10 stressors the same imbalance leads to a positive effect when Xine's threads and the X server align on different CPUs, achieving a much better performance than expected.

This case study exemplifies how KLogger can uncover migration patterns in multiprocessing environments, explaining an unexpected behaviour that at a first glance may indicate that increasing the load on a multiprocessor may actually improve performance. In fact, this result suggests that Linux's process scheduler might be inadequate for non-trivial parallel workloads.

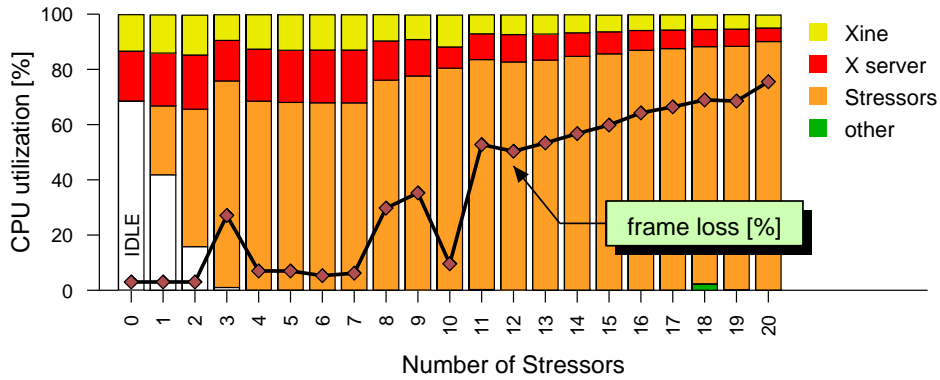


Figure 6: Aggregate CPU percentage utilized by each application, alongside the frame loss experienced by Xine.

9.3 When 8 Competitors Are (Slightly) Better Than 1

During our work we needed to evaluate the effects of multiprogramming on the overall throughput of the computation. Our testbed was the Linux 2.4.29 kernel, and the benchmark we used was a single threaded program sorting an integer array whose size is one half the L2 cache. Our throughput metric is simple: how many times was the array sorted during a specified time frame? (the array was reinitialized to the same random values after each sort). We expected that this CPU-bound benchmark would achieve lower aggregate throughput if we ran several competing copies of it, since that would require the operating system to spend CPU time on context switching (with its corresponding cache pollution).

Our results, however, showed that throughput improved slightly with more processors, and peaked at 8 — a ~0.3% improvement. In fact, this slight difference almost tempted us to dismiss it, but since it was consistent we decided to check whether KLogger can help explain this discrepancy. Using the scheduler schema, we measured the CPU scheduling overhead, only to find it has a U shape (Figure 7). In particular, the total time spent on context switches (accumulating the time between all *SCHEDOUT* events and their immediately following *SCHEDIN* events) was much greater for the single process case than for the 8 process case: 30.2ms vs. 13.24ms respectively.

Unearthing the reason for this required a careful examination of the kernel’s scheduling code. The 2.4.x scheduler linearly iterates over all the runnable processes to choose the one with the highest priority. If no runnable process exists, the scheduler iterates over all existing processes, recalculating their CPU timeslice [4]. When running the benchmark with a single process this recalculation took place at almost every scheduling point. This is inefficient, as it considers dozens of system daemons which are dormant most of the time. With more user processes the frequency of these recalculations was decreased, saving much overhead (even though the number of cycles consumed by each recalculation increased). On the other hand, the time to recalculate priorities, and the time to select a runnable process became longer — but these only grew enough to dominate the overhead at more than 8 processes — leading to 8 being the sweet spot.

Both the recalculation and the process selection loops were eliminated from the kernel as part of a complete scheduler redesign [18]

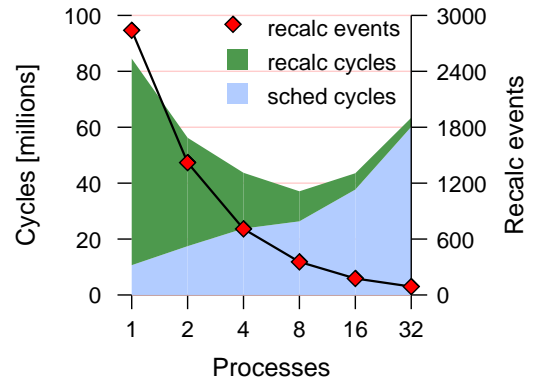


Figure 7: Analysis of the scheduling overhead into its components: choosing a new process to run (bottom) and recalculating priorities. Also shown are the number of recalculations for each run.

between 2.4.x and 2.6.x versions. Still, this case study serves as a good example of how KLogger was used to understand extremely fine inconsistencies in results which were initially attributed to common system noise. Using KLogger we were able to correctly link these results to a specific design issue.

10. CASE STUDY: INTERRUPT SCHEMA

KLogger’s *interrupt schema* measures the start and finish of all interrupts in the system, including IRQs, traps, and exceptions, as well as L1 and L2 cache misses caused by the operating system code.

10.1 Operating System Noise

Noise caused by the operating system is becoming a growing concern. Interrupts, scheduling, TLB and cache contention are all causes for computational uncertainty, affecting multimedia and HPC applications [13, 24, 31].

In order to characterize interrupt noise we designed a synthetic application, based on a calibrated loop taking 1ms on average. The only memory activity is reading the counting index from the stack, incrementing it, and writing it back to the stack. This loop is repeated 1,000,000 times, keeping track of the number of CPU cycles consumed by each repetition. We ran the application on a klogger

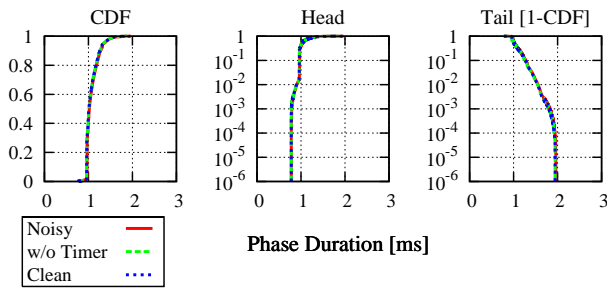


Figure 8: CDF of the iteration times for a loop calibrated to take 1ms, running with the FIFO scheduler (left), and zooming in on the the head (center) and the tail (right) of the distribution. Head and tail are shown using a log scaled Y axis. The graphs illustrate the raw intervals (“Noisy”), the intervals after removing the direct overhead of the timer ticks (“w/o Timer”), and the intervals with all interrupts’ direct overhead removed (“Clean”).

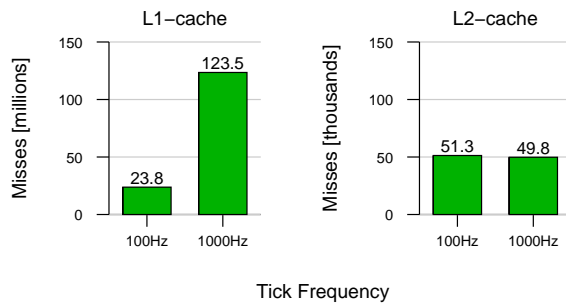


Figure 9: Number of L1 and L2 cache misses, when running a 2.6.9 Linux kernel with 100Hz and 1000Hz timer frequency.

enabled 2.6.9 Linux kernel under the Posix FIFO scheduler, so the only operating system noise that can disrupt the application is hardware interrupts.

Figure 8 shows a CDF of the repetition’s times, zooming in on the head and tail. The figure shows that over 40% of the repetitions exceed 1ms, and about 1% of them even exceed 1.5ms, reaching a maximum of 2ms. When examining the head we notice that more than $\frac{1}{1000}th$ of the iterations took less than $800\mu s$. The meaning of this is that running a specific piece of code can vary in time by a factor of over 2.5 (0.78ms vs. 1.96ms).

The only interrupts that occurred during the measurements were the timer and network interrupts. As KLogger and the application use the same cycle counter, we can identify repetitions that included specific interrupts and subtract them. However, Figure 8 also shows us that removing the direct overhead of these interrupts did not affect the measurement. Where did the cycles go, then?

The solution is apparent when measuring the cache misses caused by the interrupts. Figure 9 shows the number of cache misses caused by interrupts, when running the timer at both 100Hz and 1000Hz (100Hz is standard in the 2.4.x kernel), for perspective. It is clear that the number of cache misses caused by interrupts increases significantly with the increase in timer frequency, suggesting cache misses might cause the 1ms loop overhead. And in-

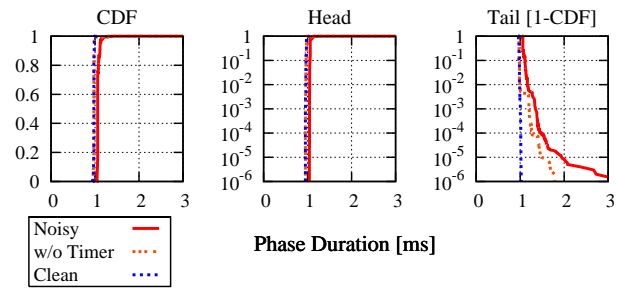


Figure 10: The measurements of Figure 8 repeated with caches disabled.

deed, when repeating the previous measurements with both the L1 and L2 caches disabled (Figure 10), subtracting the direct overhead leads to consistent measurements, indicating that the variability in the original measurements resulted from indirect overhead due to cache interference.

Identifying system noise is becoming a real problem for parallel and distributed computing [24]. This case study shows how KLogger’s tight coupling with the underlying hardware can be used to pinpoint the computational noise generated by common operating system interrupts.

10.2 Keeping Time in the Kernel

Operating systems keep track of time using the standard 8253 *programmable interrupt timer* (PIT). PIT has been used with several generations of processors for over 10 years.

In principle, whenever the kernel needs the wall clock time, it can simply access the 8253 through its I/O bus and read the data. This is done from the *do_gettimeofday* kernel function (of which the *_gettimeofday* system call is just a wrapper). Reading the time from the 8253 PIT is a relatively expensive operation, so Linux is optimized (on machines which have a hardware cycle counter) to accesses the 8253 on every timer interrupt, and interpolate using the cycle counter in *do_gettimeofday*. Accessing the hardware’s cycle counter is much faster than accessing the 8253 PIT, so this mode of operation limits the overhead incurred by the PIT to the number of timer interrupts per second. The two modes, common to both the 2.4.x and the 2.6.x kernel series, are called *PIT* mode and *TSC* mode.

Using KLogger’s interrupt schema we have measured the overhead of the timer interrupt handler in both modes, on various generations of Intel processors. The results (originally presented in [11]) are shown in Figure 11. When running the kernel in PIT mode, the timer interrupt handler does not access the 8253 PIT. It consumes roughly the same number of cycles over all the hardware generations, so its μs overhead decreases as the hardware speed increases. When running the kernel in TSC mode, however, the 8253 is accessed from the timer interrupt handler. As access time to the PIT has not changed over the years, the time consumed by the handler remain roughly the same, and the number of cycles actually grows with the increase in CPU speed [22].

Given that TSC mode is the default, the timer interrupt handler is in fact becoming a liability — the more so as the timer interrupt rate increases (2.4 used 100Hz, whereas 2.6 uses either 100Hz, 250Hz, or 1000Hz). The TSC optimization, aimed at reducing the over-

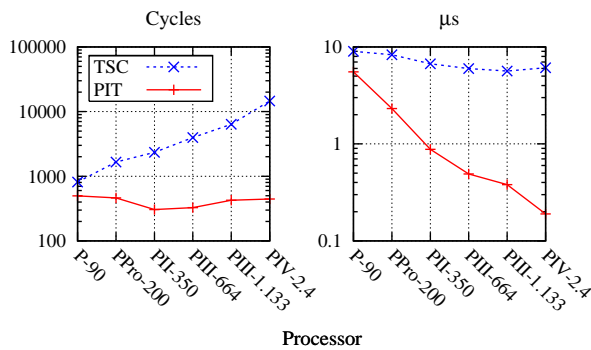


Figure 11: Overhead of the timer interrupt when using the TSC and PIT modes. Note that the Y axis is log scaled.

head of the *gettimeofday* system call, is actually workload dependent. It only helps for workloads that call *gettimeofday* at a higher frequency than the kernel’s timer frequency. The solution should be accessing the 8253 PIT on demand, but only if it was not accessed recently, and interpolating if it was.

These results demonstrate why measurements and tools are needed for kernel development. One kernel developer changed something in the kernel (the timer frequency) but is unaware of its effect on another mechanism (the *gettimeofday* optimization). A simple performance measurement tool such as KLogger can help uncover such cases, allowing for more informed design decisions.

11. CONCLUSIONS

We have presented *KLogger*, a low overhead, fine grained logging tool for the Linux kernel. *KLogger*’s prime objective is to help analyze kernel behavior, and help researchers and developers understand what is really happening under the operating system’s proverbial hood. Such support is required due to the increasing number of developers working on the Linux kernel, and situations in which modules are inter-dependent in unexpected ways.

Making efficient use of the underlying hardware features, *KLogger* is able to achieve much finer granularity and lower overheads than any other kernel logging tool. *KLogger*’s fine granularity and flexibility enables it to be used in the tightest corners of the kernel, and shed light on the operating system’s nuts and bolts. Moreover, *KLogger* allows developers to create subsystem-specific logging schemata that can be used out-of-the-box by others. Another of *KLogger* uses is for kernel debugging. Although not discussed in the case studies, it is also a very efficient debugging tool.

Using *KLogger* and its schemata in our research has helped us understand some interesting and sometime unexpected phenomena in the kernel. These case studies, which are the bulk of this paper, both demonstrate the tool’s abilities, and more importantly suggest some major design problems in the Linux kernel. We have shown how locking issues can seriously limit the kernel’s ability to handle SMP environments, and how both its scheduler and timing services’ parameters are less than optimal for modern hardware.

Kernel developers and the research community should aspire to better understand the operating system kernels’ intricacies. We hope a tool such as *KLogger* would be integrated into operating system development process by having developers write performance ana-

lyzing schemata for the subsystems they code. *KLogger*, its manual and the schemata described in this paper are available for download at www.cs.huji.ac.il/labs/parallel/klogger. It is our hope kernel researchers and developers will use this tool and create schemata for other subsystems — such as the filesystem, network, and others — through which we can all share our insights about the operating system’s kernel operation.

KLogger is currently used by a few of our research colleagues, who provide us with feedback about its interface and capabilities. The reviews so far are very encouraging.

12. REFERENCES

- [1] J. M. Anderson, W. E. Weihl, L. M. Berc, J. Dean, S. Ghemawat, M. R. Henzinger, S.-T. A. Leung, R. L. Sites, M. T. Vandevoorde, and C. A. Waldspurger. Continuous profiling: where have all the cycles gone? In *ACM Symp. on Operating Systems Principles*, pages 1–14, 1997. ACM Press.
- [2] A. C. Arpaci-Dusseau and R. H. Arpaci-Dusseau. Information and control in gray-box systems. In *ACM Symp. on Operating Systems Principles*, 2001. ACM Press.
- [3] D. L. Black, A. Tevanian, D. B. Golub, and M. W. Young. Locking and reference counting in the Mach kernel. In *Intl. Conf. on Parallel Processing*, volume 2, pages 167–173, 1991.
- [4] D. P. Bovet and M. Cesati. *Understanding the Linux Kernel*. O’Reilly & Associates, Sebastopol, CA, USA, 2001.
- [5] Z. Brown. What’s new in kernel development. *Linux Journal*, page 10, Mar 2005.
- [6] R. Bryant, R. Forester, and J. Hawkes. Filesystem performance and scalability in Linux 2.4.17. In *Usenix Annual Technical Conf. (FreeNix Track)*, 2002.
- [7] R. Bryant and J. Hawkes. Lockmeter: Highly-informative instrumentation for spin locks in the Linux kernel. In *4th Annual Linux Showcase & Conf.*, 2000.
- [8] B. M. Cantrill, M. W. Shapiro, and A. H. Leventhal. Dynamic instrumentation of production systems. In *Usenix Annual Technical Conf.*, June 2004. The Usenix Association.
- [9] J. Casmira, D. Kaeli, and D. Hunter. Tracing and characterization of NT-based system workloads. *Digital Tech. J.*, 10(1):6–21, Dec 1998.
- [10] Y. Endo and M. Seltzer. Improving interactive performance using TIPME. In *Intl. Conf. on Measurement & Modeling of Computer Systems*, pages 240–251, June 2000. ACM Press.
- [11] Y. Etsion, D. Tsafirir, and D. G. Feitelson. Effects of clock resolution on the scheduling of interactive and soft real-time processes. In *Intl. Conf. on Measurement & Modeling of Computer Systems*, pages 172–183, Jun 2003. ACM Press.
- [12] E. Frachtenberg and Y. Etsion. Hardware parallelism: Are operating systems ready? (case studies in mis-scheduling). In *Workshop on the Interaction between Operating System and Computer Architecture*, pages 7–15, June 2006. <http://www.ideal.ece.ufl.edu/workshops/wiosca06/final.pdf>.
- [13] R. Gioiosa, F. Petrini, K. Davis, and F. Lebaillif-Delamare. Analysis of system overhead on parallel computers. In *IEEE Intl. Symp. on Signal Processing and Information Technology*, Dec 2004. IEEE Computer Society Press.
- [14] Intel Corp. *IA-32 Intel Architecture Software Developer’s Manual. Vol. 3: System Programming Guide*.
- [15] M. B. Jones and J. Regehr. The problems you’re having may not be the problems you think you’re having: Results from a

- latency study of windows NT. In *7th Workshop on Hot Topics in Operating Systems*, page 96, March 1999. The Usenix Association.
- [16] M. Kravetz and H. Franke. Enhancing the Linux scheduler. In *Ottawa Linux Symp.*, 2001.
- [17] J. Lions. *Lions' Commentary on UNIX 6th Edition*. Annabooks, 1996. (reprint).
- [18] R. Love. *Linux Kernel Development*. Novell Press, Indianapolis, IN, USA, 2nd edition, 2005.
- [19] R. Love, L. Torvalds, A. Cox, and various kernel developers. LKML Thread: Improving interactivity. <http://kerneltrap.org/node/603>, Mar 2003.
- [20] J. M. Mellor-Crummey and M. L. Scott. Algorithms for scalable synchronization on shared-memory multiprocessors. *ACM Trans. on Computer Systems*, 9(1):21–65, Feb 1991.
- [21] S. W. Melvin and Y. N. Patt. The use of microcode instrumentation for development, debugging and tuning of operating system kernels. In *Intl. Conf. on Measurement & Modeling of Computer Systems*, pages 207–214, 1988. ACM Press.
- [22] J. K. Ousterhout. Why aren't operating systems getting faster as fast as hardware? In *Usenix Annual Technical Conf. (Summer)*, Jun 1990. The Usenix Association.
- [23] P. S. Panchamukhi. Kernel debugging with Kprobes. <http://www-106.ibm.com/developerworks/linux/library/l-kprobes.html>, Aug 2004.
- [24] F. Petrini, D. J. Kerbyson, and S. Pakin. The case of missing supercomputer performance: Achieving optimal performance on the 8,192 processors of ASCI Q. In *Supercomputing*, Nov 2003.
- [25] M. Rosenblum, E. Bugnion, S. Devine, and S. A. Herrod. Using the simos machine simulator to study complex computer systems. *ACM Trans. on Modelling & Computer Simulation*, 7(1):78–103, 1997.
- [26] M. Rosenblum, E. Bugnion, S. A. Herrod, E. Witchel, , and A. Gupta. The impact of architectural trends on operating system performance. In *ACM Symp. on Operating Systems Principles*, 1995. ACM Press.
- [27] C. Schimmel. *UNIX Systems for Modern Architectures*. Addison Wesley, Reading, MA, USA, 1994.
- [28] B. A. Schroeder. On line monitoring: A tutorial. *IEEE Computer*, 28(6):72–78, June 1995.
- [29] D. A. Solomon and M. E. Russinovich. *Inside Windows 2000*. Microsoft Press, Redmond, WA, USA, 3rd ed., 2000.
- [30] A. Tamches and B. P. Miller. Fine-grained dynamic instrumentation of commodity operating system kernels. In *3rd Symp. on Operating Systems Design & Impl.*, Feb 1999. The Usenix Association.
- [31] J. Torrellas, A. Gupta, and J. Hennessy. Characterizing the caching and synchronization performance of a multiprocessor operating system. In *Arch. Support for Programming Languages & Operating Systems*, pages 162–174, Oct. 1992. ACM Press.
- [32] R. Unrau, O. Krieger, B. Gamsa, and M. Stumm. Experiences with locking in a NUMA multiprocessor operating system kernel. In *Symp. on Operating Systems Design & Impl.*, Nov 1994. The Usenix Association.
- [33] Various Kernel Developers. LKML Thread: My thoughts on the "new development model". <http://www.lkml.org>, Oct 2004.
- [34] R. W. Wisniewski and B. Rosenburg. Efficient, unified, and scalable performance monitoring for multiprocessor operating systems. In *Supercomputing*, Nov 2003. IEEE Computer Society.
- [35] K. Yaghmour and M. R. Dagenais. Measuring and characterizing system behavior using kernel-level event logging. In *Usenix Annual Technical Conf.*, June 2000. The Usenix Association.
- [36] T. Zanussi, K. Yaghmour, R. W. Wisniewski, M. Dagenais, and R. Moore. An efficient unified approach for trasmitting data from kernel to user space. In *Ottawa Linux Symp.*, Jul 2003.