Enforced Deterministic Program Execution in the Linux Kernel

Chris Cotter
University of Texas

Abstract

This thesis is about designing and implementing a deterministic programming model in Linux based on Aviram et al.'s prior work, Determinator. Written from the ground up, the Determinator operating system enforces deterministic process execution by removing implicit sources of nondeterminism (e.g., data races) and enforcing strict synchronization rules that permit only "naturally deterministic" scheduling abstractions. Guided by Determinator's principles, we modify Linux's process model and kernel interface to enforce determinism on user programs, while maintaining backwards compatibility for legacy Linux applications. We implement a basic user library for writing deterministic applications and extend Determinator's in-memory file system by adding new features and giving it persistence. Evaluations of compute-bound deterministic applications against nondeterministic equivalents reveal unacceptable overheads for small inputs; for large inputs, the overhead drops to less than $2\times$ and the benchmarks begin to scale reasonably well.

1 Introduction

As processors move from single to multiple cores, more and more applications are written parallel. Today, the dominant parallel programming model is nondeterministic. In this model, threads typically share an entire address space, file system, and other globally visible system managed resources like process IDs. Operating systems schedule threads arbitrarily, and lock abstractions are neither deterministic nor predictable. This model is popular, because threads can operate on shared data "in-place" instead of having to pack and unpack data structures [5]. Unfortunately, this model is error prone and has many drawbacks [2, 25, 27]. Programmers must eliminate data races introduced by nondeterminism. Without repeatability, debugging, testing, and ensuring software quality assurance become difficult.

Existing attempts to provide determinism require special hardware [15, 16], custom programming languages [12], or specialized compilers [7]. Record-and-replay systems [24, 29] incur high overhead and do not offer any insight into the inherently nondeterministic behavior of a program. Systems that rely on a deterministic scheduler residing in user space are often arbitrary and unpredictable [32]; buggy or malicious application code can compromise the scheduler [5, 7, 14]. Some relaxed deterministic models only enforce determinism on synchronization and permit unsynchronized low-level memory accesses [32], thus leading to nondeterministic traces in programs that do not properly protect critical sections.

To overcome the challenges of nondeterminism, Aviram et al. presented a deterministic operating system called Determinator [5]. Programmers write parallel applications using a novel parallel programming model that is "naturally and pervasively deterministic" and in fact predictable. By enforcing determinism at the system level, Determinator runs programs written in general-purpose languages on conventional hardware. Determinator builds a high-level library that provides familiar parallel abstractions and an in-memory file system. Evaluations of Determinator against Linux show that such a model can be implemented to run coarse-grained parallel applications efficiently with little overhead, but fine-grained parallel applications have unacceptable overhead.

This thesis is about adapting Determinator's operating system design and programming model to Linux. We make the following contributions:

- A presentation of a deterministic Linux kernel heavily based on the Determinator kernel. This is the first known adaptation of Determinator's kernel design in a widely deployed operating system. The bulk of the work in this area is implementation of an existing model. Additionally, we identify and eliminate Linux-specific sources of nondeterminism and present the *hybrid process model*, an enhancement of Determinator's *nested workspace model*. Then end result is that we can run legacy nondeterministic applications alongside deterministic applications in Linux.
- A deterministic high-level user library for Linux applications. This library is motivated by Determinator's user library, and again the bulk of the work in this area is implementation.
- An improvement of Determinator's in-memory file system. Determinator's file system maps a fixed number of files to static locations in memory and does not provide persistence. Our file system is modeled on the BSD Fast File System [28]; thus, we support features like hard-linking and dynamic allocation of data blocks. We also provide persistence by saving the in-memory file system to Linux's disk-backed file system.
- An evaluation of deterministic Linux against traditional nondeterministic parallel Linux. We ran compute-bound parallel benchmarks and found that

small input sizes incur unacceptable overheads of up to $39.2\times$ for the smallest input size. However, the overheads drop to less than $1.3\times$ for the largest inputs when run on multiple processors. We also demonstrate a case study of the qualitative benefits of determinism by studying a buggy parallel Gaussian elimination program.

Aviram et al. were motivated by meeting the "software development, debugging, and security challenges" of writing future parallel applications [5]. Determinator was a huge step towards this. Unfortunately, Determinator has limited uptake outside the academic community. Linux is a mature and widely deployed operating system available for desktops, servers, mobile, and embedded systems. Adding determinism alongside nondeterministic Linux will be a huge next step. If we are lucky, we might be able to influence how future parallel applications are written.

2 Background

With an understanding of the goal of this thesis, we will now discuss the benefits of deterministic execution. Then, we will present the Determinator kernel and user library design.

2.1 Benefits of determinism

Determinism provides many benefits to application developers [8, 11, 32]. Bergan et al. suggest there are four main benefits in the following areas: debugging, fault-tolerance, testing, and security.

Debugging Debugging multithreaded programs can be difficult, because often bugs are not easily reproducible. Tools such as `gdb` are not always useful for tracking down heisenbugs [30]. Finding a bug's root cause becomes easier when a program's execution can be replayed.

Fault-tolerance Fault tolerance through replication relies on the assumption that running a program multiple times will always return the same output. Repeatability is a natural benefit of determinism.

Testing The difficulties in testing multithreaded applications are compounded by racy nondeterministic scheduling. Developers and automated test systems must consider the exponential blow up of possible scheduling sequences. Determinism helps alleviate this problem, since for each input, there is exactly one possible logical scheduling sequence of threads. This observation eliminates the need to consider what scheduling interactions can occur and ultimately helps developers design test strategies [8].

Since schedule sequences may still be arbitrarily deterministic, developers may still have a hard time designing test suites. Predictable programming models like Determinator allow developers to reason about code beforehand to design a more sophisticated testing strategy.

Security Processes sharing a CPU should be conscious about leaking sensitive data. A malicious thread can exploit covert timing channels to extract sensitive data from other, perhaps privileged, threads [4]. Determinism eliminates covert timing channels, because a program is purely a function of explicit inputs and cannot possibly rely subtly on the timings of hardware operations.

Whereas individual tools, like record-and-replay debuggers aid programmers in single areas, these so called "point solutions…do not compose well with one another," either interfering with each other's effectiveness or degrading performance [8]. On the other hand, determinism provides benefits in all four areas with a single mechanism without any overhead besides that inherent in the deterministic environment itself.

2.2 Determinator

Aviram et al. set out to provide

> a parallel environment that: (a) is "deterministic by default," except when we inject nondeterminism explicitly via external inputs; (b) introduces no data races, either at the memory access level or at higher semantic levels; (c) can enforce determinism on arbitrary, compromised or malicious code for security reasons; and (d) is efficient enough to use for "normal-case" execution of deployed code, not just for instrumentation during development. [5]

To this end, they presented Determinator, a novel OS written from the ground up. For the remainder of this section, we will recapitulate Aviram et al.'s work and contributions. First we will discuss aspects that influenced Determinator's design. Then, we will look at the actual kernel design itself and the accompanying user library.

The primary cause of nondeterminism is data races introduced by timing dependencies. Each source of implicit nondeterminism must be accounted for in designing a deterministic programming model. We discuss them here, and describe how Determinator handles them.

Explicit nondeterminism Often, programs rely on semantically relevant nondeterministic inputs such as network packets, user input, or clock time. A deterministic programming model must incorporate these inputs while still enforcing determinism. Determinator addresses these inputs by turning them into explicit I/O [5]. Applications have complete control over these input sources and can log the inputs for reply debugging.

Shared program state Traditional multithreaded programming models provide shared state. The popular pthreads and OpenMP APIs run parallel threads that share memory, and all Linux processes concurrently read and write to a shared file system. Even with proper synchronization on low-level memory accesses, threads still execute nondeterministically and high-level bugs may persist [2].

Determinator eliminates data races caused by shared program state by eliminating shared state altogether. In the *private workspace model*, threads only have access to private memory. When two or more threads begin executing, each has identical private virtual memory images. Threads cannot see each others' writes until explicit synchronization points. Data races become impossible: read/write races no longer exist, since only one thread may access a particular memory cell, and write/write conflicts are detected by the kernel at synchronization. [5].

Nondeterministic scheduling abstractions Traditional multithreaded synchronization abstractions are often neither deterministic nor predictable. Random hardware races determine the next thread to acquire a mutex lock. Some deterministic schedulers synthesize an arbitrarily repeatable acquisition. However, small perturbations in input can lead to radically different schedules. This approach *manages* nondeterminism instead of *removing* it.

Determinator restricts itself to only allow naturally deterministic and predictable synchronization abstractions such as fork/join [31]. In the fork/join paradigm, a main thread *forks* children threads that perform some computation. In the *join* stage, the main thread waits (in a program defined order) for each thread to finish and gathers the results.

Globally shared namespaces Operating systems and library System APIs introduce nondeterminism by using namespaces that are shared by processes (kernel APIs) or threads (library APIs). Process IDs returned by `fork()` and files created by `mktemp()` are examples. In each case, it is the resource itself and not the identifier that is semantically relevant to the program. A program does not care what process ID the OS assigns to a child created by `fork`, only that *some* child was created. Determinator disallows the system from choosing resource identifiers from globally managed namespaces and instead requires that applications choose identifiers deterministically. For example, user programs must specify the child process ID to `fork()`.

2.3 The Determinator kernel

Determinator organizes processes in a hierarchical, nested process model [5, 18]. To differentiate between the notion of processes and threads, the Determinator kernel calls all executable "tasks" *spaces*. The user library (described in the next section) uses kernel spaces to create the process and thread abstraction, but the kernel is unaware of the distinction. Spaces cannot outlive their parent and can communicate only with their parent and children. The kernel adopts a *private workspace model* and "provides no file systems, writable shared memory, or other abstractions that imply globally shared state." Only "the distinguished root space has direct access to nondeterministic inputs" [5]. It is this root space that can control explicitly nondeterministic inputs like network packets. All other spaces must communicate directly or indirectly with the root space to access I/O devices.

Kernel interface Spaces communicate with the kernel via three syscalls: `Put`, `Get`, `Ret`. Tables 1 and 2, reproduced from [5], summarize how the syscalls work and the options available to `Put` and `Get`.

Since the kernel does not manage any global namespaces, user programs specify a child space ID parameter to `Put` and `Get`. The first `Put` syscall with a previously unused child ID creates a new child space. Determinator enforces a deterministic schedule by requiring programs to explicitly define synchronization points. Parents start children by issuing a `Put` with the `Start` option. Subsequent calls to `Put` and `Get` block until the child stops. The child executes until it issues a `Ret`. Processor exceptions (e.g., divide by zero) generate an implicit `Ret` that must be acknowledged and handled by the parent.

A space's state is composed of its register values and virtual memory. The `Regs` option copies register state between a parent and child. The `Zero` option zeros a virtual memory region in a child, and the `Copy` option copies virtual memory into or out of a child.

Determinator provides a more sophisticated memory utility: snapshot/merge. `Snap` copies the calling spaces's entire virtual memory state into the specified child. Invoking `Get` with the `Merge` parameter performs a three-way diff and merge. At a high level, the kernel compares bytes in the child that have changed since the previous `Snap` invocation. Bytes that have changed in the child only are copied into the parent. Bytes that changed in the parent but not the child are not copied. Bytes that changed in both the parent and child generate a *merge conflict* exception. The snapshot-merge mechanism allows an easy library implementation of fork/join. Threads fork children with a `Put` specifying the `Snap` option and join by merging changes back into itself.

The kernel implements `Merge` efficiently by examining page table entries. Using the copy-on-write optimization, snapshot is implemented by making two copies of the page tables and saving them in the child, one copy for the child's private memory and the other as a reference for later comparison. At merge time, the kernel scans the

Call	Interacts with	Description
Put	Child	Copy register state and/or virtual memory range into child, and optionally start child executing.
Get	Child	Copy register state, virtual memory range, and/or changes since the last snapshot out of a child.
Ret	Parent	Stop and wait for parent to issue a Get or Put. Processor traps also cause implicit Ret.

Table 1: System calls comprising Determinators kernel API. [5]

Put	Get	Option	Description
X	X	Regs	PUT/GET child's register state.
X	X	Copy	Copy memory to/from child.
X	X	Zero	Zero-fill virtual memory range.
X		Snap	Snapshot child's virtual memory.
X		Start	Start child space executing.
	X	Merge	Merge childs changes into parent.
X	X	Perm	Set memory access permissions.
X	X	Tree	Copy (grand)child subtree.

Table 2: Options/arguments to the Put and Get calls. [5]

parent, child, and reference page tables. If only the child has written to a page, the page is copied via copy-on-write to the parent. If both the parent and child have written to a page, then the kernel must do a byte-granular diff.

Aviram et al. conclude their discussion of Determinator's kernel by mentioning that the three syscall primitives "reduce to blocking, one-to-one message channels, making the space hierarchy a deterministic Kahn network" [5, 22].

2.4 Determinator's user library

The Determinator kernel alone is enough to enforce deterministic program execution; to make writing deterministic programs more natural, however, Aviram et al. provide a high-level user library that wraps around the three syscall interface. In this section, we will go over the five main areas discussed by Aviram et al. in their "Emulating High-Level Abstractions" section: process API, file system, I/O, shared memory multithreading, and legacy thread APIs [5].

Process API Determinator provides an interface similar to that of `fork/exec/wait`. All of these functions are implemented in user space instead of kernel space. To `fork()` a child process, the parent invokes `dput()` to copy its register and memory state into a new child. The user library must manage a list of "free" process ID numbers, because the system itself does not manage process IDs. `waitpid()` works by entering a loop querying the status of the child; if the child needs more input to continue running (through a mechanism described below), it sets its status appropriately and issues a `dret()`. The parent gives the child more input and sets it in motion again. Once the child finishes executing, it marks itself done and issues a `dret()`. The parent collects the child's status and kills the process.

`exec()` works by forking a child process and loading the new program the new program's memory image into the child. This child is never actually run. Instead, `exec()` enters a trampoline code segment that does a `Get` to copy the new program into the existing process. The trampoline code is mapped at identical locations in both processes so that after executing the `Get`, the process begins executing valid code.

File system Since processes can only access their register set and memory, Determinator provides an in-memory file system. Each process has a private copy of the system's file system. A `fork()` copies the parent's file system state into the child. The parent and child work on private copies of the file system and merge their changes at synchronization points using file versioning techniques [33]. Two files may not be concurrently modified; such cases lead to a reconciliation conflict. The parent and child may, however, perform *append-only* changes to the same file. The file is reconciled by appending the child's additions to the end of the parent's file, and vice-versa.

The file system has limitations compared to traditional file systems. The total file system size is limited by the process's address space; on 32-bit systems, this is a serious limitation. Since the file system resides in virtual memory, buggy programs can write to the memory where the file system resides, corrupting the file system. Lastly, Determinator's implementation of the file system only supports up to 256 files, each with a max of 4MB in size [4].

I/O Since Deterministic processes have no access to external I/O, Determinator emulates I/O as a special case of the file system. Library functions like `getchar()` and `printf()` read and write from special files `stdin` and `stdout`, respectively.

A `printf()` appends output to `stdout`. When a parent merges its file system with a child, `stdout` output is forwarded to the parent and eventually reaches the root process where the root can actually write the output to the system's I/O device. A program does a `read()` to obtain the next unread character(s) in `stdin`. If the file is out of unread characters, `read()` issues a `dret()` to ask for more input from the parent.

Since the file system supports "append-only" conflicts,

the above strategy works well for handling I/O. As all processes reconcile their file systems, each process will see all other process's `printf()`s.

Legacy multithreading APIs Determinator can emulate shared memory multithreading and other legacy multithreaded APIs like pthreads. However, we will not discuss either here, since we do not use these techniques in deterministic Linux. However, we note that since Determinator emulates these legacy thread APIs using its three syscall interface, deterministic Linux could very well be extended to support these APIs. The reader is referred to Aviram et al.'s sections 4.4 and 4.5 [5].

3 Overview

We begin the discussion of adding determinism to Linux by discussing overall design goals of the project. Next we look at the challenges of making nondeterministic Linux deterministic. For the rest of this thesis, we distinguish between *legacy* Linux (the unmodified Linux kernel and its nondeterministic API) and *deterministic* Linux.

3.1 Design goals and non-goals

We wish to make 64-bit Linux deterministic, and in doing so we will present an interface similar to that of Determinator. We would like to run legacy Linux applications without modification alongside deterministic applications, for this is one of the motivating factors applying Determinator's design to Linux. We won't make any attempt to force legacy applications to run deterministically. In order to take advantage of determinism in Linux, legacy programs must be rewritten using the new operating system abstractions.

We would also like to write a user-level C library and an in-memory file system based on those of Determinator. To address Determinator's in-memory file system's limitations [4, 5], we would like to improve the file system by adapting the BSD Fast File System [28] and having deterministic applications utilize Linux's disk-backed persistent storage.

We do not wish to apply all of Determinator's features to our deterministic Linux. The Determinator kernel: (a) extends its nested process model to support deterministic distributed cluster computing; (b) supports a "tree" copy operation for `Put` and `Get`; and (c) allows threads to place an instruction limit on children threads. We have no intention of supporting these features, but we note these limitations do not detract from our goal of making Linux deterministic.

Since the primary goal is determinism, we may decide to limit or ignore Linux kernel internal features. For example, Linux supports huge 2MB pages of memory alongside "normal" 4KB pages. Since this is an internal optimization that is hidden from user applications and for reasons of implementation complexity, we may not allow deterministic applications to take advantage of certain kernel features. We would like to keep all existing functionality available to applications running in legacy mode, however.

Other useful library-level features may be unavailable to deterministic programs. For example, shared dynamic libraries require nontrivial support from the Linux kernel and standard C library. There is nothing limiting us from devising ways to support features like this, but we feel it is outside the scope of our primary goal.

3.2 Challenges

We already discussed the four categories of nondeterminism identified by Aviram et al. in section 2.2; these observations are general enough that they also apply in making Linux deterministic. The Linux kernel presents additional challenges, and we discuss them here.

In order to run legacy Linux applications, we cannot enforce Determinator's requirement that all but a single root process run in deterministic mode. Linux's process model is also too flexible: it allows reparenting and for children to outlive parents, directly opposed to Determinator's nested process model.

Since Determinator was written from scratch, the designers addressed sources of nondeterminism by simply not providing kernel support for such features. On the other hand, Linux's existing threading model is inherently nondeterministic and already provides extensive support for nondeterministic features (e.g., the `gettimeofday()` syscall). We must also address nondeterminism not covered by Aviram et al. Linux supports a wider range of process communication utilities including signals, pipes, and System V IPC.

Memory subsystem Linux supports a wide range of virtual memory features including memory-mapped files, huge pages, and swapping to disk. All of these features are layered on top of an abstraction for supporting memory management units of many different processor architectures (e.g., x86_64, Alpha). Whereas Determinator's memory management subsystem deals only with page table management for a single architecture, Linux's memory subsystem uses more complicated data structures and nontrivial algorithms. Figure 1 demonstrates the complicated data structures associated with object based reverse mapping [13] used to find all processes that map a given page of memory. From an implementation perspective, Linux's memory subsystem's complexity presents a potentially formidable challenge in implementing Determinator's three memory operations (`Zero`, `Copy`, and `Snap/Merge`).

Standard C Library Many Linux applications written in C use the standard C library. This library in large

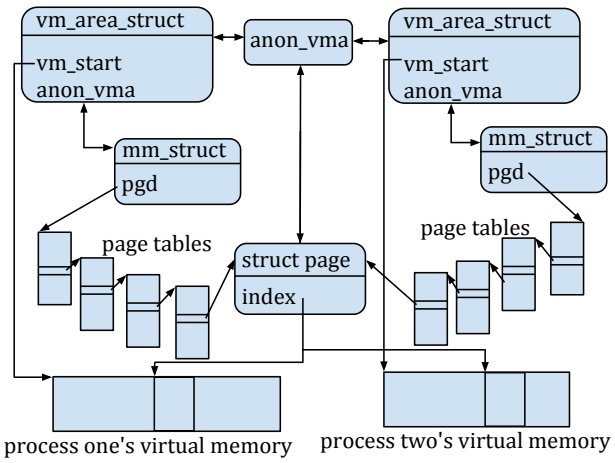

Figure 1—Data structure relationships associated with object-based reverse mapping. The `struct page` C type encapsulates information about every page frame of physical memory. Two processes map virtual memory to the same page read-only (possibly at different addresses). In order for the kernel to swap a given page to disk, object-based reverse mapping assumes each process maps the page to `vm_area_struct->vm_start + page->index` in virtual memory.

part functions as a wrapper around legacy Linux syscalls and thus is highly nondeterministic. Whereas some functions, such as `strlen()`, might not use nondeterministic syscalls, many other functions do use nondeterministic syscalls (e.g., `printf()`). Thus, deterministic programs may be forced to use a completely different library. Moreover, libraries in Linux are often linked dynamically with shared libraries, but since Determinator does not provide any native kernel support for dynamic linking, we may lose the ability to dynamically link shared libraries.

3.3 High level approach

To address concerns about Linux's more flexible process model, we present a *hybrid process model*. A Linux process invokes the `dput()` syscall (introduced below) to become a *master* process, akin to Determinator's lone root process. This master process has full access to the legacy Linux kernel API, with some minor restrictions noted below. Master processes then create *deterministic* children. We call this master process and its entire subtree a *deterministic process group* (DPG).[1] Within this process group, processes abide by Determinator's nesting rules (e.g., children cannot outlive parents). Legacy Linux applications run alongside deterministic applications with absolutely no kernel restrictions. In some sense, each deterministic application resembles an entire Determinator "virtual machine" of sorts.

We also add three new syscalls, `dput()`, `dget()`, and `dret()` and restrict deterministic processes to only use the new syscalls. These syscalls function exactly as their Determinator counterparts, excepting cluster support, the copy (grand)child subtree option, and instruction count limits. By restricting deterministic processes to these three syscalls, we can remove nearly all sources of nondeterminism; we only have to modify the kernel to ignore all signals sent to deterministic processes, and thus we have effectively blocked all sources of nondeterminism. Figure 2 illustrates the hybrid process model in a hypothetical environment.

At the expense of predictability, but without harming determinism, master processes can use nonblocking `dput()` and `dget()` (invoked with a special flag) to poll whether or not the child process has reached a synchronization point yet. Since programs are responsible for scheduling children, a CPU might become available while the scheduler thread remains blocked waiting for another thread to finish. This would be useful in a parallel `make` program [5] or a web server that needs to be responsive to incoming requests. Applications that use the nonblocking syscalls can still log schedule sequences to reproduce program output in a deterministic fashion.

We also allow signals to reach master processes. Master processes can specify a set of signals that can interrupt a blocking `dput()` or `dget()`. Allowing signals for master processes again introduces nondeterminism, but we note that the master can control the signals and write them to a log file for replay. We also note the usefulness of signals: terminal operators can send a `SIGINT` to kill an application immediately.

Once this kernel work is done, we begin work on a C user library. Deterministic applications won't use the standard C library, because many library calls invoke disallowed legacy syscalls. Unfortunately, many functions must be rewritten (e.g., `sprintf()`, `strlen()`).

To increase the usefulness of the system, we provide an in-memory file system just as Determinator does. Whereas Determinator's file system uses fixed file size [4], our file system design is based on the BSD Fast File System [28]. Our file system is organized as a rooted tree and supports hard linking, dynamic file sizes that can grow up to 1GB, and better resource management (inodes, data blocks). Since a disk-backed file system is standard on Linux systems, master processes can save and load the in-memory file system to disk to provide persistence.

[1] Our DPGs are unrelated to those in Bergan et al.'s "Deterministic Process Groups in dOS" [9].

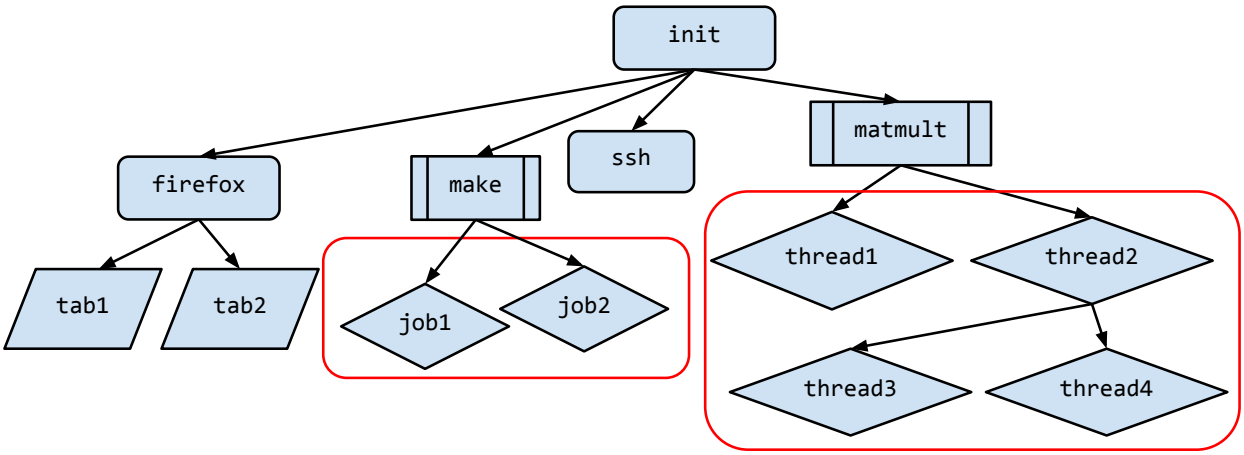

Figure 2—Illustration of the hybrid process model with two deterministic process groups. Legacy nondeterministic processes (`firefox`, `ssh`) run as children of `init`. `make` and `matmult` are masters of deterministic applications. The deterministic children (the diamond processes) are isolated from the rest of the system.

4 Implementation

With a high level design in mind, we now discuss the implementation details. We started by forking Linux from the GitHub `git` mirror repository and worked on x86_64 Linux 3.0. We developed and tested incrementally on an 8-core machine with 8BG of RAM running Arch Linux.

This section is divided into two logical parts. We first discuss the kernel implementation, then the user library and in memory file system. Our initial focus in the kernel work was adding process management related functionality; then we added the `Zero`, `Copy`, and `Snapshot/Merge` operations in that order.

4.1 Process organization

The first step was adding the three new syscalls: `dput()`, `dget()`, and `dret()`. Our initial focus was process creation and related functionality. `dput()` relies heavily on existing `do_fork()` and `do_exit()` kernel code. We added logic to enforce the requirement that deterministic processes cannot outlive their parents. We blocked all external signals generated by user applications, but allowed kernel-generated signals. We used the kernel-generated signal mechanism to trigger implicit `dret()`s on process faults like divide-by-zero (`SIGFPE`) and memory access violations (`SIGSEVG`).

We augmented Linux's `task_struct` process structure with a new *deterministic PID* and low-level synchronization primitives. `dput()` and `dget()` use these synchronization primitives to synchronize using the fork-join model. When a parent starts a child with `dput()`, any subsequent `dput()` or `dget()` call blocks until the child issues a `dret()`.

Even though master processes have direct access to kernel I/O devices, we placed some restrictions on what a master process can do. Processes become the master of a deterministic process group by invoking `dput()` with a special parameter. The kernel makes sure that the process is not the parent of any other nondeterministic process and isn't using any unsupported virtual memory features, like "kernel samepage merging" [1] or huge pages. Once a process becomes a master, it can no longer use the legacy `fork()` family of syscalls. It can only create new processes through the deterministic family of syscalls.

Master processes can use `dput()` to specify a set of signals to ignore while in a blocking `dput()` or `dget()` (i.e., waiting for a child to stop). This can be useful when a console operation wishes to kill an application with a `SIGINT`, but does not want other signals to interrupt the master process.

The last feature relating strictly to process organization is register state copying. The initial `dput()` call that creates a child automatically copies register state, since we effectively delegate work to `fork()`. Subsequent calls to `dput()` and `dget()` can put or retrieve a child's general-purpose register state by accessing fields in the process's `task_struct` descriptor.

4.2 Memory operations

As mentioned in section 3.2, Linux's memory subsystem is very complex. Before discussing how we implemented the memory operations, we will give some technical background on Linux's memory subsystem. Our implementation reuses many existing functions.

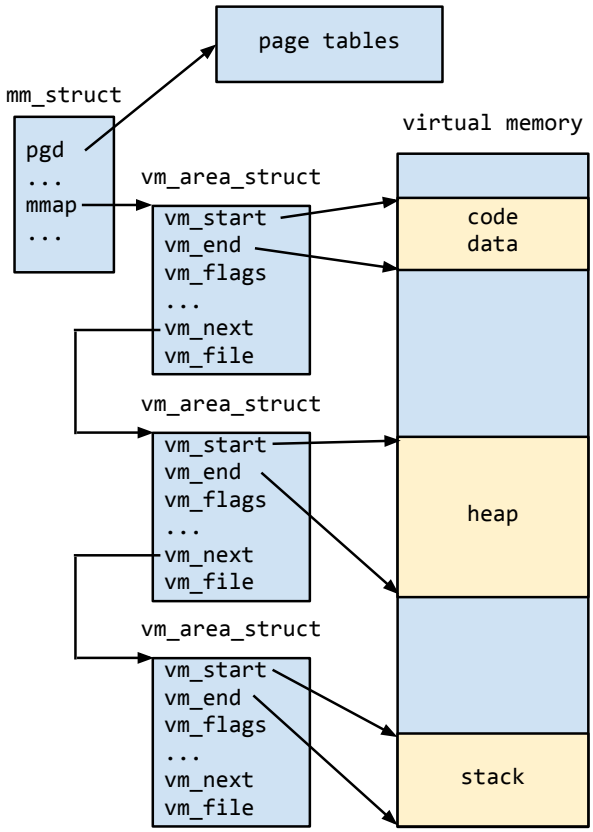

Figure 3—Illustration of Linux's memory management data structures.

Background A process's entire virtual memory is managed by a `mm_struct`. `mm_structs` maintain a list of contiguously mapped memory regions with identical permission bits (e.g., read/write/execute); these regions are stored as `vm_area_structs` (see Figure 3). Anonymous memory (memory not backed by a file) is mapped to a read only global zero page, and new pages are allocated *on demand* [26]. In order to swap a single physical page of memory to disk, Linux requires all mappings of the page have the same offset from the start of the enclosing `vm_area_struct` (see Figure 1). This requirement makes swapping space and time efficient, but it places unfortunate restrictions on how aggressively we can apply copy-on-write.

Linux provides `dup_mm()` to copy the virtual memory image of a process. `fork()` uses this to copy a process's memory via copy-on-write. `dup_mm()` copies each `vm_area_struct` of the source by invoking `copy_page_range()`. This function walks Linux's four level page table structure to perform copy-on-write by actually changing page table entries and marking the pages as read only.

Linux's memory subsystem does lack one important generality that is essential to the core of Determinator's memory operations: in Linux, most virtual memory kernel functions act on behalf of the calling process (inside a syscall, the calling process is accessed via the C macro `current`) instead of operating on *any* process's virtual memory. For example, `do_mmap()`, which maps files into memory and creates anonymous memory regions, will only perform work on `current`.

Zero Since `dput()` and `dget()` operate on a child process, not the calling process, the first step was to generalize Linux's memory subsystem. We enhanced functions like `do_mmap()` to take an extra argument specifying the target process's virtual memory structure.

Implementing the Zero operation was thus very simple. `do_mmap()` maps anonymous memory to a zeroed out region automatically, so to perform a Zero operation, the kernel unmaps then remaps the region in question. Most of the memory subsystem deals in page aligned regions, so the kernel needs to handle non page aligned begin and end regions with what amounts to a `memset()`.

Copy The copy operation was more complicated. At a high level, Copy takes an arbitrary virtual memory region from the source and copies it to the destination virtual memory, with an optional offset in the destination. To be efficient, we only allow page aligned offsets so that we can take advantage of copy-on-write.

We generalized `copy_page_range()` to map pages copy-on-write with a destination offset. The Copy operation works by unmapping the specified region in the destination, then invoking a `copy_page_range()`. To satisfy the swapper subsystem's requirement about how physical pages can be mapped, care must be taken to ensure the source and destination have their `vm_area_structs` sharing start and end boundaries (with respect to the destination offset). This can be accomplished with a helper function, `split_vma`. Finally, as before, we must handle non page aligned begin and end regions with a `memcpy()`.

Snap/Merge The Snap/Merge combination is the most complicated feature, and unfortunately we were not able to reuse as much existing code as with Zero and Copy. Performing a Snapshot is relatively easy. We destroy the target's `mm_struct`, then mimic `fork()`: we make a copy of the source's `mm_struct` and attach it to the destination. This effectively copies the source's virtual memory image into the destination. We also make a second `mm_struct` copy for use as a reference virtual memory image later. Using `dup_mm()` only has to create a copy of the `mm_struct` and page tables; pages used by the process are only copied when written to, so this method of creating a reference snapshot is space efficient.

Upon a Merge request, the kernel first must ensure `vm_area_structs` are aligned, just as for Copy. We then iterate over `vm_area_structs` and walk the page table hierarchy. Instead of doing a naive byte by byte compar-

ison, we check page table entries to quickly determine if two pages have diverged since the snapshot; if a page was written to, a new page would have been allocated via copy-on-write, thus indicating the kernel must do a byte by byte comparison. Pages that have changed only in the source are copied via copy-on-write to the destination, when a byte by byte combination must be used, changed bytes are also copied over. Writes by the source and destination to the same byte location generate an exception. The child can no longer run, and the parent must acknowledge that exception by killing the child. As usual, we handle non page aligned start and end regions manually by doing a direct byte comparison.

4.3 User library

We require that deterministic Linux applications use a custom user library. We model the library design and API on the C standard library. Many simple functions must be rewritten, like `strlen()`, and indeed we borrowed many header and implementation source files from the instructional JOS operating system [21].

Designing and implementing what amounts to be a replacement for the C standard library is no easy task, and indeed a fully functional library merits an entirely separate discussion. Thus, we did not set out with a specific plan or set of functionality to implement. Much of the library was constructed in a reactive manner where new functionality was added only when necessary for building applications.

Aviram et al. devote an entire section, "Emulating High-Level Abstractions," discussing how to implement a traditional Unix API [5]. We do not have any new insight in this area, so we will limit our discussion here to Linux-specific details as they apply to writing the user library.

Linux specific considerations Before executing `main()`, the library runtime detects whether the executing process is the master or deterministic and sets up internal variables. The in-memory file system is initiated (described in detail below), and special `stdin` and `stdout` files are created so that processes can emulate functions like `getchar()`. When returning from main, or performing an `exit()`, the file system is cleaned up and the process signals to the parent that it has finished by settings its status code and issuing a `dret()`. The parent must acknowledge the child's death before proceeding.

Many functions have dual roles depending on whether the executing process is the master or deterministic. For example, since master processes have direct access to nondeterministic resources like file I/O, functions like `printf()` write directly to the system's `stdout` via the `write()` syscall. On the other hand, when deterministic processes invoke `printf()`, the output is buffered in the special `stdout` file in the in-memory file system. At synchronization points, the file system forwards this `stdout` to the parent; eventually, the output reaches the master space and is directed to the system's `stdout`. The library runtime choses the appropriate action by checking if the executing process is the master.

Component	Lines of code added
Primary syscall implementation	1187
Memory subsytsem	1081
Kenrel miscellaneous	296
New user library	2492
Borrowed user library	1727
Total	7135

Table 3: Count of number of lines of code added.

4.4 File system

The in memory file system for deterministic Linux has significant improvements over that of Determinator. Whereas Determinator uses a fixed file size and all files are mapped to a known location in memory, we chose to implement the more general BSD Fast File System design. We since implemented deterministic Linux on x86_-64, we do not run into the address space limitations imposed by 32-bit systems.

The file system is divided into 4096-byte *blocks*. The first block is a *superblock* containing metatata about the file system. A region of fixed size following the superblock is reserved for *inodes* and a bitmap for managing which blocks are used. The rest of the blocks are data blocks. Each `inode` represents a traditional Unix file object and contains ten direct block pointers and a singly and doubly indirect block pointer. A file may be up to 1GB on 64-bit systems.

Deterministic applications can also take advantage of the master process's access to the system file system. When an application starts up, it can read an in-memory file system image from permanent storage. When the application finishes, it can save the in-memory image back to permanent storage for use later.

4.5 Implementation Statistics

We began with a `git` fork of Linux 3.0. Table 3 lists lines of code added (including comments but excluding newlines) for various kernel and user library components. The "new user library" category counts code added by us, and the "borrowed user library" category counts code that was reused from `JOS`. In total, the kernel required 2564 additional lines of code.

The kernel[2] and user library[3] are available on GitHub

[2] `https://github.com/ccotter/linux-deterministic`
[3] `https://github.com/ccotter/libdeterm`

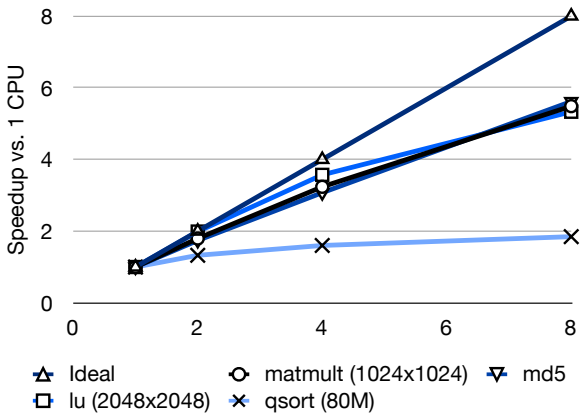

Figure 4—Deterministic speedup for the parallel benchmarks.

Processors	Pthread	Deterministic
1	56.4 ± 0.0	65.5 ± 0.0
2	34.4 ± 0.3	37.6 ± 0.1
4	20.3 ± 0.3	21.3 ± 0.1
8	11.1 ± 0.1	11.7 ± 0.1

Table 4: Run times (seconds) for *md5*. Times are averaged over five runs.

as two separate repositories. The deterministic Linux kernel is maintained as a separate branch `v3.0-det`.

5 Evaluation

This section evaluates deterministic Linux by running compute-bound applications using deterministic Linux and nondeterministic pthreads. We conclude by considering a case study demonstrating the qualitative debugging benefits of determinism.

5.1 Empirical experiments

Aviram et al. already demonstrated coarse-grained applications in Determinator performed comparable to nondeterministic Linux equivalents, but fine-grained applications did not scale nearly as well, incurring high performance costs owing to memory synchronization costs.

The primary goal of our evaluations is to determine if Linux can efficiently run applications using Determinator's programming model. Since the deterministic Linux kernel reuses a lot of existing code, we expect reasonable performance compared to the equivalent nondeterministic application. However, applications that heavily rely on memory-intensive kernel operations like Snapshot/Merge might incur performance penalties.

5.1.1 Methodology

We evaluated four benchmark programs. *md5* searches for a string whose md5 hash yields a particular hash value (i.e., a brute force password cracker). *md5* creates $N + 1$ threads to perform parallel work where N is the number of processors available. The *qsort* benchmark recursively sorts an integer array by forking children threads going $log_2(2 \cdot N)$ levels deep, after which the recursive computation is carried out sequentially. *matmult* multiplies two square matrices by dividing the input into blocks and forking off threads to multiply the individual blocks. *matmult* creates a block for each available core. *lu* performs LU decomposition of a square matrix by breaking the input into $4N^2$ blocks and creating a thread for each block.[4] Our LU algorithm breaks the blocks up into discontiguous blocks of memory (i.e., we represent our matrix in row-major order).

All benchmarks are designed so the deterministic and pthread versions operate on identical inputs generated from a pseudorandom number generator (*matult*, *lu*, *qsort*) or "hardcoded" input (*md5*). For a given input size, *qsort* runs ten tests on different inputs and averages the run times. Both matrix benchmarks run ten tests on different inputs and average the results. *md5* searches for ten fixed ASCII strings whose values are distributed randomly over a fixed alphabet. The run times for all ten runs are averaged to a single value, and this is done five times. Both matrix benchmarks were tested before running the actual evaluations; we compared output from both versions of each benchmark and found them to be identical.

The *matmult*, *qsort*, and *md5* benchmarks are modified versions of benchmarks found in Determinator's source on GitHub.[5] The *lu* benchmark was written from scratch using an algorithm described on the Internet [19]. We tested on a 2 socket $\times$ 4 core 2.33 GHz Intel Xeon PC running Arch Linux with 8GB of RAM. All benchmarks use the same modified x86_64 Linux 3.0 kernel.

5.1.2 Run time data

Table 4 shows the average of five runs for *md5*. Run times for the other benchmarks are reported at the end of this document. Table 8 shows average run times for the pthread *lu* and *matmult*, and table 9 shows average run times for the deterministic versions. Table 10 shows the time spent in the kernel doing a virtual memory merge for *lu* and *matmult*. Table 11 reports the average run times for the pthread and deterministic *pqsort* benchmark.

[4] Threads are created only when data dependencies are satisfied. In our program, each row of blocks has a data dependency on the row above, so at any given point in the program, *lu* creates at most $2N$ threads.

[5] `https://github.com/bford/Determinator`

	lu				matmult			
Dimension	$N = 1$	$N = 2$	$N = 4$	$N = 8$	$N = 1$	$N = 2$	$N = 4$	$N = 8$
16×16	13.1 (41.5%)	45.0 (46.7%)	46.3 (45.8%)	30.9 (31.5%)	9.6 (48.2%)	21.8 (45.0%)	37.8 (45.2%)	16.0 (26.5%)
32×32	8.5 (34.0%)	37.3 (45.5%)	45.9 (46.1%)	29.1 (31.1%)	3.3 (37.7%)	13.4 (42.0%)	21.1 (42.2%)	17.5 (24.1%)
64×64	2.6 (19.5%)	20.6 (41.6%)	42.1 (44.2%)	32.1 (30.9%)	1.3 (13.4%)	3.8 (26.3%)	7.8 (34.0%)	13.2 (25.8%)
128×128	1.4 (2.3%)	6.0 (32.8%)	22.4 (39.0%)	30.8 (31.0%)	1.0 (0.3%)	1.9 (1.7%)	4.5 (13.3%)	6.7 (18.2%)
256×256	1.1 (0.5%)	2.1 (11.0%)	7.0 (25.9%)	18.8 (31.1%)	1.0 (0.0%)	1.2 (1.0%)	1.8 (1.6%)	2.3 (5.0%)
512×512	1.0 (0.1%)	1.2 (1.4%)	2.3 (9.4%)	5.9 (19.4%)	1.0 (0.0%)	1.0 (0.5%)	1.1 (0.9%)	1.5 (3.0%)
1024×1024	1.4 (0.0%)	1.0 (0.3%)	1.2 (1.5%)	1.9 (8.4%)	1.0 (0.0%)	0.9 (0.0%)	1.0 (0.1%)	1.2 (0.2%)
2048×2048	1.4 (0.0%)	1.0 (0.0%)	1.0 (0.1%)	1.1 (1.0%)	-	-	-	-

Table 5: Deterministic overhead for *lu* and *matmult*. Overhead is deterministic run time divided by pthread run time. The numbers in parentheses indicate time spent in the kernel performing a virtual memory merge as a percentage of overall runtime.

	qsort			
Input size	$N = 1$	$N = 2$	$N = 4$	$N = 8$
1K	37.0	48.9	32.3	20.1
4K	15.3	44.8	44.5	25.4
8K	8.7	26.1	33.3	22.0
10K	7.3	23.9	31.2	25.7
40K	2.5	7.8	14.8	17.0
80K	1.8	4.5	8.1	10.7
100K	1.6	4.0	6.5	10.1
400K	1.2	1.7	2.3	2.8
800K	1.1	1.4	1.6	2.0
1M	1.1	1.5	1.7	1.9
4M	1.0	1.2	1.3	1.4
8M	1.0	1.2	1.5	1.6
10M	1.0	1.1	1.1	1.3
40M	1.0	1.1	1.1	1.2
80M	1.0	1.1	1.1	1.3

Table 6: Deterministic overhead for *qsort*.

5.1.3 Results

Applications that require a lot of memory synchronization show considerable overhead for small inputs but have much more acceptable overheads for large inputs. Table 5 shows deterministic overheads for the *lu* and *matmult* benchmarks. Execution on small inputs spends up to 48.2% of their run time in the kernel doing a memory merge. Small inputs in general have a hard time seeing parallel benefits: the pthread versions of both benchmarks did not see parallel speedup benefits until the input matrix reached at least 32×32 (*matmult*) or 128×128 (*lu*). Still, it isn't until even larger input sizes (at least 1024×1024 for *lu* and 256×256 for *matmult*) when we begin to see more "acceptable" deterministic overheads of at most $2.3\times$. The matmult benchmark shows overhead of .9 for $N = 2$ on input of size 1024×1024, indicating the deterministic version ran faster than the pthread version, but we have no explanation for this behavior.

The *qsort* benchmark also shows unacceptable overhead for small inputs, but once the input array size becomes at least 800K, overheads stayed under $2\times$ (Table 6). The embarrassingly parallel *md5* benchmark exhibited very little overhead; it had overhead of $1.16\times$ for $N = 1$ and $1.05\times$ for $N = 8$. This can likely be attributed to the little amount of information transfered back and forth between threads (a single boolean and the matching string when it is found).

Figure 4 shows speedup over one CPU for the deterministic benchmarks. We show the results when run on the largest input size available (e.g., 80M for *qsort*). The *md5*, *lu*, and *matmult* benchmarks scale well, and *qsort* scales poorly, not even reaching $2\times$ when $N = 8$. Figure 5 compares scalability of the pthreads benchmarks against the deterministic versions. In all four benchmarks, we see that the deterministic programs have about the same ability to scale as the pthread programs. For *lu* and *md5*, the deterministic versions scale better due to high overheads in the $N = 1$ case that diminish as we add more CPU cores.

5.1.4 Fine-tuning the benchmarks

We may attribute some of the exceptionally high overhead for small inputs to the cost of merging. For simplicity, our deterministic thread join merges the entire static data segment. *qsort*, *matmult*, and *lu* all declare static arrays with sizes as large as the maximum expected input. For example, *lu* declares two `long double` arrays of size 2048×2048 for a total of 128MB on our 64-bit machine. When we join on all 256 threads for the 16×16 input, we merge over all 128MB. Page table optimizations (Section 4.2) enable the kernel to only have to do a byte-by-byte merge for the single page containing the input matrix, but the kernel must still check at least 32767 other page table entries for each merge![6]

Thankfully, the syscall API allows us to specify a range of virtual memory to merge; we could fine-tune our programs to merge only what we know has changed. For the *qsort* benchmark, we could opt to use the `Copy` option to achieve the same effect. `Copy` copies page tables directly without checking page table entries and possibly the bytes themselves as `Merge` does. By having our high-level join function merge the entire static data segment,

[6] Assuming 4KB pages.

our join method is very general in the types of programs it can be used with. We could likely trade this generality for performance gains.

5.1.5 Qualitative evaluation

This section attempts to give a qualitative understanding of how the benchmarks work. Each of our benchmarks uses the fork/join paradigm. The pthread applications fork and join threads with `pthread_create()` and `pthread_join()`. The deterministic benchmarks create threads by issuing a `dput()` specifying the `Snap` option; we join thread results by performing a `Merge` on the program's entire static data segment (initialized and uninitialized). At the kernel level, creating deterministic threads incurs the additional cost of creating two copies of the parent's page tables and associated kernel data structures. Since pthreads threads share memory, the kernel merely copies a pointer to the parent's memory management structure to the child. Joining blocks the parent thread until the child finishes execution. The deterministic join, however, incurs the cost of merging a potentially large region of memory.

We should also consider an otherwise overlooked overhead associated with copy-on-write. When a deterministic child thread writes to page marked COW, the kernel intervenes to allocate a new page, copy the page contents, and invalidate at least one TLB entry.[7] In fact, tracing through the page fault code and manually counting the number of semicolons gives a conservative estimate of 84. A similar trace through Determinator's COW code counts only 23 semicolons. Thus, for all inputs, and especially the smaller input sizes, we suspect that COW contributes a nontrivial amount of overhead.

5.2 Finding bugs deterministically

To demonstrate one of the key benefits of deterministic execution (Section 2.1), we consider a Gaussian elimination program written using the deterministic API and pthreads. Figures 6a and 6b show nondeterministic and deterministic portions of Gaussian elimination code. There is a crucial synchronization bug, however: both algorithms create `nrows - k` worker threads but only join `nrows - k - 1` threads. We purposefully inserted this bug, but the reader can imagine a programmer making this typo by mistake.

We ran both versions multiple times and examined the resulting matrix. The deterministic program *always* produced the wrong answer. `djoin()` merges changes back into the main thread for all but the last worker thread. When the last worker thread finishes, its changes are private and never seen by the main thread.

[7] TLB invalidation is a costly operation, especially if the CPU only supports invalidating the entire TLB.

Abstraction	Occurrence
Fork/Join	17.9%
Barrier	14.8%
Work Sharing Constructs	32.8%
Deterministic through nondeterministic primitives	26.7%
Nondeterministic	8.4%

Table 7: Synchronization abstractions used in the SPLASH, NPB-OMP, and PARSEC benchmark programs.

On the other hand, the pthreads program executes nondeterministically. We observed three different output matrices, one of which was the correct result. If the final worker thread finishes before we observe the final result, the output will be correct. If the final worker thread does not finish by the time we examine the output, we get an intermediate result. The output is nondeterministic, owing to a race condition, since all threads share memory.

Determinism provides a clear benefit in debugging, since the deterministic version always behaves the same, whereas the buggy behavior might exhibit itself rarely in the pthread version. We also consider testing our application to ensure its correctness. When testing our Gaussian elimination application, determinism allows us to see right away that, in general, all inputs give incorrect answers. On the other hand, it may be the case that the nondeterministic version always gives the correct answer for matrices we test on a development machine, but as soon as we ship the program to a client's machines with a different number of processors (for example), the bug exhibits itself.

6 Limitations and future work

While our user library supports the fork/join construct, there are other parallel abstractions we could support. The barrier abstraction blocks thread execution until all threads in in some program defined group reach a certain point, upon which all threads in the group proceed to the next barrier. The work sharing construct, used in the OpenMP parallel API, divides work among a "team" of threads and blocks until all threads finish work. Our user library support neither of these.

In his PhD thesis, Aviram classified the types of synchronization abstractions used in the SPLASH, NBP-OMP, and PARSEC benchmark programs [3], and the results are summarized in Table 7. The first three rows shows uses of naturally deterministic library constructs; these account for 65.5% of the total. Uses of nondeterministic primitives to enforce deterministic behavior by the program itself are shown in row four. The last row shows uses nondeterministic abstractions like mutex locks and condition variables. Future versions of our user library could implement barriers and the

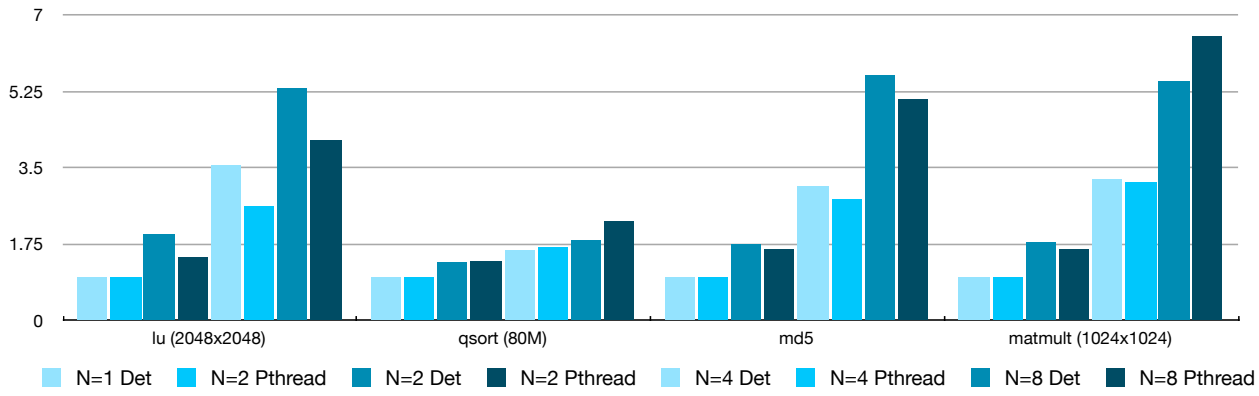

Figure 5—Comparing the speedup over $N = 1$ for the deterministic and pthread versions of the benchmarks. This figure demonstrates the ability of both versions to scale as we add more CPU cores.

```
pthread_t thread[MAXTHREADS];
struct thread_data data[MAXTHREADS];
void pthread_reduce(void) {
   for (i = 1; k <= nrows - 1; ++k) {
      for (i = k + 1; i <= nrows; ++i) {
         data[i] = /* Setup worker. */;
         pthread_create(&thread[i], NULL, worker,
            &data[i]);
      }

      /* Bug! Should be i <= nrows */
      for (i = k + 1; i < nrows; ++i)
         pthread_join(thread[i], NULL);
   }
}
```

Figure 6a—pthread Gaussian elimination.

```
/* Forks a deterministic child. Returns 0 into the
 * child and 1 into the parent. */
int dfork(pid_t childid);
/* Merges a child's changes into the parent after
 * the child issues a dret(). */
void djoin(pid_t childid);

void det_reduce(void) {
   for (i = 1; k <= nrows - 1; ++k) {
      for (i = k + 1; i <= nrows; ++i) {
         data[i] = /* Setup worker. */;
         if (!dfork(i)) { worker(&data[i]); dret(); }
      }

      /* Bug! Should be i <= nrows */
      for (i = k + 1; i < nrows; ++i)
         djoin(i);
   }
}
```

Figure 6b—Deterministic Gaussian elimination.

work sharing parallel abstractions, building upon the `dput()`/`dget()`/`dret()` syscall primitives. Doing so would appear to extend the types of parallel programs our deterministic Linux could support.[8]

Efficiency and overhead are important considerations in the attractiveness of our system. As discussed in Section 5.1.4, merging on a huge region of memory is costly, even if we know at most a page or two of memory could possibly have changed. We could enrich our library API to, for example, narrow down the region that must be merged. If our kernel implementation allowed deterministic processes to use huge pages, we could cut the number of page table entries that must be examined by a factor of 500-1000.[9]

Since the kernel API provides no direct file system support, our library uses an in-memory file system. Our choice to use a Unix-like implementation may not be well suited for reconciliation in user-space. If a file's data blocks do not reside contiguously in memory, copying a file from a child into the parent may require a `dget()` call for each data block. Instead, we could move the file system reconciliation logic into the kernel and extend the syscall API. A more radical step would be to implement the entire file system itself in the kernel, perhaps as part of Linux's Virtual File System API, the generic interface to all file systems in Linux [20].

Our deterministic Linux intentionally did not implement all of Determinator's features. Determinator's `Put` can limit the number of instructions a child process can execute before an implicit `Ret` is generated. Since deterministic Linux does not support this, malicious or buggy children can run in an infinite loop, blocking the parent indefinitely. The instruction count feature is also crucial in emulating legacy thread APIs. Section 4.5 in Aviram et al.'s Determinator paper describes how to implement a

[8] Assuming the three benchmark suites reflect real world parallel programs, as Aviram also assumes [3].

[9] Depending on the huge page size (2MB or 4MB).

deterministic version of the pthreads API [5]; a deterministic scheduler thread uses the instruction count limit to ensure a deterministic (though arbitrary) schedule.

7 Related Work

Many systems attempt to alleviate and control the effects of nondeterminism. The Velodrome [17] and SingleTrack [34] tools dynamically check atomicity and determinism constraints in user programs to report possible race conditions or nondeterministic behavior. Replay debuggers like Instant Replay [24] record all relevant inputs and thread interactions to execute a program in a repeatable fashion for debugging purposes. However, these tools are expensive in terms of run time overhead and storage. None of these tools do anything to fix the inherently nondeterministic environment in which developers write programs.

Some systems provide determinism at the expense of using nonstandard programming languages, compilers, or hardware. Bocchino et al.'s DPJ extends Java with a deterministic type system [12]. Environments like RCDC [16] and DMP [15] provide determinism with transactional memory hardware support. CoreDet [7] builds on DMP by using the LLVM compiler [23] and a runtime to provide determinism.

CoreDet, DMP, and Grace [10] provide user space deterministic schedulers for C/C++ programs. However, wild pointer writes can corrupt the scheduler. Kendo [32] provides a relaxed determinism model: programs that correctly protect critical sections of code execute deterministically by synthesizing an arbitrary lock acquisition order based on instruction counting. Unprotected access to shared variables, however, will lead to nondeterministic traces, owing to data races. None of these systems are able to enforce determinism on *arbitrary* user programs as Determinator does.

dOS [9] adds *deterministic process groups* (DPGs) to the Linux kernel. A deterministic scheduler in the kernel enforces deterministic thread interactions within a DPG. To interact with external objects (e.g., network packets), DPGs use a user space service called a *shim*. The shim gives DPGs complete control over external nondeterministic inputs, similar to how Determinator's root space controls external I/O. dOS does not use a "clean-slate" approach as Determinator does [5], and thus has better backwards compatibility with Linux. However, like the other deterministic environments mentioned, dOS merely masks the effects of nondeterminism instead of removing nondeterminism entirely as Determinator does with its novel programming model.

8 Reflections on my research experience

Before we conclude, this section will reflect on the undergraduate research journey as experienced by the author, Chris Cotter. I initially read Aviram et al.'s "Efficient System-Enforced Deterministic Parallelism" paper in July of 2011 and wrote my first line of code in the Linux kernel that August. After many implementation iterations, my final implementation of deterministic Linux took up the month of September 2012, and I wrote this thesis soon after.

Learning the Kernel There is no "Linux Kernel 101" course at the University, and most existing documentation and comments in kernel code are written for seasoned kernel programmers. Thus, I very often found myself lost and frustrated. It wasn't until I spent a year (August 2011 - 2012) of kernel hacking until I finally felt comfortable implementing my third and final iteration of deterministic Linux.

8.1 First Iteration

In August 2011, I began by downloading Linux 2.6.32 source and learned to compile and run the kernel with `QEMU` [6]. I ran `QEMU` with a ramdisk containing a single program to run as *init*. My advisor Mike Walfish gave me my first goal: to implement a new syscall in Linux. After scouring the Internet, I learned how to do this, and I wrote skeleton code for my three syscalls: `dput()`, `dget()`, and `dret()`.

I iteratively implemented Determinator's functionality in these syscalls, starting with process organization and moving to memory operations. I had a particularly difficult time with memory operations. Even though I was a wizard with my operating system's instructional `JOS` OS and page table management, I had no idea how to maneuver in Linux's memory subsystem.

After fumbling around with countless kernel panics, I set out with a simple goal: to change a single page table entry. After accomplishing this goal, I was ready to implement Determinator's Zero and Copy operations. In fact, I eventually found I could reuse and adapt a lot of existing code for these operations. Implementing Merge took considerably more effort, since Merge in Linux is an entirely novel operation.

Unfortunately, this version of my implementation was buggy, primarily due to misuse of internal kernel API and race conditions in my kernel code.

8.2 Second Iteration

In November 2011, I downloaded Linux 2.6.38 source and began rewriting my code, copying and pasting most of my first iteration code. I also started running an Arch Linux distribution on QEMU, since I was able to run

more sophisticated tests with an actual Linux distribution running my kernel. With a few months of kernel hacking knowledge under my belt, I identified many logic bugs and had a better understanding of how things worked "under the hood". Unfortunately, I encountered many setbacks.

New Memory Features Moving from Linux 2.6.32 to 2.6.38 introduced new memory subsystem features. Notable among these was transparent huge pages (THP). When processes map a large enough region of virtual memory (e.g., at least 4MB), the kernel will sometimes fulfill demand paging requests with huge pages without the user knowing.

Since my original code did not account for THP, a lot of my existing kernel code broke, and I spent days trying to understand what went wrong and perhaps a week devising a solution. In the end, I came to realize I still lacked a great deal of knowledge about Linux's memory subsystem, and this lack of knowledge would continue to plague my second iteration's quality.

Condition Variable Usage Violation My operating systems professor Mike Walfish taught us to always surround the testing of condition variables with a while-loop and not an if-statement. Unfortunately, I completely disregarded this lesson at some point in my implementation of `dput()` synchronization, and I found threads being woken up prematurely.

Snapshot/Merge Bug When I ran a stress test to fork hundreds of processes then did a simple Snapshot and Merge, I encountered a kernel panic that caused unrelated processes to crash (e.g., `bash`). Through the course of a month, I never identified the issue except to say that my lack of a thorough understanding of the memory subsystem was at fault. This, and a general lack of organization in my code lead me to write a third iteration.

8.3 Final Iteration

In September 2012, I forked a copy of the Linux `git` repository and started with Linux 3.0. I started running my code on an 8-core machine with 8GB of RAM; the 8 cores maximized parallelism to help expose concurrency bugs in my kernel code. I also decided to do a complete rewrite — no old code from previous iterations would be copied and pasted.

After a year of kernel hacking, I had never felt more confident in the code I wrote; whereas in my second iteration I was not confident in my code's correctness, in my third iteration I could explain nearly every part of the kernel that my code interacted with.

General Success As I wrote and tested code, I often found that my code worked on the first or second try. This was primarily due to careful and thoughtful reasoning about anything I wrote. In previous iterations, I often wrote code and ran it without fully knowing what to expect.

Squashing the Snapshot Bug Through the process of rewriting, I identified the Snapshot/Merge bug described above: I did not acquire a spinlock when operating on sensitive kernel data structures in my Snapshot code path. Unfortunately, this spinlock had very little accompanying documentation, and it was only through many months of kernel hacking that I even knew to use the spinlock.

9 Conclusion

Determinator introduced a novel parallel programming model that completely removes nondeterminism to provide predictability to user programs. Our work implemented Determinator's programming model in Linux and showed it is possible to efficiently run deterministic parallel applications on large input sizes. Improvements to our kernel implementation, enriching our user library, and upcoming hardware support for transactional memory may reduce the unacceptable overhead for parallel applications when run on small inputs.

References

[1] A. Arcangeli, I. Eidus, and C. Wright. Increasing memory density by using KSM. In *Proceedings of the Linux symposium*, pages 19–28, 2009.

[2] C. Artho, K. Havelund, and A. Biere. High-level data races. *Software Testing, Verification and Reliability*, 13(4):207–227, 2003.

[3] A. Aviram and B. Ford. Deterministic OpenMP for race-free parallelism. *3rd HotPar*, 2011.

[4] A. Aviram, S. Hu, B. Ford, and R. Gummadi. Determinating timing channels in compute clouds. In *CCSW*, 2010.

[5] A. Aviram, S. Weng, S. Hu, and B. Ford. Efficient system-enforced deterministic parallelism. In *OSDI*, 2010.

[6] F. Bellard. QEMU open source processor emulator. `http://www.qemu.org`.

[7] T. Bergan, O. Anderson, J. Devietti, L. Ceze, and D. Grossman. CoreDet: a compiler and runtime system for deterministic multithreaded execution. *ACM SIGARCH Computer Architecture News*, 38(1):53–64, 2010.

[8] T. Bergan, J. Devietti, N. Hunt, and L. Ceze. The deterministic execution hammer: How well does it actually pound nails? In *WoDet*, 2011.

[9] T. Bergan, N. Hunt, L. Ceze, and S. Gribble. Deterministic process groups in dOS. In *Proceedings of the 9th USENIX conference on Operating systems design and implementation*, pages 1–16. USENIX Association, 2010.

[10] E. Berger, T. Yang, T. Liu, and G. Novark. Grace: Safe multithreaded programming for c/c++. In *ACM SIGPLAN Notices*, volume 44, pages 81–96. ACM, 2009.

[11] R. Bocchino, V. Adve, S. Adve, and M. Snir. Parallel programming must be deterministic by default. In *First USENIX workshop on hot topics in parallelism (HotPar)*, 2009.

[12] R. Bocchino Jr, V. Adve, D. Dig, S. Adve, S. Heumann, R. Komuravelli, J. Overbey, P. Simmons, H. Sung, and M. Vakilian. A type and effect system for deterministic parallel Java. In *ACM SIGPLAN Notices*, volume 44, pages 97–116. ACM, 2009.

[13] J. Corbet. The case of the overly anonymous anon_vma. `http://lwn.net/Articles/383162/`.

[14] H. Cui, J. Wu, C. Tsai, and J. Yang. Stable deterministic multithreading through schedule memoization. *9th OSDI*, 2010.

[15] J. Devietti, B. Lucia, L. Ceze, and M. Oskin. DMP: Deterministic shared memory multiprocessing. In *ASPLOS*, 2009.

[16] J. Devietti, J. Nelson, T. Bergan, L. Ceze, and D. Grossman. RCDC: a relaxed consistency deterministic computer. *ACM SIGARCH Computer Architecture News*, 39(1):67–78, 2011.

[17] C. Flanagan, S. Freund, and J. Yi. Velodrome: a sound and complete dynamic atomicity checker for multithreaded programs. *ACM SIGPLAN Notices*, 43(6):293–303, 2008.

[18] B. Ford, M. Hibler, J. Lepreau, P. Tullmann, G. Back, and S. Clawson. Microkernels meet recursive virtual machines. In *OSDI*, 1996.

[19] M. T. Heath. Parallel numerical algorithms. `http://www.cse.uiuc.edu/courses/cs554/notes/06_lu.pdf`.

[20] M. K. Johnson. A rout of the Linux VFS. `http://www.tldp.org/LDP/khg/HyperNews/get/fs/vfstour.html`.

[21] F. Kaashoek et al. 6.828: Operating system engineering. `http://pdos.csail.mit.edu/6.828`.

[22] G. Kahn. The semantics of a simple language for parallel programming. In *Information Processing*, pages 471–475, 1974.

[23] C. Lattner and V. Adve. LLVM: A compilation framework for lifelong program analysis & transformation. In *Code Generation and Optimization, 2004. CGO 2004. International Symposium on*, pages 75–86. IEEE, 2004.

[24] T. LeBlanc and J. Mellor-Crummey. Debugging parallel programs with Instant Replay. *Computers, IEEE Transactions on*, 100(4):471–482, 1987.

[25] E. Lee et al. The problem with threads. *Computer*, 39(5):33–42, 2006.

[26] K. Li and P. Hudak. Memory coherence in shared virtual memory systems. *ACM Transactions on Computer Systems (TOCS)*, 7(4):321–359, 1989.

[27] S. Lu, S. Park, E. Seo, and Y. Zhou. Learning from mistakes: a comprehensive study on real world concurrency bug characteristics. In *ACM Sigplan Notices*, volume 43, pages 329–339. ACM, 2008.

[28] M. McKusick, W. Joy, S. Leffler, and R. Fabry. A fast file system for UNIX. *ACM Transactions on Computer Systems (TOCS)*, 2(3):181–197, 1984.

[29] P. Montesinos, L. Ceze, and J. Torrellas. Delorean: Recording and deterministically replaying shared-memory multiprocessor execution efficiently. In *Computer Architecture, 2008. ISCA'08. 35th International Symposium on*, pages 289–300. IEEE, 2008.

[30] M. Musuvathi, S. Qadeer, T. Ball, G. Basler, P. A. Nainar, and I. Neamtiu. Finding and reproducing heisenbugs in concurrent programs. In *OSDI*, 2008.

[31] R. Nelson and A. Tantawi. Approximate analysis of fork/join synchronization in parallel queues. *Computers, IEEE Transactions on*, 37(6):739–743, 1988.

[32] M. Olszewski, J. Ansel, and S. Amarasinghe. Kendo: efficient deterministic multithreading in software. In *14th ASPLOS*, Mar. 2009.

[33] D. Parker Jr, G. Popek, G. Rudisin, A. Stoughton, B. Walker, E. Walton, J. Chow, D. Edwards, S. Kiser, and C. Kline. Detection of mutual inconsistency in distributed systems. *Software Engineering, IEEE Transactions on*, (3):240–247, 1983.

[34] C. Sadowski, S. Freund, and C. Flanagan. SingleTrack: A dynamic determinism checker for multithreaded programs. *Programming Languages and Systems*, pages 394–409, 2009.

	lu				matmult			
Dimension	$N = 1$	$N = 2$	$N = 4$	$N = 8$	$N = 1$	$N = 2$	$N = 4$	$N = 8$
16×16	0.1 ± 0.1	0.8 ± 0.2	3.1 ± 0.3	15.1 ± 0.5	0.06 ± 0.01	0.1 ± 0.01	0.1 ± 0.01	0.5 ± 0.02
32×32	0.2 ± 0.03	0.9 ± 0.07	3.1 ± 0.2	15.9 ± 0.6	0.2 ± 0.01	0.2 ± 0.01	0.2 ± 0.01	0.5 ± 0.03
64×64	1.4 ± 0.07	1.9 ± 0.2	3.6 ± 0.2	14.7 ± 0.3	1.7 ± 0.01	1.3 ± 0.4	1.0 ± 0.01	0.6 ± 0.04
128×128	11.2 ± 0.08	8.8 ± 0.2	7.7 ± 0.2	16.0 ± 1.1	14.8 ± 0.02	8.7 ± 0.02	4.5 ± 0.02	2.5 ± 0.01
256×256	88.9 ± 0.2	65.6 ± 0.5	41.3 ± 1.0	29.6 ± 0.5	118 ± 0.04	69.0 ± 0.01	34.6 ± 0.01	17.6 ± 0.02
512×512	722 ± 25.3	497 ± 9.6	302 ± 6.5	180 ± 2.3	963 ± 25.0	558 ± 0.1	303 ± 49.6	138 ± 0.03
1024×1024	8,382 ± 35.2	4,136 ± 64.4	2,506 ± 41.7	1,373 ± 8.9	62,822 ± 14.4	38,405 ± 3,958	19,785 ± 1,331	9,646 ± 604
2048×2048	105,074 ± 42.6	72,694 ± 412	39,971 ± 405	25,395 ± 293	-	-	-	-

Table 8: Run times in milliseconds for pthread *lu* and *matmult*. Times are averaged over ten runs with standard deviations shown.

	lu				matmult			
Dimension	$N = 1$	$N = 2$	$N = 4$	$N = 8$	$N = 1$	$N = 2$	$N = 4$	$N = 8$
16×16	1.7 ± 0.06	33.7 ± 2.1	145 ± 6.9	465 ± 11.1	0.6 ± 0.02	2.4 ± 0.01	5.0 ± 0.4	8.2 ± 0.8
32×32	2.1 ± 0.04	34.4 ± 2.2	142 ± 5.5	463 ± 16.8	0.8 ± 0.00	2.6 ± 0.01	4.9 ± 0.6	8.0 ± 0.6
64×64	3.7 ± 0.02	39.5 ± 1.9	150 ± 6.6	471 ± 10.5	2.2 ± 0.00	4.9 ± 0.4	7.6 ± 0.7	8.3 ± 0.7
128×128	15.3 ± 0.2	53.1 ± 2.6	172 ± 5.4	492 ± 7.6	15.2 ± 0.2	16.3 ± 2.5	20.0 ± 2.1	17.0 ± 1.2
256×256	98.9 ± 0.3	138 ± 9.7	288 ± 6.4	555 ± 5.5	119 ± 0.10	80.0 ± 6.0	61.1 ± 4.9	41.0 ± 4.8
512×512	745 ± 25.9	605 ± 20.7	701 ± 23.5	1059 ± 22.3	957 ± 2.1	584 ± 41.0	346 ± 46.2	206 ± 33.1
1024×1024	11,863 ± 270	4,324 ± 133	3,013 ± 94.8	2,569 ± 63.6	63,480 ± 72.3	35,344 ± 1994	19,606 ± 542	11,569 ± 571
2048×2048	144,053 ± 209	72,420 ± 798	40,339 ± 1,362	27,018 ± 383	-	-	-	-

Table 9: Run times in milliseconds for deterministic *lu* and *matmult*. Times are averaged over ten runs with standard deviations shown.

	lu				matmult			
Dimension	$N = 1$	$N = 2$	$N = 4$	$N = 8$	$N = 1$	$N = 2$	$N = 4$	$N = 8$
16×16	0.7 ± 0.03	15.8 ± 1.0	66.2 ± 3.1	146 ± 6.8	0.3 ± 0.00	1.1 ± 0.01	2.3 ± 0.2	2.2 ± 0.8
32×32	0.7 ± 0.01	15.7 ± 1.1	65.4 ± 2.1	144 ± 7.0	0.3 ± 0.00	1.1 ± 0.01	2.1 ± 0.3	1.9 ± 0.5
64×64	0.7 ± 0.01	16.4 ± 0.8	66.1 ± 3.1	145 ± 4.4	0.3 ± 0.00	1.3 ± 0.5	2.6 ± 0.5	2.1 ± 0.5
128×128	0.4 ± 0.01	17.4 ± 0.9	67.0 ± 3.0	152 ± 2.9	0.04 ± 0.00	0.3 ± 0.04	2.7 ± 0.9	3.1 ± 0.7
256×256	0.5 ± 0.02	15.2 ± 2.5	74.5 ± 3.5	173 ± 5.3	0.05 ± 0.00	0.8 ± 0.1	1.0 ± 0.04	2.0 ± 0.2
512×512	0.6 ± 0.02	8.7 ± 0.7	65.5 ± 6.0	206 ± 11.1	0.06 ± 0.00	2.6 ± 0.4	3.0 ± 0.2	6.3 ± 0.6
1024×1024	1.0 ± 0.02	10.9 ± 2.1	44.3 ± 8.6	215 ± 25.9	0.1 ± 0.00	9.6 ± 1.7	11.1 ± 0.5	22.3 ± 1.3
2048×2048	2.2 ± 0.1	22.5 ± 3.6	60.1 ± 11.3	261 ± 52.7	-	-	-	-

Table 10: Time (milliseconds) spent in the kernel doing a virtual memory merge for the deterministic *lu* and *matmult* benchmarks.

	Pthread				Deterministic			
Input size	$N = 1$	$N = 2$	$N = 4$	$N = 8$	$N = 1$	$N = 2$	$N = 4$	$N = 8$
1K	0.2 ± 0.04	0.4 ± 0.1	1.2 ± 0.3	2.4 ± 0.4	5.7 ± 1.1	20.6 ± 5.0	40.0 ± 12.5	48.9 ± 15.5
4K	0.4 ± 0.03	0.5 ± 0.06	1.1 ± 0.1	2.4 ± 0.4	6.2 ± 0.01	24.3 ± 2.3	49.2 ± 6.8	60.2 ± 5.1
8K	0.8 ± 0.02	0.9 ± 0.2	1.4 ± 0.1	2.6 ± 0.3	6.6 ± 0.03	23.9 ± 2.4	45.3 ± 7.0	56.2 ± 4.4
10K	0.9 ± 0.01	1.0 ± 0.1	1.4 ± 0.2	2.3 ± 0.2	6.7 ± 0.03	24.4 ± 3.2	44.0 ± 6.8	59.9 ± 4.1
40K	4.0 ± 0.02	3.5 ± 0.6	3.1 ± 0.9	3.5 ± 0.9	9.9 ± 0.02	27.6 ± 3.5	45.6 ± 8.2	59.9 ± 7.2
80K	8.3 ± 0.02	6.8 ± 1.1	5.9 ± 1.0	5.4 ± 0.8	14.5 ± 0.6	30.5 ± 4.0	47.9 ± 8.6	57.7 ± 3.9
100K	10.5 ± 0.09	8.4 ± 1.1	7.3 ± 1.2	6.2 ± 0.9	16.9 ± 0.3	33.4 ± 2.7	48.0 ± 6.2	62.6 ± 6.2
400K	46.5 ± 0.1	37.4 ± 6.1	30.4 ± 5.5	27.0 ± 8.0	55.0 ± 0.6	62.5 ± 8.9	69.8 ± 8.9	74.9 ± 6.8
800K	97.6 ± 0.4	78.3 ± 7.3	66.6 ± 13.1	54.3 ± 12.3	108 ± 0.3	108 ± 17.5	106 ± 12.3	109 ± 11.3
1M	124 ± 0.7	91.6 ± 14.2	78.8 ± 15.4	68.0 ± 8.2	136 ± 0.4	134 ± 21.7	133 ± 16.7	131 ± 8.9
4M	549 ± 9.8	392 ± 38.8	316 ± 42.6	249 ± 30.4	574 ± 2.8	456 ± 52.1	406 ± 47.7	361 ± 33.3
8M	1,143 ± 11.1	844 ± 164	688 ± 164	526 ± 171	1194 ± 18.5	990 ± 205	1001 ± 211	865 ± 213
10M	1,450 ± 15.7	1,058 ± 172	867 ± 192	704 ± 113	1,510 ± 14.4	1,178 ± 231	1,000 ± 84.0	909 ± 70.4
40M	6,269 ± 21.4	4,788 ± 692	3,745 ± 819	2,969 ± 660	6,491 ± 16.6	5,059 ± 585	4,095 ± 1,119	3,691 ± 1,256
80M	13,060 ± 54.2	9,652 ± 1,852	7,743 ± 2,288	5,741 ± 1,110	13,474 ± 57.9	10,154 ± 1,859	8,412 ± 844	7,293 ± 538

Table 11: Run times (in milliseconds) for *pqsort*. Times are averaged over ten runs.