Architectural and Compiler Support for Strongly Atomic Transactional Memory

by

C. Scott Ananian

M.Sc Electrical Engineering and Computer Science,
Massachusetts Institute of Technology, 1999;
B.S.E. Electrical Engineering,
Princeton University, 1997.

Submitted to the Department of Electrical Engineering and Computer Science in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Electrical Engineering and Computer Science at the Massachusetts Institute of Technology.

May 25, 2007

Author .....
Department of Electrical Engineering and Computer Science
May 25, 2007

Certified by .....
Martin Rinard
Professor of Computer Science and Engineering
Thesis Supervisor

Accepted by .....
Arthur C. Smith
Chairman, Department Committee on Graduate Theses

Architectural and Compiler Support for Strongly Atomic Transactional Memory

by

C. Scott Ananian

Submitted to the
Department of Electrical Engineering and Computer Science
May 25, 2007
In partial fulfillment of the requirements for the Degree of Doctor of Philosophy in Electrical Engineering and Computer Science.

Abstract

Transactions are gaining ground as a programmer-friendly means of expressing concurrency, as microarchitecture trends make it clear that parallel systems are in our future. This thesis presents the design and implementation of four efficient and powerful transaction systems: ApeX, an object-oriented software-only system; UTM and LTM, two scalable systems using custom processor extensions; and HyApeX, a hybrid of the software and hardware systems, obtaining the benefits of both.

The software transaction system implements *strong atomicity*, which ensures that transactions are protected from the influence of nontransactional code. Previous software systems use weaker atomicity guarantees because strong atomicity is presumed to be too expensive. In this thesis strong atomicity is obtained with minimal slowdown for nontransactional code. Compiler analyses can further improve the efficiency of the mechanism, which has been formally verified with the Spin model-checker.

The low overhead of ApeX allows it to be profitably combined with a hardware transaction system to provide fast execution of short and small transactions, while allowing fallback to software for large or complicated transactions. I present UTM, a hardware transactional memory system allowing unbounded virtualizable transactions, and show how a hybrid system can be obtained.

Thesis Supervisor: Martin Rinard

Title: Professor of Computer Science and Engineering

Acknowledgments

I would like to thank my advisor Martin Rinard for his infinite patience, and for his guidance and support. I would like to thank Tom Knight for pointing me in the direction of transactional memory many years ago, and Charles Leiserson, Bradley Kuszmaul, Krste Asanović, and Sean Lie for the opportunity to collaborate on our hardware transactional memory research.

A list of thanks would not be complete without my "officemates" (real or virtual) who made my years at LCS and CSAIL so enjoyable: Maria-Cristina Marinescu (#2), Radu Rugina (#3), Darko Marinov (#5), Brian Demsky (#6), Alex Salcianu (#7), Karen Zee (#8), Patrick Lam (#10), Viktor Kuncak (#11), Amy Williams (#13), and Angelina Lee (who somehow has no number).

My mother, Cyndy Bernotas, waited a long time for this thesis to be complete. My brother, Chris Ananian, profited by the delay and became the first Dr. Ananian—but didn't rub it in too much. Jessica Wong and Kathy Peter put up with my indefinite studenthood. Thank you.

C. Scott Ananian
Cambridge, Massachusetts
May 2007

Contents

List of Figures

[Parallel programming] is hard.

Teen Talk Barbie
(apocryphal)

Chapter 1

Introduction

How can we fully utilize the coming generation of parallel systems? The primitives available to today's programmers are largely inadequate. ***Transactions*** have been proposed as an alternative to the existing mechanisms for expressing concurrency and synchronization, but current implementations of transactions are either too limited or too inefficient for general use. This thesis presents the design and implementation of efficient and powerful transaction systems to help address the challenges posed by current trends in computing hardware.

1.1 The rising challenge of multicore systems

Processor technology is nearing its limits: even though transistor quantities continue to grow exponentially, we are now unable to effectively harness those vast quantities of transistors to create speedier single processors. The smaller transistors yield relatively slower signal propagation times, dooming attempts to create a single synchronized processor from all of those resources. Instead, hardware manufacturers are providing tightly integrated ***multicore*** systems which integrate multiple parallel processors on one chip.

The widespread adoption of parallel systems creates problems: how can we ensure that operations occur in an appropriate order? How can we

ensure certain operations occur *atomically*, so that other components of the parallel system only observe data structures in well-defined states?

Atomicity in shared-memory multiprocessors is conventionally provided via mutual-exclusion ***locks*** (see, for example, [93, p. 35]). Although locks are easy to implement using test-and-set, compare-and-swap, or load-linked/store-conditional instructions, they introduce a host of difficulties. Protocols to avoid deadlock when locking multiple objects often involve acquiring the locks in a consistent linear order, making programming with locks error-prone and introducing significant time and space overheads in the resulting code. The granularity of each lock must also be explicitly chosen, as locks that are too fine introduce unnecessary space and time overhead, while locks that are too coarse sacrifice attainable parallelism (or may even deadlock). Every access to a shared object must hold some lock protecting that object, regardless of whether another thread is actually attempting to access the same object.

1.2 Advantages of transactions

Transactions are an alternative means of providing concurrency control. A transaction can be thought of as a sequence of loads and stores performed as part of a program which either ***commits*** or ***aborts***. If a transaction commits, then all of the loads and stores appear to have run atomically with respect to other transactions. That is, the transaction's operations are not interleaved with those of other transactions. If a transaction aborts, then none of its stores take effect and the transaction may be restarted, with some mechanism to ensure forward progress.

By structuring concurrency at a high level with transactions, human programmers no longer need to manage the details required to ensure atomicity. A full mental model of the global concurrency structure must be kept in mind when writing synchronization code, which programmers have difficulty correctly maintaining. The simpler "global atomicity" guaranteed

under the transactional model eliminates potential errors and simplifies the conceptual model of the system, making future modifications safer as well.

The transaction primitives presented in this thesis can exploit optimistic concurrency, provide fault tolerance, and prevent delays by using nonblocking synchronization. Although transactions can be implemented using mutual exclusion (locks), our algorithms utilize nonblocking synchronization [40, 54, 56, 65, 73] to exploit optimistic concurrency among transactions. Nonblocking synchronization offers several advantages; from the system builder's perspective the principle advantage is fault tolerance. In a traditionally constructed system, a process that fails or pauses while holding a lock within a critical region can prevent all other processes from making progress. It is in general not possible to restore locked data structures to a consistent state after such a failure. Nonblocking synchronization offers a graceful solution, as non-progress or failure of any one thread does not affect the progress or consistency of other threads or the system.

Implementing transactions using nonblocking synchronization offers performance benefits as well. When mutual exclusion is used to enforce atomicity, page faults, cache misses, context switches, I/O, and other unpredictable events may result in delays to the entire system. Nonblocking synchronization allows undelayed processes or processors to continue to make progress. In real-time systems, the use of nonblocking synchronization can prevent *priority inversion* in the system [60], although naive implementations may result in starvation of low-priority tasks (see Section 8.2 for a discussion).

1.3 Unlimited transactions

The ***transactional memory*** abstraction [49, 50, 63, 81, 87, 89], has been proposed as a general and flexible way to allow programs to read and modify disparate primary memory locations as a single operation (atomically), much as a database transaction can atomically modify many records on disk.

Hardware transactional memory (HTM) supports atomicity through architectural means, whereas ***software transactional memory*** (STM) supports atomicity through languages, compilers, and libraries. I will present both software and hardware implementations of the transaction model.

Researchers of both HTM and STM commonly express the opinion that transactions need never touch many memory locations, and hence it is reasonable to put a (small) bound on their size [49, 50].[1] For HTM implementations, they conclude that a small piece of additional hardware—typically in the form of a fixed-size content-addressable memory and supporting logic—should suffice. For STM implementations, some researchers argue additionally that transactions occur infrequently, and hence the software overhead would be dwarfed by the other processing done by an application. In contrast, this thesis supports transactions of arbitrary size and duration. Just as most programmers need not pay any attention to the exact size and replacement policy of their system's cache, programmers ought not concern themselves with limits or implementation details of their transaction system.

My goal is to make concurrent and fault-tolerant programming easier, without incurring excessive overhead. This thesis advocates unbounded transactions because neither programmers nor compilers can easily cope when an architecture imposes a hard limit on transaction size. An implementation might be optimized for transactions below a certain size, but must still operate correctly for larger transactions. The size of transactional hardware should be an implementation parameter, like cache size or memory size, which can vary without affecting the portability of binaries.

In Chapter 7 I show how a fast hardware implementation for frequent short transactions can gracefully fail over to a software implementation designed to efficiently execute large long-lived transactions. The hybrid ap-

[1] For example, [49, section 5.2] states, "Our [HTM] implementation relies on the assumption that transactions have short durations and small data sets"; while the STM described in [50] has quadratic slowdown when transactions touch many objects: performance is $O((R+W)R)$, where R and W are the number of objects opened for reading and writing, respectively.

proach allows more sophisticated transaction models to be implemented, while allowing a simpler hardware transaction mechanism to provide speed in the common case.

1.4 Strong atomicity

Blundell, Lewis, and Martin [17] distinguish between ***strongly atomic*** transaction systems, which protect transactions from interference from nontransactional code, and ***weakly atomic*** transaction systems which do not afford this protection. Nearly all current software transaction systems are weakly atomic,[2] however, despite the pitfalls thus opened for the unwary programmer, because of the perceived difficulty in efficiently implementing the required protection.

Strong atomicity is clearly preferable, as the programmer will inevitably overlook some nontransactional references to shared data; we wish to preserve correctness in this common case. Blundell *et al.* point out that programs written for a weakly atomic model (to run on a current software transaction system, say) may deadlock when run under strong atomicity (for example, on a hardware transaction system). The transaction systems considered in this thesis preserve the correct atomic behavior of transactions even in the face of unsynchronized accesses from outside the transaction.

1.5 Summary of contributions

In summary, this thesis makes the following contributions:

- I provide efficient implementations of *strongly atomic* transaction primitives to enable their general use. My experiments have not

[2] There are some systems which use type systems to disallow nontransactional access to objects with a "shared" type [45, 50, 84], but to my knowledge all systems which allow (or cannot prevent) access to shared objects from nontransactional code do so unsafely [26, 31, 44, 51, 53, 87].

shown an overhead of more than 150% in real applications, although proper compiler support should reduce that worst case considerably.

- The transaction primitives presented in this thesis can exploit optimistic concurrency, provide fault tolerance, and prevent delays by using nonblocking synchronization.

- This thesis proposes systems which can support transactions of arbitrary size and duration, sparing the programmer from detailed knowledge of the system's implementation details.

- I present both software and hardware implementations of the transaction model. The software transaction system runs real programs written in Java; I discuss the practical implementation details encountered. The hardware transaction systems require only small changes to the processor core and cache.

- I show how a fast hardware implementation for frequent short transactions can gracefully fail over to a software implementation designed to efficiently execute large long-lived transactions.

In Chapter 2 I provide some concurrent programming examples that illustrate the limitations of current lock-based methodologies. I also provide examples illustrating the uses (some novel) of a transaction system, and conclude with a brief caution about the current limits of transactions.

In Chapter 3 I present the design of ApeX, an efficient software-only implementation of transactions. Software-only transactions can be implemented on current hardware, and can easily accommodate many different transaction and nesting models. Software transactions excel at certain long-lived transactions, where the overhead is small compared to the transaction execution time. An early version of ApeX was published as [4]. I conclude by presenting a microbenchmark that demonstrates the performance limits of the design.

In Chapter 4 I discuss the practical implementation of ApeX. I present details of the compiler analyses and transformations performed, as well as solutions to problems that arise when implementing Java. I then present benchmark results using real applications, discuss the benefits and limitations revealed, and describe how the limits could be overcome.

Chapter 6 explores one such limitation in depth, describing how large arrays fit into an object-oriented transaction scheme. I present a potential solution to the problem based on fast functional arrays.

In Chapter 7 I present LTM and UTM, hardware systems that enable fast short transactions. The transaction model is more limited, but short committing transactions may execute with no overhead. The additional hardware is small and easily added to current processor and memory system designs. This portion of the chapter is joint work with Krste Asanović, Bradley C. Kuszmaul, Charles E. Leiserson, and Sean Lie, and has been previously published as [6, 7].

At the end of the chapter, I present HyApeX, a ***hybrid*** transaction implementation that builds on the strengths of simple hardware support while allowing software fallback to support a robust and capable transaction mechanism. Unlike the extended hardware scheme, the transaction model is still easy to change and update; the hardware primarily supports fast small transactions and conflict checking in the common case. I discuss some ways compilers can further optimize software and hybrid transaction systems. These opportunities may not be available to pure-hardware implementations.

In Chapter 8 I discuss remaining challenges to the use of ubiquitous transactions for synchronization, and present some ideas toward solutions. Chapter 9 discusses related work, and my final chapter summarizes my findings and draws conclusions.

You have a hardware or a software problem.

Service manual for Gestetner 3240

Chapter 2

Transactional programming: The good, bad, and the ugly

Before diving into the design of an efficient transaction system, I motivate the transactional programming model by presenting four common scenarios that are needlessly difficult using lock-based concurrency. I then present four novel applications that a transactional model facilitates. To ground the discussion in reality, I conclude by enumerating a few cases where transactions may *not* be the best solution.

2.1 Four old things you can't (easily) do with locks

Locks engender poor modularity and composability, an inability to deal gracefully with asynchronous events, and fragile and complex safety protocols that are often expressed externally to their implementations. These limitations of locks are well-known [52]. I present four common tasks which locks make unnecessarily difficult: making localized changes to synchronization, performing atomic operations on containers, creating a thread-safe double-ended queue, and handling asynchronous exceptions.

Tweak performance with localized changes

Preventing deadlocks and races requires global protocols and non-local reasoning. It is not enough to simply specify a lock of a certain granularity protecting certain data items; the order or circumstances in which the lock may be acquired and released must also be specified globally, in the context of all locks in the system, in order to prevent deadlocks or unexpected races. This requirement prevents the programmer from easily tuning the system using localized changes: every small change must be re-verified against the protocols specified in the whole-program context in order to prevent races and/or deadlock.

Furthermore, this whole-program protocol is not typically expressed directly in code. With common programming languages, acquire/release ordering and guarantees must be expressed externally, often as comments in the source code that easily drift out of sync with the implementation. For example, in [6] we counted the comments in the Linux filesystem layer, and found that about 15% of these relate to locking protocols; often describing global invariants of the program which are difficult to verify. Many reported kernel bugs involve races and deadlocks.

Atomically move data between thread-safe containers

Another common programming pitfall with locks is their ***non-composability***. For example, given two thread-safe container classes implemented with locks, it is impossible to safely compose the *get* function of one with the *put* function of the other to create an atomic move. We must peek inside the implementation of the containers to synthesize an appropriate locking mechanism for such an action—for example, to acquire the appropriate container, element, or other locks on *both* containers—and even then, we need to resort to some global lock ordering to guard against deadlock. Modularity must be broken in order to synthesize the appropriate composed function, if it is possible at all.

In the Jade programming language, Rinard presents a partial solution using "implicitly synchronized" objects [83, p14]. Lock acquisition for each module is exposed in the module's API as an executable "access declaration." Operation composition is accomplished by creating an access declaration for the composed operation which invokes the appropriate access declarations for the components. The runtime system orders the lock acquisitions to prevent deadlock. This process suffices for conservative mutual exclusion, but pre-declaration makes it difficult to express optimistic locking protocols, where one or more locks are only rarely required.

Transactions do not suffer from the composability problem [45]. Because transactions only specify the atomicity properties, not the locks required, the programmer's job is made easier and implementations are free to optimistically synchronize in any way that preserves atomicity.[1]

Create a thread-safe double-ended queue

Herlihy suggests creating a thread-safe double-ended queue using locks as "sadistic homework" for computer science students [52]. Although double-ended queues are a simple data structure, creating a scalable locking protocol is a non-trivial exercise. One wants dequeue and enqueue operations to complete concurrently when the ends of the queue are "far enough" apart, while safely handling the interference in the small-queue case. In fact, the solution to this assignment was a publishable result, as Michael and Scott demonstrated in 1996 [75].

The simple "one lock" solution to the double-ended queue problem, ruled out as unscalable in the locking case, is scalable and efficient for nonblocking transactions [52, 56].

[1] Section 2.2 presents a concrete example.

Handle asynchronous exceptions

Properly handling asynchronous events is difficult with locks, because it is impossible to safely go off to handle the event while holding an arbitrary set of locks—and it is impossible to safely drop the locks. The solution implemented in the Real-Time Specification for Java [19] and similar systems (see [71, section 9]) is to generally forbid asynchronous events within locked regions, allowing the programmer to explicitly specify certain points within the region at which execution can be interrupted, dropping all locks in order to do so. Maintaining the correctness in the face of even explicitly declared interruption points is still difficult.

Transactional atomic regions handle asynchronous exceptions gracefully, by aborting the transaction to allow an event to occur.

2.2 Four new things transactions make easy

I present four examples in this section, illustrating how transactions can support fault tolerance and backtracking, simplify locking, and provide a more intuitive means for specifying thread-safety properties. I first examine a destructive traversal algorithm, showing how a transaction implementation can be treated as an exception-handling mechanism. I show how a variant of this mechanism can be used to implement backtracking search. Using a network flow example, I then show how the transaction mechanism can be used to simplify the locking discipline required when synchronizing concurrent modifications to multiple objects. Finally, I show an existing race in the Java standard libraries (in the class `java.lang.StringBuffer`). "Transactification" of the existing class corrects this race.

Destructive traversal

Many recursive data structures can be traversed without the use of a stack by using pointer reversal. This technique is widely used in garbage collectors,

```
// destructive list traversal.
void traverse(List l) {
  List last = null, t;

  /* zip through the list, reversing links */
  for (int i=0; i<2; i++) {
    do {
      if (i==0) visit(l); // visit node
      t = l.next;
      l.next = last;
      last = l;
      l = t;
    } while (l!=null);
    l = last;
    // now do again, backwards. (restoring links)
  }
}
```

Figure 2.1: Destructive linked-list traversal.

and was first demonstrated in this context by Schorr and Waite [86]. An implementation of a pointer-reversal traversal of a simple singly-linked list is shown in Figure 2.1.

The `traverse()` function traverses the list, visiting nodes in order and then reversing the `next` pointer. When the end of the list is reached, the reversed links are traversed to restore the list's original state.

Of course, I have chosen the simplest possible data structure here, but the technique works for trees and graphs—and the reader may mentally substitute their favorite hairy update on a complicated data structure.

In normal execution, the data structure is left complete and intact after the operation. But imagine that an exception or fault occurs inside the `visit()` method at some point during the traversal: an assertion fires, an exception occurs, the hardware hiccups, or a thread is killed. Control may leave the `traverse()` method, but the data structure is left in shambles. What is needed is some exception-handling procedure to restore the proper state of the list. This handler can be manually coded with Java's existing `try/catch` construct, but the exception-handling code must be tightly-coupled to the traversal if it is going to undo the list mutations.

Instead, I can provide a non-deterministic choice operator, `try/else`, and write the recovery code at a higher level as:

```
try {
  traverse(list);
} else { // try-else construct
  throw new Error();
}
```

The `try/else` block appears to make a non-deterministic choice between executing the `try` or the `else` clause, depending on whether the `try` would succeed or not. This construct can be straightforwardly implemented with a transaction around the traversal, always initially attempting the `try`. Exceptions or faults cause the transaction to abort; when it does so all the heap side-effects of the `try` block disappear.

Backtracking search

Introducing an explicit `fail` statement allows us to use the same `try/else` to facilitate backtracking search. Backtracking search is used to implement practical[2] regular expressions, parsers, logic programming languages, Scrabble-playing programs [9], and (in general) any problem with multiple solutions or multiple solution techniques.

As a simple example, let us consider a recursive-decent parser such as that shown in Figure 2.2. We don't know whether to apply the `sum()` or `difference()` production until after we've parsed some common left prefix. We can use backtracking to attempt one rule (`sum`) and `fail` out of it inside the `eat()` method, in the process undoing any data structure updates performed on this path, and then attempt the other possible production.

Optimistic synchronization

Let's now turn our attention now to parallel codes, the more conventional application of transaction systems. Consider a serial program for computing

[2] As opposed to the limited regular expressions demonstrated in theory classes which are always neatly compiled to deterministic finite automata [33].

```
char buffer[];
int  pos;

void eat(char token) {
  if (buffer[pos++] != token)
    fail;
}

int expr() {
   try {
     return sum();
   } else {
     return difference();
   }
}

int sum() {
   int a = number();
   eat(ADD);
   int b = number();
   return a+b;
}

int difference() {
   int a = number();
   eat(MINUS);
   int b = number();
   return a-b;
}
```

Figure 2.2: A simple backtracking recursive-decent parser, using transactional `try/else/fail`.

network flow (see, for example, [24, Chapter 26]). The inner loop of the code pushes flow across an edge by increasing the "excess flow" on one vertex and decreasing it by the same amount on another vertex. One might see the following Java code:

```
void pushFlow(Vertex v1, Vertex v2, double flow) {
  v1.excess += flow; /* Move excess flow from v1 */
  v2.excess -= flow; /* to v2.                   */
}
```

To parallelize this code, one must preclude multiple threads from modifying the excess flow on those two vertices at the same time. Locks provide one way to enforce this mutual exclusion:

```
void pushFlow(Vertex v1, Vertex v2, double f) {
  Object lock1, lock2;
  if (v1.id < v2.id) {       /* Avoid deadlock. */
    lock1 = v1; lock2 = v2;
  } else {
    lock1 = v2; lock2 = v1;
  }
  synchronized(lock1) {
    synchronized(lock2) {
      v1.excess += f; /* Move excess flow from v1 */
      v2.excess -= f; /* to v2.                   */
    } /* unlock lock2 */
  } /* unlock lock1 */
}
```

This code is surprisingly complicated and slow compared to the original. Space for each object's lock must be reserved. To avoid deadlock, the code must acquire the locks in a consistent linear order, resulting in an unpredictable branch in the code. In the code shown, I have required the programmer to insert an `id` field into each vertex object to maintain a total ordering. The time required to acquire the locks may be an order of magnitude larger than the time to modify the excess flow. What's more, all of this overhead is rarely needed! For a graph with thousands or millions of vertices, the number of threads operating on the graph is likely to be less than a hundred. Consequently, the chances are quite small that two different threads actually conflict. Without the locks to implement mutual exclusion, however, the program would occasionally fail.

Software transactions (and some language support) allow the programmer to parallelize the original code using an `atomic` keyword to indicate that the code block should appear to execute atomically:

```
void pushFlow(Vertex v1, Vertex v2, double flow) {
  atomic { /* Transaction begin. */
    v1.excess += flow; /* Move excess flow from v1 */
    v2.excess -= flow; /* to v2.                   */
  } /* Transaction end. */
}
```

This `atomic` region can be implemented as a transaction, and with an appropriately nonblocking implementation, it will scale better, execute faster, and use less memory than the locking version [6, 40, 44, 49, 73, 87]. From the programmer's point of view, I have also eliminated the convoluted locking protocol which must be observed rigorously everywhere the related fields are accessed, if deadlock and races are to be avoided.

Further, I can implement `atomic` using the `try/else` exception-handling mechanism I have already introduced:

```
for (int b=0; ; b++) {
  try {
    // atomic actions
  } else {
    backOff(b);
    continue;
  }
  break; // success!
}
```

I non-deterministically choose to execute the body of the `atomic` block if and only if it will be observed by all to execute atomically. The same linguistic mechanism I introduced for fault tolerance and backtracking provides atomic regions for synchronization as well.

Bug fixing

The existing ***monitor synchronization*** methodology for Java, building on such features in progenitors such as Emerald [16, 61],[3] implicitly associates

[3] See Section 9.4.

an lock with each object. Data races are prevented by requiring a thread to acquire an object's lock before touching the object's shared fields. The lack of races is not sufficient to prevent unanticipated parallel behavior, however.

Flanagan and Qadeer [29] demonstrated this insufficiency with an actual bug they discovered in the Sun JDK 1.4.2 Java standard libraries. The `java.lang.StringBuffer` class, which implements a mutable string abstraction, is implemented as follows:

```
public final class StringBuffer ... {
  private char value[];
  private int count;
  ...
  public synchronized
  StringBuffer append(StringBuffer sb) {
    ...
A:  int len = sb.length();
    int newcount = count + len;
    if (newcount > value.length)
      expandCapacity(newcount);
    // next statement may use stale len
B:  sb.getChars(0, len, value, count);
    count = newcount;
    return this;
  }
  public synchronized int length() { return count; }
  public synchronized void getChars(...) { ... }
}
```

The library documentation indicates that the methods of this class are meant to execute atomically, and the `synchronized` modifiers on the methods are meant to accomplish this.

The `append()` method is *not* atomic, however. Another thread may change the length of the parameter `sb` (by adding or removing characters) between the call to `sb.length()` at label A and the call to `sb.getChars(...)` at label B. This non-atomicity may cause incorrect data to be appended to the target or a `StringIndexOutOfBoundException` to be thrown. Although the calls to `sb.length()` and `sb.getChars()` are individually atomic, they do not compose to form an atomic implementation of `append()`.

Replacing `synchronized` with `atomic` in this code gives us the semantics we desire: the atomicity of nested `atomic` blocks is guaranteed by the

atomicity of the outermost block, ensuring that the entire operation appears atomic.

Both the network flow example and this `StringBuffer` example require synchronization of updates to more than one object. Monitor synchronization is not well-suited to this task. Atomic regions implemented with transactions can be used to simplify the locking discipline required to synchronize multiobject mutation and provide a more intuitive specification for the desired concurrent properties. Further, the `StringBuffer` example shows that simply replacing `synchronized` with `atomic` provides a alternative semantics that may in fact correct existing synchronization errors. For many Java programs, the semantics of `atomic` and `synchronized` are identical; see Section 8.3.

2.3 Some things we still can't (easily) do

The transaction mechanism presented here is not a universal replacement for all synchronization. In particular, transactions cannot replace mutual exclusion required to serialize I/O, although the needed locks can certainly be built with transactions. The challenge of integrating I/O within a transactional context is discussed in Section 8.4. Large programs—the Linux kernel, for example—have been written such that locks are never held across context switches or I/O operations, however. Transactions provide a complete solution for this limited form of synchronization.

Blocking producer-consumer queues, or other options that require a transaction to wait upon a condition variable, may introduce complications into a transaction system: transactions cannot immediately retry when they fail, but instead must wait for some condition to become true. Section 4.4.4 describes some solutions to this problem, ranging from naive (keep retrying and aborting with exponential backoff until the condition is finally true) to clever; the clever solutions require additional transaction machinery.

Easy things should stay easy,
hard things should get easier,
and impossible things should get
hard.

Motto for Perl 6 development

Chapter 3

Designing a software transaction system

In this chapter I detail the design of ApeX, a high-performance software transaction system. I first present a methodology for isolating likely transactions from benchmarks. I then describe the system goals, and use quantitative data from benchmarks to buttress the design choices. I formalize an implementation meeting those goals in the modeling language Promela. The correctness of this implementation can be model-checked using the Spin tool; the details of this effort are in Appendix A.

After outlining the design of ApeX in this chapter, Chapter 4 discusses its practical implementation.

3.1 Finding transactions

One of the difficulties of proposing a novel language feature is the lack of benchmarks for its evaluation. Although there is no existing body of code that uses transactions, there is a substantial body of code that uses Java (locking) synchronization. This thesis utilizes the Flex compiler [3] to substitute `atomic` blocks (methods) for `synchronized` blocks (methods) in order to evaluate the properties Java transactions are likely to have. The

semantics are not precisely compatible: the existing Java memory model allows unsynchronized updates to shared fields to be observed within a synchronized block, while such updates are never visible to a transaction expressed with an `atomic` block.[1] Despite the differences in semantics, the automatic substitution of `atomic` for `synchronized` does, in fact, preserve the correctness of the benchmarks I examine here.

The initial results of this chapter explore the implications of exposing the transaction mechanism to user-level code through a compiler. I compiled the SPECjvm98 benchmark suite with the FLEX Java compiler, modified to turn synchronized blocks and methods into transactions, in order to investigate the properties of the transactions in such "automatically converted" code. FLEX performed method cloning to distinguish methods called from within transactions, and implemented nested locks as a single transaction around the outermost.[2] I instrumented this transformed program to produce a trace of memory references and transaction boundaries for analysis. I found both large transactions (involving millions of cache lines) and frequent transactions (up to 45 million of them). We will show that these properties are not unusual for typical applications.

The SPECjvm98 benchmark suite represents a variety of typical Java applications that use the capabilities of the Java standard library. Although the SPECjvm98 benchmarks are largely single-threaded, they contain synchronized code within the thread-safe Java standard libraries which is transformed into transactions. Because in this evaluation I am looking at transaction properties only, the multithreaded `227_mtrt` benchmark is identical to its serialized version, `205_raytrace`. For consistency, I present only the latter.

[1] The Java 5.0 revision of the Java memory model [68–70] narrows the semantic gap. The full memory model includes details—such as `volatile` fields—which I do not treat in this work. Section 8.3 discusses the challenges involved in automatic transactification, and [41] contains a fuller discussion of the interactions between transactions and high-level memory models.

[2] See Section 4.2 for more details on this transformation.

program	total memory ops	transactions	transactional memory ops	biggest transaction
201_compress	2,981,777,890	2,272	<0.1%	2,302
202_jess	405,153,255	4,892,829	9.1%	7,092
205_raytrace	420,005,763	4,177	1.7%	7,149,099
209_db	848,082,597	45,222,742	23.0%	498,349
213_javac	472,416,129	668	99.9%	118,041,685
222_mpegaudio	2,620,818,169	2,991	<0.1%	2,281
228_jack	187,029,744	12,017,041	34.2%	14,266

Figure 3.1: Transactification of SPECjvm98 benchmark suite: resulting transaction counts and sizes, compared to total number of memory operations (loads and stores). These numbers represent full input size runs.

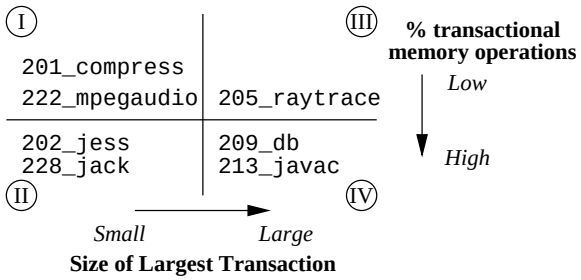

Figure 3.2: Classification of SPECjvm98 benchmarks into quadrants based on transaction properties.

Figure 3.1 shows the raw sizes and frequency of transactions in the transactified SPECjvm98 suite. Because the run times of the benchmarks are roughly comparable,[3] transaction count is also an indicator of transaction rate. Figure 3.2 proposes a taxonomy for Java applications with transactions, grouping the SPECjvm98 applications into quadrants based on the number and size of the transactions that they perform. Applications in Quads II and IV require an efficient transaction implementation, because they contain many transactional operations. Quads III and IV contain at least some large transactions, which pose difficulties for currently proposed hardware transactional memory schemes. We now examine the benchmarks in each quadrant to determine why its program logic caused it to be classified in that quadrant.

Quad I applications perform few (up to about 2000) small transactions. These applications include `201_compress`, an implementation of gzip compression, and `222_mpegaudio`, an MP3 decoder. Both of these applications perform inherently serial tasks. They perform quite well with locks, and would likely execute with acceptable performance even with a naive software implementation of transactions, as long as the impact on nontransactional operations was minimal.

Quad II applications perform a large number of small transactions. The expert system `202_jess` falls in this category, as does the parser generator `228_jack` and small input sizes of `209_db`, a database. These benchmarks perform at least an order of magnitude more transactions than Quad I applications, and all of the transactions are small enough to comfortably fit the known hardware transactional memory schemes [49, etc], if one were to be implemented.

Quad III includes `205_raytrace`, a ray-tracing renderer. A small number of transactions are performed, but they may grow large. Existing bounded

[3] The canonical run times used for computation of the Spec ratio range from 380 to 1175 seconds. Sorted by run time: `202_jess`, 380 s; `213_javac`, 425 s; `228_jack`, 455 s; `227_mtrt`, 460 s; `209_db`, 505 s; `222_mpegaudio`, 1100 s; `201_compress`, 1175 s.

hardware transactional schemes do not suffice. The large transactions may account for a large percentage of total memory operations, which may make software schemes impractical.

Finally, Quad IV applications such as 209_db and the 213_javac Java compiler application perform a large number of transactional memory operations with at least a few large transactions.

The 213_javac Java compiler application and the large input size of the 209_db benchmark illustrate that some programs contain *extremely* large transactions. When 213_javac is run on its full input set, it contains 4 huge transactions, each of which contains over 118 million transactional memory operations. Closer examination reveals that the method `Javac.compile()`, which implements the entire compilation process, is marked as synchronized: the programmer has explicitly requested that the entire compilation occur atomically.

The large transactions in Quad III and IV may be, as in this case, a result of overly coarse-grained locking, but my goal is to relieve the programmer from the burden of specifying correct atomic regions of smaller granularity. Performance may benefit from narrowing the atomic regions, but execution with coarse regions should be possible and not prohibitively slow.

The 209_db benchmark suffers from a different problem: at one point the benchmark atomically scans an index vector and removes an element, creating a potentially large transaction if the index is large. The size of this index is correlated in these benchmarks with the input size, but it need not be: a large input could still result in a small index, and (to some degree) vice-versa.

A similar situation arises in the `java.lang.StringBuffer` code shown in Section 2.2: a call to the synchronized `sb.getChars()` method means that the size of the transaction for this method grows like the length of the parameter `sb`. In other words, the transaction can be made arbitrarily large by increasing the length of `sb`; or, equivalently, there is no bound on transaction size without a bound on the size of the string `sb`.

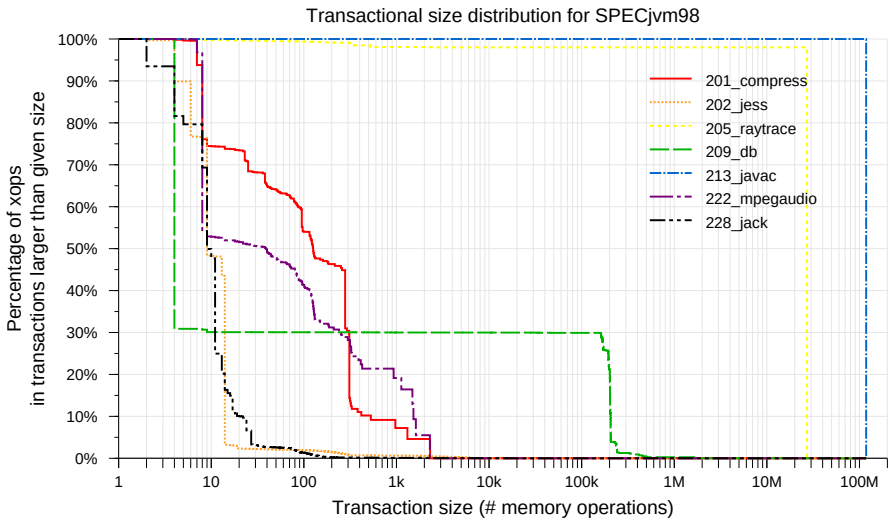

Figure 3.3: Distribution of transaction size in the SPECjvm98 benchmark suite. The x-axis uses a logarithmic scale.

Any transaction system that allows the programmer free reign over specifying desired transaction and/or atomicity properties will result in some applications in each of these categories. Large transactions, for example, are a side-effect of modularity: when a cross-module call occurs inside a transaction, the abstraction boundary prevents precise knowledge of the memory accesses implicitly included. Existing hardware transactional memory schemes only efficiently handle relatively short-lived and small transactions (Quad I or II), although they are efficient for these transactions. Object-based transaction systems can asymptotically approach that efficiency for long-lived transactions; the existence of which is shown in Figure 3.3, which plots the distribution of transaction sizes in SPECjvm98 on a semi-log scale.

These initial results indicate that real applications can be transactified with modest effort, yielding significant gains in concurrency. In other work [6] we have shown that a factor of 4 increase in concurrency can be obtained by doing nothing more than converting locks to transactions. Since the

transactified applications may contain large transactions, prior proposed hardware support for transactions is inadequate.

3.2 Design goals

In this section I briefly describe the desired properties of the ApeX software transaction system: strong atomicity, object-orientation, low bloat, and fast reads. Where possible I justify these desiderata using quantitative data obtained from analyses of the SpecJVM98 benchmarks, which I implemented using the Flex Java compiler framework.

3.2.1 Weak vs. strong atomicity

As previously discussed in Section 1.4, ***strongly atomic*** transaction systems protect transactions from interference from nontransactional code, while ***weakly atomic*** transaction systems do not afford this protection. Consider unsynchronized code directly altering the length field of the `StringBuffer` class, the example discussed in Section 2.2. In a weakly atomic system, an unsynchronized decrement to the `count` field between labels A and B in `StringBuffer.append()` on page 30 causes a `StringIndexOutOfBoundException` to be thrown in the call to `getChars()` at label B. This exception should never be thrown by an atomic execution of `StringBuffer.append()`.

Current software transaction systems implement only weak atomicity because of the assumed expense of implementing strong atomicity. The ApeX system achieves strong atomicity without adding excessive overhead, so that correct operation is assured even despite concurrent nontransactional operations on locations involved in a transaction.

3.2.2 Object-oriented vs. flat TM

The ApeX transaction system, unlike most current proposals [44, 49] (including the UTM and LTM hardware systems presented in Chapter 7), uses

an "object-oriented" design. Much contemporary research is focused on implementing a flat (transactional) memory abstraction in software, primarily because flat systems side-step the large object issues presented in Chapter 6. The object-oriented approach, however, offers several benefits:

Efficient execution of long-running transactions. As discussed briefly in Sections 9.2 and 9.3, flat word-oriented transaction schemes typically require overhead proportional to the number of words read/written in the transaction, even if these locations have been accessed before inside the transaction. Object-oriented schemes impose a cost proportional to the number of *objects* touched by the transaction—but once the cost of "opening" those objects is paid, the transaction can continue to work indefinitely upon those objects without paying any further penalty. Object-oriented schemes are thus seen to be more efficient for *long-running transactions*.

Preservation of optimization opportunities. Furthermore, transaction-local objects can be identified (statically or dynamically) and creation/updates to these objects can be done without any transaction tax at all. Word oriented schemes discard the high-level information required to implement these optimizations.

I contend that the problems with previous object-oriented schemes can be solved while preserving the inherent benefits of an object-oriented approach, and the current thesis presents one such solution.

3.2.3 Tolerable limits for object expansion

An object-oriented transaction scheme requires transaction state information about each object. In ApeX, this information is added directly to the objects, to eliminate indirection cost. I measured the slowdown caused by various amounts of object "bloat" to determine reasonable bounds on the size of this extra information. Figure 3.4 presents these results for the

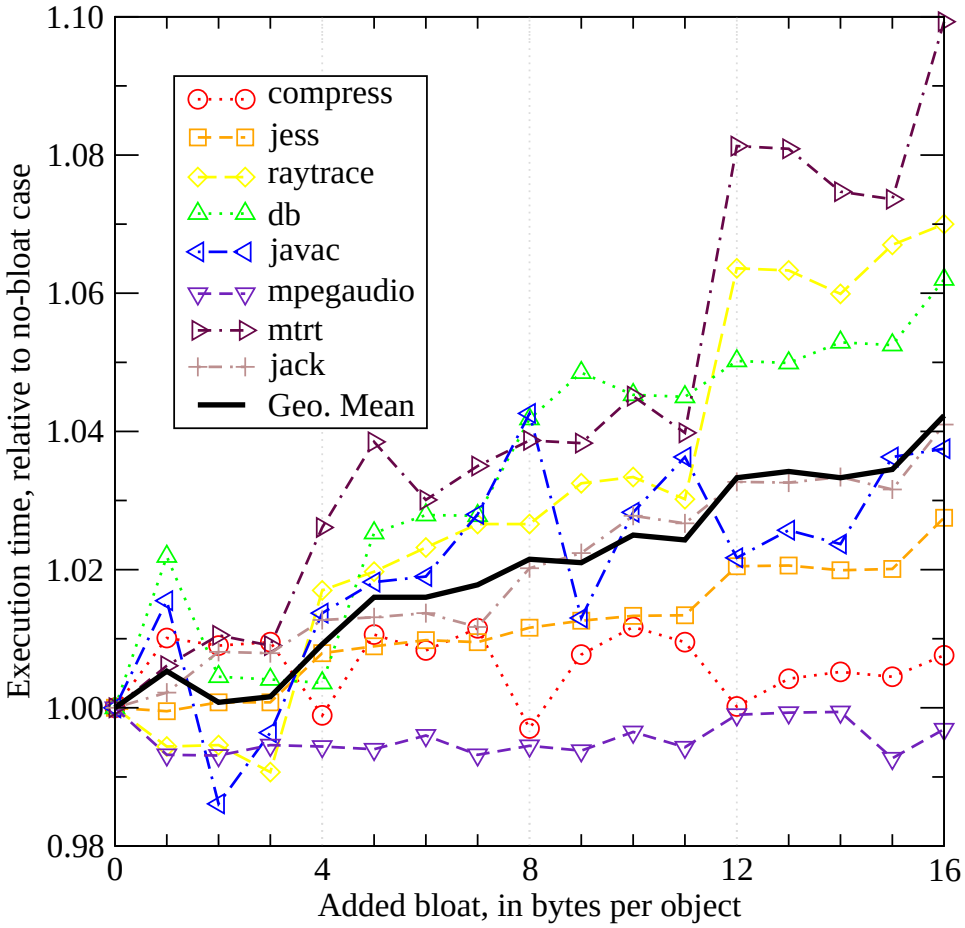

Figure 3.4: Application slowdown with increasing object bloat for the SPECjvm98 benchmark applications.

program	transactional memory ops	transactional stores %
201_compress	50,029	26.2%
202_jess	36,701,037	0.6%
205_raytrace	7,294,648	23.2%
209_db	195,374,420	6.3%
213_javac	472,134,289	22.9%
222_mpegaudio	41,422	18.6%
228_jack	63,912,386	17.0%

Figure 3.5: Comparison of loads and stores inside transactions for the SPECjvm98 benchmark suite, full input runs.

SPECjvm98 applications; I determined that two words (eight bytes) of additional storage per object would not impact performance unreasonably. This amount of bloat causes a geometric mean of 2% slowdown on these benchmarks.

3.2.4 Reads vs. writes

I also measured the number and types of reads and writes for the SpecJVM98 benchmarks. Figure 3.5 shows that transactional reads typically outnumber transactional writes by at least 4 to 1; in some cases reads outnumber writes by over 100 to 1.[4] It is worthwhile, therefore, to make reads more efficient than writes. In particular, since the flag-overwrite technique discussed in Section 3.2.5 requires us to allocate additional memory to store the "real" value of the field, I wish to avoid this process for transactional reads, reserving the extra allocation effort for transactional writes.

[4] The typical ratio roughly matches the 3:1 average observed in Hennessy and Patterson [46, pp. 105, 379].

3.2.5 The big idea: Waving FLAGs

I would like nontransactional code to execute with minimal overhead, but transactions should still appear atomic to nontransactional code. My basic mechanism is loosely based on the distributed shared memory implementation of Scales and Gharachorloo [85]. I pick a special "flag" value, and "cross-out" locations currently involved in a transaction by overwriting them with the flag value. Reading or attempting to overwrite a flagged value indicates to nontransactional code that exceptional processing is necessary; all other nontransactional operations proceed as usual. The flagged field either is currently involved in an active transaction, belongs to a committed or aborted transaction and requires copy back, or indicates a "false flag", a field with a true value equal to the flag value (see page 46).

This technique explicitly allows safe access to fields involved in a transaction from nontransactional code, which provides strong atomicity, a design goal of the system.

3.3 Specifying the basic mechanism

I now present algorithms that have these desired properties. The ApeX algorithms provide strongly-atomic object-oriented transactions with low bloat and fast reads. Further, the ApeX algorithms are completely nonblocking, which allows good scaling and proper fault-tolerant behavior. Specifically, one faulty or slow processor cannot hold up the remaining good processors.

I implement the synchronization required by the ApeX algorithms using load-linked/store-conditional instructions. I require a particular variant of these instructions that allows the location of the load-linked to be different from the target of the store-conditional. This variant is supported on many chips in the PowerPC processor family, although it has been deprecated in version 2.02 of the PowerPC architecture standard.[5] This disjoint location

[5] Version 2.02 of the PowerPC architecture standard says, "If a reservation exists but the storage location specified by the `stwcx.` is not the same as the location specified

capability is essential to allow us to keep a finger on one location while modifying another: a poor man's "Double Compare And Swap" instruction.

I describe the APEX algorithms in the Promela modeling language [57], which I used to allow mechanical model checking of the race-safety and correctness of the design. Portions of the model have been abbreviated for this presentation; the full Promela model is given in Appendix A, along with a brief primer on Promela syntax and semantics.

3.3.1 Object structures

Figure 3.6 illustrates the basic data structures of the APEX software transaction implementation. Objects are extended with two additional fields. The first field, `versions`, points to a singly linked list of object versions. Each one contains field values corresponding to a committed, aborted, or in-progress transaction, identified by its `owner` field. There is a single unique transaction object for each transaction.

The other added field, `readers`, points to a singly linked list of transactions that have read from this object. Committed and aborted transactions are pruned from this list. The `readers` field is used to ensure that a transaction does not operate with out-of-date values if the object is later written

by the *Load And Reserve* instruction that established the reservation... it is undefined whether [the operand is] stored into the word in storage addressed by [the specified effective address]" and states that the condition code indicating a successful store is also undefined in this circumstance [58, p 25]. The user manual for the MPC7447/7457 ("G4") PowerPC chips states, however, "The `stwcx.` instruction does not check the reservation for a matching address. The `stwcx.` instruction is only required to determine whether a reservation exists. The `stwcx.` instruction performs a store word operation only if the reservation exists," [32, Section 3.3.3.6] which is the behavior we require. I believe version 1.10 of the PowerPC Architecture specification required this behavior, although I have not been able to locate a copy to confirm this requirement. The Cell architecture specification follows version 2.02 of the PowerPC specification, although it adds cache-line reservation operations that can also be used to implement our algorithm for reasonably-sized objects aligned within cache lines; see [98] for an implementation of Java I created that observes the appropriate alignments.

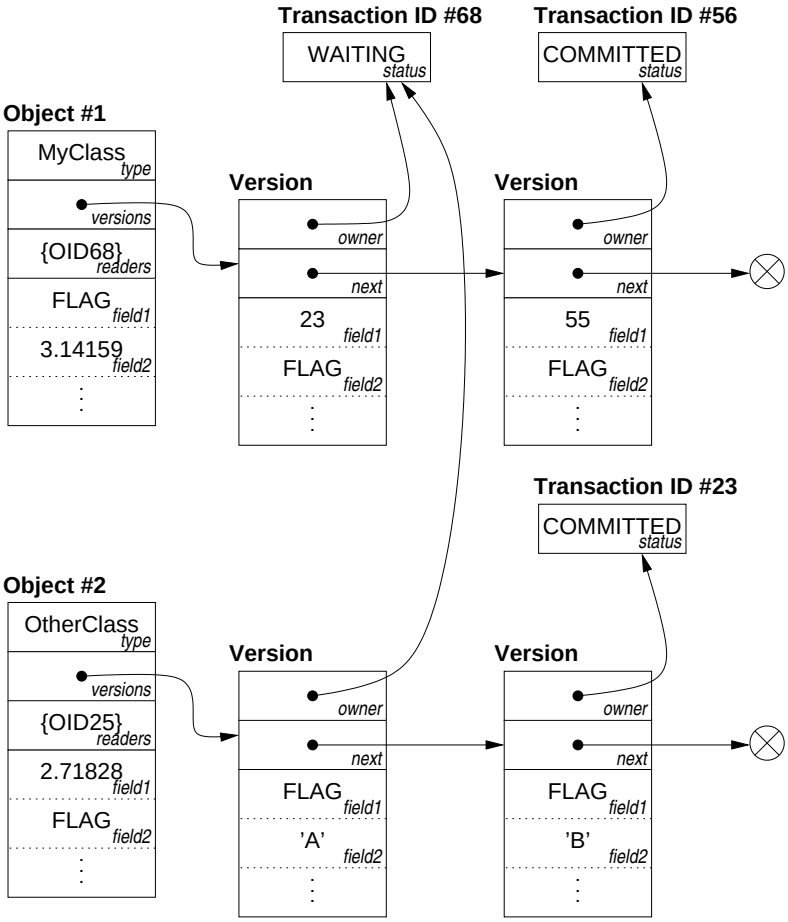

Figure 3.6: Implementing software transactions with version lists. A transaction object consists of a single field *status*, which indicates if it has COMMITTED, ABORTED, or is WAITING. Each object contains two extra fields: *readers*, a singly-linked list of transactions that have read this object; and *versions* a linked list of version objects. If an object field is FLAG, then the value for the field is obtained from the appropriate linked version object.

```
#define FLAG 202 /* special value to represent 'not here' */

typedef Object {
  byte version;
  byte readerList; /* we do LL and CAS operations on this field */
  pid fieldLock[NUM_FIELDS]; /* we do LL operations on fields */
  byte field[NUM_FIELDS];
};
typedef VersiOn { /* 'Version' misspelled because SPIN #define's it. */
  byte owner;
  byte next;
  byte field[NUM_FIELDS];
};
typedef ReaderList {
  byte transid;
  byte next;
};
mtype = { waiting, committed, aborted };
typedef TransID {
  mtype status;
};
```

Figure 3.7: Declaring objects with version lists in Promela. The `byte` datatype encodes pointers. The `fieldLock` field assists in the implementation of the load-linked/store-conditional pair of operations in Promela.

nontransactionally.

There is a special flag value, here denoted by `FLAG` . It should be an uncommon value, i.e. not a small positive or negative integer constant, nor zero. In my implementation, I have chosen the byte `0xCA` to be the flag value, repeated as necessary to fill out the width of the appropriate type. The semantic value of an object field is the value in the original object structure, *unless that value is FLAG* , in which case the field's value is the value of the field in the first committed transaction in the object's version list. A "false flag" occurs when the application wishes to "really" store the value `FLAG` in a field. To do so we create a fully-committed version attached to the object and store `FLAG` in that version, as well as in the object field.

Figure 3.7 declares these object structures in Promela.

3.3.2 Operations

I support transactional read/write and nontransactional read/write as well as transaction begin, transaction abort, and transaction commit. Transaction begin simply involves the creation of a new transaction identifier object. Transaction commit and abort are simply compare-and-swap operations that atomically set the transaction object's `status` field appropriately if and only if it was previously in the WAITING state. The simplicity of commit and abort are appealing: the ApeX algorithm requires no complicated processing, delay, roll-back or validate procedure to commit or abort a transaction.

I could support nontransactional read and write (that is, reads and writes that take place outside of any transaction) by creating a new short transaction that encloses only the single read or write. Since nontransactional accesses to objects can be frequent, I provide more efficient implementations with the same semantics.

I now present the operations one-by-one.

Nontransactional read

The `readNT` function performs a nontransactional read of field `f` from object `o`, putting the result in v. In the common case, the only overhead is to check that the read value is not `FLAG` . If the value read *is* `FLAG`, however, we copy back the field value from the most-recently committed transaction (aborting all other transactions) and try again. The copy-back procedure notifies the caller if this is a "false flag", in which case the value of this field really is `FLAG` . We pass the `kill_writers` constant to the copy-back procedure to indicate that only transactional writers need be aborted, not transactional readers. All possible races are confined to the copy-back procedure, which I describe on page 49.

The top of Figure 3.8 specifies the nontransactional read operation in Promela.

```
inline readNT(o, f, v) {
  do
  :: v = object[o].field[f];
     if
     :: (v!=FLAG) -> break /* done! */
     :: else
     fi;
     copyBackField(o, f, kill_writers, _st);
     if
     :: (_st==false_flag) ->
        v = FLAG;
        break
     :: else
     fi
  od
}

inline writeNT(o, f, nval) {
  if
  :: (nval != FLAG) ->
     do
     :: atomic {
          if /* this is a LL(readerList)/SC(field) */
          :: (object[o].readerList == NIL) ->
             object[o].fieldLock[f] = _thread_id;
             object[o].field[f] = nval;
             break /* success! */
          :: else
          fi
        }
        /* unsuccessful SC */
        copyBackField(o, f, kill_all, _st)
     od
  :: else -> /* create false flag */
     /* implement this as a short *transactional* write. */
     /* start a new transaction, write FLAG, commit the */
     /* transaction; repeat until successful. */
     /* Implementation elided. */
     ...
  fi;
}
```

Figure 3.8: Promela specification of nontransactional read and write operations.

Nontransactional write

The `writeNT` function performs a nontransactional write of new value *nval* to field `f` of object `o`. For correctness, we must ensure that the reader list is empty before we do the write. I implement this check with a load-linked/store-conditional pair, which is modeled in Promela slightly differently, ensuring that the write only succeeds so long as the reader list remains empty.[6] If it is not empty, we call the copy-back procedure (as in `readNT`), passing the constant `kill_all` to indicate that both transactional readers and writers should be aborted during the copy-back. The copy-back procedure leaves the reader list empty.

If the value to be written is actually the `FLAG` value, things get a little bit trickier. This case does not occur often, and so the simplest correct implementation is to treat this nontransactional write as a short transactional write, creating a new transaction for this one write, and attempting to commit it immediately after the write. This implementation is slow, but adequate for this uncommon case.

The bottom of Figure 3.8 specifies the nontransactional write operation in Promela.

Field Copy-Back

Figures 3.9 and 3.10 present the field copy-back routine. We create a new version owned by a pre-aborted transaction which serves as a reservation on the head of the version list. We then write to the object field with a load-linked/store-conditional pair if and only if our version is still at the head of the versions list.[7] This addresses the major race possible in this routine.

[6] A standard CAS would not suffice, as the load-linked targets a different location than the store-conditional.

[7] Again, a CAS does not suffice.

```
inline copyBackField(o, f, mode, st) {
  _nonceV=NIL; _ver = NIL; _r = NIL; st = success;
  /* try to abort each version.  when abort fails, we've got a committed
   * version. */
  do
  :: _ver = object[o].version;
     if
     :: (_ver==NIL) ->
        st = saw_race; break /* someone's done the copyback for us */
     :: else
     fi;
      /* move owner to local var to avoid races
       * (owner set to NIL behind our back) */
     _tmp_tid=version[_ver].owner;
     tryToAbort(_tmp_tid);
     if
     :: (_tmp_tid==NIL || transid[_tmp_tid].status==committed) ->
        break /* found a committed version */
     :: else
     fi;
     /* link out an aborted version */
     assert(transid[_tmp_tid].status==aborted);
     CAS_Version(object[o].version, _ver, version[_ver].next, _);
  od;
  /* okay, link in our nonce.  this will prevent others from doing the
   * copyback. */
  if
  :: (st==success) ->
     assert (_ver!=NIL);
     allocVersion(_retval, _nonceV, aborted_tid, _ver);
     CAS_Version(object[o].version, _ver, _nonceV, _cas_stat);
     if
     :: (!_cas_stat) ->
        st = saw_race_cleanup
     :: else
     fi
  :: else
  fi;
  /* check that no one's beaten us to the copy back, then kill readers. */
  checkCopyAndKill(o,f,mode,st);
  /* done */
}
```

Figure 3.9: First half of the field copy-back routine. The final steps have been split off into `checkCopyAndKill`, which is presented in Figure 3.10.

```
inline checkCopyAndKill(o, f, mode, st) {
  /* check that no one's beaten us to the copy back */
  if
  :: (st==success) ->
     if
     :: (object[o].field[f]==FLAG) ->
        _val = version[_ver].field[f];
        if
        :: (_val==FLAG) -> /* false flag... */
           st = false_flag /* ...no copy back needed */
        :: else -> /* not a false flag */
           d_step { /* LL/SC */
             if
             :: (object[o].version == _nonceV) ->
                object[o].fieldLock[f] = _thread_id;
                object[o].field[f] = _val;
             :: else /* hmm, fail.  Must retry. */
                st = saw_race_cleanup /* need to clean up nonce */
             fi
           }
        fi
     :: else /* may arrive here because of readT, which doesn't set _val=FLAG*/
        st = saw_race_cleanup /* need to clean up nonce */
     fi
  :: else /* !success */
  fi;
  /* always kill readers, whether successful or not.  This ensures that we
   * make progress if called from writeNT after a readNT sets readerList
   * non-null without changing FLAG to _val (see immediately above; st will
   * equal saw_race_cleanup in this scenario). */
  if
  :: (mode == kill_all) ->
     killAllReaders(o, _r); /* see Appendix for details */
  :: else /* no more killing needed. */
  fi;
  /* done */
}
```

Figure 3.10: Final portion of the field copy-back routine of Figure 3.9.

Transactional Read

A transactional read is split into two parts. The first part, which occurs before the read, is encapsulated in a routine named `ensureReader`. Its primary task is to ensure that our transaction is on the reader list for the object. This check is straightforward to do in a nonblocking manner as long as we always add readers to the head of the list. The `ensureReader` routine must also walk the versions list to abort any uncommitted transaction other than our own. Since the `ensureReader` depends only on the object involved, not the precise field, redundant read checks for different fields in the object can be combined and hoisted so that `ensureReader` is performed once before the first read from an object and not repeated.

At read time, we initially read from the original object. If the value read is not `FLAG`, we use it. Otherwise, we look up the version object associated with our transaction (our version is typically at the head of the version list) and read the appropriate value from that version. The initial read-and-check can be omitted if we know that we have already written to this field inside this transaction.

The top of Figure 3.11 specifies read-time portion of the transactional read operation in Promela. The initial `ensureReader` portion is straightforward; it is included in Appendix A.

Transactional Write

Again, writes are split in two. Once for each object, we must traverse the version list, aborting other versions and locating or creating a version corresponding to our transaction. We must also traverse the reader list, aborting all transactions on the list except ourself. This portion of the algorithm is shown in the `ensureWriter` routine in Figure 3.12.

Once for each field we intend to write, we must perform a copy-through: copy the object's field value into all the versions and then write `FLAG` to the object's field. We use load-linked/store-conditional to update versions

```
inline readT(tid, o, f, ver, result) {
  do
  ::
     /* we should always either be on the readerList or
      * aborted here */
     result = object[o].field[f];
     if
     :: (result==FLAG) ->
        if
        :: (ver!=NIL) ->
           result = version[ver].field[f];
           break /* done! */
        :: else ->
           findVersion(tid, o, ver);
           if
           :: (ver==NIL) ->/*use val from committed vers.*/
              assert (_r!=NIL);
              result = version[_r].field[f];/*false flag?*/
              moveVersion(_r, NIL);
              break /* done */
           :: else /* try, try, again */
           fi
        fi
     :: else -> break /* done! */
     fi
  od
}

inline writeT(ver, f, nval) {
  /* easy enough: */
  version[ver].field[f] = nval;
}
```

Figure 3.11: Promela specification of transactional read and write operations.

```
/* per-object, before write. */
inline ensureWriter(tid, o, ver) {
  assert(tid!=NIL);
  ver = NIL; _r = NIL; _rr = NIL;
  do
  :: assert (ver==NIL);
     findVersion(tid, o, ver);
     if
     :: (ver!=NIL) -> break /* found a writable version for us */
     :: (ver==NIL && _r==NIL) ->
        /* create and link a fully-committed root version, then
         * use this as our base. */
        allocVersion(_retval, _r, NIL, NIL);
        CAS_Version(object[o].version, NIL, _r, _cas_stat)
     :: else ->
        _cas_stat = true
     fi;
     if
     :: (_cas_stat) ->
        /* so far, so good. */
        assert (_r!=NIL);
        assert (version[_r].owner==NIL ||
                transid[version[_r].owner].status==committed);
        /* okay, make new version for this transaction. */
        assert (ver==NIL);
        allocVersion(_retval, ver, tid, _r);
        /* want copy of committed version _r.  No race because
         * we never write to a committed versions. */
        version[ver].field[0] = version[_r].field[0];
        version[ver].field[1] = version[_r].field[1];
        assert(NUM_FIELDS==2); /* else ought to initialize more fields */
        CAS_Version(object[o].version, _r, ver, _cas_stat);
        moveVersion(_r, NIL); /* free _r */
        if
        :: (_cas_stat) ->
           /* kill all readers (except ourself) */
           /* all changes have to be made from the front of the
            * list, so we unlink ourself and then re-add us. */
           do
           :: moveReaderList(_r, object[o].readerList);
              if
              :: (_r==NIL) -> break
              :: (_r!=NIL && readerlist[_r].transid!=tid)->
                 tryToAbort(readerlist[_r].transid)
              :: else
              fi;
              /* link out this reader */
              CAS_Reader(object[o].readerList, _r, readerlist[_r].next, _)
           od;
           /* okay, all pre-existing readers dead & gone. */
           assert(_r==NIL);
           /* link us back in. */
           ensureReaderList(tid, o);
           break
        :: else
        fi;
        /* try again */
     :: else
     fi;
     /* try again from the top */
     moveVersion(ver, NIL)
  od;
  /* done! */
  assert (_r==NIL);
}
```

Figure 3.12: The per-object version-setup routine for transactional writes.

```
/* per-field, before write. */
inline checkWriteField(o, f) {
  _r = NIL; _rr = NIL;
  do
  ::
     /* set write flag, if not already set */
     _val = object[o].field[f];
     if
     :: (_val==FLAG) ->
        break; /* done! */
     :: else
     fi;
     /* okay, need to set write flag. */
     moveVersion(_rr, object[o].version);
     moveVersion(_r, _rr);
     assert (_r!=NIL);
     do
     :: (_r==NIL) -> break /* done */
     :: else ->
        object[o].fieldLock[f] = _thread_id;
        if
        /* this next check ensures that concurrent copythroughs don't stomp on each other's versions,
         * because the field will become FLAG before any other version will be written. */
        :: (object[o].field[f]==_val) ->
           if
           :: (object[o].version==_rr) ->
              atomic {
                if
                :: (object[o].fieldLock[f]==_thread_id) ->
                   version[_r].field[f] = _val;
                :: else -> break /* abort */
                fi
              }
           :: else -> break /* abort */
           fi
        :: else -> break /* abort */
        fi;
        moveVersion(_r, version[_r].next) /* on to next */
     od;
     if
     :: (_r==NIL) ->
        /* field has been successfully copied to all versions */
        atomic {
          if
          :: (object[o].version==_rr) ->
             assert(object[o].field[f]==_val ||
                    /* we can race with another copythrough and that's okay; the locking strategy
                     * above ensures that we're all writing the same values to all the versions
                     * and not overwriting anything. */
                    object[o].field[f]==FLAG);
             object[o].fieldLock[f]=_thread_id;
             object[o].field[f] = FLAG;
             break; /* success!  done! */
          :: else
          fi
        }
     :: else
     fi
     /* retry */
  od;
  /* clean up */
  moveVersion(_r, NIL);
  moveVersion(_rr, NIL);
}
```

Figure 3.13: The per-field copy-through routine for transactional writes.

only if the object's field has not already been set to `FLAG` behind our backs by another copy-through. The `checkWriteField` routine is shown in Figure 3.13.

Then, for each write, we simply write to the identified version, as shown at the bottom of Figure 3.11.

Implementation

Some details are not yet covered by our model. For example, how are fields of different byte lengths handled? How does transactional Java code interact with the garbage collector, or with native code in the runtime system? Chapter 4 addresses the real-world implementation details of our software transaction system.

In theory, there is no difference between theory and practice.

Jan L. A. van de Snepscheut

Chapter 4

ApeX implementation: Efficient software transactions

This chapter discusses the challenges involved in transforming the design of Chapter 3 into a practical Java compiler. I describe the foundation of the implementation, the Flex compiler infrastructure and its "Precise C" backend. I then outline the transformations and analyses Flex performs. The runtime system must also be modified to support transactions; Flex modifies the Java Native Interface to allow most native libraries to work seamlessly with ApeX. I list some limitations of the present implementation, including access granularity and support for static fields and condition variables, and discuss how they may be overcome in a production-quality compiler. I conclude by listing additional optimizations which could be implemented to improve ApeX performance.

4.1 The Flex compiler infrastructure

In 1998 I began implementing the Flex compiler infrastructure for Java [3], which I used to implement the transaction systems in this thesis. Flex

is a whole-program static compiler for Java with a runtime system built around the GNU Classpath implementation of the Java standard libraries [36]. FLEX takes as input Java bytecode, generated by any Java 1.0-1.6 compiler from user code and the GNU Classpath 0.08 library implementation, and it emits either assembly code for MIPS, Sparc, or StrongARM, or else "portable assembly language" written in gcc 3.4's variant of C. The emitted code is compiled and linked against the FLEX runtime system and the GNU Classpath 0.08 native libraries to produce a stand-alone binary for the target system. The FLEX compiler and analysis code is written in Java, while the FLEX runtime system is written in C. FLEX has been used in over 20 published papers [2, 4–7, 13, 14, 20, 25, 28, 30, 34, 35, 62, 82, 90–92, 95–99, 101], on a wide range of topics.

Using C as a target

The experiments in this thesis were conducted on either the x86 or PowerPC architectures, for which FLEX does not have a native assembly backend. The "Precise C" backend was thus used, so named because, aside from emitting C code, it's original purpose was to investigate the possibility of precise garbage collection while emitting high-level code. To that end, the backend contains code to maintain a separate stack for values of non-primitive types and to push and pop all live variables to this stack at gc points. Since FLEX's low-level IR may create derived pointers that point inside objects, for example during loop optimizations, the Precise C backend also reconstructs the derived pointers after the gc point, in case the objects involved have moved.

Experiments showed that this mechanism has minimal impact on performance, because variables pushed onto the explicitly managed "object stack" are then dead across the call from the perspective of the underlying C compiler; in effect our explicit stack management replaced the implicit stack save/restore which the C compiler would otherwise have performed to maintain its calling convention.

Java exceptions presented another difficulty when implementing a C backend. The mechanisms used in our assembly backends (separate return addresses for "returning" an exception, either derived by rule from the normal return address or stored in a sorted table keyed by the normal return address) cannot be implemented in C. FLEX supports two alternate translations: the first uses `setjmp` and `longjmp` to branch directly from the site where the exception is thrown to an appropriate handler, and the second returns a pair value from every function call. In the `setjmp` method FLEX emits a `setjmp` call when a new exception-handling region (`try/catch` block) is entered and a `longjmp` when an exception is thrown, but `setjmp` and `longjmp` are rather expensive. The pair-return method returns a C `struct` from every call consisting of the "real" return value along with an exception value.[1] When the function returns, the caller must first test the exception value; if it is non-`NULL`, then an exception has been thrown and the caller must handle it (if it can) or re-throw it. The mandatory test in the pair-return method adds overhead to every function call, but it is minimal.

Experiments on x86 indicated that a small benchmark consisting of 50,000 method calls executed in 2 milliseconds (ms) using pair-return, regardless of whether the calls returned normally or threw an exception. The same benchmark using the `setjmp` method ran in only 1ms when all calls returned normally, but took 73ms when each call threw an exception. As a result, `setjmp` is dramatically slower for applications that heavily use exceptions. For example, the SPECjvm98 benchmark `228_jack` executes in 460 seconds with pair return, but 703 seconds with `setjmp`. Figure 4.1 presents full results for the SPECjvm98 benchmark suite.

The experiments in this thesis all use pair-return to implement exceptions. All experiments also use conservative garbage collection [18], making the "Precise C" stack unnecessary.

[1] Methods declared as `void` return only the exception value, of course.

	execution time (seconds)		
benchmark	assembly	pair return	`setjmp`
`201_compress`	223.2	193.0	183.7
`202_jess`	-	446.1	463.3
`205_raytrace`	265.8	854.8	858.3
`209_db`	591.3	578.1	597.7
`222_mpegaudio`	1,210.3	3,035.0	3,014.0
`227_mtrt`	277.0	866.0	867.9
`228_jack`	895.7	460.1	703.0

Figure 4.1: Speed comparison of exception return techniques for SPECjvm98 benchmarks on a Netwinder platform. The Netwinder is a small, low power computing platform designed by Corel Computer around the StrongARM 110 processor. For the presented results, lower execution time is better. The "assembly" column uses FLEX's StrongARM backend and uses a lookup table keyed on the procedure return address to find an appropriate exception handler. The "pair return" and "`setjmp`" columns use FLEX's "Precise C" backend, and use the implementation strategies described in the text.

```
                                t = CommitRecord.newTransaction();
                                while(true) {
                                  try {
                                    ...
                                    v = ensureReader(o, t);
synchronized {                      x = readT(o, FIELD_F, &v, t);
    ...                             ...
    x = o.f;                        v = ensureWriter(o, t);
    ...                             checkWriteField(o, FIELD_F);
    o.f = y;            ⇒           writeT(o, FIELD_F, y, v); /* v.f = y */
    ...                             ...
    z = foo(a, b, c);               z = foo$$withtrans(t, a, b, c);
    ...                             ...
}                                   t.commitTransaction();
                                    break;
                                  } catch (TransactionAbortException tex) {
                                    t = t.retryTransaction();
                                  }
                                }
```

Figure 4.2: Software transaction transformation. The code on the left is the original Java source; the transformed source is on the right.

4.2 Transforming synchronization

Aside from implementing the read and write mechanisms of the software transaction design presented in the previous chapter, several other transformations and analyses of our Java benchmarks must be performed: transactions must be synthesized from the benchmark's monitor synchronization, methods must be cloned according to the context (whether inside a transaction or not) of their call site, and analyses must be performed in order to reduce redundant checks in the transformed code. In addition, some minor desugaring is done to ease implementation.

4.2.1 Method transformation

I automatically created transactions in the benchmarks I present from Java synchronized blocks and methods. Figure 4.2 illustrates the transformation applied.

Entering a top-level synchronized block creates a new transaction object and starts a new exception handler context. Exiting the block causes a call

to the transaction's commit method. If the commit, or any other operation inside the block, throws `TransactionAbortException` indicating that the transaction must be aborted, we recreate the transaction object and loop to retry the transaction's operations. The `retryTransaction` method performs backoff. Although not implemented in this work, `retryTransaction` may also perform livelock detection or other transaction management functions.

Inside the transaction context, reads and writes must be transformed. Before the read, the `ensureReader` algorithm must be invoked (see Section 3.3.2). This is only done once for each object read in this transaction. Section 4.2.2 describes how these checks are hoisted and combined to reduce their number. The `ensureReader` routine returns a pointer to a ***version object*** containing this transaction's field values for the object. The pointer may be NULL, however, if this transaction has not written the object. The actual read is done via the `readT` algorithm described in Figure 3.11 and Section 3.3.2; it may update the cached version object for the given object (for example, if the object has been written within this transaction since the point at which `ensureReader` was invoked).[2]

Before each write, `ensureWriter` and `checkWriteField` must be executed. Like `ensureReader`, `ensureWriter` (Figure 3.12) need only be performed once for each object written in the transaction, and it is hoisted and combined in the same way. The `checkWriteField` algorithm (Figure 3.13) need only be executed once for each object field written; if the same field in the same object is written multiple times, the subsequent `checkWriteField` invocations can be eliminated. The actual write is performed via `writeT` (Figure 3.11), which is a simple store to the version object.

Nested transactions are implemented via subsumption; that is, nested

[2]Since the FLEX infrastructure cannot create pointers to temporaries, or alternately return multiple values from function calls, the version object in my current implementation cannot be updated by `readT`. This limitation impairs the ability to hoist and combine `ensureReader` and requires redundant calls in our current implementation.

synchronized blocks inside a transaction context are ignored: the inner transaction is subsumed by the outermost.

Method invocations inside transaction context must be transformed, since read and write operations are implemented differently depending on whether or not the current execution is inside a transaction. FLEX creates a transactional clone from each method, named with a `$$withtrans` suffix, which is invoked when executing inside a transaction. The current transaction is passed as the first parameter of the cloned method.[3]

Nontransactional code inside a method must be transformed to use the `readNT` and `writeNT` mechanisms to perform object reads and writes (Figure 3.8); the implementation is similar to that shown in Figure 5.2 and Figure 5.3.

4.2.2 Analyses

Performance improves if FLEX can identify objects or fields that are guaranteed not to be concurrently mutated and replace the checks and read/write protocols with direct accesses. I performed two analyses of this sort; Section 4.5 describes additional analyses which we have not implemented.

Our first analysis classifies all fields in the program according to their use in transactional and nontransactional regions. This analysis is encapsulated in a FLEX class named `GlobalFieldOracle`. Fields that can never be written within a transaction do not need the special `readNT` mechanism from nontransactional code; their values can be loaded directly without testing for `FLAG`. The only way that the field can have the `FLAG` value, if it is not written within a transaction, is if it is a false flag, in which case its value really is `FLAG`. Similarly, fields that can never be read or written within a transaction do not need to use the `writeNT` mechanism; they can just store directly to the field.[4] Thus, `GlobalFieldOracle` can make nontransactional

[3] Cloning must take virtual dispatch into account: cloning a method in a superclass must also create appropriate cloned subclass implementations.

[4] It is not enough for the field not to be written; we must also protect against write-

code more efficient by eliminating the overhead of the software transaction system in some cases.

While `GlobalFieldOracle` targets nontransactional code, the `CheckOracle` analysis classes make transactional code more efficient by removing unneeded read and write checks. As mentioned above, the `ensureReader` and `ensureWriter` checks need only be invoked once on each unique object read/written by the transaction. Similarly, `checkWriteField` need only be performed once on each object field written by the transaction. The `CheckOracle` classes hoist each check to its immediate dominator [67, 77, 94] if and only if this new location for the check is postdominated by the current location and the new location is dominated by the definition of the variable referenced in the check. This condition ensures that all paths to the original location of the check must pass through the new location. Moreover, it ensures that the object involved in the check is still defined appropriately at the new check location, since the analysis is performed on a single-assignment intermediate representation [2]. The check-hoisting process is repeated until no checks can be moved higher, and then FLEX eliminates any check that is dominated by an identical check.

4.2.3 Desugaring

FLEX also desugars some Java idioms to ease implementation. Java specifies a `clone()` method of `java.lang.Object`, from which every object is derived. The top-level `clone()` method is somewhat magical, as it creates an exact copy of its subclass, including all its fields, which are otherwise hidden from methods of a superclass. The transactional translation of this method is also somewhat baroque: it would have to mark all fields of the original method as "read" by the transaction, and then construct a versioned object written by the transaction. As an end-run around this complexity, we desugar `clone()` into individual implementations in appropriate classes,

after-write conflicts.

each of which reads all fields of the class, and constructs a new object bypassing its constructors, and then individually sets all the fields to their new values. Arrays get a similar `clone()` implementation that reads and writes the elements of the array. These new implementations can then be transformed by the transaction pass like any other piece of code that reads and writes fields.

The FLEX infrastructure also contains a mechanism for fast initialization of arrays. This mechanism is desugared into individual array set operations so that it, too, can be uniformly transformed by the transaction pass.

4.3 Runtime system implementation

The FLEX runtime system was extended to support transactions. The runtime uses the standard Java Native Interface (JNI) [76] for native methods (written in C) called from Java. Certain parts of the transaction system were written as native methods invoked via JNI, but others interact with the runtime system at a much lower level.

The runtime system is parameterized to allow a large number of memory allocation strategies and collectors, but for the experiments reported here, I used the Boehm-Demers-Weiser conservative garbage collector [18]. Ideally the garbage collector would know enough about the transaction implementation to be able to automatically prune unneeded committed or aborted versions from the object's version list during collection. FLEX relies on opportunistically nulling out tail pointers during version list traversal,[5] however, which may result in incomplete collections (and more memory usage than strictly necessary).

Each transaction had a `CommitRecord` object storing its state, whether `COMMITTED`, `ABORTED`, or still `WAITING`. Most `CommitRecord` methods were written in Java, including object creation, exponential backoff on retry, and throwing the appropriate `TransactionAbortException` on abort. Only the

[5] For example, while looking for the last committed version.

crucial `commit()` and `abort()` operations, which atomically set the transaction status if and only if it is still `WAITING`, were written in C, as JNI methods.

4.3.1 Implementing the Java Native Interface

No Java transaction implementation is complete without some mechanism for executing native methods, that is, C code. One cannot banish all native methods, since the Java standard library is built on them. If one cannot make native code observe the `readNT` and `writeNT` protocols, one must ensure that the native code never reads any object written in a transaction, or writes any object read or written in a transaction.[6] These restrictions would be inconvenient.

My solution ensures not only that (separately compiled) native code properly uses the `readNT` and `writeNT` protocols outside a transaction, but also that reads and writes inside a transaction are handled properly. This ability allows the use of "safe" native methods (those without external side effects) inside a transaction.

I am able to transform native methods because FLEX uses the Java Native Interface [76] to interact with native code. The JNI, a portion of which is shown in Figure 4.3, abstracts field accesses and Java method invocations from native code, so I can substitute implementations that behave appropriately whether in a transaction or not.

I distinguish among three classes of native methods. The first are native methods that are "safe" in a transactional context. That is, they have no external side effects (which would need to be undone if the transaction aborted) and behave correctly in a transaction if all reads, writes, and method calls are simply replaced by the appropriate transactional protocol. The second are methods that have a different implementation in a transactional context. One example would be the native methods that implement

[6] If one can implement `writeNT` but not `readNT`, one need only ensure that native code does not write fields that are also written inside a transaction, as discussed in Section 4.2.2.

```
struct JNINativeInterface {
  ...
  /* Calling instance methods */
  jmethodID (*GetMethodID)
    (JNIEnv *env, jclass clazz, const char *name, const char *sig);
  jobject (*CallObjectMethod)
    (JNIEnv *env, jobject obj, jmethodID methodID, ...);
  ...
  jboolean (*CallBooleanMethod)
    (JNIEnv *env, jobject obj, jmethodID methodID, ...);
  ...
  jbyte (*CallByteMethod)
    (JNIEnv *env, jobject obj, jmethodID methodID, ...);
  ...
  /* Accessing fields of objects */
  jfieldID (*GetFieldID)
    (JNIEnv *env, jclass clazz, const char *name, const char *sig);
  jobject (*GetObjectField)
    (JNIEnv *env, jobject obj, jfieldID fieldID);
  jboolean (*GetBooleanField)
    (JNIEnv *env, jobject obj, jfieldID fieldID);
  ...
  void (*SetObjectField)
    (JNIEnv *env, jobject obj, jfieldID fieldID, jobject value);
  void (*SetBooleanField)
    (JNIEnv *env, jobject obj, jfieldID fieldID, jboolean value);
  ...
  /* Array Operations */
  jobject (*GetObjectArrayElement)
    (JNIEnv *env, jobjectArray array, jsize index);
  void (*SetObjectArrayElement)
    (JNIEnv *env, jobjectArray array, jsize index, jobject value);
  ...
  jboolean* (*GetBooleanArrayElements)
    (JNIEnv *env, jbooleanArray array, jboolean *isCopy);
  jbyte* (*GetByteArrayElements)
    (JNIEnv *env, jbyteArray array, jboolean *isCopy);
  ...
};
```

Figure 4.3: A portion of the Java Native Interface for interacting with the Java runtime from C native methods [76]. There are function variants for all of the basic Java types: boolean, byte, char, short, int, long, float, double, and Object. The `jobject` and `j*Array` types are not direct references to the heap objects, but rather opaque wrappers that preserve the garbage collector's invariants and protect the runtime system's private implementation.

file I/O; in a transactional context, these methods should instead interface with a underlying transactional file system and arrange for the I/O operations to commit if and only if the current transaction commits. Another example is the `Object.wait()` method; the appropriate alternate implementation is described in Section 4.4.4. The last class of native methods are those that are inherently impossible in a transactional context, usually due to irreversible external I/O (Section 8.4 discusses I/O interactions in more detail). A compiler configuration file identifies the safe native methods.

For "safe" native methods, the FLEX compiler creates a thunk that stores the transaction object in the JNI environment structure (`JNIEnv`) which is passed to every JNI method. When that native method invokes the field or array operations in the JNI (for example, `Get*Field`, `Set*Field`, `Get*ArrayElements`, etc.) the runtime checks the environment structure to determine whether to use the appropriate transactional or nontransactional read or write protocol. When Java methods are invoked, via the `Call*Method` family of methods, in a transactional context the runtime looks up the `$$withtrans` suffixed version of the method (see Section 4.2.1) and invokes it, passing the transaction object from the environment as the first parameter.

For unsafe native methods, FLEX generates a call to a `$$withtrans`-suffixed JNI method, passing the transaction as the first parameter as is done for pure Java calls inside a transaction context. The implementer is then responsible for correct transactional behavior.

4.3.2 Preprocessor specialization

Part of the challenge in engineering a practical implementation is properly accounting for the wide range of data types in a real programming language. Java contains 8 primitive types in addition to types deriving from `java.lang.Object`, which are themselves divided into array and non-array

```
/* Framework for use of preprocessor specialization. */
/* (from readwrite.c) */

#define VALUETYPE jboolean
#define VALUENAME Boolean
#include "transact/readwrite-impl.c"
#define ARRAY
#include "transact/readwrite-impl.c"
#undef ARRAY
#undef VALUENAME
#undef VALUETYPE

/* ... etc ... */

#define VALUETYPE jdouble
#define VALUENAME Double
#include "transact/readwrite-impl.c"
#define ARRAY
#include "transact/readwrite-impl.c"
#undef ARRAY
#undef VALUENAME
#undef VALUETYPE

#define VALUETYPE struct oobj *
#define VALUENAME Object
#include "transact/readwrite-impl.c"
#define ARRAY
#include "transact/readwrite-impl.c"
#undef ARRAY
#undef VALUENAME
#undef VALUETYPE
```

Figure 4.4: Specializing transaction primitives by field size and object type: `readwrite.c`. This figure shows how `readwrite.c` specializes the code in `readwrite-impl.c` (Figure 4.5) through the use of multiple `#include` statements.

```
/* Implementing generic functions. */
/* (from readwrite-impl.c) */
#include "transact/preproc.h" /* Defines 'T()' and 'TA'() macros. */

VALUETYPE TA(EXACT_readNT)(struct oobj *obj, unsigned offset);

VALUETYPE TA(EXACT_readNT)(struct oobj *obj, unsigned offset) {
  do {
    VALUETYPE f = *(VALUETYPE*)(FIELDBASE(obj) + offset);
    if (likely(f!=T(TRANS_FLAG))) return f;
    if (unlikely(SAW_FALSE_FLAG ==
                 TA(copyBackField)(obj, offset, KILL_WRITERS)))
      return T(TRANS_FLAG); // "false" transaction: field really is FLAG.
    // okay, we've done a copy-back now.  retry.
  } while(1);
}

/* ... undefine macros at end of readwrite-impl.c ... */
```

Figure 4.5: Specializing transaction primitives by field size and object type: `readwrite-impl.c`. In this snippet the `readNT` algorithm is implemented using macros (defined in Figure 4.6, `preproc.h`) to implement generic field types and naming conventions. The macros `likely` and `unlikely` communicate static branch prediction information to the C compiler.

```
/* Preprocessor specialization macros */
/* (from preproc.h) */
#define x1(x,v) x2(x,v)
#define x2(x,v) x ## _ ## v

#if defined(ARRAY)
# define A(x) x1(x,Array)
# define OBJ_OR_ARRAY(x,y) (y)
#else
# define A(x) x
# define OBJ_OR_ARRAY(x,y) (x)
#endif

#define T(x) x1(x,VALUENAME)
#define TA(x) T(A(x))

#define FIELDBASE(obj) \
         OBJ_OR_ARRAY(obj->field_start,((struct aarray *)obj)->element_start)
```

Figure 4.6: Specializing transaction primitives by field size and object type: `preproc.h`. In this portion of the file, we define the specialization macros used in Figure 4.5.

types.[7]

Figures 4.4, 4.5, and 4.6 demonstrate how field size and object type specialization are implemented in the FLEX transaction runtime. The main header and source files do nothing but repeatedly `#include` an `-impl`-suffixed version of the file with `VALUETYPE`, `VALUENAME`, and `ARRAY` defined to range over the possible primitive and array types. This technique allows compact naming and definition of specializable functions to account for array/object (among other) differences. For example, the `FIELDBASE` macro (defined in Figure 4.5, used in Figure 4.4) allows the uniform treatment of object fields and array elements, even though the first array element starts at a different offset from the first field (since array data starts with an immutable length field).

Section 4.4.3 discusses some of the other challenges involved in synthesizing atomic operations on subword and multiword datatypes.

4.4 Limitations

The present APEX implementation contains a few non-fundamental limitations with well-understood solutions which, nonetheless, would add additional implementation complexity. These limitations involve static fields, coarse granularity of LL/SC instructions, subword and multiword fields, and condition variables (`Object.wait()`). In addition, there are limitations related to the manipulation of large objects (usually arrays); we put off to Chapter 6 a full discussion of the large object problem and its solutions.

[7] The FLEX flag value was carefully chosen so as not to be a NaN for either of the floating-point types, since comparisons against NaN can be surprising. The flag value consists of a repeated byte value so that a new object can be initialized with all fields `FLAG` without knowledge of the exact types and locations of each field.

```
                          class C$$static {
                            int f;
                          }
class C {                 class C {
  static int f;             final static C$$static $static = new C$$static();
  ...                       ...
  void foo() {              void foo() {
    ... = C.f;      ⇒         ... = C.$static.f;
    C.f = ...;                C.$static.f = ...;
  }                         }
}                         }
```

Figure 4.7: Static field transformation. Static fields of class `C` have been hoisted into a new class `C$$static`. Since the new field `$static` is write-once during initialization, it is safe to read/write directly, unlike the old field `f`.

4.4.1 Static fields

Static fields are not encapsulated in an object as other fields are; they are really a form of controlled-visibility global variable. The present implementation ignores static fields when performing the synchronization transformation; reads and writes to static fields are always direct. In theory this could cause races, but in our benchmarks most static fields are written only during initialization. An analysis that proved that writes only occurred during a single-threaded class initialization phase of execution could prove that these direct accesses are safe in most cases. Proper handling of static fields that do not meet this condition is straightforward: new singleton classes would be created encapsulating these fields, and access would be via an extra indirection. Figure 4.7 illustrates this transformation.

4.4.2 Coarse-grained LL/SC

In our description of the software transaction algorithms we have so far neglected to state the size of the reservation created by the platform load-linked instruction. The implicit assumption is that the reservation is roughly word-sized: our algorithm uses load-linked to watch the pointer-valued `readerList` (Figures 3.8, 3.12) and `version` (Figures 3.9, 3.10, 3.12, 3.13)

fields, and to perform the compare-and-swap operation on the transaction status field necessary for aborting and committing transactions. Implementations of load-linked and store-conditional on most architectures, however, usually place a reservation on a specific cache line, which spans many words. In the case of our MPC7447 (G4) PowerPC processor, a reservation (and cache line) is 32 bytes.

In our present implementation, we have not taken any special action to account for this difference. In general, a larger-than-expected reservation may cause false conflicts between independent transactions, and the false conflicts can lead to deadlock or livelock if contention is not managed properly. Our implementation uses randomized[8] exponential backoff, which ensures that every transaction eventually has an opportunity to execute in isolation, which will moot any false conflict.

Other mechanisms to resolve contention can also be used to address false conflicts. In an extreme case, a contention manager could invoke a copying collector to relocate objects involved in a false conflict, although this tactic is likely merited only when the conflicts are extremely frequent or when long-lived transactions make the copying cost less than the cost of serializing the transactions involved.

Alternatively, false conflicts can be managed with a careful allocation policy. The allocator in the present implementation ensures 8-byte alignment of objects. As we have done in prior work [98], the Java allocator can be modified to align objects to cache line (32-byte) boundaries, to ensure there are no false conflicts between objects. Some of the wasted space can be reclaimed by using a smaller alignment for objects known not to escape from their thread, or by packing multiple objects into a cache line if they are known never to participate in transactions that might conflict.

[8] The randomization is via the varying latency of the Unix `sleep` syscall; stronger randomization could be implemented if it mattered.

4.4.3 Handling subword and multiword reads/writes

The granularity of the store-conditional instruction is more problematic. Although we have specified our algorithms assuming that a store-conditional of any field is possible, in practice store-conditional operations are only provided for a restricted set of data types: the PowerPC architecture defines only 32-bit and 64-bit stores, and 32-bit implementations (like our G4) don't implement the 64-bit variant. We need to carefully consider how to safely implement our algorithms with only a word-sized store-conditional.

First, let us consider the case where the field size is larger than a word, which for a Java implementation is limited to fields with `double` and `long` types. Within a transaction we can simply write two `int`s (for example) to implement the store of a `long`; the transaction mechanism already ensures that the multiple writes appear atomic.

This partial solution leaves only large writes outside of a transaction. One solution is to decompose these fields into multiple smaller fields as we did in the transactional case; we don't give up strong atomicity but we allow other readers (both transactional and nontransactional) to see the partial writes. The Java memory model actually permits this behavior (Java Language Specification chapter 17.7, "Non-atomic Treatment of double and long" [37]). The specification continues, "VM implementers are encouraged to avoid splitting their 64-bit values where possible." We can easily implement this behavior as an alternative: nontransactional writes of doubles and longs can be converted to small transactions encompassing nothing but the multiple word-size writes.

Now let us consider subword-sized writes. Section 17.6 of the Java Language Specification specifically says:

> One implementation consideration for Java virtual machines is that every field and array element is considered distinct; updates to one field or element must not interact with reads or updates of any other field or element. In particular, two threads that

> update adjacent elements of a byte array separately must not interfere or interact and do not need synchronization to ensure sequential consistency.
>
> Some processors do not provide the ability to write to a single byte. It would be illegal to implement byte array updates on such a processor by simply reading an entire word, updating the appropriate byte, and then writing the entire word back to memory. This problem is sometimes known as *word tearing*, and on processors that cannot easily update a single byte in isolation some other approach will be required.

Again, writes within a transaction are straightforward: reading an entire word, updating it, and rewriting will appear atomic due to the transaction mechanism, and no observable word tearing will occur. A minor difficulty is the possibility of "false" conflicts between updates to adjacent fields of an object; these are handled as described in Section 4.4.2.

The final case is then nontransactional writes to subword values, which occur frequently in real programs during string manipulations. We could address this by using a short transaction to do the read-modify-write of the word surrounding the subword field, but this is likely too expensive, considering the number of subword manipulations in most programs. A better algorithm to address this problem is presented in Figure 4.8; compare this to the earlier expression of the `writeNT` algorithm in Figure 5.3. The revised algorithm steals the lower two bits of the `readerList` pointer to serve as subword reservations, ensuring that racing writes to different subwords in the same word are prevented. The cost of this algorithm is mostly extra masking when the readerList is read, and an extra store (to set the `readerList` reservation) on every subword write.[9]

[9] It would be nice if we could avoid updating the `readerList` pointer on every write when using common access patterns, like stepping linearly through the array, but I don't see a way to do that.

```
void writeNT(struct oobj *obj, int byte_idx, small_field_t val) {
  if (unlikely(val==FLAG))
    unusualWrite(obj,byte_idx,val); // do a short transactional write
  else {
    do {
      int r = (int) LL(&(obj->readerList));
      if (unlikely((r & ~3) != 0)) {
        // readerList is not "NULL"; have to kill readers, etc.
        unusualWrite(obj,idx,val);
        return;
      } else if (r == (byte_idx & 3)) {
        // okay to write to this subword
        word_t w = obj->word[byte_idx>>2];
        if (SC(&(obj->word[byte_idx>>2]), combine(w, val, byte_idx)))
          return;
      } else
        // last write was to some other subword; switch to this one.
        SC(&(obj->readerList), byte_idx & 3);
    } while (1);
  }
}
```

Figure 4.8: Nontransactional write to a small (subword) field. Compare to Figure 5.3. The `byte_idx` value is the offset of the field within the objects, in bytes, and the `combine` function does the appropriate shifting and masking to combine the new subword value with the read value of the surrounding word.

The present implementation does not specially handle subword-sized writes, resulting in programmer-visible word tearing.

4.4.4 Condition variables

A condition variable is a synchronization device that allows threads to suspend execution and relinquish the processors until some predicate on shared data is satisfied. When the predicate becomes true, a thread signals the condition (`Object.notify()` and `Object.notifyAll()` in Java); other threads can suspend themselves waiting for the signal (`Object.wait()`). In order to prevent races between notification and waiting, condition variables are associated with a mutex (the object monitor, in Java) and `wait` and `notify` require the lock to be held when they are called. The `wait` operation atomi-

```
public class Drop {
    //Message sent from producer to consumer.
    private String message;
    //True if consumer should wait for producer to send message, false
    //if producer should wait for consumer to retrieve message.
    private boolean empty = true;

    public synchronized String take() {
        //Wait until message is available.
        while (empty) {
            try {
                wait();
            } catch (InterruptedException e) {}
        }
        //Toggle status.
        empty = true;
        //Notify producer that status has changed.
        notifyAll();
        return message;
    }

    public synchronized void put(String message) {
        //Wait until message has been retrieved.
        while (!empty) {
            try {
                wait();
            } catch (InterruptedException e) {}
        }
        //Toggle status.
        empty = false;
        //Store message.
        this.message = message;
        //Notify consumer that status has changed.
        notifyAll();
    }
}
```

Figure 4.9: Drop box example illustrating the use of condition variables in Java, from [100].

cally releases the lock and waits for the variable; when notification is received `wait` re-acquires the lock before resuming. An example of the use of these operations in Java is provided in Figure 4.9.

Implementing these semantics properly in a transactional context is challenging; a better solution to the fundamental problem is a mechanism like Harris and Fraser's ***guarded transactions*** [44]. A reasonable implementation is possible, however.

The JDK1.6 specification for the `Object.wait()` method says:

> A thread can also wake up without being notified, interrupted, or timing out, a so-called spurious wakeup. While this will rarely occur in practice, applications must guard against it by testing for the condition that should have caused the thread to be awakened, and continuing to wait if the condition is not satisfied. In other words, waits should always occur in loops...

This allowance simplifies the implementation of the notification methods. It is always safe to do the notify immediately—we needn't wait until the transaction doing the notify is sure to commit. If we end up getting aborted we'll just redo the notify later when we retry. So, we acquire a special mutex (either per-object as with standard Java or global; the lock is not held long), do the notification, and release the mutex.

The `wait` method, however, is a commit point, since it releases and reacquires the monitor lock. Thus, it splits the transaction containing it into two.[10] The fundamental mechanism is straightforward. We first acquire our special mutex, then attempt to commit the transaction, and check the result. If we don't commit successfully, we unlock the mutex and throw the usual `TransactionAbortException`. If we have committed successfully, we can safely wait (releasing our mutex atomically). When we're woken, we create

[10] Which arguably calls for a special type annotation on methods that may call `wait()`, so that users are not misled into thinking a block is atomic that may invoke something that calls `wait()`.

a new transaction, release our mutex, and continue. The mutex prevents notification from occurring between the commit and the wait.[11]

Implementation in a practical system introduces some difficulties. As described in Section 4.2, the transaction object is passed as the first argument when calling methods within a transaction context. Since `Object.wait()` commits the initial transaction and begins a new one, we must either factor out a continuation after the call to `Object.wait()` which can be invoked with the new transaction after the wait is complete, or else provide a means to update the active transaction pointer and the retry point. One way to update the active transaction pointer is to use indirection: instead of keeping a direct pointer to the transaction object, we can keep a pointer to a cell containing the pointer; the cell can then be updated to point to the new transaction by `Object.wait()`. Alternatively, an extra return value can be added to methods called within a transaction, so that method can return an updated transaction object as well as their usual return value. Either of these transformations can be applied selectively to only those methods that could possibly invoke, directly or indirectly, `Object.wait()`. The more difficult challenge is to update the location at which we'll resume for another attempt when the transaction is aborted; this retry point should be immediately following the `wait()`. The retry mechanism becomes much more complicated, and some means to reconstruct the appropriate call-stack is necessary.

For code like the example given in Figure 4.9 the "correct" resumption behavior is actually identical to restarting the transaction from the original start point (i.e. the beginning of the `put()` or `take()` method), assuming

[11] In the drop box example in Figure 4.9 the standard monitor lock is present to prevent (for example) `put()` from setting `empty` to false and doing the `notifyAll()` *after* `take()` has seen that `empty` is true but *before* the `wait()` has occurred (hence the notification would be lost). With the transaction transformation, the lost notification is prevented because put's write to `empty` will abort the `take` transaction if it occurs before `take` is committed (causing the `wait()` to be retried), and the special mutex will prevent the notification from occurring between the commit and the `wait`.

that the `put()` or `take()` isn't in a subsumed nested transaction. It is probably worth identifying this situation through compile-time analysis, as it simplifies the resumption transformation considerably.

In the present implementation I implement correct `notify` semantics, but I do not allow `wait` to be a commit point. Instead FLEX takes the special mutex and sleeps on the condition variable without attempting to commit the transaction, and resumes executing in the same transaction when it awakes. This is sufficient for the limited uses of `wait` in the benchmarks of this thesis.

4.5 Additional optimizations

This section discusses some additional optimizations that could improve performance on transactional code; they have not been implemented in the present compiler. The improvements include version passing, escape analysis, object immutability analysis, and fast allocation.

As described in Section 4.2, the basic transaction transformation inserts a call to `ensureReader/ensureWriter` once per object prior to a read/write of a field in the object. These methods return a version object (the "current version") which is then passed to the `readT/writeT` methods. A straightforward improvement would avoid redundant calls to `ensureReader/ensureWriter` by passing the "current version" of the various parameters when a method is invoked. A strictness analysis should be performed, to avoid marking fields as read/written that are not guaranteed to be read/written by the transaction, and the "current version" added as an extra argument for the strict parameters of the method. This allows the caller to combine redundant `ensureReader/ensureWriter` invocations for the argument.[12]

[12] One could also consider adding "current version" arguments even for non-strict parameters, with the caveat that the caller pass in NULL if it had not already read/written the non-strict parameter. All variables used in a method or the method's callees would then be potentially eligible for version passing, however. Some more sophisticated heuris-

Escape analysis [97] is a standard technique to reduce unnecessary synchronization in Java programs. In a typical application, monitor locks on objects that do not escape their thread are eliminated; these lock eliminations can translate to transaction eliminations as well. Further, escape analysis can identify individual objects that do not escape their method, thread, or call context, and replace the `readT`/`writeT` or `readNT`/`writeNT` protocol with direct access.

Another orthogonal analysis can identify object immutability [103]. Immutable objects can be read directly and their field values can be safely cached across transaction boundaries. Further, the initialization of immutable objects typically can be done with direct writes as well.

Finally, a substantial portion of the cost of small transactions is spent allocating the version object, especially for the conservative collector used in the present implementation. Fast allocators could alleviate this bottleneck [8]. Alternatively, a free list of versions could be hung from objects involved in many transactions: as soon as a version commits, the previous version can be added to the free list for the use of the next transaction.

tic would be needed to determine for which objects version passing is worthwhile.

All you owe the public is a good performance.

Humphrey Bogart

Chapter 5

ApeX performance

This chapter presents measurements and analysis of the performance of the ApeX software transaction system. Since ApeX adds read and write checks to even nontransactional code, we begin by examining a microbenchmark to discover the performance limits in the absence of transactions. We then move on to full transactional application benchmarks. We formulate a linear model to explain ApeX overhead, and describe how a programmer can use this model to predict and improve the performance of their code running under ApeX.

Since previous work has adequately demonstrated the scalability benefits of nonblocking synchronization [6, 40, 43, 64, 73, 80, 81], I concentrate on single-threaded efficiency measures. All experiments in this chapter were performed on a 1GHz MPC7447 (G4) PowerPC processor, with 512kB unified L2 cache and 512MB of main memory, running Ubuntu 6.10 on a Linux 2.6.17 kernel. The Flex compiler and runtime system sources were from the head CVS revision on 2007-05-27. C sources were compiled using `gcc-3.4`. Appendix B contains the full Flex command-lines and configurations used to obtain the results in this chapter.

5.1 Performance limits for nontransactional code

The ApeX software transaction system depends on the insertion of simple read and write checks in nontransactional code. The performance of read and write barriers has been well-studied in the garbage collection community, but our checks are slightly different: in particular, they are checks on the *contents* of a memory cell, rather than on its address. Thus our checks introduce a more direct dependency which may affect performance. Further, our write check involves a LL/SC instruction pair, which may behave differently from the standard loads and stores used in barriers.

In this section I use a simple counter microbenchmark to evaluate the "best worst case" nontransactional performance of ApeX. It is an idealized "best case" in that I don't benchmark the effects of "false flags" or other forays into the transactional code path, nor do I account for double-word or subword writes, cache effects due to code duplication, or other details of a real implementation. I investigate these effects with full benchmarks in Section 5.2. In this section I also freely hand-optimize down to the assembly level to better determine the fundamental performance limits.

The tight read/write dependency makes counter increment a "worst case" benchmark, however. In general, modern compilers are good at separating reads from writes in "real" code to mask load latency, but this particular microbenchmark cannot be reasonably unrolled to accomplish the separation. My conclusions about the fundamental costs of read and write costs should thus be tempered by the knowledge that some of these costs are in practice masked by the same standard compiler techniques used to mitigate memory-access latencies.

Figure 5.1 presents the basic structure of the counter microbenchmark. The `read()` and `write()` methods have appropriate definitions inlined for each variant of the benchmark. The `readerList` and `field` fields of the `struct oobj` are marked `volatile` to prevent the C compiler from optimizing away the accesses we are interested in benchmarking. Without these

```
typedef int32_t field_t;

struct oobj {
  struct version *version;
  struct readerList * volatile readerList;
  volatile field_t field[NUM_FIELDS];
};
void do_bench(struct oobj *obj) {
  int i;
  for (i=0; i<REPETITIONS; i++) {
    field_t v = read(obj, 0);
    v++;
    write(obj, 0, v);
  }
}
```

Figure 5.1: Counter microbenchmark to evaluate read- and write-check overhead for nontransactional code.

```
#if !defined(WITH_READ_CHECKS)

static inline field_t read(struct oobj *obj, int idx) {
  field_t val = obj->field[idx];
  return val;
}

#else

static inline field_t read(struct oobj *obj, int idx) {
  field_t val = obj->field[idx];
  if (unlikely(val==FLAG))
    return unusualRead(obj,idx);
  else return val;
}

#endif
```

Figure 5.2: C implementation of read checks for counter microbenchmark.

```
#if !defined(WITH_WRITE_CHECKS)

static inline void write(struct oobj *obj, int idx, field_t val) {
  obj->field[idx] = val;
}

#else

static inline void write(struct oobj *obj, int idx, field_t val) {
  if (unlikely(val==FLAG))
    unusualWrite(obj,idx,val); // never called
  else {
    do {
      if (unlikely(NULL != LL(&(obj->readerList)))) {
        unusualWrite(obj,idx,val); // never called
        break;
      }
    } while (unlikely(!SC(&(obj->field[idx]), val)));
  }
}

#endif
```

Figure 5.3: C implementation of write checks for counter microbenchmark.

declarations, gcc 4.1.2 optimizes away the entire benchmark loop and replaces it with a direct addition of REPETITIONS.

Figure 5.2 shows the baseline read() implementation, which inlines to a single lwz instruction on PowerPC, along with the C implementation of the read check necessary for nontransactional code under our transaction system. We use the unlikely() macro, implemented using the gcc extension __builtin_expect(), to apply the appropriate static prediction bits indicating that FLAG is expected to be an unusual value. In our microbenchmark, this prediction is always correct. The C compiler must still save whatever registers are necessary to allow the call to unusualRead(), although this function is never actually called during the microbenchmark. The unusualRead function would be expected to perform the remainder of the readNT algorithm from Figure 3.8, including looking up and copying back the most recently-committed value for this field.

Figure 5.3 shows the equivalent implementation pairs for a nontrans-

actional write. The basic implementation inlines to a single `stw` instruction. The write-check version at the bottom must perform three tests. First, it must ensure that the value to be written is not `FLAG`. If it is, we must use the "false flag" mechanism to write it, modeled here by a call to `unusualWrite()`. This check can be optimized away when writing a constant,[1] but the `volatile` declarations ensure that our benchmark always performs the check. Second, we must perform a load-linked of the object's `readerList` to check that there are not any current transactional readers of this object. Last, we perform a store-conditional of the value we wish to write and check that it was successful. Unlike the other two tests, this check fails occasionally[2] during the benchmark run due to context switches that may occur between the load-linked and the store-conditional instructions.

Figure 5.4 shows the performance of this benchmark on the PowerPC hardware fully described in Chapter 5. Read checks, even in this worst case where the value read is demanded immediately, only add 14% overhead. Write checks are more costly, due to the load-linked and store-conditional pair. Specifically, they add 200% overhead to the benchmark. Combining read and write checks yields the expected 214% overhead, as our microbenchmark was carefully constructed to avoid the opportunities for instruction-level parallelism one might expect in real-world applications.

The base benchmark performs 10^9 reads and 10^9 writes in an elapsed time of 7.06 seconds. Our read (write) rate is thus 142×10^6 reads (writes) per second. Read checks cost 1.0 ns/read, and write checks cost 14.1 ns/write. We can predict the overhead fraction o for an arbitrary program with the linear equation:

$$o = (1.0 \times 10^9)r + (14.1 \times 10^9)w \tag{5.1}$$

[1] Hennessy and Patterson indicate that 35% of integer instructions use an immediate operand [46, p. 78]. Their methodology does not allow breaking out the percentage of stores specifically.

[2] Experimentally, about 3,600 times in the 1 billion repetitions of the write during the counter microbenchmark.

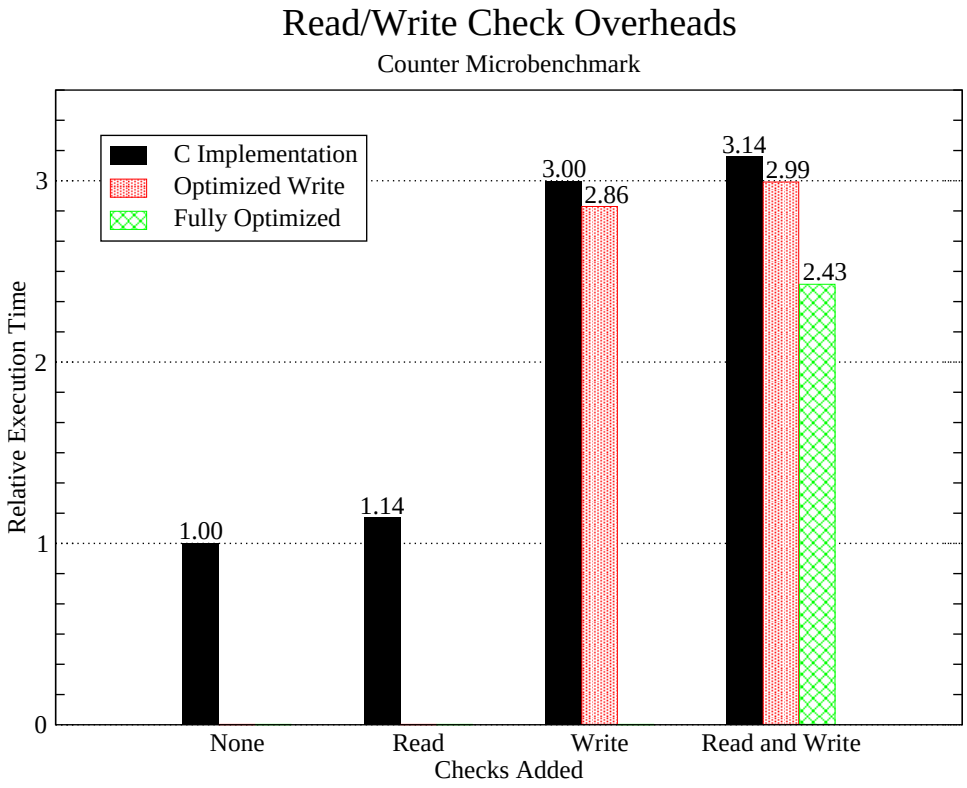

Figure 5.4: Time overhead of read checks, write checks, and both read and write checks for nontransactional code, in both a pure C implementation and optimized assembly.

where r (w) is the read (write) rate.

Examination of the assembly code generated for this simple benchmark seems to indicate substantial scope for improvement. The inline assembly mechanism of gcc/C provides no inherent support for instructions like the PowerPC's "store-conditional" (`stwcx.`) that leave their results in a condition code register. Figure 5.5 shows the assembly emitted for the write-check version of the microbenchmark, with the cumbersome mechanism required to move the condition code to a register so that it can then be retested. We can easily hand-code a better write-check mechanism, shown in the right-hand column of the figure.[3] The optimized store-conditional test reduces write-check overhead to 186% (13.1 ns/write) and combined read- and write-check overhead to 199%, as shown in Figure 5.4.

Further performance improvement is possible by optimizing the benchmark with both read and write checks as a whole. Figure 5.6 shows an optimized assembly version of our `do_bench()` function. I've primarily rescheduled the code here to separate dependent instructions as much as possible, but I've also ensured repeatable instruction cache alignment,[4] replaced our canonical `FLAG` value with `0xFFFFCACA`, which can be represented by the PowerPC's 16-bit signed immediate instruction field (eliminating the need for a register), and combined the flag checks in the read and write routines by masking and comparing against `0xFFFFCACB`.[5] The read-and-write check overhead is improved to 143% with these optimizations. The remaining overhead is due mostly to the irreducible dependency between

[3] This version has a stub where a function call to `unusualWrite` would usually go; the correct code here would branch to a small thunk that performs the frame operations and register saves necessary to adhere to the C calling convention; gcc makes it difficult to ensure that this thunk is "near enough" for a direct branch (within a 24-bit signed displacement) without duplicating it needlessly.

[4] The MPC7447 (G4) PowerPC has a 32-byte cache line, although fetches occur in 16-byte (4 instruction) chunks.

[5] Again, the branches to `unusualRead()` and `unusualWrite()` are difficult to express in this hybrid of C and assembly, but small stubs would be written to save registers and perform a branch with the appropriate calling conventions.

```
do_bench:
  mflr 0
  stwu 1,-16(1)
  lis 5,0x3b9a
  li 8,0
  ori 5,5,51712
  addi 6,3,4
  addi 7,3,8
  stw 0,20(1)
  b .L4
.L22:
  mr 10,6
  mr 11,7
.L5:
  lwarx 0,0,10
  cmpwi 7,0,0
  bne- 7,.L19
  stwcx. 9,0,11
  li 0,0
  bne- 0f
  li 0,1
0:
  cmpwi 7,0,0
  beq- 7,.L5
  addi 8,8,1
  cmpw 7,8,5
  beq- 7,.L21
.L4:
  lwz 9,8(3)
  cmpwi 7,9,-13623
  addi 9,9,1
  bne+ 7,.L22
.L21:
  lwz 0,8(3)

  cmpw 7,0,8
  bne- 7,.L23
  lwz 0,20(1)
  addi 1,1,16
  mtlr 0
  blr
```

```
do_bench:
  mflr 0
  stwu 1,-16(1)
  addi 10,3,8
  addi 11,3,4
  stw 0,20(1)
  lis 0,0x3b9a
  ori 0,0,51712
  mtctr 0
  b .L4

.L5:
 0:
   lwarx 0,0,11
   cmpwi 0,0
   beq+ 1f
   b . # stub for unusualWrite branch
 1:stwcx. 9,0,10
   bne 0b

   bdz .L14

.L4:
  lwz 9,8(3)
  cmpwi 7,9,-13623
  addi 9,9,1
  bne+ 7,.L5
.L14:
  lwz 0,8(3)
  xoris 9,0,0xc465
  cmpwi 7,9,-13824
  bne 7,.L15
  lwz 0,20(1)
  addi 1,1,16
  mtlr 0
  blr
```

Figure 5.5: PowerPC assembly for counter microbenchmark with write checks. The right-hand column is generated from the C implementation. The left-hand column has an optimized version of the write check inlined into the code using the `asm` keyword. Italicized code is essentially unchanged. The optimized section is shown in boldface.

```
    # repetition count is in the count register to start
    b 0f
    .balign 32 # align to 32-byte cache-line boundary
    nop
    nop
    nop
    nop
    nop
    nop
0:  lwarx 5,0,0
    lwz 6, 0(9)
1:  ori 8, 6, 3
    addi 7, 6, 1
    cmpwi 1, 5,0
    cmpwi 2, 8, 0xFFFFCACB
    bne- 1, 0b # stub: unusualWrite
    beq- 2, 0b # stub: unusualRead or unusualWrite
    stwcx. 7,0,9
    lwarx 5,0,0
    lwz 6, 0(9)
    bne- 0, 1b
    bdnz 1b
```

Figure 5.6: Optimized PowerPC assembly for counter microbenchmark with both read and write checks.

the load-linked and store-conditional instructions.

Hennessy and Patterson's measurements indicate that load and store instructions comprise respectively 26% and 9% of dynamic instruction counts in SPECint92 on a RISC microarchitecture (DLX) [46, p. 105]. As a rough estimate, using dynamic instruction count as a proxy for instruction time, the 14% read overhead and 186% write overhead measured in this microbenchmark thus translate into 20% net overhead for nontransactional code.[6] If we conservatively assume that most of the improvement in Figure 5.6's optimized checks should be credited to reduced write-check overhead, giving the same 14% read overhead but only 129% write overhead, the same metric gives only 15% net overhead, as shown in Figure 5.7. This figure should be considered a rough lower bound for the time overhead, possible with aggressive optimization and scheduling.[7]

[6] $0.65 + (1.14 \times 0.26) + (2.86 \times 0.09) = 1.20$

[7] This calculation combines over- and underestimation, making this number only a

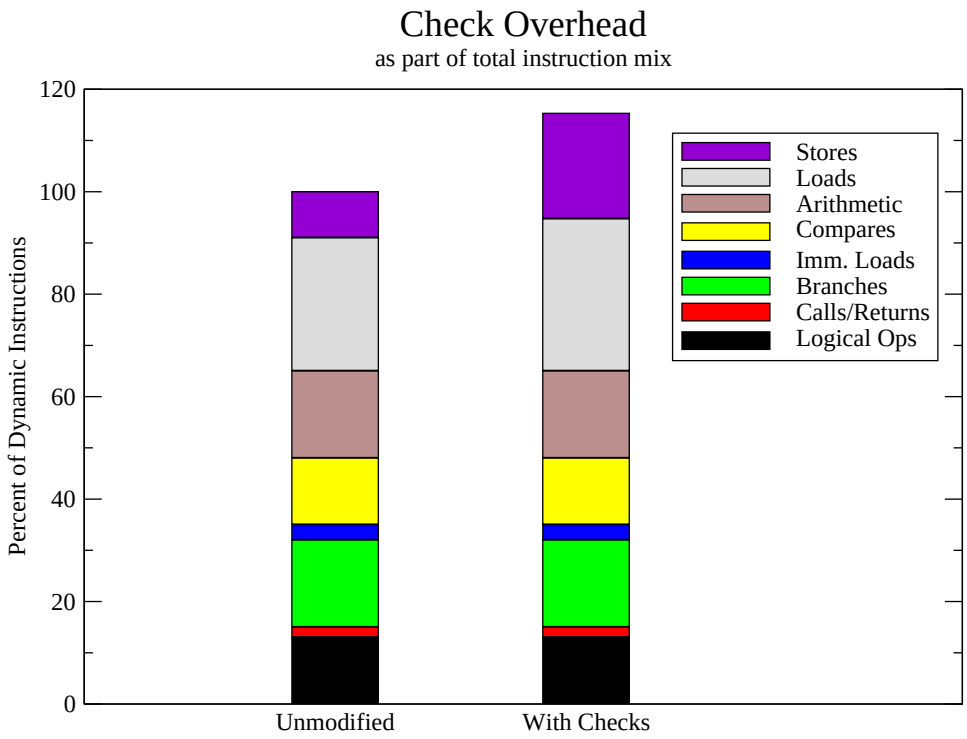

Figure 5.7: Check overhead as percentage of total dynamic instruction count. The bar on the left is the DLX instruction mix for SPECint92 from [46, p.105]. On the right is the same instruction mix after scaling the loads and stores according to Figure 5.4.

5.2 Full application benchmarks

In this section I evaluate the performance of the ApeX implementation on full applications. The limitations noted in Section 4.4 counsel that the numbers I obtain ought to be considered guidelines to potential performance, not fundamental limits.

I examine the SPECjvm98 benchmark suite, a collection of practical Java programs including an expert system, a simple database, a Java compiler, an audio encoder, and a parser generator. These benchmarks were previously described and characterized in Section 3.1. Figure 3.1 lists the 7 benchmarks in SPECjvm98. As in the previous section, we use the single-threaded 205_raytrace benchmark in the place of its multithreaded variant 227_mtrt. We omit the 201_compress benchmark since a poor interaction with the blacklist mechanism of the Boehm-Demers-Weiser conservative garbage collector causes the large array it allocates to be leaked. Runs of 201_compress are distorted by the resulting large memory working set.

To clarify the measurements, the baseline for all comparisons will be a version of the benchmarks compiled with the same method cloning and desugaring (Section 4.2) that is done for the transaction transformation. In addition to the full "100%" input for each benchmark, SPECjvm98 defines two smaller inputs, sized to roughly correspond to 1% and 10% of the runtime of the full benchmark. Unless otherwise specified, all benchmarks were run on the full "100%" input.

5.2.1 Nontransactional check overhead

In Section 5.1 I examined a simple counter microbenchmark to discover the lower limits on our nontransactional overhead. These are the fundamental

rough bound. The use of dynamic instruction count, rather than dynamic instruction time, likely underestimates the contribution of loads and stores to the run time. Scheduling, prefetching, and store buffers can pipeline the cost of the loads and stores, however, which might cause our overhead estimate to exceed the actual cost.

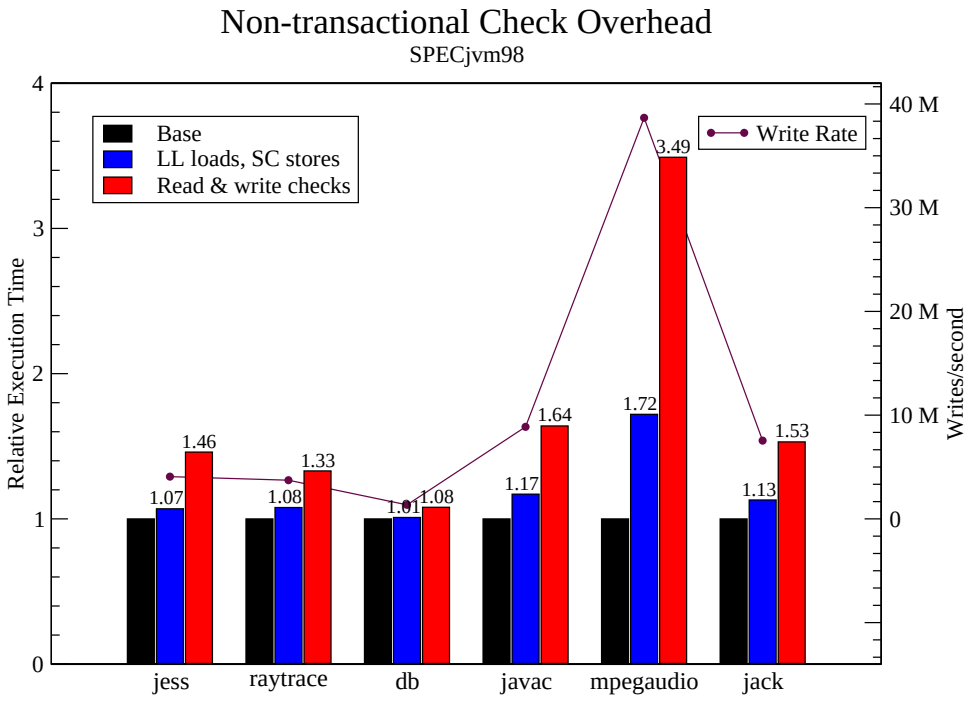

Figure 5.8: Nontransactional check overhead for SPECjvm98. All benchmarks are compiled with the transformations in Section 4.2 (method cloning, desugaring) and all synchronization is removed. The left bar in each group performs direct field access. The middle bar performs all reads with load-linked, and performs all writes with a load-linked/store-conditional pair. The right bar has all reads and writes transformed using the `readNT` and `writeNT` protocols. The line graph in the background shows the write rate (stores performed divided by base runtime) for each benchmark.

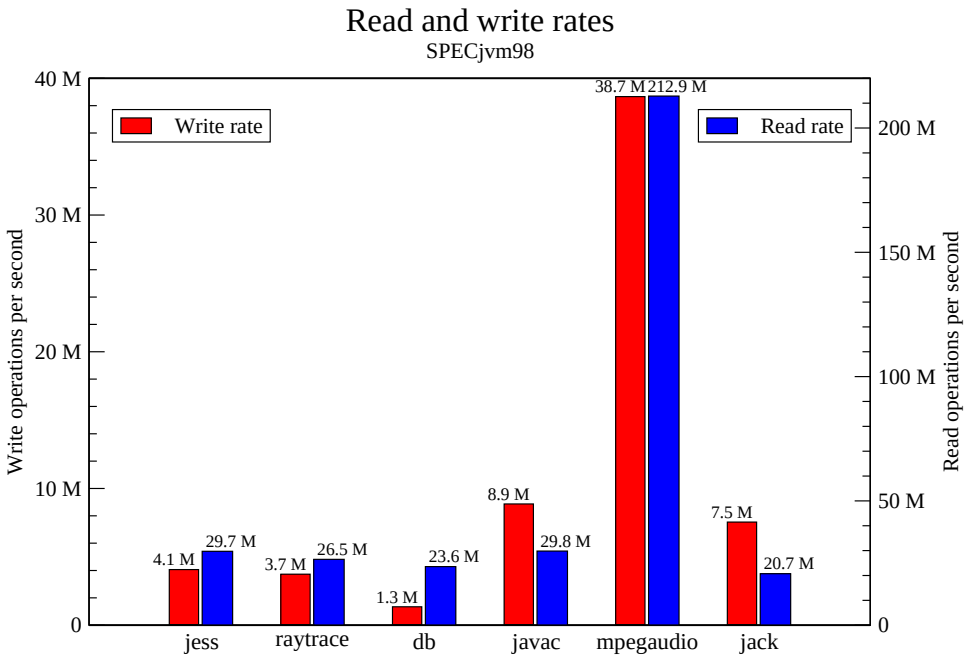

Figure 5.9: Read and write rates for SPECjvm98 benchmarks, in operations per second.

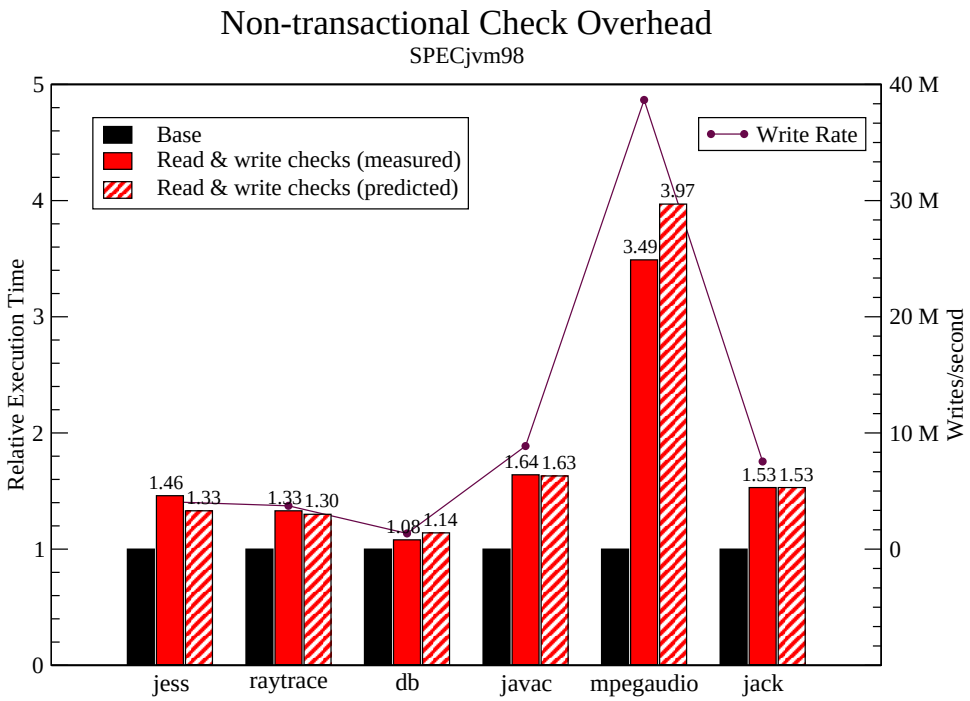

Figure 5.10: Actual (solid) versus predicted (striped) nontransactional check overhead for SPECjvm98. Measurement methodology is identical to Figure 5.8. Overhead is predicted as 2.4 ns per read, and 63.5 ns per write.

Benchmark	reads	false flags read	writes	false flags written
jess	25,583,878	0 (0%)	703,303	0 (0%)
raytrace	25,439,376	0 (0%)	3,605,463	0 (0%)
db	11,078,177	0 (0%)	912,965	0 (0%)
javac	10,315	0 (0%)	2,312	0 (0%)
mpegaudio	239,928,276	2,056 (<0.001%)	42,164,416	9,132 (.02%)
jack	9,882,846	0 (0%)	4,595,774	0 (0%)

Figure 5.11: Number of false flags read/written in SPECjvm98 benchmarks. To compile this table the applications were run with the 10% input size.

costs of strong atomicity: the penalty we must pay even if we are not using transactions at all.[8] Figure 5.8 shows equivalent measurements on full Java applications. FLEX has removed all the synchronization in these benchmarks (safe because they are single-threaded) leaving only the costs of the read and write checks required by the transaction protocols.

In the figure, the rightmost bar in each group shows the costs of performing the `readNT` and `writeNT` protocols on every read and write in the application. The earlier microbenchmark hinted that we might obtain nontransactional overheads as low as 15%-20%. In fact, we see a 19% overhead on the `209_db` benchmark, although the other benchmarks show that an overhead of 40%–50% is more typical.

The middle bar in the figure shows the overhead incurred by simply replacing the loads and stores with load-linked and store-conditional instructions, as our protocol must do. The PowerPC G4 is not optimized for frequent "synchronization" instructions such as load-linked and store-conditional, and store buffers and other architectural features are bypassed for their execution. For most applications, the performance impact is not significant. Applications with large write rates, such as `mpegaudio`, are affected to a greater degree.

The line graph in the background of Figure 5.8 shows the write rate for each benchmark. Read and write rates for the benchmark are shown in more detail in Figure 5.9. Write rate is a good predictor for overall check overhead. The largest measured overhead, 249%, comes from a `mpegaudio`, which has a write rate 4 times as large as any other benchmark.

Using the measured benchmark performance, we can compute a minimum norm solution to the linear equation

$$o_{NT} = k_{r_{NT}} r_{NT} + k_{w_{NT}} w_{NT} \tag{5.2}$$

where o_{NT} is the overhead fraction due to nontransactional checks, r and w

[8] Although, of course, in practice one would turn off the transaction transformation completely if static or dynamic analysis indicated a program is single-threaded, for example if no `Thread` objects are every created.

are (nontransactional) read and write rates, respectively, and $k_{r_{NT}}$ and $k_{w_{NT}}$ are the unknown coefficients. The solution yields overhead contributions:

$$\begin{aligned} k_{r_{NT}} &= 2.4 \text{ ns/read} \\ k_{w_{NT}} &= 63.5 \text{ ns/write} \end{aligned}$$

These values are comparable to the 1.0 ns/read and 13.1 ns/write overheads of the optimized checks we wrote for the counter microbenchmark of Section 5.1. The ApeX check implementation is slower than the hand-tuned assembly code.

Close inspection of the assembly code emitted by the C compiler for these benchmarks indicates that, as in the microbenchmark of Section 5.1, compiler limitations account for some of the difference. It is hard to obtain maximum performance with a C backend (such as used by Flex) due to expressive limits, and the emitted assembly indicates that the compiler is not optimally acting on the hints we can give it. In particular, we want as little extraneous code as possible on the fast path through the common case, but the compiler adds register spills and moves on the fast path that are necessary only in unusual cases. Further, because the fundamental load-linked and store-conditional instructions are inside inline assembly blocks and thus opaque to the `gcc-3.4` backend compiler, many opportunities for code motion and scheduling are being missed.

Slower checks are also partially the result of subword and multiword accesses in the full benchmarks, which are not present in our microbenchmark.

The nontransactional versions of the benchmarks must still properly handle false flag values. Figure 5.11 shows that false flags are not a significant percentage of values read or written, and thus not a major contribution to overhead.

Figure 5.10 shows the performance predicted by Equation 5.2 overlaid on actual performance. The `jess` benchmark has 39% higher overhead than `raytrace`, although its write rate is only 9% higher. Similarly, `db` has a 64% lower write rate than `raytrace` but a 76% lower overhead. The read rates

for these three benchmarks are too similar to explain the difference. The overhead nonlinearity can't be captured by the linear model. Furthermore, actual performance is affected by unmodeled factors such as the programs' read and write dependencies. The jess benchmark contains tight dependencies which make its overhead more sensitive to a small increase in write rate, while db has loose dependencies that lower its overhead. Nevertheless, the model matches actual performance within 14% for all benchmarks.

5.2.2 Transaction overhead

Figure 5.12 presents the performance of the transactional versions of the SPECjvm98 benchmarks. As in Figure 5.10, the leftmost bar represents the original performance. The next two bars represent versions of the benchmarks compiled with only part of the transaction transformation enabled. These are used to evaluate different components of the overhead. The rightmost solid bar shows the overhead of the full transactified benchmark. The final, striped, bar shows predicted overhead for the application, broken down by source.

The second bar in Figure 5.12 shows the performance of the transactified benchmark after all transactional read and write operations have been replaced by direct reads and writes. We label this the "no check" variant. Four changes to the application remain: transaction records are created at the start of every atomic region, method calls inside transactions are transformed to pass the transaction record and to check for abort on an exceptional return, method cloning occurs to separate transactional from nontransactional code, and a few live variables are passed around inside the method body, increasing register pressure.

The transaction rate is the largest contributor to overhead for the "no check" variant. The number of transactional method calls is a secondary factor. Creation of a transaction record at the start of every atomic region can be a considerable fraction of the overhead if the number of transactions

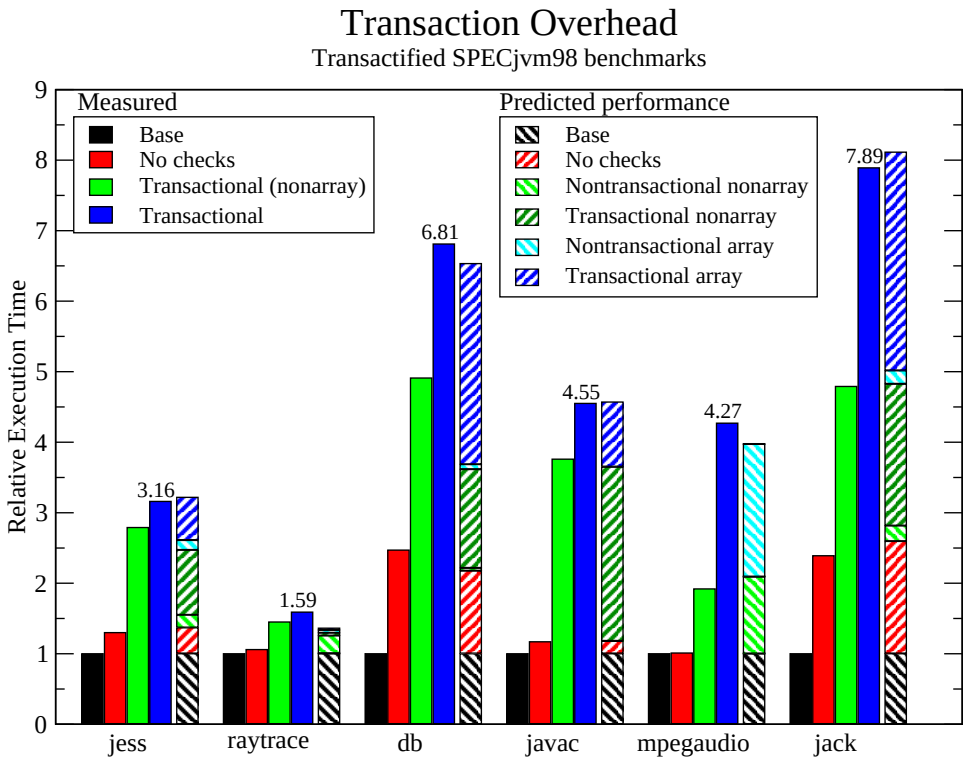

Figure 5.12: Transaction overhead for SPECjvm98. The leftmost "Base" bar is compiled with method cloning and desugaring, all transactions removed, and direct reads and writes. The "No checks" bar transforms methods, creates commit records at the start of each transaction, and checks for abort after every call, but uses direct loads and stores for all accesses. The "Transactional (nonarray)" bar is fully transactional for object accesses, but uses direct loads and stores for reads and writes of arrays. The "Transactional" bar is fully transactional. The striped final bar shows predicted performance for the benchmark, broken down by the modeled source of the overhead.

is large. The "no check" contribution is modeled as follows:

$$o_t = k_t t + k_{c_T} c_T \tag{5.3}$$

where o_t is the overhead fraction due to this portion of the transaction transformation, t is the transaction rate, c_T is the number of method calls which are inside transactions, and k_t and k_{c_T} are unknown constants. The least squares solution is:

$$\begin{aligned} k_t &= 855 \text{ ns/transaction} \\ k_{c_T} &= 18.2 \text{ ns/method call within transaction} \end{aligned}$$

The high value for k_t represents the high cost of memory allocation under the Boehm-Demers-Weiser conservative collector. In Figure 5.12 the red striped component represents o_t, computed according to Equation 5.3. The prediction tracks the actual measured values (red solid bars) within 21% of the actual overhead (ignoring `raytrace` and `mpegaudio`, where o_t is within the measurement noise).

I can use Equation 5.2 from the previous section to compute the overhead contribution from nontransactional read and write checks. The light green and light blue striped portions of the predicted overhead in Figure 5.12 represent o_{NT}, computed using Equation 5.2 from the rate of nontransactional reads and writes and the previously computed $k_{r_{NT}}$ and $k_{w_{NT}}$ values.

The final component of the overhead is from the transactional read and write protocols on accesses occurring inside transactions. Modeling this component is more difficult, due to the number of factors involved. Transactional read cost depends on whether the object involved was previously written inside the transaction. Transactional write cost depends on whether this is the first time this particular object has been written inside a transaction, and if so, on the size of the object. Both read and write costs are also affected by the effectiveness of check hoisting and coalescing (Section 4.2.2).

The model I use for transactional access costs distinguishes ***virgin writes*** (***virgin reads***), which are writes (reads) to an object which has not previously been written to (read from) in the current transaction. These accesses

involve transaction bookkeeping which is not required for subsequent accesses. I also accumulate the object sizes involved in a virgin write, to model the costs of creating a transactional clone. The resulting model is:

$$o_T = k_{w_v} w_v + k_{ws} w_v s_v + k_{r_v} r_v + k_{r_T} r_T \tag{5.4}$$

where o_T is the overhead fraction due to transactional accesses, w_v is the (transactional) virgin write rate, s_v is the average size of an object targeted by a virgin write, r_v is the (transactional) virgin read rate, r_T is the transactional read rate, and k_{w_v}, k_{ws}, k_{r_v}, and k_{r_T} are unknown constants. Solving for the constants yields:

$$\begin{aligned} k_{w_v} &= 680 \text{ ns/virgin write} \\ k_{ws} &= 85 \text{ ns/byte targeted by a virgin write} \\ k_{r_v} &= 660 \text{ ns/virgin read} \\ k_{r_T} &= 54 \text{ ns/transactional read} \end{aligned}$$

The k_{w_v} (virgin write) and k_{r_v} (virgin read) costs are comparable to k_t, since the primary component of all of these costs is memory allocation (of a clone object, a reader list entry, and a transaction commit record, respectively). When the object touched by the virgin write is large, k_{ws} accounts for the more expensive allocation and initialization. The transactional read cost, k_{r_T} is comparable to $k_{w_{NT}}$, the nontransactional write cost. The operations involved are similar: a read of the head of the reader list, a comparison, and then typically an access to the field location. The light green, green, light blue, and blue striped bars in Figure 5.12 reflects these component of the cost prediction. The overall prediction is within 10% of actual costs for every application except for `raytrace`, where the prediction underestimates actual costs by 38%.

The constants in our model were derived from the transactional performance. I can validate the model by deriving predictions for the object and array components of the overhead separately. The green bar in Figure 5.12 represents variants of the benchmarks which replaced array reads and writes

Benchmark	t	c_T	r_{NT}	w_{NT}
jess	408,416	1,006,376	26,700,326	4,054,334
raytrace	301	133,686	26,104,154	3,608,662
db	1,328,518	2,015,319	18,187,254	987,484
javac	55	9,817,721	15,091	7,983
mpegaudio	287	9,618	212,850,162	38,663,921
jack	1,815,263	2,236,720	12,695,066	5,902,722

Figure 5.13: Transaction, call, and nontransactional read and write rates for SPECjvm98 benchmarks.

with direct accesses. The resulting overhead includes o_t, the "no checks" overhead, as well as both transactional and nontransactional overheads for nonarray objects. This breakdown was motivated by the observation that s_v, the average size of an object touched by a virgin write, in some cases varies quite a lot between array and nonarray objects, as Figure 5.14 shows. The design of the ApeX transaction system necessitates an object copy for every unique object written to inside a transaction; if the objects are large, then this cost becomes excessive. Chapter 6 will discuss this problem at length and propose a solution based on functional array data structures.

The light green and dark green bars in Figure 5.12 show our predictions for the nonarray variants of the benchmarks, based on the w_v, s_v, r_T, and r_v values for nonarray accesses. These track the solid green bar indicating actual performance fairly well.

5.3 Performance recommendations

Our predictive model of ApeX performance uses 8 parameters, ordered here by their contribution to overall overhead:

- r_{NT}, the number of nontransactional reads. Each nontransactional read adds 2.4 ns overhead.
- c_T, the number of method calls within transactions. Each call adds

Benchmark	w_ν	s_ν	r_ν	r_T	nonarray s_ν	array s_ν
jess	5,291	710	1,573,712	3,046,653	30	851
raytrace	10,833	44	1,421	403,442	29	67
db	90,649	263	2,832,970	5,378,123	24	4,247
javac	464,294	37	4,630	29,804,501	32	50
mpegaudio	700	35	2,285	3,236	13	283
jack	412,891	64	3,268,924	8,016,197	32	69

Figure 5.14: Transaction performance model inputs for SPECjvm98 benchmarks. Each benchmark lists w_ν, the virgin write rate, s_ν, the average object size touched by a virgin write, r_ν, the virgin read rate, and r_T, the transactional read rate. The nonarray and array values of s_ν are broken out in the final two columns.

18.2 ns overhead.

- r_T, the number of transactional reads. Each transactional read adds 54 ns overhead.
- w_{NT}, the number of nontransactional writes. Each nontransactional write adds 63.5 ns overhead.
- $w_\nu s_\nu$, the product of the virgin write rate with the average object size touched in a virgin write. This contributes 85 ns overhead per byte in an object touched by a virgin write.
- r_ν, the number of virgin reads. Each virgin read adds 660 ns overhead.
- w_ν, the number of virgin writes. Each virgin write adds 680 ns overhead.
- t, the number of transactions. Each transaction adds 855 ns overhead.

The costs associated with t, w_ν, and r_ν might be reduced with the use of a faster allocator.

A programmer seeking to improve the performance of their code should first reduce the number of transactions created. ApeX is suited to small numbers of large transactions. Chapter 7 presents an alternative transaction system which is better suited for large numbers of small transactions.

The programmer can also reduce the number of virgin reads and writes by structuring transactions to group fields involved in a transaction inside a single object. Combining smaller transactions will also tend to reduce the number of expensive virgin reads and writes. An alternative to combining small transactions is to reduce the number of memory accesses covered by the transaction, since nontransactional reads are twenty times cheaper than transactional reads.

Decreasing the number of memory operations also directly reduces overhead. There are well-known compiler optimizations which cache memory contents in local variables or registers. Traditionally, their widespread use is avoided because the increase in register pressure makes caching counterproductive. ApeX's more-expensive memory accesses should encourage use of these optimizations.

Reducing the size of objects written by a transaction will reduce the $w_v s_v$ term and improve performance. In particular, several SPECjvm98 benchmarks allocate a large array and then perform multiple small transactions which modify a single element of the array. This access pattern causes the large array to be copied multiple times, increasing overhead. Chapter 6 describes a modification to ApeX which avoids this dependence on object size, removing the $w_v s_v$ term from the performance model.

Well, you go in and you ask for some toothpaste—the small size—and the man brings you the large size. You tell him you wanted the small size but he says the large size is the small size. I always thought the large size was the largest size, but he says that the family size, the economy size and the giant size are all larger than the large size—that the large size is the smallest size there is.

Charade (1963)

Chapter 6

Arrays and large objects

This chapter presents a solution to the "large object problem." Cloning objects to store rollback or transactional state becomes impractical when the objects are large. My solution involves the ***functional array*** datatype, which I first review. I then recast the basic ApeX design of the previous chapters as a "small-object protocol" using naive functional arrays. I extend the improved functional array implementations of Baker [10] and Chuang [21] to obtain a lock-free variant of the data structure. Substituting this improved implementation for the naive functional arrays in our recast transaction system creates Large ApeX, which solves the large object problem. I conclude by evaluating the performance of lock-free functional arrays on a simple microbenchmark, showing that they become worthwhile when the object being accessed is larger than 32 words.

The basic ApeX software transaction system clones objects on transactional writes so that the previous state of the object can be restored if the transaction aborts. Figure 6.1 shows the object size distribution of transactional writes for SPECjvm98, and indicates that over 10% of writes may be to large objects. As we've seen in Section 5.2, the copying cost can become excessive.

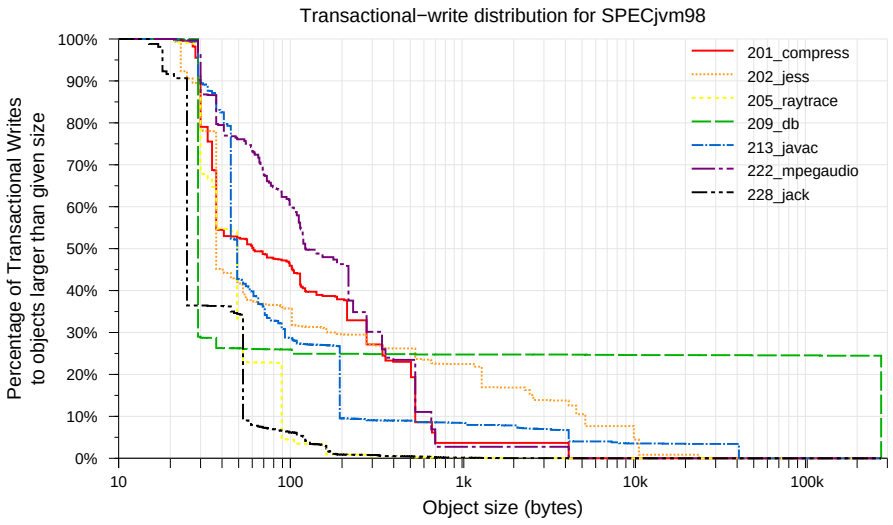

Figure 6.1: Proportion of transactional writes to objects equal to or smaller than a given size.

6.1 Basic operations on functional arrays

Let us begin by reviewing the basic operations on functional arrays. Functional arrays are *persistent*; that is, after an element is updated, both the new and the old contents of the array are available for use. Since arrays are simply maps from integers (indexes) to values, any functional-map datatype (for example, a functional balanced tree) can be used to implement functional arrays. O'Neill and Burton [79] give a fairly inclusive overview of functional array algorithms.

In contrast, an imperative array—such as those in imperative languages such as Java—is mutable. Elements of the array are updated "in place", destroying the old value of the element. Only the updated array is available for use after an update. The distinguishing characteristic of an imperative array is its time complexity: $O(1)$ time to access or update any element. Implementing functional arrays with a functional balanced tree yields $O(\lg n)$

worst-case access or update.[1]

For concreteness, a functional array defines the following three operations:

- FA-CREATE(n): Return an array A of size n. The contents of the array are initialized to 0.
- FA-UPDATE(A, i, v): Return an array A' that is functionally identical to array A except that FA-READ(A', i) $= v$. Array A is not destroyed and can be accessed further.
- FA-READ(A, i): Return $A(i)$ (that is, the value of the ith element of array A).

We allow any of these operations to *fail*. Failed operations can be safely retried, as all operations are idempotent by definition.

For the moment, consider the following naive implementation:

- FA-CREATE(n): Return an ordinary imperative array of size n.
- FA-UPDATE(A, i, v): Create a new imperative array A' and copy the contents of A to A'. Set $A'[i] = v$. Return A'.
- FA-READ(A, i): Return $A[i]$.

Since this implementation costs $O(1)$ to read and $O(n)$ to update, it matches the performance of imperative arrays only when $R = O(U \cdot n)$, where R is the number of reads and U is the number of updates. In most code, $R \approx 3U$ [46, p. 105] and the constant factors hidden by the big-O notation are small, so the performance equivalence is only valid when n is relatively small. I therefore call these *small-object functional arrays*. Operations in this implementation never fail. Every operation is nonblocking and no synchronization is necessary, since the imperative arrays are never mutated after they are created. Section 6.4 reviews better functional array implementations and presents a new lock-free variant.

[1] I return to a discussion of operational complexity in Section 6.4.

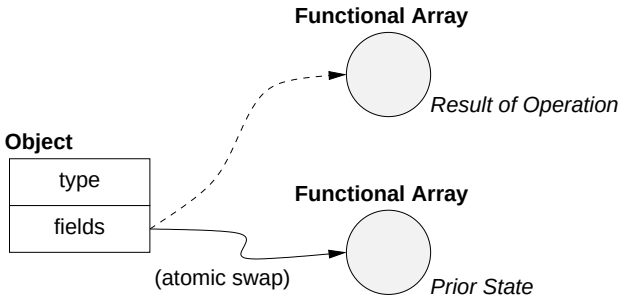

Figure 6.2: Implementing nonblocking single-object concurrent operations with functional arrays.

6.2 A single-object protocol

Given a nonblocking implementation of functional arrays, we can construct a transaction implementation for single objects. In this implementation, fields of at most one object may be referenced during the execution of the transaction.

Consider the following two operations on objects:

- READ(o, f): Read field f of o. Assume that there is a constant mapping function, which, given a field name f, returns an integer index, `f.index`. For simplicity and without loss of generality, assume that all field sizes are equal.

- WRITE(o, f, v): Write value v to field f of o.

All other operations on Java objects, such as method dispatch and type interrogation, can be performed using the immutable `type` field in the object. Because the `type` field never changes after object creation, implementing nonblocking operations on the `type` field is straightforward.

As Figure 6.2 shows, our single-object transaction implementation represents objects as a pair, combining `type` and a reference to a functional array. When not inside a transaction, object reads and writes are implemented using the corresponding functional array operation, with the array

reference in the object being updated appropriately:

- READ(o, f): Return FA-READ(o.fields, f.index).
- WRITE(o, f, v): Replace o.fields with the result of FA-UPDATE(o.fields, f.index, v).

The interesting cases are reads and writes inside a transaction. At entry to a transaction that will access (only) object o, the single-object version of LARGE APEX stores o.fields in a local variable u. We create another local variable u′ initialized to u. Then, the read and write operations are implemented as follows:

- READT(o, f): Return FA-READ(u′, f.index).
- WRITET(o, f, v): Update variable u′ to the result of FA-UPDATE(u′, f.index, v).

At the end of the transaction, we use Compare-And-Swap to atomically set o.fields to u′ if and only if it contained u. If the CAS fails, the transaction is aborted (we simply discard u′) and retried.

With our naive "small object" functional arrays, this implementation is exactly the "small-object protocol" of Herlihy [48]. Herlihy's protocol is rightly criticized for an excessive amount of copying. I address this criticism with a better implementation of functional arrays in Section 6.4. First, however, I remove the restriction that only one object may be referenced within a transaction.

6.3 Extension to multiple objects

I extend the implementation to allow the fields of any number of objects to be accessed during the transaction. Figure 6.3 shows our new object representation. Compare this figure to Figure 3.6; we've successfully recast the basic APEX design in terms of operations on an array datatype.

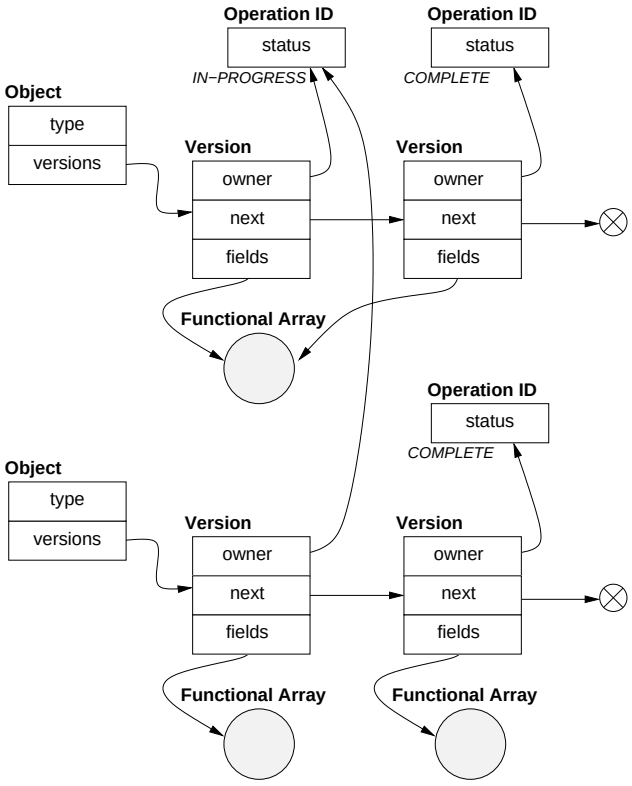

Figure 6.3: Data structures to support nonblocking multiobject concurrent operations. Objects point to a linked list of versions, which reference transaction identifiers. Versions created within the same execution of a transaction share the same transaction identifier. Version structure also contain pointers to functional arrays, which record the values for the fields of the object. If no modifications have been made to the object, multiple versions in the list may share the same functional array. (Compare this model of a transaction system to our concrete design in Figure 3.6.)

```
READ(o, f):
begin
retry:
  u ← o.versions
  u′ ← u.next
  s ← u.owner.status
  if (s = DISCARDED)                          [Delete DISCARDED?]
    CAS(u, u′, &(o.versions))
    goto retry
  else if (s = COMPLETE)
    a ← u.fields                                [u is COMPLETE]
    u.next ← null                               [Trim version list]
  else
    a ← u′.fields                               [u′ is COMPLETE]
  return FA-READ(a, f.index)                    [Do the read]
end

READT(o, f):
begin
  u ← o.versions
  if (oid = u.owner)                          [My OID should be first]
    return FA-READ(u.fields, f.index)           [Do the read]
  else                                          [Make me first!]
    u′ ← u.next
    s ← u.owner.status
    if (s = DISCARDED)                        [Delete DISCARDED?]
      CAS(u, u′, &(o.versions))
    else if (oid.status = DISCARDED)            [Am I alive?]
      fail
    else if (s = IN-PROGRESS)                 [Abort IN-PROGRESS?]
      CAS(s, DISCARDED, &(u.owner.status))
    else                                        [Link new version in:]
      u.next ← null                             [Trim version list]
      u′ ← new Version(oid, u, null)            [Create new version]
      if (CAS(u, u′, &(o.versions)) ≠ FAIL)
        u′.fields ← u.fields                    [Copy old fields]
    goto retry
end
```

Figure 6.4: READ and READT implementations for the multiobject protocol.

```
WRITE(o, f, v):
begin
retry:
  u ← o.versions
  u′ ← u.next
  s ← u.owner.status
  if (s = DISCARDED)                          [Delete DISCARDED?]
    CAS(u, u′, &(o.versions))
  else if (s = IN-PROGRESS)                   [Abort IN-PROGRESS?]
    CAS(s, DISCARDED, &(u.owner.status))
  else                                                [u is COMPLETE]
    u.next ← null                                     [Trim version list]
    a ← u.fields
    a′ ← FA-UPDATE(a, f.index, v)
    if (CAS(a, a′, &(u.fields)) ≠ FAIL)               [Do the write]
      return                                                [Success!]
  goto retry
end

WRITET(o, f, v):
begin
  u ← o.versions
  if (oid = u.owner)                          [My OID should be first]
    u.fields ← FA-UPDATE(u.fields, f.index, v)[Do write]
  else                                                [Make me first!]
    u′ ← u.next
    s ← u.owner.status
    if (s = DISCARDED)                        [Delete DISCARDED?]
      CAS(u, u′, &(o.versions))
    else if (oid.status = DISCARDED)                  [Am I alive?]
      fail
    else if (s = IN-PROGRESS)                 [Abort IN-PROGRESS?]
      CAS(s, DISCARDED, &(u.owner.status))
    else                                        [Link new version in:]
      u.next ← null                                   [Trim version list]
      u′ ← new Version(oid, u, null)                  [Create new version]
      if (CAS(u, u′, &(o.versions)) ≠ FAIL)
        u′.fields ← u.fields                          [Copy old fields]
    goto retry
end
```

Figure 6.5: WRITE and WRITET implementations for the multiobject protocol.

Objects consist of two slots, and the first represents the immutable type, as before. The second field, versions, points to a linked list of Version structures. The Version structures contain a pointer fields to a functional array, and a pointer owner to an *transaction identifier*. The transaction identifier contains a single field, status, which can be set to one of three values: *COMMITTED*, *IN-PROGRESS*, or *ABORTED*. When the transaction identifier is created, the status field is initialized to *IN-PROGRESS*, and it will be updated exactly once thereafter to either *COMMITTED* or *ABORTED*. A *COMMITTED* transaction identifier never later becomes *IN-PROGRESS* or *ABORTED*, and a *ABORTED* transaction identifier never becomes *COMMITTED* or *IN-PROGRESS*.

We create an transaction identifier when we begin or restart a transaction and place it in a local variable *tid*. At the end of the transaction, we use CAS to set *tid*.status to *COMMITTED* if and only if it was *IN-PROGRESS*. If the CAS is successful, the transaction has also executed successfully; otherwise *tid*.status = *ABORTED* (which indicates that our transaction has been aborted), and we must back off and retry. All Version structures created while in the transaction reference *tid* in their owner field.

Semantically, the current field values for the object are given by the first version in the versions list whose transaction identifier is *COMMITTED*. These semantics allow us to link *IN-PROGRESS* versions in at the head of multiple objects' versions lists and atomically change the values of all these objects by setting the one common transaction identifier to *COMMITTED*. We only allow one *IN-PROGRESS* version on the versions list, and it must be at the head. Thus, before we can link a new version at the head, we must ensure that every other version on the list is *ABORTED* or *COMMITTED*.

Since we never look past the first *COMMITTED* version in the versions list, we can free all versions past that point. In our presentation of the algorithm, we do this deallocation by explicitly setting the next field of every *COMMITTED* version we see to **null**; overwriting the reference allows the versions past that point to be garbage collected. An optimization is for

the garbage collector to do the list trimming for us when it does a collection.

Because we don't want to inadvertently chase the null next pointer of a *COMMITTED* version, we always load the next field of a version *before* we load owner.status. Since the writes occur in the reverse order (*COMMITTED* to owner.status, then **null** to next), we have ensured that our next pointer is valid whenever the status is not *COMMITTED*.[2]

We begin an atomic method with TransStart and attempt to complete an atomic method with TransEnd. They are defined as follows:

- TransStart: create a new transaction identifier with its status initialized to *IN-PROGRESS*. Assign it to the thread-local variable *tid*.
- TransEnd: If

  $$\text{CAS}(\textit{IN-PROGRESS}, \textit{COMMITTED}, \&(\textit{tid}.\texttt{status}))$$

  is successful, the transaction as a whole has completed successfully and can be linearized at the location of the CAS. Otherwise, the transaction has been aborted. Back off and retry from TransStart.

Pseudocode describing Read, Write, ReadT, and WriteT is presented in Figures 6.4 and 6.5. In the absence of contention, all operations take constant time plus an invocation of FA-Read or FA-Update.

6.4 Lock-free functional arrays

This section presents a lock-free implementation of functional arrays with O(1) performance for both read and update in the absence of contention. The crucial operation is a rotation of a *difference node* with the main body of the array. Using this implementation of functional arrays in the multiobject transaction protocol of the previous section creates Large ApeX, our

[2] Memory barriers will be necessary here if the architecture does not support sequential consistency.

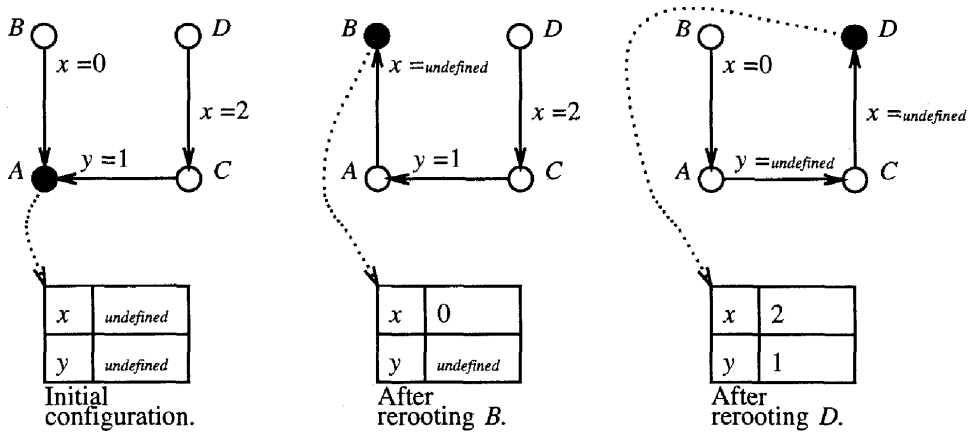

Figure 6.6: Shallow binding scheme for functional arrays [21, Figure 1]. The array is of size 2 and is indexed by x and y. The initial array A is undefined, and B is defined as an update to A at index x by value 0. Similarly for C and D. The dark node is the root node which has the cache. White nodes are differential nodes which must first be rerooted before being read. Note that only the root node has the cache.

reimplementation of nonblocking transactions which solves the large-object problem.

Let's begin by reviewing the well-known functional array implementations. As mentioned previously, O'Neill and Burton [79] give an inclusive overview. Functional array implementations fall generally into one of three categories: *tree-based*, *fat-elements*, or *shallow-binding*.

Tree-based implementations typically have a logarithmic term in their complexity. The simplest is the persistent binary tree with $O(\ln n)$ look-up time; Chris Okasaki [78] has implemented a purely functional random-access list with $O(\ln i)$ expected lookup time, where i is the index of the desired element.

Fat-elements implementations have per-element data structures indexed by a master array. Cohen [23] hangs a list of versions from each element in the master array. O'Neill and Burton [79], in a more sophisticated technique, hang a splay tree off each element and achieve $O(1)$ operations for single-threaded use, $O(1)$ amortized cost when accesses to the array are "uniform", and $O(\ln n)$ amortized worst case time.

Shallow binding was introduced by Baker [11] as a method to achieve fast variable lookup in Lisp environments. Baker clarified the relationship to functional arrays in [10]. Shallow binding is also called *version tree arrays*, *trailer arrays*, or *reversible differential lists*. A typical drawback of shallow binding is that reads may take $O(u)$ worst-case time, where u is the number of updates made to the array. Tyng-Ruey Chuang [21] uses randomized cuts to the version tree to limit the cost of a read to $O(n)$ in the worst case. Single-threaded accesses are $O(1)$.

Our use of functional arrays is single-threaded in the common case, when transactions do not abort. Chuang's scheme is attractive because it limits the worst-case cost of an abort with little added complexity. In this section I will present a lock-free version of Chuang's randomized algorithm.

In shallow binding, only one version of the functional array (the *root*) keeps its contents in an imperative array (the *cache*). Each of the other

versions is represented as a path of *differential nodes*, where each node describes the differences between the current array and the previous array. The difference is represented as a pair $\langle index, value \rangle$, representing the new value to be stored at the specified index. All paths lead to the root. An update to the functional array is simply implemented by adding a differential node pointing to the array it is updating.

The key to constant-time access for single-threaded use is provided by the read operation. A read to the root simply reads the appropriate value from the cache. A read to a differential node, however, triggers a series of rotations that swap the direction of differential nodes and result in the current array acquiring the cache and becoming the new root. This sequence of rotations is called *rerooting*, and is illustrated in Figure 6.6. Each rotation exchanges the root nodes for a differential node pointing to it, after which the differential node becomes the new root and the root becomes a differential node pointing to the new root. The cost of a read is proportional to its rerooting length, but after the first read accesses to the same version are $O(1)$ until the array is rerooted again.

Shallow binding performs badly if read operations ping-pong between two widely separated versions of the array, as we will continually reroot the array from one version to the other. Chuang's contribution is to provide for *cuts* to the chain of differential nodes: once in a while we clone the cache and create a new root instead of performing a rotation. Since this operation takes $O(n)$ time, we amortize it over n operations by randomly choosing to perform a cut with probability $1/n$.

Figure 6.7 shows the data structures used for the functional array implementation, as well as the series of atomic steps used to implement a rotation. The `Array` class, which represents a functional array, consists of a `size` for the array and a pointer to a `Node`. There are two types of nodes: a `CacheNode` stores a value for every index in the array, whereas a `DiffNode` stores a single change to an array. `Array` objects that point to `CacheNodes` are roots.

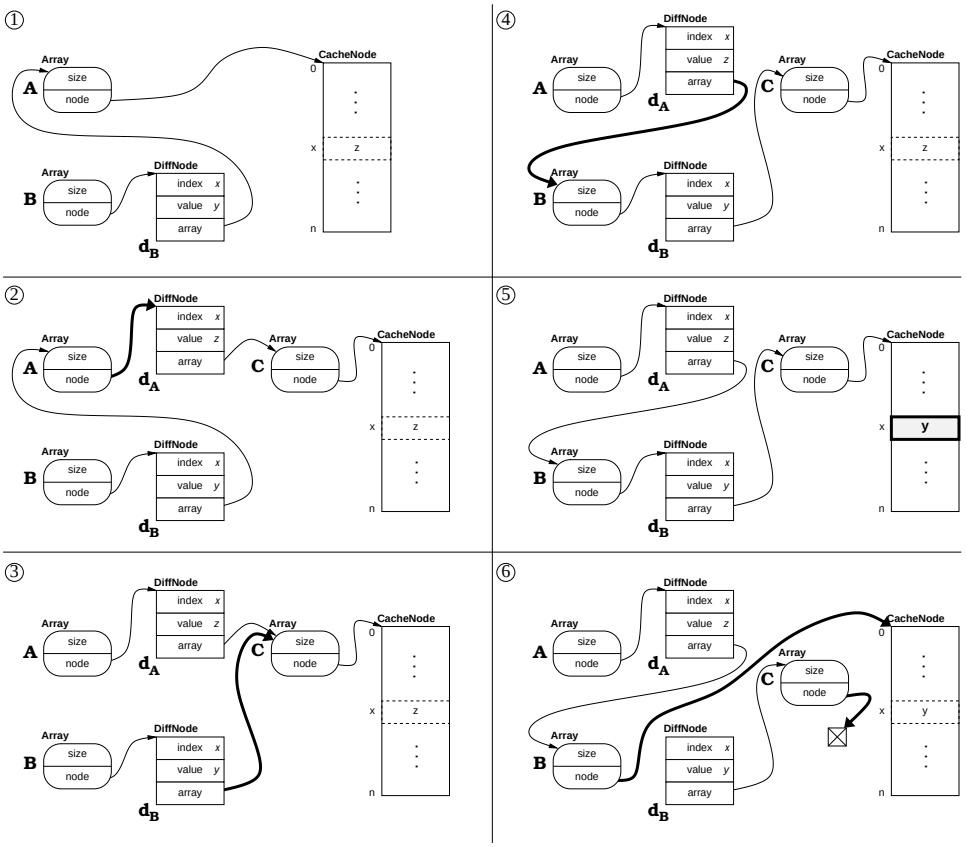

Figure 6.7: Atomic steps in FA-ROTATE(B). Time proceeds top-to-bottom on the left hand side, and then top-to-bottom on the right. Array A is a root node, and FA-READ$(A, x) = z$. Array B has the almost the same contents as A, but FA-READ$(B, x) = y$.

```
FA-UPDATE(A, i, v):
begin
  d ← new DiffNode(i, v, A)
  A′ ← new Array(A.size, d)
  return A′
end

FA-READ(A, i):
begin
retry:
  d_C ← A.node
  if d_C is a cache, then
    v ← A.node[i]
    if (A.node ≠ d_C)[consistency check]
      goto retry
    return v
  else
    FA-ROTATE(A)
    goto retry
end
```

Figure 6.8: Implementation of lock-free functional array using shallow binding and randomized cuts (part 1).

```
FA-ROTATE(B):
begin
retry:
  d_B ← B.node    [step (1): assign names as per Figure 6.7.]
  A ← d_B.array
  x ← d_B.index
  y ← d_B.value
  z ← FA-READ(A, x)                    [rotates A as side effect]

  d_C ← A.node
  if d_C is not a cache, then
    goto retry

  if (0 = (random mod A.size))                              [random cut]
    d'_C ← copy of d_C
    d'_C[x] ← y
    s ← DCAS(d_C, d_C, &(A.node), d_B, d'_C, &(B.node))
    if (s ≠ SUCCESS) goto retry
    else return

  C ← new Array(A.size, d_C)
  d_A ← new DiffNode(x, z, C)

  s ← CAS(d_C, d_A, &(A.node))                              [step (2)]
  if (s ≠ SUCCESS) goto retry
  s ← CAS(A, C, &(d_B.array))                               [step (3)]
  if (s ≠ SUCCESS) goto retry
  s ← CAS(C, B, &(d_A.array))                               [step (4)]
  if (s ≠ SUCCESS) goto retry
  s ← DCAS(z, y, &(d_C[x]), d_C, d_C, &(C.node))            [step (5)]
  if (s ≠ SUCCESS) goto retry
  s ← DCAS(d_B, d_C, &(B.node), d_C, nil, &(C.node))        [step (6)]
  if (s ≠ SUCCESS) goto retry
end
```

Figure 6.9: Implementation of lock-free functional array using shallow binding and randomized cuts (part 2).

In step 1 of the figure, we have a root array A and an array B whose differential node d_B points to A. The functional arrays A and B differ in one element: element x of A is z, while element x of B is y. We are about to rotate B to give it the cache, while linking a differential node to A.

Step 2 shows our first atomic action. We have created a new `DiffNode` d_A and a new `Array` C and linked them between A and its cache. The `DiffNode` d_A contains the value for element x contained in the cache, z, so there is no change in the value of A.

We continue swinging pointers until step 5, when we can finally set the element x in the cache to y. We perform this operation with a DCAS operation that checks that C.node is still pointing to the cache as we expect. A concurrent rotation would swing C.node in its step 1. In general, therefore, the location pointing to the cache serves as a reservation on the cache.

Thus, in step 6 we need to again use DCAS to simultaneously swing C.node away from the cache as we swing B.node to point to the cache.

Figures 6.8 and 6.9 present pseudocode for FA-ROTATE, FA-READ, and FA-UPDATE. Like FA-ROTATE, FA-READ procedure also uses the cache pointer as a reservation, double-checking the cache pointer after it finishes its read to ensure that the cache hasn't been stolen from it.

Let us now consider cuts, where FA-READ clones the cache instead of performing a rotation. Cuts also check the cache pointer to protect against concurrent rotations. But, what if the cut occurs while a rotation is mutating the cache in step 5? In this case, because the only array adjacent to the root is B, the cut must be occurring during an invocation of FA-ROTATE(B), in which case the differential node d_B will be applied after the cache is copied, thereby safely overwriting the mutation we were concerned about.

With hardware support for small transactions [49] we could cheaply perform the entire rotation atomically, instead of using this six-step approach.

```
#define REPETITIONS 100000
typedef int32_t field_t;
typedef int32_t index_t;

void do_bench(struct aarray *obj, index_t len) {
  int i, j;

  /** Initialize the array */
  for (i=0; i<len; i++)
    obj = write(obj, i, i);
  /** Now reverse the array many times. */
  for (j=0; j<(REPETITIONS*2); j++) {
#if defined(SINGLETHREAD) // single-threaded access
    for (i=0; i<len/2; i++) {
      field_t v1 = read(obj, i);
      field_t v2 = read(obj, len-i-1);
      obj = write(obj, i, v2);
      obj = write(obj, len-i-1, v1);
    }
#else // multithreaded access
    struct aarray *robj = obj;
    for (i=0; i<len; i++)
      obj = write(obj, len-i-1, read(robj, i));
#endif
  }
  /** Check that the array has the expected values */
  for (i=0; i<len; i++)
    assert(read(obj, i)==i);
}
```

Figure 6.10: Array reversal microbenchmark to evaluate performance of functional array implementations.

6.5 Performance of functional array implementations

This section presents performance measurements for our lock-free functional array implementation using a simple read/update microbenchmark. We compare our lock-free functional arrays with imperative arrays, naive functional arrays (Section 6.1), shallow-binding functional arrays [10], and Chuang's randomized-cut shallow-binding functional arrays [21].

Figure 6.10 shows the basic structure of the microbenchmark. The `read()` and `write()` methods have appropriate definitions inlined for each

variant of the benchmark. This microbenchmark is patterned after that used in [79, p.507]. With SINGLETHREAD defined, accesses are imperative or "single-threaded"—only the latest version of the array is referenced. The algorithm corresponds to the typical imperative array reversal algorithm. We swap the leftmost and rightmost element of the array and move inward as we continue to swap. Without SINGLETHREAD defined, the reversal algorithm always reads from the original array. This pattern of access corresponds to a scenario where aborts are frequent. In our experiments the single-threaded and multithreaded variants of the benchmark had almost identical performance, contrary to the claims of [79]. Standard shallow-binding will have poor performance only if reads of the half-reversed array were to occur in the multithreaded variant. These reads are not necessary for array reversal.

Figure 6.11 shows the performance of various array implementations using the single-threaded variant of the benchmark. I measured the number of microseconds required for a read and update pair on the array, as the size of the array ranged from 8 to 4096 elements. Benchmarks were executed on the PowerPC hardware described in Section 5.2.[3]

Standard imperative arrays averaged 5.0 nanoseconds per read-update, mostly invariant with array size. The naive functional array implementation, described in Section 6.1, ranged from 475 ns for 8-element arrays up to 47,629 ns for 4096-element arrays. The exponential growth in run time with increasing array size demonstrates the "large object problem" which motivated our investigation of lock-free functional arrays in this chapter.

Shallow-binding functional arrays [10] averaged 327 nanoseconds per read-update. This is over 60 times slower than standard imperative arrays, but is invariant with array size. Shallow binding shows poor performance when accesses are not single-threaded—in a transactional application, when transactions abort often.

Chuang's randomized-cut shallow-binding functional arrays [21], labeled "random cut" in Figure 6.11, allow good performance even when transac-

[3] Results on a 1.4GHz Pentium M were similar.

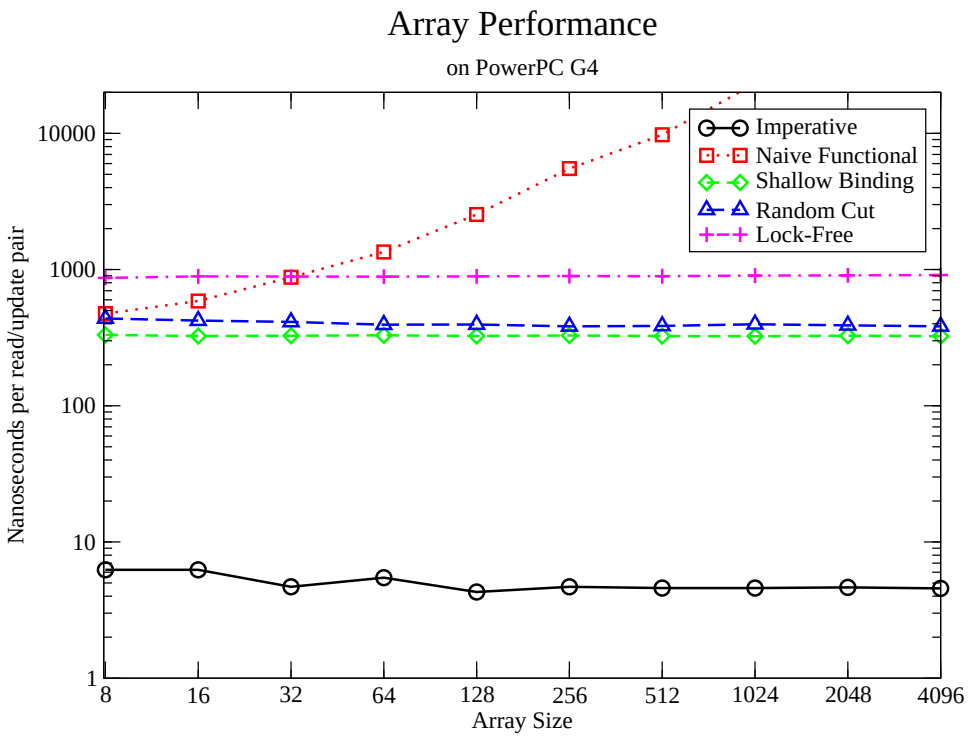

Figure 6.11: Functional array performance on an array reversal microbenchmark as the size of the array is varied. Both axes are logarithmic. The y axis shows the average time, in nanoseconds, required to do an array read followed by an update. The benchmark initialized the array and then reversed its contents 200,000 times. The reversal swapped the first and last elements, then the second and second-to-last elements, etc.

tions abort often. My implementation averaged 400 ns per read-update. The root-cloning which allows multithreaded access imposes an additional 22% overhead when compared to the standard shallow binding functional array implementation.

Finally, my lock-free version of Chuang's functional arrays, labeled "lock-free" in the figure, averages 894 ns per read-update. The penalty for performing the lock-free algorithm is 124%, resulting in read-updates which are over 175 times slower than imperative array read-updates. Figure 6.11 shows that the lock free algorithm is still faster than copying objects on update when the objects are larger than 32 words long.

In a hybrid transaction system which used small hardware transaction support for the functional array implementation, one could expect performance similar to the standard randomized-cut shallow-binding implementation. One would simply make the rotation operations atomic using the hardware transaction system.

People who are really serious about software should make their own hardware. Remember, it's all software, it just depends on when you crystallize it.

Alan Kay

Chapter 7

Transactions in hardware: Unbounded Transactional Memory

With hardware support, we can construct transaction systems that are more efficient than ApeX for certain types of transactions. This chapter presents UTM, an implementation of unbounded transactional memory [6, 7] which fully virtualizes transactions. I also present LTM, a much simpler design which can be pin-compatible with today's processors. Although LTM supports more limited transactions, I show how LTM can be combined with the software ApeX transaction system to yield HyApeX, a hybrid system with a great deal of power and flexibility.[1]

7.1 The UTM architecture

UTM, a system that implements unbounded transactional memory in hardware, allows transactions to grow (nearly) as large as virtual memory. It also supports a semantics for nested transactions, where interior transac-

[1] Portions of this chapter are adapted from [6, 7], co-written with Krste Asanović, Bradley C. Kuszmaul, Charles E. Leiserson, and Sean Lie.

tions are subsumed into the atomic region represented by the outer transaction. Unlike previous schemes that tie a thread's transactional state to a particular processor and/or cache, UTM maintains bookkeeping information for a transaction in a memory-resident data structure, the ***transaction log***. This log enables transactions to survive timeslice interrupts and process migration from one processor to another. We first present the software interface to UTM and then describe the implementation details.

7.1.1 New instructions

UTM adds two new instructions to a processor's instruction set architecture:

`XBEGIN pc`: Begin a new transaction. The `pc` argument to `XBEGIN` specifies the address of an ***abort handler*** (e.g., using a PC-relative offset). If at any time during the execution of a transaction the hardware determines that the transaction must fail, it immediately rolls back the processor and memory state to what it was when `XBEGIN` was executed, then jumps to `pc` to execute the abort handler.

`XEND`: End the current transaction. If `XEND` completes, then the transaction is committed, and all of its operations appear to be atomic with respect to any other transaction.

Semantically, we can think of an `XBEGIN` instruction as a conditional branch to the abort handler. The `XBEGIN` for a transaction that fails has the behavior of a mispredicted branch. Initially, the processor executes the `XBEGIN` as a not-taken branch, falling through into the body of the transaction. Eventually the processor realizes that the transaction cannot commit, at which point it reverts all processor and memory state back to the point of misprediction and branches to the abort handler.

In the same manner as the ApeX software implementation (Section 4.2.1), UTM supports the nesting of transactions by subsuming the inner transaction. For example, within an outer transaction, a subroutine that contains

an inner transaction may be called. UTM simply treats the inner transaction as part of the atomic region defined by the outer one. This strategy is correct, because it maintains the property that the inner transaction executes atomically. Subsumed nested transactions are implemented by using a counter to keep track of nesting depth. If the nesting depth is positive, then XBEGIN and XEND simply increment and decrement the counter, respectively, and perform no other transactional bookkeeping.

7.1.2 Rolling back processor state

The branch mispredict mechanism in conventional superscalar processors can roll back register state only for the small window of recent instructions that have not graduated from the reorder buffer. To circumvent the window-size restriction and allow arbitrary rollback for unbounded transactions, the processor must be modified to retain an additional snapshot of the architectural register state. A UTM processor saves the state of its architectural registers when it graduates an XBEGIN. The snapshot is retained either until the transaction aborts, at which point the snapshot is restored into the architectural registers, or until the matching XEND graduates indicating that the transaction has committed.

UTM's modifications to the processor core are illustrated in Figure 7.1. We assume a machine with a unified physical register file, and so rather than saving the architectural registers themselves, UTM saves a snapshot of the register-renaming table and ensures the corresponding physical registers are not reused until the transaction commits. The rename stage maintains an additional "saved" bit for each physical register to indicate which registers are part of the working architectural state, and it takes a snapshot as each branch or XBEGIN is decoded and renamed. When an XBEGIN instruction graduates, activating the transaction, the associated "S bit" snapshot has bits set on exactly those registers holding the graduated architectural state. Physical registers are normally freed on graduation of a later instruction that overwrites the same architectural register. If the S bit on the snapshot

for the active transaction is set, the physical register is added to a FIFO called a ***Register Reserved List*** instead of the normal ***Register Free List***, thereby preventing physical registers containing saved data from being overwritten during a transaction. When the transaction's `XEND` commits, the active snapshot's `S` bits are cleared and the Register Reserved List is drained into the regular Register Free List. In the event that the transaction aborts, the saved register-renaming table is restored and the reorder buffer is rolled back, as in an exception. After restoring the architectural register state, the branch is taken to the abort handler. Even though the processor can internally speculatively execute ahead through multiple transactions, transactions only affect the global memory system as instructions graduate, and hence UTM requires only a single snapshot of the architectural register state.

The current transaction abort handler address, nesting depth, and register snapshot are part of the transactional state. They are made visible to the operating system (as additional processor control registers) to allow them to be saved and restored on context switches.

7.1.3 Memory state

Previous HTM systems [49, 63] represent a transaction partly in the processor and partly in the cache, taking advantage of the coincidence between the cache-consistency protocol and the underlying consistency requirements of transactional memory. Unlike those systems, UTM transactions are represented by a single ***xstate*** data structure held in the memory of the system. The cache in UTM is used to gain performance, but the correctness of UTM does not depend on having a cache. In the following paragraphs, we first describe the xstate and how the system uses it assuming there is no caching. Then, we describe how caching accelerates xstate operations.

The xstate is illustrated in Figure 7.2. The xstate contains a transaction log for each active transaction in the system. A transaction log is allocated by the operating system for each thread, and two processor control registers

hold the base and bounds of the currently active thread's log. Each log consists of a ***commit record*** and a vector of ***log entries***. The commit record maintains the transaction's status: `PENDING`, `COMMITTED`, or `ABORTED`. Each log entry corresponds to a block of memory that has been read or written by the transaction. The entry provides a pointer to the block and the old (backup) value for the block so that memory can be restored in case the transaction aborts. Each log entry also contains a pointer to the commit record and pointers that form a linked list of all entries in all transaction logs that refer to the same block.

The final part of the xstate consists of a ***log pointer*** and one ***RW bit*** for each block in memory (and on disk, when paging). If the RW bit is `R`, any transactions that have accessed the block did so with a load; otherwise, if it is `W`, the block may have been the target of a transaction's store. When a processor running a transaction reads or writes a block, the block's log pointer is made to point to a transaction log entry for that block. Further, if the access is a write, the RW bit for the block is set to `W`. Whenever another processor references a block that is already part of a pending transaction, the system consults the RW bit and log pointer to determine the correct action, for example, to use the old value, to use the new value, or to abort the transaction.

When a processor makes an update as part of a transaction, the new value is stored in memory and the old value is stored in an entry in the transaction log. In principle, there is one log entry for every load or store performed by the transaction. If the memory allocated to the log is not large enough, the transaction aborts and the operating system allocates a larger transaction log and retries the transaction. When operating on the same block more than once in a transaction, the system can avoid writing multiple entries into the transaction log by checking the log pointer to see whether a log entry for the block already exists as part of the running transaction.

By following the log pointer to the log entry and then following the log entry pointer to the commit record, one can determine the transaction status

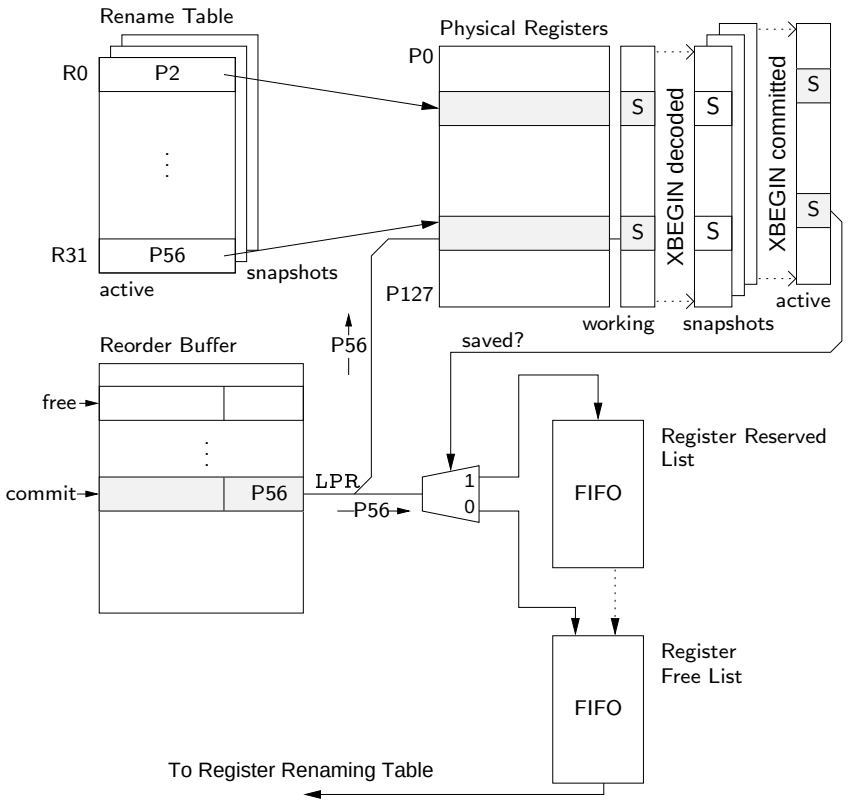

Figure 7.1: UTM processor modifications. The S bit vector tracks the active physical registers. For each rename table snapshot, there is an associated S bit vector snapshot. The Register Reserved List holds the otherwise free physical registers until the transaction commits. The LPR field is the next physical register to free (the last physical register referenced by the destination architectural register).

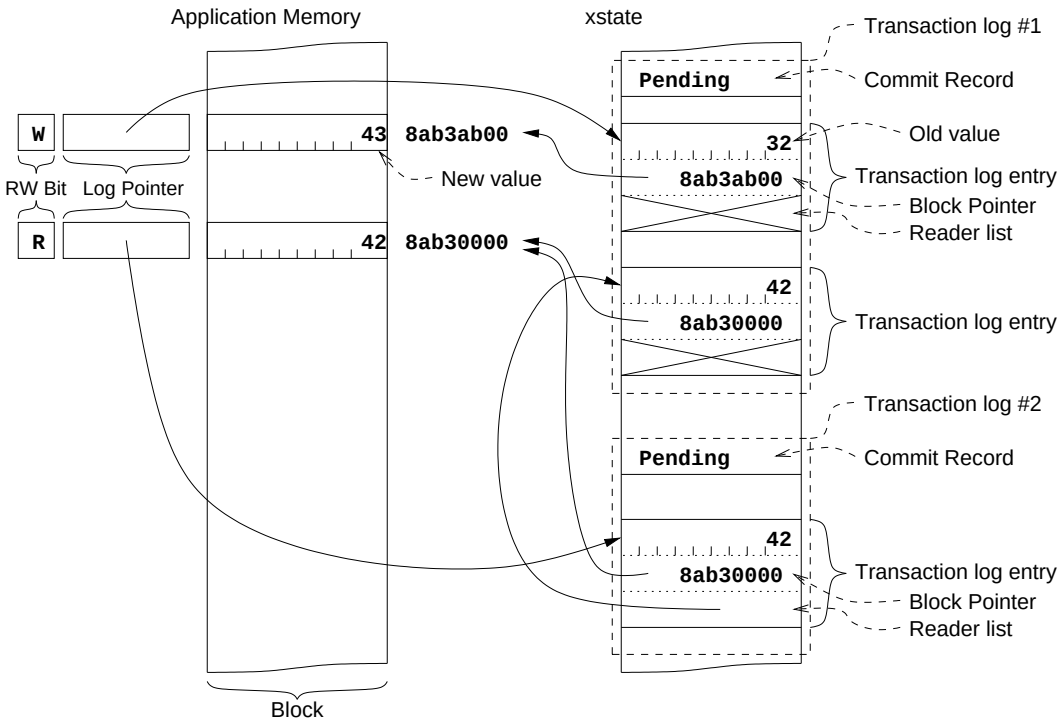

Figure 7.2: The xstate data structure. The transaction log for a transaction contains a commit record and a vector of log entries. The log pointer of a block in memory points to a log entry, which contains the old value of the block and a pointer to the transaction's commit record. Two transaction logs are shown here; generally, the xstate includes the active transaction logs for the entire system.

(pending, committed, or aborted) of each block. To commit a transaction, the system simply changes the commit record from PENDING to COMMITTED. At this point, a reference to the block produces the new value stored in memory, albeit after some delay in chasing pointers to discover that the transaction has been committed. To avoid this delay, as well as to free the transaction log for reuse, the system must clean up after committing. It does so by iterating through the log entries, clearing the log pointer for each block mentioned, thereby finalizing the contents of the block. Future references to that block will continue to produce the new value stored in memory, but without the delay of chasing pointers. To abort a transaction, the system changes the commit record from PENDING to ABORTED. To clean up, it iterates through the entries, storing the old value back to memory and then clearing the log pointer. We chose to store the old value of a block in the transaction log and the new value in memory, rather than the reverse, to optimize the case when a transaction commits. No data copying is needed to clean up after a commit, only after an abort.

When two or more pending transactions have accessed a block and at least one of the accesses is a store, the transactions conflict. Conflicts are detected during operations on memory. When a transaction performs a load, the system checks that either the log pointer refers to an entry in the current transaction log, or else that the RW bit is R (additionally creating an entry in the current log for the block if needed). When a transaction performs a store, the system checks that no other transaction is referenced by the log pointer (i.e., that the log pointer is cleared or that the linked list of log entries corresponding to this block are all contained in the current transaction log). If the conflict check fails, then some of the conflicting transactions are aborted. To guarantee forward progress, UTM writes a timestamp into the transaction log the first time a transaction is attempted. Then, when choosing which transactions to abort, older transactions take priority. As an alternative, a backoff scheme [74] could also be used.

When a writing transaction wins a conflict, there may be multiple read-

ing transactions that must be aborted. These transactions are found efficiently by following the block's log pointer to an entry and traversing the linked list found there, which enumerates all entries for that block in all transaction logs.

7.1.4 Caching

Although UTM can support transactions of unbounded size using the xstate data structure, multiple memory accesses for each operation may be required. Caching is needed to achieve acceptable performance. In the common case of a transaction that fits in cache, UTM, like the earlier proposed HTM systems [49, 63], monitors the cache-coherence traffic for the transaction's cache lines to determine if another processor is performing a conflicting operation. For example, when a transaction writes to a memory location, the cache protocol obtains exclusive ownership on the whole cache block. New values can be stored in cache with old values left in memory. As long as nothing revokes the ownership of any block, the transaction can succeed. Since the contents of the transaction log are undefined after the transaction commits or aborts, in many cases the system does not even need to write back a transaction log. Thus, for a small transaction that commits without intervention from another transaction, no additional interprocessor communication is required beyond the coherence traffic for the nontransactional case. When the transaction is too big to fit in cache or interactions with other transactions are indicated by the cache protocol, the xstate for the transaction overflows into the ordinary memory hierarchy. Thus, the UTM system does not actually need to create a log entry or update the log pointer for a cached block unless it is evicted. After a transaction commits or aborts, the log entries of unspilled cached blocks can be discarded and the log pointer of each such block can be marked clean to avoid write-back traffic for the log pointer, which is no longer needed. Most of the overhead is borne in the uncommon case, allowing the common case to run fast.

The in-cache representation of transactional state and the xstate data

structure stored in memory need not match. The system can optimize the on-processor representation as long as, at the cache interface, the view of the xstate is properly maintained. For convenience, the transaction block size can match the cache line size.

7.1.5 System issues

The goal of UTM is to support transactions that can run for an indefinite length of time (surviving time slice interrupts), can migrate from one processor to another along with the rest of a process's state, and can have footprints bigger than the physical memory. Several system issues must be solved for UTM to achieve that goal. The "big idea" of UTM is to treat the xstate as a system-wide data structure that uses global virtual addresses.

Treating the xstate as data structure directly solves part of the problem. For a transaction to run for an indefinite length of time, it must be able to survive a time-slice interrupt. By adding the log pointer to the processor state and storing everything else in a data structure, it is easy to suspend a transaction and run another thread with its own transaction. Similarly, transactions can be migrated from one processor to another. The log pointer is simply part of the thread or process state provided by the operating system.

UTM can support transactions that are even larger than physical memory. The only limitation is how much virtual memory is available to store both old and new values. To page the xstate out of main memory, the UTM data structures might employ global virtual addresses for their pointers. Global virtual addresses are system-wide unique addresses that remain valid even if the referenced pages are paged out to disk and reloaded in another location. Typically, systems that provide global virtual addresses provide an additional level of address translation, compared to ordinary virtual memory systems. Hardware first translates a process's virtual address into a global virtual address. The global virtual address is then translated into a physical address. Multics [15] provided user-level global virtual addressing

using segment-offset pairs as the addresses. The HP Precision Architecture [66] supports global virtual addresses in a 64-bit RISC processor.

The log pointer and state bits for each user memory block, while typically not visible to a user-level programmer, are themselves stored in addressable physical memory to allow the operating system to page this information to disk. The location of the memory holding the log pointer information for a given user data page is kept in the page table and cached in the TLB.

During execution of a single load or store instruction, the processor can potentially touch a large number of disparate memory locations in the xstate, any of which may be paged out to disk. To ensure forward progress, either the system must allow load or store instructions to be restarted in the middle of the xstate traversal, or, if only precise interrupts are allowed, the operating system must ensure that all pages required by an xstate traversal can be resident simultaneously to allow the load or store to complete without page faults.

UTM assumes that each transaction is a serial instruction stream beginning with an `XBEGIN` instruction, ending with a `XEND` instruction, and containing only register, memory, and branch instructions in between. A fault occurs if an I/O instruction is executed during a transaction.

7.2 The LTM architecture

UTM is an idealized design for HTM which requires significant changes to both the processor and the memory subsystem of a contemporary computer architecture. By scaling back on the degree of "unboundedness," however, a compromise between programmability and practicality can be achieved. This section presents such an architectural compromise, called LTM, for which we have implemented a detailed cycle-level simulation using UVSIM [102]. The limited transactions supported by LTM are still powerful enough to serve as the basis for a hybrid system, as we will show in Section 7.4.

LTM's design is easier to implement than UTM, because it does not support transactions of virtual-memory size. Instead, LTM avoids the intricacies of virtual memory by limiting the footprint of a transaction to (nearly) the size of physical memory. In addition, the duration of a transaction must be less than a time slice, and transactions cannot migrate between processors. With these restrictions, LTM can be implemented by only modifying the cache and processor core and without making changes to the main memory, the cache-coherence protocols, or even the contents of the cache-coherence messages. Unlike a UTM processor, an LTM processor can be pin-compatible with a conventional processor. The design presented here is based on the SGI Origin 3000 shared-memory multiprocessor, with memory distributed among the processor nodes and cache coherency maintained using a directory-based write-invalidate protocol.

The UTM and LTM schemes share many ideas. Like UTM, LTM maintains data about pending transactions in the cache and detects conflicts using the cache-coherency protocol in much the same way as previous HTM proposals [55, 63]. LTM also employs an architectural state-save mechanism in hardware. Unlike UTM, LTM does not treat the transaction as a data structure. Instead, it binds a transaction to a particular cache. Transactional data overflows from the cache into a hash table in main memory, which allows LTM to handle transactions too big to fit in the cache without the full implementation complexity of the xstate data structure.

LTM has similar semantics to UTM, and the format and behavior of the `XBEGIN` and `XEND` instructions are the same. The information that UTM keeps in the transaction log is kept partly in the processor, partly in the cache, and partly in an area of physical memory allocated by the operating system.

LTM requires only a few small modifications to the cache, as shown in Figure 7.3. For small transactions, the cache is used to store the speculative transactional state. For large transactions, transactional state is spilled into an overflow data structure in main memory. An additional bit (T) is added

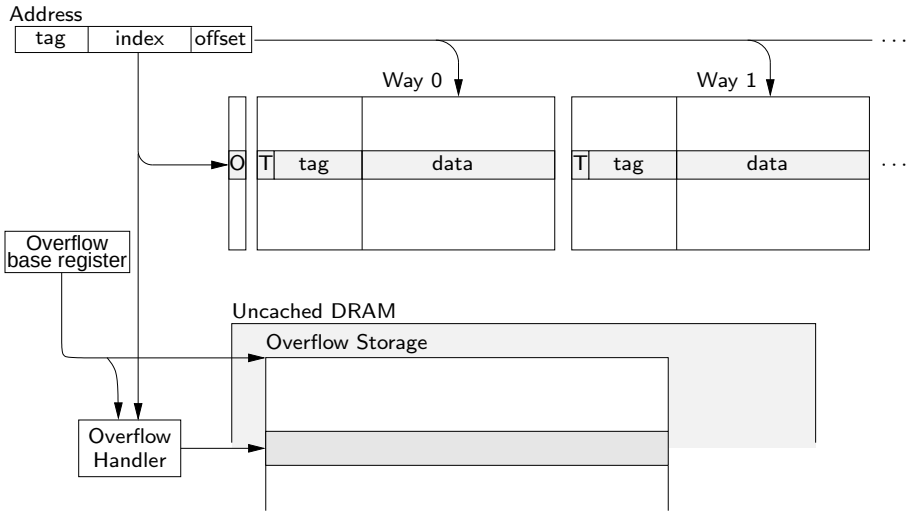

Figure 7.3: LTM cache modifications. The T bit indicates if the line is transactional. The O bit indicates if the set has overflowed. Overflowed data is stored in a data structure in uncached DRAM.

per cache line to indicate if the data has been accessed as part of a pending transaction. When a transactional-memory request hits a cache line, the T bit is set. An additional bit (O) is added per cache set to indicate if it has overflowed. When a transactional cache line is evicted from the cache for capacity reasons, the O bit is set.

In LTM, the main memory always contains the original state of any data being modified transactionally, and all speculative transactional state is stored in the cache and overflow hash table. A transaction is committed by simply clearing all the T bits in cache and writing all overflowed data back to memory. Conflicts are detected using the cache-coherency protocol. When an incoming cache intervention hits a transactional cache line, the running transaction is aborted by simply clearing all the T bits and invalidating all modified transactional cache lines.

The overflow hash table in uncached main memory is maintained by hardware, but its location and size are set up by the operating system. If a

request from the processor or a cache intervention misses on the resident tags of an overflowed set, the overflow hash table is searched for the requested line. If the requested cache line is found, it is swapped with a line in the cache set and handled like a hit. If the line is not found, it is handled like a miss. While handling overflows, all incoming cache interventions are stalled using a NACK-based network protocol.

The LTM overflow data structure uses the low-order bits of the address as the hash index and uses linear probing to resolve conflicts. When the overflow data structure is full, the hardware signals an exception so that the operating system can increase the size of the hash table and retry the transaction.

LTM was designed to be a first step towards a truly unbounded transactional memory system such as UTM. LTM has most of the advantages of UTM while being much easier to implement. LTM's more practical implementation of quasi-unbounded transactional memory suffices for many real-world concerns. Moreover, as Section 7.4 shows, LTM can be symbiotically paired with the more flexible ApeX software transaction system to achieve truly unbounded transactions at minimal hardware cost.

7.3 Evaluation

This section evaluates the UTM and LTM designs, demonstrating low overhead and scalability. We examine overflow behavior, providing motivation for the hybrid system proposed in Section 7.4.

7.3.1 Scalability

We used a parallel microbenchmark to examine program behavior for small transactions with high contention. Our results show that the extremely low overhead of small hardware transactions enable them to almost always complete even when contention is high.

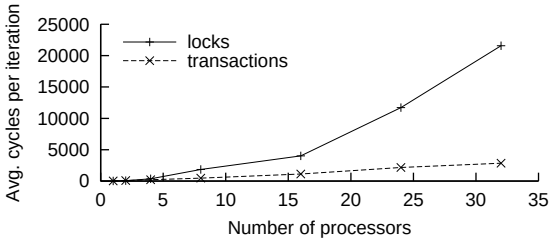

Figure 7.4: `Counter` performance on UVSIM.

The `Counter` microbenchmark has one shared variable that each processor atomically increments repeatedly with no backoff policy; the basic idea is identical to the microbenchmark we used in Section 5.1. Each transaction is only a few instructions long and every processor attempts to write to the same location repeatedly. Both a locking and a transactional version of `Counter` were run on UVSIM with LTM, and the results are shown in Figure 7.4. In the locking version, there is a global spin-lock that each processor obtains using a load-linked/store-conditional (LLSC) sequence from the SGI synchronization libraries.

The locking version scales poorly, because the LLSC causes many cache interventions even when the lock cannot be obtained. On the other hand, the transactional version scales much better, despite having no backoff policy. When a transaction obtains a cache line, it is likely to be able to execute a few more instructions before receiving an intervention since the network latency is high. Therefore, small transactions can start and complete (and perhaps even start and complete the next transaction) before the cache line is taken away. Similar behavior is expected from UTM, and other transactional-memory systems that use the cache and cache-coherency protocol to store transactional state, since small transactions effectively use the cache the same way.

7.3.2 Overhead

A main goal of LTM and UTM is to run the common case fast. As shown in Section 3.1, the common case is when transactions are small and fit in the cache. Therefore, by using the cache and cache coherency mechanism to handle small transactions, LTM is able to execute with almost no overhead over serial code in the common case. In this section, we discuss qualitatively how the LTM implementation is optimized for the common case and how similar techniques are used in UTM. The discussion is broken into the following three cases: starting, running, committing a transaction.

Starting a transaction in LTM requires virtually no overhead in the common case since the hardware only needs to record the abort handler address. No communication with the cache or other external hardware is necessary. There is the added overhead of decoding the `XBEGIN` however that overhead is generally insignificant compared to the cost of the transaction. Further, instruction decode overhead is much lower in LTM than with locks. Even schemes where the lock is not actually held such as SLE [80] have higher decode overhead since they have more instructions. LTM's low transaction startup overhead is a good indicator of the corresponding overhead in UTM, since transaction start up in UTM is virtually the same.

Running a transaction in LTM requires no more overhead than running the corresponding non-synchronized code in the common case. In LTM, the T bit is simply set on each transactional cache access. LTM's low overhead in this case unfortunately does not translate directly to UTM since UTM modifies the transaction record on each memory request. In the common case, however, the transaction record entry is also in the cache. Thus, all operations are local and no external communication is needed. Also, in some cases, the cache can respond to the memory request once the requested data is found. If the request requires data from the transaction record before it can be serviced, however, an additional cache lookup is necessary, but the lookup is local and can be done relatively quickly. Therefore, the common case overhead of running a transaction can be minimal even in UTM.

Committing a transaction in LTM has virtually no overhead in the common case, since it can be done in one clock cycle. LTM transaction commits only requires a simple flash clear of all the transaction bits in the cache. Similarly, UTM transaction commits only require a single change of the cached transaction record to "committed." Although UTM transaction commit also writes the updated values from the transaction record back to memory, this write-back can be done lazily in the background. Therefore, since transaction commit requires only a single change in the cache for both LTM and UTM, the overhead is minimal in both cases.

7.3.3 Overflows

Although overflows occur only in the uncommon case, our studies show that it is important to have a scalable data structure even though it is used infrequently.

For evaluation, we compiled three versions of the SPECjvm98 benchmark suite to run under UVSIM using FLEX. We compiled a ***Base*** version which uses no synchronization, a ***Locks*** version which uses spin-locks for synchronization, and a ***Trans*** version which uses LTM transactions for synchronization. To measure overheads, we ran these versions of the SPECjvm98 benchmark suite on one processor of UVSIM.

As described in Section 4.2, our transactional version uses method cloning to flatten transactions. We performed the same cloning on the other compiled versions so that performance improvements due to the specialization would not be improperly attributed to transactionification. The three different benchmark versions were built from a common code-base using method inlining in gcc[2] to remove or replace all invocations of lock and transaction entry and exit code with appropriate implementations. No garbage collection was performed during these benchmark runs.

[2]We compiled the files generated by the "Precise C" backend (Section 4.1) of FLEX with -O9 for a -mips4 target using the n64 API to generate fully static binaries executable by UVSIM.

Our initial results from Section 3.1 suggested that overflows ought to be infrequent, implying that the efficiency of the overflow data structure would have a negligible effect on overall performance. Consequently, our first LTM implementation used an unsorted array which required a linear search on each miss to an overflowed set. The unsorted array was effective for most of our test cases, as they had less overhead than locks. Using LTM with the unsorted array, however, the transactional version of 213_javac was 14 times slower than the base version. Virtually all of the overhead came from handling overflows, which is not surprising, since the entire application is enclosed in one large transaction. The large transaction touches 13K cache lines with 9K lines overflowed. So, even though only 0.5% of the transactional memory operations miss in the cache, each one incurs a huge search cost. This unexpected slowdown indicated that a naive unsorted array is insufficient as an overflow data structure. Therefore, LTM was redesigned to use a hash table to store overflows.

Since the entire application was enclosed in a transaction, the 213_javac application was clearly not written to be a parallel application. It is important, however, that an unbounded transactional memory system be able to support even such applications with reasonable performance. Therefore, we redesigned LTM to use hash table as described in Section 7.2.

Using LTM with the hash table, the SPECjvm98 application overheads were much more reasonable as shown in Figure 7.5. The hash table data structure decreased the overhead from a 14x slowdown to under 15% in 213_javac. Using the hash table, LTM transactional overhead is less than locking overhead in all cases.

7.4 A hybrid transaction implementation

We've seen that UTM and LTM can operate with little overhead, but hardware schemes encounter difficulties when scaling to large or long-lived transactions. We have overcome some of the difficulties with an overflow cache

Benchmark application	Base time (cycles)	Locks time (% of Base time)	Trans time (% of Base time)	Time in trans (% of Trans time)	Time in overflow (% of Trans time)
200_check	8.1M	124.0%	101.0%	32.5%	0.0085%
202_jess	75.0M	140.9%	108.0%	59.4%	0.0072%
209_db	11.8M	142.4%	105.2%	54.0%	0%
213_javac	30.7M	169.9%	114.2%	84.2%	10%
222_mpegaudio	99.0M	100.3%	99.6%	0.8%	0%
228_jack	261.4M	175.3%	104.3%	32.1%	0.0056%

Figure 7.5: SPECjvm98 performance on a 1-processor UVSIM simulation. The *Time in trans* and *Time in overflow* are the times spent actually running a transaction and handling overflows respectively. The input size is 1. The overflow hash table is 128MB.

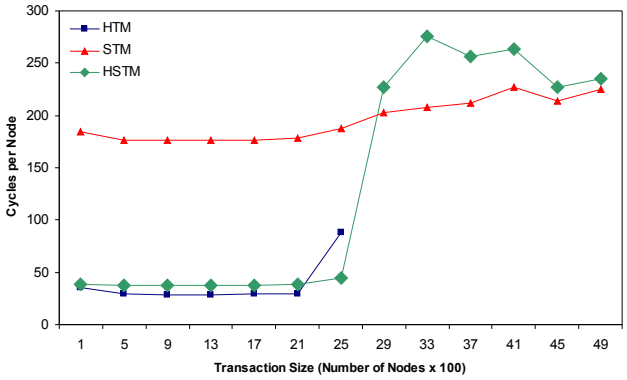

Figure 7.6: Performance (in cycles per node push on a simple queue benchmark) of LTM [6] (HTM), the object-based system presented in this paper (STM) and a hybrid scheme (HSTM).

(LTM), or by virtualizing transactions and dumping their state to a data structure (UTM). It is worth considering, however, whether this extra complexity is worthwhile: why not combine the strengths of our object-based software transaction system (explicit transaction state, unlimited transaction size, flexibility) with the fast small transactions at which a hardware system naturally excels?

Figure 7.6 presents the results of such a combination. In the figure, combining the systems is done in the most simple-minded way: all transactions are begun in LTM, and after any abort the transaction is restarted in the object-based software system. The field flag mechanism described in Section 3.2.5 ensures that software transactions properly abort conflicting hardware transactions: when the software scribbles `FLAG` over the original field, the hardware detects the conflict. Hardware transactions must perform the `ReadNT` and `WriteNT` algorithms to ensure they interact properly with concurrent software transactions, although these checks need not be part of the hardware transaction mechanism. In the figure, the checks were done in software, with an implementation similar to that described in Section 5.1.

The figure shows the performance of a simple queue benchmark as the transaction size increases. The hardware transaction mechanism is fastest, as one would expect, but its performance falters and then fails at a transaction size of around 2500 nodes pushed. At this transaction size, the hardware scheme ran out of cache; in a more realistic system it might also have run out of its timeslice, aborting (LTM) or spilling (UTM) the transaction at the context switch.

Above HTM in the figure is the performance of the software transaction system: about 4x slower, which is a pessimistic figure. No special effort was made to tune code or otherwise minimize slowdown, and the processor simulated had limited ability to exploit ILP (2 ALUs and 4-instruction issue width). The software scheme, however, is unaffected by increasing transaction size.

The hybrid scheme successfully combines the best features of both. It

is only about 20% slower than the basic hardware scheme, due to the read and write barriers it must implement, but at the point where the hardware stops working well, it smoothly crosses over to track the performance of the software transaction system.

There are many fortuitous synergies in such an approach. Hardware support for small transactions may be used to implement the software transaction implementation's Load Linked/Store Conditional sequences, which may not otherwise be available on a target processor. The small transaction support can also facilitate a functional-array solution to the large-object problem, as we saw in Section 6.4. We might further improve performance by adding a bit of hardware support for the `readNT`/`writeNT` barriers [22].

I believe this hybrid approach is the most promising direction for transaction implementations in the near future. It preserves the flexibility to investigate novel solutions to the outstanding challenges of transactional models, which we review in Chapter 8, and it solves an important chicken-and-egg problem preventing the development of transactional software and the deployment of transactional hardware. Since the speed of the hardware mechanism is tempered by the cooperation protocol of the software transaction system, high-efficiency software transaction mechanisms, such as the one presented in this thesis, are the key enabler for hybrid systems.

The thing is to remember that the future comes one day at a time.

Dean Acheson

Chapter 8

Challenges

This chapter reviews objections that have been raised to straightforward or naive transaction implementations. Although some of these objections do not apply to our implementation, discussing them may further illuminate our design choices. Other objections apply to certain situations, and should be kept in mind when creating applications. Some of the problems raised remain unsolved and are the subject of future work. For these problems we attempt to sketch research directions.

8.1 Performance isolation

Zilles and Flint [104] identified ***performance isolation*** as a potential issue for transaction implementations. In a system with performance isolation, the execution of one task (process, thread, transaction) should complete in an amount of time which is independent of the execution of other tasks (processes, threads, transactions) in the system. For a system with N processors, it is ideal if a task is guaranteed to complete at least as quickly as it would running alone on 1 processor.

Most common systems do not provide any guarantee of performance isolation. On a typical multiuser system, the execution of a given task can

be made arbitrarily slow by the concurrent execution of competing tasks.[1] A nontransactional system can nevertheless be constructed with a good deal of performance isolation by appropriately restricting the processes that can be run and the resources they consume.

Zilles and Flint object that many transactional systems are constructed such that a single large transaction may monopolize the atomicity resources such that no other transactions may commit. By opening a transaction, touching a large number of memory locations, and then never closing the transaction, a malicious application may deny service to all concurrent applications in a transaction system.

For this reason, it is important that there are no global resources required to complete a transaction. ApeX and the UTM hardware implementation achieve this end, but the LTM design uses a per-processor overflow table. If an LTM design is implemented with a snoopy bus for coherence traffic, overflows on one processor can impact the performance of all other processors on the bus. A directory-based coherence protocol (as we have described in this thesis) eliminates this problem. Hybrid schemes based on LTM also eliminate the problem, because an overflowing transaction can be aborted and retried in software, which requires no global resources.

Concerns about performance isolation are not limited to transaction systems. Transaction systems provide a solution not available to systems with lock-based concurrency, however: the offending transaction can be safely aborted at any point to allow the other transactions to progress.

8.2 Progress guarantees

Aborting troublesome transactions raises another potential pitfall: how do we guarantee that our system makes forward progress? Zilles and Flint [104] note that transaction systems are subject to an "all-or-nothing" problem: it's fine to abort a troublesome large transaction to allow other work to

[1] Grunwald and Ghiasi [42] call this a "microarchitectural denial of service" attack.

complete, but then we throw away any progress in that transaction. The operating system is forced either to allocate for a large transaction all the resources it requires, or to refuse to make any progress on the transaction. There is no middle ground.

This criticism applies to the LTM hardware scheme and other unvirtualized transaction implementation. In an LTM system, it is the programmer's responsibility to structure transactions such that the application is likely to complete. The operating system can deny progress when necessary to prevent priority inversion.

The UTM, hybrid, and software-only implementations do not suffer the same problem. UTM and software-only implementations can virtualize the transaction, as all transaction state resides in memory, thereby allowing the resources required to be paged in incrementally as needed. The hybrid scheme, even when built on an unvirtualized mechanism such as LTM, can abort and fail-over to a virtualized software system if sufficient resources are not available.

8.3 The semantic gap

In a vein of optimism, transactions are often casually said to be "compatible" with locks: transform your lock acquisition and release statements to begin-transaction and end-transaction, respectively, and your application is transformed. Some even claim that your application will suddenly be faster/more parallel and that some lingering locking bugs and race conditions will be cured by the change as well.

It is the latter part of the claim that draws first scrutiny. If you've "fixed" my race conditions, haven't you altered the semantics of my program? What other behaviors might have changed?

Blundell, Lewis, and Martin [17] describe the "semantic gap" opened between the locking and naively transactified code. They point out that programs with data races—even "benign" ones—may deadlock when con-

verted to use transactions, since the data race will never occur.[2] One may even wrap every access with a location-specific lock to "remove" the race (for some definitions) without altering the behavior of the locking code or the deadlock for the transactional version.

This concern is valid, and claims of automatic transactification should not be taken too lightly. Nevertheless, most "best-practices" concurrent code *will* behave as expected, and (unlike timing-dependent code with races) deadlocks make it obvious where things go wrong. Further, type systems are capable of detecting the deadlocks in transactified code and alerting the programmer of the problem.

Blundell, Lewis, and Martin also point out that some transaction implementations ignore "nontransactional" accesses—even if they access locations that are currently involved in a transaction. This oversight leads to additional alterations in the semantics of the code. In the implementations described in this thesis, we are careful to ensure that "nontransactional" code still executes as if each statement is its own individual transaction, what Blundell, Lewis, and Martin term ***strong atomicity***.

8.4 I/O mechanisms

To be useful, computing systems must be effectively connected to the outside world. "Reality" creates a discontinuity in the transactional model: real-world events cannot be rolled back in the same way as can changes to program state. This section presents four mechanisms to accommodate I/O in a transactional model. We can forbid I/O inside a transaction, via runtime check or type system. If we must have I/O, we can either create uninterruptible transactions to perform the I/O, or move the I/O to the

[2] For example, one thread can acquire lock L1 and then loop waiting for flag F1 to be set. Another thread can acquire lock L2, set flag F1 and loop waiting for F2. The first thread sees the update to F1, sets F2, and then releases L1. The second thread sees the update to F2 and releases L2. Neither thread will be able to complete if these locks are turned into atomic regions.

start or end of the transaction. The most general mechanism is to integrate programmer-specified compensating actions to achieve "rollback" of irrevocable actions within transactions.

8.4.1 Forbidding I/O

The most straightforward means to accommodate I/O in the transactional framework is to forbid it: I/O may only happen outside of a transaction. A runtime check or simple type system can be used to enforce this restriction.

A useful programmer technique in this model is to create a separate concurrent thread for I/O. A transaction can interact with a queue to request I/O, and the I/O thread dequeues requests and enqueues responses. This strategy works well for unidirectional communication. Since round-trip communication with the I/O thread cannot be accomplished within a single transaction,[3] transactions still must be broken between a request and reply. Thus, some forms of interaction cannot be accomplished atomically.

8.4.2 Mutual exclusion

Another alternative is to integrate mutual exclusion into the transaction model. Once we start an I/O activity within a transaction, the transaction becomes ***uninterruptible***: it may no longer be aborted and must be successfully executed through commit. Only a single uninterruptible transaction may execute at a given time (although other interruptible transactions may be concurrent). Effectively there is a single global mutual exclusion lock for I/O activity. Transactions attempting I/O while the lock is already held are either delayed or aborted.

This scheme is reasonable as long as I/O is infrequent in transactions. A single debugging print-statement, however, is sufficient to serialize a trans-

[3] Deadlock would result, since I can't send a message and then wait for a reply atomically—my communication partner won't see the sent message and know to reply until the transaction is committed.

action, and efforts to make the single global I/O lock more fine-grained may ultimately dissipate the gains in simplicity and concurrency afforded by transactions in the first place.

8.4.3 Postponing I/O

A hybrid approach attempts to anticipate or postpone I/O operations so that they run only at transaction start or end. Only the I/O operations then need to be run serially, and the remainder of the transaction may still execute concurrently. Input must be moved to the start of a transaction, and once the input has been consumed, the transaction must run uninterrupted. It may be aborted only if a push-back buffer can be constructed for the input, which is not always reasonable. Output is moved to the end of the transaction, but only the actual I/O must be performed uninterrupted: the transaction can still be aborted at any time prior to commit.

If output and input need to be interleaved, or the input occurs after an output and thus cannot be moved to the start of the transaction, uninterruptible transactions are still required. This approach thus works around the disadvantages of mutual exclusion in some cases, but a single misplaced debugging statement can still force serialization.

It is worth noting that modern interface hardware is often designed such that it works well with this approach. For example, a GPU or network card takes commands from or delivers input to a buffer. A single operation is sufficient to hand over a buffer to the card to commence I/O. This single I/O action may be made atomic with the transaction commit.

8.4.4 Integrating do/undo

The most sophisticated integration of I/O with transactions allows the programmer to specify "undo" code for I/O which cannot otherwise be rolled back. In the database community, undo code is referred to as a ***compensating transaction***.

Again, I/O is forbidden within "pure" transactions, but do/undo blocks may be nested within transactions. A do block executes uninterruptibly. If a transaction aborts before it reaches a do block, rollback occurs conventionally. If it aborts after it has executed a do block, then the undo block is executed uninterruptibly as part of transactional rollback. Mutual exclusion must still be used in portions of the transaction processing, but ideally the critical regions are short and infrequently invoked.

The do/undo behavior allows sophisticated exception processing: an undo block may emit a backspace to the terminal after do emits a character, or it may send an apology email after an email is sent irrevocably in a do block. The do/undo is invisible to clients outside the transaction. Sophisticated libraries can be built up using this mechanism. For example, disk I/O can be made transactional using file versioning and journalling.

The undo blocks may be difficult to program. The most straightforward implementation prevents the undo from accessing any transactional state, and the programmer must take special care if she is to maintain a consistent view of program state. A friendlier implementation presents a view of program state such that it appears that the undo block executes immediately after the do block in the same lexical environment (regardless of what ultimately aborted transactional code has executed in the interim). The programmer is then able to naturally write code such as the following:

```
String from = ..., to = ..., subject = ...;
do {
  sendEmail(from, to, subject);
} undo {
  sendApology(from, to, "Re: "+subject);
}
/* code here can modify from, to, subject */
/* before transaction commit */
```

Presenting a time-warped view to the undo block can be difficult or impossible, depending on the history mechanism involved. In particular, if the undo reads a new location previously untouched by the transaction, the value of this location at the (previous) time immediately following the do

might be unavailable. Presenting an inconsistent view of memory may be undesirable.

8.5 OS interactions

While transaction-style synchronization has been successfully used to structure interactions within an operating system [73], transactions crossing the boundary between operating system and application present additional challenges. Some operating system requests are either I/O operations or can be handled with the same mechanism used to handle I/O within transactions, as discussed in the previous section. Using memory allocation as an example of an OS request, transactions can be forbidden to allocate memory, required to take a lock, forced to preallocate all required memory,[4] or use a compensation mechanism to deallocate requested memory in case of abort.

Since the role of an operating system is to administer shared resources, special care must be taken that transactions involving operating system requests do not "contaminate" other processes. In particular, if data in kernel space is to be included in a transaction, the following challenges arise:

- If mutual exclusion may used to implement any part of the transaction semantics (as in the various I/O schemes above), then it may be possible to tie up the entire system (including unrelated processes) until a transaction touching kernel structures commits.
- If transaction state if to be tracked in the cache, the kernel address space must be reserved from the application memory map on architectures with virtually addressed caches.
- It may be desirable to include some loophole mechanism so that kernel data structures can be released from the transaction. Similarly, if

[4] A retry mechanism can be used to incrementally increase the preallocation until the transaction can successfully complete.

the operating system wishes to use transactions within itself, it may be desirable for these transactions to be independent from the application's invoking transaction. Motivating examples include fault handlers and the paging mechanism, which ought to be transparent to the application's transaction.

It is possible to handle these challenges simply, for example by forbidding OS calls within a transaction and aborting transactions if necessary on context switches or faults. Such an approach raises hurdles for the application programmer but simplifies the operating system's task considerably. In unvirtualized transaction implementations such as LTM, this approach also limits the maximum duration of a transaction to a single time slice, although the OS may be able to stretch a processes time slice when necessary.

A more sophisticated approach with explicit OS management of transactions may be able to provide better transparency of OS/transaction interactions for the application programmer and improved performance.

8.6 Recommendations for future work

Based on the discussion in this section, a virtualizable transaction mechanism is recommended: if LTM is implemented in hardware, a hybrid scheme like HyApeX should back it up in order to provide performance isolation and progress guarantees. The ApeX and UTM systems are already virtualizable.

A do/undo mechanism for transactions allows better management of critical regions where mutual exclusion must be integrated with the transaction mechanism to support I/O and certain OS interactions. All OS interactions can then be performed in do block so that mutual exclusion need not be extended across the OS boundary. The complexity of do/undo probably argues for ApeX or HyApeX instead of a pure hardware scheme.

Everything in the universe relates to [transactions], one way or another, given enough ingenuity on the part of the interpreter.

Principia Discordia (amended)

Chapter 9

Related work

Many researchers have been investigating transactional memory systems. This chapter discusses their related work and distinguishes the work of this thesis. In particular, this thesis is unique in presenting a hybrid hardware/software model-checked nonblocking object-oriented system that allows co-existence of nontransactional and transactional accesses to a dynamic set of object fields.

9.1 Nonblocking synchronization

Lamport [65] presented the first alternative to synchronization via mutual exclusion for a limited situation involving a single writer and multiple readers. Lamport's technique relies on reading guard elements in an order opposite to that in which they are written, guaranteeing that a consistent data snapshot can be recognized. The writer always completes its part of the algorithm in a constant number of steps, but readers are guaranteed to complete only in the absence of concurrent writes.

Herlihy [54] formalized *wait-free* implementations of concurrent data objects. A wait-free implementation guarantees that any process can complete any operation in a finite number of steps regardless of the activities of other processes. Lamport's algorithm is not wait-free, because readers can

be delayed indefinitely.

Massalin and Pu [73] introduced the term *lock-free* to describe algorithms with weaker progress guarantees. A lock-free implementation guarantees only that *some* process completes in a finite number of steps. Unlike a wait-free implementation, lock-freedom allows starvation. Since other simple techniques can be layered to prevent starvation (for example, exponential backoff), simple lock-free implementations are usually seen as worthwhile practical alternatives to more complex wait-free implementations.

An even weaker criterion, *obstruction-freedom*, was introduced by Herlihy, Luchangco, and Moir [56]. Obstruction-freedom only guarantees progress for threads executing in isolation; that is, although other threads may have partially completed operations, no other thread may take a step until the isolated thread completes. Obstruction-freedom not only allows starvation of a particular thread, it allows contention among threads to halt all progress in all threads indefinitely. External mechanisms are used to reduce contention (thus, achieve progress) including backoff, queueing, or timestamping.

I use the term *nonblocking* to describe generally any synchronization mechanism that doesn't rely on mutual exclusion or locking, including wait-free, lock-free, and obstruction-free implementations. I consider mainly lock-free algorithms.[1]

9.2 Efficiency

Herlihy [47, 54] presented the first *universal* method for wait-free concurrent implementation of an arbitrary sequential object. This original method was based on a *fetch-and-cons* primitive, which atomically places an item

[1] Some authors use "nonblocking" and "lock-free" as synonyms, usually meaning what we here call *lock-free*. Others exchange our definitions for "lock-free" and "nonblocking", using lock-free as a generic term and nonblocking to describe a specific class of implementations. As there is variation in the field, we choose to use the parallel construction *wait-free*, *lock-free*, and *obstruction-free* for our three specific progress criteria, and the dissimilar *nonblocking* for the general class.

on the head of a list and returns the list of items following it. Herlihy showed that all concurrent primitives capable of solving the n-process consensus problem—*universal* primitives—are powerful enough to implement *fetch-and-cons*. In Herlihy's method, every sequential operation is translated into two steps. In the first, *fetch-and-cons* is used to place the name and arguments of the operation to be performed at the head of a list, returning the other operations on the list. Since the state of a deterministic object is completely determined by the history of operations performed on it, applying the operations returned in order from last to first is sufficient to locally reconstruct the object state prior to the operation. The prior state can now be used to compute the result of the operation without requiring further synchronization with the other processes.

This first universal method was not very practical, a shortcoming which Herlihy soon addressed [48]. In addition, his revised universal method can be made lock-free, rather than wait-free, resulting in improved performance. In the lock-free version of this method, objects contain a shared variable holding a pointer to their current state. Processes begin by loading the current state pointer and then copying the referenced state to a local copy. The sequential operation is performed on the copy, and then if the object's shared state pointer is unchanged from its initial load, it is atomically swung to point at the updated state.

Herlihy called this the "small object protocol" because the object copying overhead is prohibitive unless the object is small enough to be copied efficiently (in, say, $O(1)$ time). He also presented a "large object protocol" which requires the programmer to manually break the object into small blocks, after which the small object protocol can be employed. (This trouble with large objects is common to many nonblocking implementations; a solution is presented in Chapter 6.)

Barnes [12] provided the first universal nonblocking implementation method that avoids object copying. He eliminates the need to store "old" object state in case of operation failure by having all threads cooperate to ap-

ply operations. For example, if the first processor begins an operation and then halts, another processor will complete the operation of the first before applying its own. Barnes proposes to accomplish the cooperation by creating a parallel state machine for each operation so that each thread can independently try to advance the machine from state to state and thus advance incomplete operations.[2] Although this strategy avoids copying state, the lock-step cooperative process is extremely cumbersome and does not appear to have ever been implemented. Furthermore, it does not protect against errors in the implementation of the operations, which could cause *every* thread to fail in turn as one by one they attempt to execute a buggy operation.

Alemany and Felten [1] identified two factors hindering the performance of nonblocking algorithms to date: resources wasted by operations that fail, and the cost of data copying. Unfortunately, they proceeded to "solve" these problems by ignoring short delays and failures and using operating system support to handle delays caused by context switches, page faults, and I/O operations. This strategy works in some situations, but it obviously suffers from a bootstrapping problem as the means to implement an operating system.

Although lock-free implementations are usually assumed to be more efficient that wait-free implementations, LaMarca [64] showed experimental evidence that Herlihy's simple wait-free protocol scales well on parallel machines. When more than about twenty threads are involved, the wait-free protocol becomes faster than Herlihy's lock-free small-object protocol [48], three OS-aided protocols of LaMarca [64] and Alemany and Felten [1], and a *test-and-Compare&Swap* spin-lock.

[2] It is interesting that Barnes' cooperative method for nonblocking situation plays out in a real-time system very similarly to priority inheritance for locking synchronization.

9.3 Transactional memory systems

Transactions are described in the database context by Gray [38], and Gray and Reuter [39] contains a thorough treatment of database issues. Hardware Transactional Memory (HTM) was first proposed by Knight [63], and Herlihy and Moss coined the term "transactional memory" and proposed HTM in the context of lock-free data structures [49, 55]. The BBN Pluribus [88, Ch. 23] provided transactions with an architectural limit on the size of a transaction. Experience with Pluribus showed that the headaches of programming correctly with such limits can be at least as challenging as using locks. The ***Oklahoma Update*** [89] is another variation on transactional operations with an architectural limit on the number of values in a transaction.

Transactional memory is sometimes described as an extension of Load-Linked/Store-Conditional [59] and other atomic instruction sequences. In fact, some CISC machines, such as the VAX [27], had complex atomic instructions such as enqueue and dequeue.

Of particular relevance are ***Speculative Lock Elision*** (SLE) [80] and ***Transactional Lock Removal*** (TLR) [81], which speculatively identify locks and use the cache to give the appearance of atomicity. SLE and TLR handle mutual exclusion through a standard programmer interface (locks), dynamically translating locks into transactional regions. My research thrust differs from theirs in that I hope to free programmers from the protocol complexities of locking, not just optimize existing practice. The quantitative results presented in Figure 7.5 of this thesis confirm their finding that transactional hardware can be more efficient than locks.

Martinez and Torrellas proposed ***Speculative Synchronization*** [72], which allows some threads to execute atomic regions of code speculatively, using locks, while guaranteeing forward progress by maintaining a nonspeculative thread. These techniques gain many of the performance advantages of transactional memory, but they still require new code to obey a locking

protocol to avoid deadlock.

The recent work on ***Transactional memory Coherence and Consistency*** (TCC) [43] is also relevant to our work. TCC uses a broadcast bus to implement the transaction protocols, performing all the writes of a particular transaction in one atomic bus operation. This strategy limits scalability, whereas both the UTM and LTM proposals in Chapter 7 can employ scalable cache-consistency protocols to implement transactions. TCC affirms the conclusion we draw from our own Figure 3.3: most transactions are small, but some are very large. TCC supports large transactions by locking the broadcast bus and stalling all other processors when any processor buffer overflows, whereas UTM and LTM allow overlapped execution of multiple large transactions with local overflow buffers. TCC is similar to LTM in that transactions are bound to processor state and cannot extend across page faults, timer interrupts, or thread migrations.

Several software transaction systems have been proposed [44, 50, 84, 87]. Some constrain the programmer and make transactions difficult to use. All have relatively high overheads, which make transactions unattractive for uniprocessor and small SMP systems. (Once the number of processors is large enough, the increased parallelism that can be provided by optimistic transactions may cancel out the performance penalty of their use.)

Software transactional memory was first proposed by Shavit and Touitou [87]. Their system requires that all input and output locations touched by a transaction be known in advance, which limits its application. It performs at least 10 fetches and 4 stores per location accessed (not counting the loads and stores directly required by the computation). The benchmarks presented were run on between 10 and 64 processors.

Rudys and Wallach [84] proposed a copying-based transaction system to allow rollback of hostile codelets. They show an order of magnitude slowdown for field and array accesses, and 6x to 23x slowdown on their benchmarks.

Herlihy, Luchangco, Moss, and Scherer [50] present a scheme that allows

transactions to touch a dynamic set of memory locations. The user still must explicitly *open* every object touched, however, before it can be used in a transaction. This implementation is based on object copying, and so it has poor performance for large objects and arrays. Not including work necessary to copy objects involved in writes, they require $O(R(R+W))$ work to open R objects for reading and W objects for writing, which may be quadratic in the number of objects involved in the transaction. A list insertion benchmark that they present shows 9x slowdown over a locking scheme, although they beat the locking implementation when more than 5-10 processors are active. They present benchmark data with up to 576 threads on 72 processors.

Harris and Fraser [44] built a software transaction system on a flat word-oriented transactional memory abstraction, roughly similar to simulating Herlihy's original hardware transactional memory proposal in software. The flat memory abstraction avoids problems with large objects. Performing m memory operations touching l distinct locations costs at least $m+l$ extra reads and $l+1$ CAS operations, in addition to the reads and writes required by the computation. They appear to execute about twice as slowly as a locking implementation on some microbenchmarks. They benchmark on a 4-processor as well as a 106-processor machine. Their crossover point (at which the blocking overhead of locks matches the software transaction overhead) is around 4 processors. Harris and Fraser do not address the problem of concurrent nontransactional operations on locations involved in a transaction, however. Java synchronization allows such concurrent operations, with semantics given by the Java memory model [68–70]. The mechanisms presented in Chapter 3 support these operations safely.

Herlihy and Moss [49] suggested that small transactions might be handled in cache with overflows handled by software. These software overflows must interact with the transactional hardware in the same way that the hardware interacts with itself, however. Section 7.4 presented just such a system.

9.4 Language-level approaches to synchronization

Our work on integrating transactions into the Java programming language is related to prior work on integrating synchronization mechanisms for multiprogramming, and in particular, to prior work on synchronization in an object-oriented framework.

The Emerald system [16, 61] introduced *monitored objects* for synchronization. Emerald code to implement a simple directory object is shown in Figure 9.1. Each object is associated with a Hoare-style monitor, which provides mutual exclusion and process signaling. Each Emerald object is divided into a monitored part and a non-monitored part. Variables declared in the monitored part are shared, and access to them from methods in the non-monitored part is prohibited—although non-monitored methods may call monitored methods to effect the access. Methods in the monitored part acquire the monitor lock associated with the receiver object before entry and release it on exit, providing for mutual exclusion and safe update of the shared variables. Monitored objects naturally integrate synchronization into the object model.

Unlike Emerald monitored objects, where methods can only acquire the monitor of their receiver and where restricted access to shared variables is enforced by the compiler, Java implements a loose variant where any monitor may be explicitly acquired and no shared variable protection exists. As a default, however, Java methods declared with the `synchronized` keyword behave like Emerald monitored methods, ensuring that the monitor lock of their receiver is held during execution.

Java's synchronization primitives arguably allow for more efficient concurrent code than Emerald's—for example, Java objects can use multiple locks to protect disjoint sets of fields, and coarse-grain locks can be used which protect multiple objects—but Java is also more prone to programmer error. Even Emerald's restrictive monitored objects, however, are not sufficient to prevent data races. As a simple example, imagine that an object

```
const myDirectory == object oneEntryDirectory
  export Store, Lookup
  monitor
    var name : String
    var AnObject : Any

    operation Store [ n : String, o : Any ]
      name ← n
      AnObject ← o
    end Store
    function Lookup [ n : String ] → [ o : Any ]
      if n = name
        then o ← AnObject
        else o ← nil
      end if
    end Lookup

    initially
      name ← nil
      AnObject ← nil
    end initially

  end monitor
end oneEntryDirectory
```

Figure 9.1: A directory object in Emerald [16], illustrating the use of monitor synchronization.

```
class Account {

  int balance = 0;

  atomic int deposit(int amt) {
    int t = this.balance;
    t = t + amt;
    this.balance = t;
    return t;
  }

  atomic int readBalance() {
    return this.balance;
  }

  atomic int withdraw(int amt) {
    int t = this.balance;
    t = t - amt;
    this.balance = t;
    return t;
  }

}
```

Figure 9.2: A simple bank account object, adapted from [29], illustrating the use of the `atomic` modifier.

provides two monitored methods `read` and `write` which access a shared variable. Non-monitored code can call `read`, increment the value returned, and then call `write`, creating a classic race-condition scenario. The atomicity of the parts is not sufficient to guarantee atomicity of the whole [29].

This example suggests that a better model for synchronization in object-oriented systems is *atomicity*. Figure 9.2 shows Java extended with an `atomic` keyword to implement an object representing a bank account. Rather than explicitly synchronizing on locks, I simply require that the methods marked `atomic` execute atomically with respect to other threads in the system. To execute atomically, every execution of the program must compute the same result as some execution where all atomic methods were run *in isolation* at a certain point in time[3] between their invocation and return. Atomic methods invoked directly or indirectly from an atomic method are subsumed by it: if the outermost method appears atomic, then by definition all inner method invocations also appear atomic. Flanagan and Qadeer [29] provide a more formal semantics. Atomic methods can be analyzed using sequential reasoning techniques, which significantly simplifies reasoning about program correctness.

Atomic methods can be implemented using locks. A simple if inefficient implementation would simply acquire a single global lock during the execution of every atomic method. Flanagan and Qadeer [29] present a more sophisticated technique to prove that a given implementation using standard Java monitors correctly guarantees method atomicity.

The transaction implementations presented in this thesis all use non-blocking synchronization to implement atomic methods.

[3] The *linearization point*.

"Begin at the beginning," the King said, very gravely, "and go on till you come to the end: then stop."

Lewis Carroll, *Alice's Adventures in Wonderland*

Chapter 10

Conclusion

There is no escape: parallel systems are in our future. Programming them does not have to be as fraught as it is presently, however. I believe that transactions provide a programmer-friendly model of concurrency which eliminates many potential pitfalls of our current locking-based methodologies.

In this thesis I have demonstrated that it is possible to implement an efficient strongly atomic software transaction system, ApeX. I have shown that nonblocking transactions can be used in applications beyond synchronization, including fault tolerance and backtracking search. I have presented implementation details to address the practical problems of building such a system.

I have argued the transactions should not be subject to limits on size or duration. I have presented both software and hardware implementations free of such restrictions.

Since the low overhead of the ApeX software system allows it to be profitably combined with a hardware transaction system, I have shown how to build a hybrid, HyApeX, which yields fast execution of short and small transactions while allowing fallback to software for large or complicated transactions.

This thesis presents several designs for efficient transaction systems that enable the transactional programming model. The ApeX software-only sys-

tem runs on current hardware, and LTM and UTM indicate possible directions for future hardware. There are challenges and design decisions remaining, however. How should I/O be handled? What are the proper semantics for nested transactions? What loophole mechanisms are necessary to allow information about a transaction's progress to escape?

I believe hybrid systems such as HyApeX offer the best answer to these challenges. They combine the inherent speed of hardware systems with the flexibility of software, allowing novel solutions to be attempted without requiring that design decisions be cast in silicon. The flag-based ApeX software transaction system described in this thesis imposes low overhead, allowing transactional programming to get off the ground without hardware support in the near term, while later supporting the development of new transactional models and methodologies as part of a hybrid system.

Designing correct transaction systems is not easy, however. In the appendix you will find a Promela model of my software transaction system. Automated verification was essential when designing and debugging the system, uncovering via exhaustive enumeration race conditions much too subtle for me to discover by other means. I believe any credible transaction system must be buttressed by formal verification.

Essentially, all models are
wrong, but some are useful.

George E. P. Box, *Empirical Model-Building and Response Surfaces*

Appendix A

Model-checking the software implementation

My work on both software and hardware transaction systems has reiterated the difficulty of creating correct implementations of concurrent and fault-tolerant constructs. Automatic model checking is a prerequisite to achieving confidence in the design and implementation. Versions of the software transaction system have been modeled in Promela using SPIN 4.1.0 and verified on an SGI 64-processor MIPS machine with 16G of main memory.

Sequences of transactional and nontransactional load and store operations were checked using two concurrent processes, and all possible interleavings were found to produce results consistent with the semantic atomicity of the transactions. Several test scripts were run against the model using separate processors of the verification machine (SPIN cannot otherwise exploit SMP). Some representative costs include:

- testing two concurrent `writeT` operations (including "false flag" conditions) against a single object required 3.8×10^6 states and 170M memory;

- testing sequences of transactional and nontransactional reads and writes against two objects (checking that all views of the two objects were

consistent) required 4.6×10^6 states and 207M memory; and

- testing a pair of concurrent increments at values bracketing the FLAG value to 99.8% coverage of the state space required 7.6×10^7 states and 4.3G memory. Simultaneously model-checking a range of values caused the state space explosion in this case.

SPIN's unreachable code reporting was used to ensure that my test cases exercised every code path, although exercising all code paths does not guarantee that every interesting interaction was checked.

In the process one bug in SPIN was discovered,[1] and several subtle race conditions in the model were discovered and corrected. These race conditions included several modeling artifacts. In particular, I was extremely aggressive about reference-counting and deallocating objects in order to control the size of the state space, which proved difficult to do correctly. I also discovered some subtle-but-legitimate race conditions in the transactions algorithm. For example:

- A race allowed conflicting readers to be created while a writer was inside `ensureWriter` creating a new version object.
- Allowing already-committed version objects to be mutated when `writeT` or `writeNT` was asked to store a "false flag" produced races between `ensureWriter` and `copyBackField`. The code that was expected to manage these races had unexpected corner cases.
- Using a bitmask to provide per-field granularity to the list of readers proved unmanageable as there were three-way races among the bitmask, the readers list, and the version tree.

In addition, the model-in-progress proved a valuable design tool, as portions of the algorithm could readily be mechanically checked to validate (or discredit) the designer's reasoning about the concurrent system. Humans do

[1] Breadth-first search of atomic regions was performed incorrectly in SPIN 4.0.7; after I reported this bug, a fix was incorporated in SPIN 4.1.0.

not excel at exhaustive state-space exploration. In fact, after discovering the these (and other) bugs and fixing them in the model, I replaced my original ad-hoc implementation of ApeX with new code which more closely followed the model.

Spin is not particularly suited to checking models with dynamic allocation and deallocation. In particular, it considers the location of objects as part of the state space, and allocating object A before object B produces a different state than if object B were allocated first. This behavior artificially enlarges the state space. A great deal of effort was expended tweaking the model to approach a canonical allocation ordering. The `#ifdef REFCOUNT` bracketed portions of the model are evidence of this; they can be safely ignored when studying the semantics of the model, but are necessary for mechanical verification. A better solution to this problem would allow larger model instances to be checked.

A.1 Promela primer

A concise Promela reference is available at http://spinroot.com/spin/Man/Quick.html. Here, I attempt to summarize just enough of the language to allow the model I've presented in this thesis to be understood.

Promela syntax is C-like, with the same lexical and commenting conventions. Statements are separated by either a semicolon, or, equivalently, an arrow. The arrow is typically used to separate a guard expression from the statements it is guarding.

The program counter moves past a statement only if the statement is *enabled*. Most statements, including assignments, are always enabled. A statement consisting only of an expression is enabled if and only if the expression is true (non-zero). Our model uses three basic Promela statements: selection, repetition, and `atomic`.

The selection statement

```
if
:: guard -> statements
...
:: guard -> statements
fi
```

selects one among its options and executes it. An option can be selected if and only if its first statement (the guard) is enabled. The special guard `else` is enabled if and only if all other guards are not.

The repetition statement

```
do
:: statements
...
:: statements
od
```

is similar: one among its enabled statements is selected and executed, and then the process is repeated (with a different statement possibly being selected each time) until control is explicitly transferred out of the loop with a `break` or `goto`.

Finally,

```
atomic { statements }
```

executes the given statements in one indivisible step. For the purposes of this model, a `d_step` block

```
d_step { statements }
```

is functionally identical. Outside `atomic` or `d_step` blocks, Promela allows interleaving before and after every statement, but statements are indivisible.

Functions as specified in this model are similar to C macros: every parameter is potentially both an input *and* an output. Calls to functions with names starting with `move` are simple assignments, but I've turned them into macros so that reference counting can be performed.

A.2 Spin model for software transaction system

The complete SPIN 4.1.0 model for the FLEX software transaction system is presented here. It may also be downloaded from http://flex-compiler.csail.mit.edu/Harpoon/swx.pml.

```
/*************************************************************************
 * Detailed model of software transaction code.
 * Checking for safety and correctness properties.  Not too worried about
 * liveness.
 *
 * (C) 2006 C. Scott Ananian <cananian@alumni.princeton.edu>
 *************************************************************************/

/* CURRENT ISSUES:
 *  none known.
 */
/* MINOR ISSUES:
 * 1) use smaller values for FLAG and NIL to save state space?
 */

/* Should use Spin 4.1.0, for correct nested-atomic behavior. */

#define REFCOUNT

#define NUM_OBJS 2
#define NUM_VERSIONS 6 /* each obj: committed and waiting version, plus nonce
                        * plus addition nonce for NT copyback in test3 */
#define NUM_READERS 4 /* both 'read' trans reading both objs */
#define NUM_TRANS 5 /* two 'real' TIDs, plus 2 outstanding TIDs for
                     * writeNT(FLAG) [test3], plus perma-aborted TID. */
#define NUM_FIELDS 2

#define NIL 255 /* special value to represent 'alloc impossible', etc. */
#define FLAG 202 /* special value to represent 'not here' */

typedef Object {
  byte version;
  byte readerList; /* we do LL and CAS operations on this field */
  pid fieldLock[NUM_FIELDS]; /* we do LL operations on fields */
  byte field[NUM_FIELDS];
};
typedef VersiOn { /* 'Version' misspelled because spin #define's it. */
  byte owner;
  byte next;
  byte field[NUM_FIELDS];
#ifdef REFCOUNT
  byte ref; /* reference count */
#endif /* REFCOUNT */
};
```

```
typedef ReaderList {
  byte transid;
  byte next;
#ifdef REFCOUNT
  byte ref; /* reference count */
#endif /* REFCOUNT */
};
mtype = { waiting, committed, aborted };
typedef TransID {
  mtype status;
#ifdef REFCOUNT
  byte ref; /* reference count */
#endif /* REFCOUNT */
};
Object object[NUM_OBJS];
VersiOn version[NUM_VERSIONS];
ReaderList readerlist[NUM_READERS];
TransID transid[NUM_TRANS];
byte aborted_tid; /* global variable; 'perma-aborted' */

/* --------------------------- alloc.pml ----------------------- */
mtype = { request, return };

inline manager(NUM_ITEMS, allocchan) {
  chan pool = [NUM_ITEMS] of { byte };
  chan client;
  byte nodenum;
  /* fill up the pool with node identifiers */
  d_step {
    i=0;
    do
    :: i<NUM_ITEMS -> pool!!i; i++
    :: else -> break
    od;
  }
end:
  do
  :: allocchan?request(client,_) ->
     if
     :: empty(pool) -> assert(0); client!NIL /* deny */
     :: nempty(pool) ->
        pool?nodenum;
        client!nodenum;
        nodenum=0
     fi
  :: allocchan?return(client,nodenum) ->
     pool!!nodenum; /* sorted, to reduce state space */
     nodenum=0
  od
}

chan allocObjectChan = [0] of { mtype, chan, byte };
```

```
active proctype ObjectManager() {
  atomic {
    byte i;
    manager(NUM_OBJS, allocObjectChan)
  }
}
chan allocVersionChan = [0] of { mtype, chan, byte };
active proctype VersionManager() {
  atomic {
    byte i=0;
    d_step {
      do
      :: i<NUM_VERSIONS ->
         version[i].owner=NIL; version[i].next=NIL;
         version[i].field[0]=FLAG; version[i].field[1]=FLAG;
         assert(NUM_FIELDS==2);
         i++
      :: else -> break
      od;
    }
    manager(NUM_VERSIONS, allocVersionChan)
  }
}
chan allocReaderListChan = [0] of { mtype, chan, byte };
active proctype ReaderListManager() {
  atomic {
    byte i=0;
    d_step {
      do
      :: i<NUM_READERS ->
         readerlist[i].transid=NIL; readerlist[i].next=NIL;
         i++
      :: else -> break
      od;
    }
    manager(NUM_READERS, allocReaderListChan)
  }
}
chan allocTransIDChan = [0] of { mtype, chan, byte };
active proctype TransIDManager() {
  atomic {
    byte i=0;
    d_step {
      do
      :: i<NUM_TRANS -> transid[i].status=waiting; i++
      :: else -> break
      od;
    }
    manager(NUM_TRANS, allocTransIDChan)
  }
}
```

```
inline alloc(allocchan, retval, result) {
  result = NIL;
  do
  :: result != NIL -> break
  :: else -> allocchan!request(retval,0) ; retval ? result
  od;
  skip /* target of break. */
}
inline free(allocchan, retval, result) {
  allocchan!return(retval,result)
}
inline allocObject(retval, result) {
  atomic {
    alloc(allocObjectChan, retval, result);
    d_step {
      object[result].version = NIL;
      object[result].readerList = NIL;
      object[result].field[0] = 0;
      object[result].field[1] = 0;
      object[result].fieldLock[0] = _thread_id;
      object[result].fieldLock[1] = _thread_id;
      assert(NUM_FIELDS==2); /* else ought to initialize more fields */
    }
  }
}
inline allocTransID(retval, result) {
  atomic {
    alloc(allocTransIDChan, retval, result);
    d_step {
      transid[result].status = waiting;
#ifdef REFCOUNT
      transid[result].ref = 1;
#endif /* REFCOUNT */
    }
  }
}
inline moveTransID(dst, src) {
  atomic {
#ifdef REFCOUNT
    _free = NIL;
    if
    :: (src!=NIL) ->
       transid[src].ref++
    :: else
    fi;
    if
    :: (dst!=NIL) ->
       transid[dst].ref--;
       if
       :: (transid[dst].ref==0) -> _free=dst
       :: else
       fi
```

```
    :: else
    fi;
#endif /* REFCOUNT */
    dst = src;
#ifdef REFCOUNT
    /* receive must be last, as it breaks atomicity. */
    if
    :: (_free!=NIL) -> run freeTransID(_free, _retval); _free=NIL; _retval?_
    :: else
    fi
#endif /* REFCOUNT */
  }
}
proctype freeTransID(byte result; chan retval) {
  chan _retval = [0] of { byte };
  atomic {
#ifdef REFCOUNT
    assert(transid[result].ref==0);
#endif /* REFCOUNT */
    transid[result].status = waiting;
    free(allocTransIDChan, _retval, result)
    retval!0; /* done */
  }
}
inline allocVersion(retval, result, a_transid, tail) {
  atomic {
    alloc(allocVersionChan, retval, result);
    d_step {
#ifdef REFCOUNT
      if
      :: (a_transid!=NIL) -> transid[a_transid].ref++;
      :: else
      fi;
      if
      :: (tail!=NIL) -> version[tail].ref++;
      :: else
      fi;
      version[result].ref = 1;
#endif /* REFCOUNT */
      version[result].owner = a_transid;
      version[result].next = tail;
      version[result].field[0] = FLAG;
      version[result].field[1] = FLAG;
      assert(NUM_FIELDS==2); /* else ought to initialize more fields */
    }
  }
}
inline moveVersion(dst, src) {
  atomic {
#ifdef REFCOUNT
    _free = NIL;
    if
```

```
    :: (src!=NIL) ->
       version[src].ref++
    :: else
    fi;
    if
    :: (dst!=NIL) ->
       version[dst].ref--;
       if
       :: (version[dst].ref==0) -> _free=dst
       :: else
       fi
    :: else
    fi;
#endif /* REFCOUNT */
    dst = src;
#ifdef REFCOUNT
    /* receive must be last, as it breaks atomicity. */
    if
    :: (_free!=NIL) -> run freeVersion(_free, _retval); _free=NIL; _retval?_
    :: else
    fi
#endif /* REFCOUNT */
  }
}
proctype freeVersion(byte result; chan retval) {
  chan _retval = [0] of { byte };
  byte _free;
  atomic { /* zero out version structure */
#ifdef REFCOUNT
    assert(version[result].ref==0);
#endif /* REFCOUNT */
    moveTransID(version[result].owner, NIL);
    moveVersion(version[result].next, NIL);
    version[result].field[0] = FLAG;
    version[result].field[1] = FLAG;
    assert(NUM_FIELDS==2);
    free(allocVersionChan, _retval, result)
    retval!0; /* done */
  }
}

inline allocReaderList(retval, result, head, tail) {
  atomic {
    assert(head!=NIL);
    alloc(allocReaderListChan, retval, result);
    d_step {
#ifdef REFCOUNT
      readerlist[result].ref = 1;
      transid[head].ref++;
      if
      :: (tail!=NIL) -> readerlist[tail].ref++
      :: else
```

```
      fi;
#endif /* REFCOUNT */
      readerlist[result].transid = head;
      readerlist[result].next = tail;
    }
  }
}
inline moveReaderList(dst, src) {
  atomic {
#ifdef REFCOUNT
    _free = NIL;
    if
    :: (src!=NIL) ->
       readerlist[src].ref++
    :: else
    fi;
    if
    :: (dst!=NIL) ->
       readerlist[dst].ref--;
       if
       :: (readerlist[dst].ref==0) -> _free=dst
       :: else
       fi
    :: else
    fi;
#endif /* REFCOUNT */
    dst = src;
#ifdef REFCOUNT
    /* receive must be last, as it breaks atomicity. */
    if
    :: (_free!=NIL) -> run freeReaderList(_free, _retval); _free=NIL; _retval?_
    :: else
    fi
#endif
  }
}
proctype freeReaderList(byte result; chan retval) {
  chan _retval = [0] of { byte };
  byte _free;
  atomic {
#ifdef REFCOUNT
    assert(readerlist[result].ref==0);
#endif /* REFCOUNT */
    moveTransID(readerlist[result].transid, NIL);
    moveReaderList(readerlist[result].next, NIL);
    free(allocReaderListChan, _retval, result)
    retval!0; /* done */
  }
}

/* -------------------------- atomic.pml ----------------------- */
inline DCAS(loc1, oval1, nval1, loc2, oval2, nval2, st) {
```

```
  d_step {
    if
    :: (loc1==oval1) && (loc2==oval2) ->
       loc1=nval1;
       loc2=nval2;
       st=true
    :: else ->
       st=false
    fi
  }
}
inline CAS(loc1, oval1, nval1, st) {
  d_step {
    if
    :: (loc1==oval1) ->
       loc1=nval1;
       st=true
    :: else ->
       st=false
    fi
  }
}
inline CAS_Version(loc1, oval1, nval1, st) {
  atomic {
    _free = NIL;
    if
    :: (loc1==oval1) ->
#ifdef REFCOUNT
       if
       :: (nval1!=NIL) -> version[nval1].ref++;
       :: else
       fi;
       if
       :: (oval1!=NIL) -> version[oval1].ref--;
          if
          :: (version[oval1].ref==0) -> _free = oval1
          :: else
          fi
       :: else
       fi;
#endif /* REFCOUNT */
       loc1=nval1;
       st=true
    :: else ->
       st=false
    fi;
#ifdef REFCOUNT
    /* receive must be last, as it breaks atomicity. */
    if
    :: (_free!=NIL) -> run freeVersion(_free, _retval); _free=NIL; _retval?_
    :: else
    fi
```

```
#endif /* REFCOUNT */
  }
}
inline CAS_Reader(loc1, oval1, nval1, st) {
  atomic {
    /* save oval1, as it could change as soon as we leave the d_step */
    _free = NIL;
    if
    :: (loc1==oval1) ->
#ifdef REFCOUNT
       if
       :: (nval1!=NIL) -> readerlist[nval1].ref++;
       :: else
       fi;
       if
       :: (oval1!=NIL) -> readerlist[oval1].ref--;
          if
          :: (readerlist[oval1].ref==0) -> _free = oval1
          :: else
          fi
       :: else
       fi;
#endif /* REFCOUNT */
       loc1=nval1;
       st=true
    :: else ->
       st=false
    fi;
#ifdef REFCOUNT
    /* receive must be last, as it breaks atomicity. */
    if
    :: (_free!=NIL) -> run freeReaderList(_free, _retval); _free=NIL; _retval?_
    :: else
    fi
#endif /* REFCOUNT */
  }
}

/* ---------------- end atomic.pml ------------------ */

mtype = { kill_writers, kill_all };
mtype = { success, saw_race, saw_race_cleanup, false_flag };

inline tryToAbort(t) {
  assert(t!=NIL);
  CAS(transid[t].status, waiting, aborted, _);
  assert(transid[t].status==aborted || transid[t].status==committed)
}
inline tryToCommit(t) {
  assert(t!=NIL);
  CAS(transid[t].status, waiting, committed, _);
  assert(transid[t].status==aborted || transid[t].status==committed)
```

```
}
inline copyBackField(o, f, mode, st) {
  _nonceV=NIL; _ver = NIL; _r = NIL; st = success;
  /* try to abort each version.  when abort fails, we've got a
   * committed version. */
  do
  :: moveVersion(_ver, object[o].version);
     if
     :: (_ver==NIL) ->
        st = saw_race; break /* someone's done the copyback for us */
     :: else
     fi;
      /* move owner to local var to avoid races (owner set to NIL behind
       * our back) */
     _tmp_tid=NIL;
     moveTransID(_tmp_tid, version[_ver].owner);
     if
     :: (_tmp_tid==NIL) ->
        break; /* found a committed version */
     :: else
     fi;
     tryToAbort(_tmp_tid);
     if
     :: (transid[_tmp_tid].status==committed) ->
        moveTransID(_tmp_tid, NIL);
        moveTransID(version[_ver].owner, NIL); /* opportunistic free */
        moveVersion(version[_ver].next, NIL); /* opportunistic free */
        break /* found a committed version */
     :: else
     fi;
     /* link out an aborted version */
     assert(transid[_tmp_tid].status==aborted);
     CAS_Version(object[o].version, _ver, version[_ver].next, _);
     moveTransID(_tmp_tid, NIL);
  od;
  /* okay, link in our nonce.  this will prevent others from doing the
   * copyback. */
  if
  :: (st==success) ->
     assert (_ver!=NIL);
     allocVersion(_retval, _nonceV, aborted_tid, _ver);
     CAS_Version(object[o].version, _ver, _nonceV, _cas_stat);
     if
     :: (!_cas_stat) ->
        st = saw_race_cleanup
     :: else
     fi
  :: else
  fi;
  /* check that no one's beaten us to the copy back */
  if
  :: (st==success) ->
```

```
   if
   :: (object[o].field[f]==FLAG) ->
      _val = version[_ver].field[f];
      if
      :: (_val==FLAG) -> /* false flag... */
         st = false_flag /* ...no copy back needed */
      :: else -> /* not a false flag */
         d_step { /* could be DCAS */
           if
           :: (object[o].version == _nonceV) ->
              object[o].fieldLock[f] = _thread_id;
              object[o].field[f] = _val;
           :: else /* hmm, fail.  Must retry. */
              st = saw_race_cleanup /* need to clean up nonce */
           fi
         }
      fi
   :: else /* may arrive here because of readT, which doesn't set _val=FLAG*/
      st = saw_race_cleanup /* need to clean up nonce */
   fi
:: else /* !success */
fi;

/* always kill readers, whether successful or not.  This ensures that we
 * make progress if called from writeNT after a readNT sets readerList
 * non-null without changing FLAG to _val (see immediately above; st will
 * equal saw_race_cleanup in this scenario). */
if
:: (mode == kill_all) ->
   do /* kill all readers */
   :: moveReaderList(_r, object[o].readerList);
      if
      :: (_r==NIL) -> break
      :: else
      fi;
      tryToAbort(readerlist[_r].transid);
      /* link out this reader */
      CAS_Reader(object[o].readerList, _r, readerlist[_r].next, _);
   od;
:: else /* no more killing needed. */
fi;

/* finally, clean up our mess. */
moveVersion(_ver, NIL);
if
:: (st == saw_race_cleanup || st == success || st == false_flag) ->
   CAS_Version(object[o].version, _nonceV, version[_nonceV].next, _);
   moveVersion(_nonceV, NIL);
   if
   :: (st==saw_race_cleanup) -> st=saw_race
   :: else
   fi
```

```
  :: else
  fi;
  /* done */
  assert(_nonceV==NIL);
}

inline readNT(o, f, v) {
  do
  :: v = object[o].field[f];
     if
     :: (v!=FLAG) -> break /* done! */
     :: else
     fi;
     copyBackField(o, f, kill_writers, _st);
     if
     :: (_st==false_flag) ->
        v = FLAG;
        break
     :: else
     fi
  od
}
inline writeNT(o, f, nval) {
  if
  :: (nval != FLAG) ->
     do
     ::
        atomic {
          if /* this is a LL(readerList)/SC(field) */
          :: (object[o].readerList == NIL) ->
             object[o].fieldLock[f] = _thread_id;
             object[o].field[f] = nval;
             break /* success! */
          :: else
          fi
        }
        /* unsuccessful SC */
        copyBackField(o, f, kill_all, _st)
        /* ignore return status */
     od
  :: else -> /* create false flag */
     /* implement this as a short *transactional* write.  this may be slow,
      * but it greatly reduces the race conditions we have to think about. */
     do
     :: allocTransID(_retval, _writeTID);
        ensureWriter(_writeTID, o, _tmp_ver);
        checkWriteField(o, f);
        writeT(_tmp_ver, f, nval);
        tryToCommit(_writeTID);
        moveVersion(_tmp_ver, NIL);
        if
        :: (transid[_writeTID].status==committed) ->
```

```
            moveTransID(_writeTID, NIL);
            break /* success! */
         :: else ->/* try again */
            moveTransID(_writeTID, NIL)
         fi
      od
  fi;
}
inline readT(tid, o, f, ver, result) {
  do
  ::
     /* we should always either be on the readerlist or aborted here */
     atomic { /* complicated assertion; evaluate atomically */
       if
       :: (transid[tid].status == aborted) -> skip /* okay then */
       :: else ->
          assert (transid[tid].status == waiting);
          _r = object[o].readerList;
          do
          :: (_r==NIL || readerlist[_r].transid==tid) -> break
          :: else -> _r = readerlist[_r].next
          od;
          assert (_r!=NIL); /* we're on the list */
          _r = NIL /* back to normal */
       fi
     }
     /* okay, sanity checking done -- now let's get to work! */
     result = object[o].field[f];
     if
     :: (result==FLAG) ->
        if
        :: (ver!=NIL) ->
           result = version[ver].field[f];
           break /* done! */
        :: else ->
           findVersion(tid, o, ver);
           if
           :: (ver==NIL) -> /* use value from committed version */
              assert (_r!=NIL);
              result = version[_r].field[f]; /* false flag? */
              moveVersion(_r, NIL);
              break /* done */
           :: else /* try, try, again */
           fi
        fi
     :: else -> break /* done! */
     fi
  od
}
inline writeT(ver, f, nval) {
  /* easy enough: */
  version[ver].field[f] = nval;
```

```
}

/* make sure 'tid' is on reader list. */
inline ensureReaderList(tid, o) {
  /* add yourself to readers list. */
  _rr = NIL; _r = NIL;
  do
  :: moveReaderList(_rr, object[o].readerList); /* first_reader */
     moveReaderList(_r, _rr);
     do
     :: (_r==NIL) ->
        break /* not on the list */
     :: (_r!=NIL && readerlist[_r].transid==tid) ->
        break /* on the list */
     :: else ->
        /* opportunistic free? */
        if
        :: (_r==_rr && transid[readerlist[_r].transid].status != waiting) ->
           CAS_Reader(object[o].readerList, _r, readerlist[_r].next, _)
           if
           :: (_cas_stat) -> moveReaderList(_rr, readerlist[_r].next)
           :: else
           fi
        :: else
        fi;
        /* keep looking */
        moveReaderList(_r, readerlist[_r].next)
     od;
     if
     :: (_r!=NIL) ->
        break /* on the list; we're done! */
     :: else ->
        /* try to put ourselves on the list. */
        assert(tid!=NIL && _r==NIL);
        allocReaderList(_retval, _r, tid, _rr);
        CAS_Reader(object[o].readerList, _rr, _r, _cas_stat);
        if
        :: (_cas_stat) ->
           break /* we're on the list */
        :: else
        fi
        /* failed to put ourselves on the list, retry. */
     fi
  od;
  moveReaderList(_rr, NIL);
  moveReaderList(_r, NIL);
  /* done. */
}

/* look for a version read/writable by 'tid'.  Returns:
 *   ver!=NIL -- ver is the 'waiting' version for 'tid'. _r == NIL.
 *   ver==NIL, _r != NIL -- _r is the first committed version in the chain.
```

```
 *   ver==NIL, _r == NIL -- there are no commited versions for this object
 *                           (i.e. object[o].version==NIL)
 */
inline findVersion(tid, o, ver) {
  assert(tid!=NIL);
  _r = NIL; ver = NIL; _tmp_tid=NIL;
  do
  :: moveVersion(_r, object[o].version);
     if
     :: (_r==NIL) -> break /* no versions. */
     :: else
     fi;
     moveTransID(_tmp_tid, version[_r].owner);/*use local copy to avoid races*/
     if
     :: (_tmp_tid==tid) ->
        ver = _r; /* found a version: ourself! */
        _r = NIL; /* transfer owner of the reference to ver, without ++/-- */
        break
     :: (_tmp_tid==NIL) ->
        /* perma-committed version.  Return in _r. */
        moveVersion(version[_r].next, NIL); /* opportunistic free */
        break
     :: else -> /* strange version.  try to kill it. */
        /* ! could double-check that our own transid is not aborted here. */
        tryToAbort(_tmp_tid);
        if
        :: (transid[_tmp_tid].status==committed) ->
           /* committed version.  Return this in _r. */
           moveTransID(version[_r].owner, NIL); /* opportunistic free */
           moveVersion(version[_r].next, NIL); /* opportunistic free */
           break /* no need to look further. */
        :: else ->
           assert (transid[_tmp_tid].status==aborted);
           /* unlink this useless version */
           CAS_Version(object[o].version, _r, version[_r].next, _)
           /* repeat */
        fi
     fi
  od;
  moveTransID(_tmp_tid, NIL); /* free tmp transid copy */
  assert (ver!=NIL -> _r == NIL : 1)
}

inline ensureReader(tid, o, ver) {
  assert(tid!=NIL);
  /* make sure we're on the readerlist */
  ensureReaderList(tid, o)
  /* now kill any transactions associated with uncommitted versions, unless
   * the transaction is ourself! */
  findVersion(tid, o, ver);
  /* don't care about which committed version to use, at the moment. */
  moveVersion(_r, NIL);
```

```
}

/* per-object, before write. */
/* returns NIL in ver to indicate suicide. */
inline ensureWriter(tid, o, ver) {
  assert(tid!=NIL);
  /* Same beginning as ensureReader */
  ver = NIL; _r = NIL; _rr = NIL;
  do
  :: assert (ver==NIL);
     findVersion(tid, o, ver);
     if
     :: (ver!=NIL) -> break /* found a writable version for us */
     :: (ver==NIL && _r==NIL) ->
        /* create and link a fully-committed root version, then
         * use this as our base. */
        allocVersion(_retval, _r, NIL, NIL);
        CAS_Version(object[o].version, NIL, _r, _cas_stat)
     :: else ->
        _cas_stat = true
     fi;
     if
     :: (_cas_stat) ->
        /* so far, so good. */
        assert (_r!=NIL);
        assert (version[_r].owner==NIL ||
                transid[version[_r].owner].status==committed);
        /* okay, make new version for this transaction. */
        assert (ver==NIL);
        allocVersion(_retval, ver, tid, _r);
        /* want copy of committed version _r.  Race here because _r can be
         * written to under peculiar circumstances, namely: _r has
         * non-flag value, non-flag value is copied back to parent,
         * flag_value is written to parent -- this forces flag_value to
         * be written to committed version. */
        /* IF WRITES ARE ALLOWED TO COMMITTED VERSIONS, THERE IS A RACE HERE.
         * But our implementation of false_flag writes at the moment does
         * not permit *any* writes to committed versions. */
        version[ver].field[0] = version[_r].field[0];
        version[ver].field[1] = version[_r].field[1];
        assert(NUM_FIELDS==2); /* else ought to initialize more fields */
        CAS_Version(object[o].version, _r, ver, _cas_stat);
        moveVersion(_r, NIL); /* free _r */
        if
        :: (_cas_stat) ->
           /* kill all readers (except ourself) */
           /* note that all changes have to be made from the front of the
            * list, so we unlink ourself and then re-add us. */
           do
           :: moveReaderList(_r, object[o].readerList);
              if
              :: (_r==NIL) -> break
```

```
            :: (_r!=NIL && readerlist[_r].transid!=tid)->
               tryToAbort(readerlist[_r].transid)
            :: else
            fi;
            /* link out this reader */
            CAS_Reader(object[o].readerList, _r, readerlist[_r].next, _)
         od;
         /* okay, all pre-existing readers dead & gone. */
         assert(_r==NIL);
         /* link us back in. */
         ensureReaderList(tid, o);
         break
      :: else
      fi;
      /* try again */
   :: else
   fi;
   /* try again from the top */
   moveVersion(ver, NIL)
 od;
 /* done! */
 assert (_r==NIL);
}
/* per-field, before read. */
inline checkReadField(o, f) {
 /* do nothing: no per-field read stats are kept. */
 skip
}
/* per-field, before write. */
inline checkWriteField(o, f) {
 _r = NIL; _rr = NIL;
 do
 ::
   /* set write flag, if not already set */
   _val = object[o].field[f];
   if
   :: (_val==FLAG) ->
      break; /* done! */
   :: else
   fi;
   /* okay, need to set write flag. */
   moveVersion(_rr, object[o].version);
   moveVersion(_r, _rr);
   assert (_r!=NIL);
   do
   :: (_r==NIL) -> break /* done */
   :: else ->
      object[o].fieldLock[f] = _thread_id;
      if
      /* this next check ensures that concurrent copythroughs don't stomp
       * on each other's versions, because the field will become FLAG
       * before any other version will be written. */
```

```
            :: (object[o].field[f]==_val) ->
               if
               :: (object[o].version==_rr) ->
                  atomic {
                    if
                    :: (object[o].fieldLock[f]==_thread_id) ->
                       version[_r].field[f] = _val;
                    :: else -> break /* abort */
                    fi
                  }
               :: else -> break /* abort */
               fi
            :: else -> break /* abort */
            fi;
            moveVersion(_r, version[_r].next) /* on to next */
         od;
         if
         :: (_r==NIL) ->
            /* field has been successfully copied to all versions */
            atomic {
              if
              :: (object[o].version==_rr) ->
                 assert(object[o].field[f]==_val ||
                        /* we can race with another copythrough and that's okay;
                         * the locking strategy above ensures that we're all
                         * writing the same values to all the versions and not
                         * overwriting anything. */
                        object[o].field[f]==FLAG);
                 object[o].fieldLock[f]=_thread_id;
                 object[o].field[f] = FLAG;
                 break; /* success!  done! */
              :: else
              fi
            }
         :: else
         fi
         /* retry */
      od;
      /* clean up */
      moveVersion(_r, NIL);
      moveVersion(_rr, NIL);
   }
```

This configuration of Fleetwood Mac's been together 10 years, and it's taken us entire years to make a single record.

Christine McVie

Appendix B

Compiler and runtime system configurations

This section contains details on the compiler and runtime system configurations used to obtain the results in Section 5.2.

```
Four variants of the spec benchmarks were compiled, labeled T, TNOA,
NT, and TUNIQ.

T: Basic transaction transformation
$ SPECSUFFIX=T \
  CLASSPATH_HOME=.../classpath-0.08-install \
  bin/build-spec-precisec -T 1 10

TNOA: "No array" transformation
$ SPECSUFFIX=TNOA \
  CLASSPATH_HOME=.../classpath-0.08-install \
  bin/build-spec-precisec -Dharpoon.synctrans.noarraymods=true -T 1 10

NT: Strip all transactions so we can discover non-transactional overhead.
$ SPECSUFFIX=NT \
  CLASSPATH_HOME=.../classpath-0.08-install \
  bin/build-spec-precisec -Dharpoon.synctrans.removetrans=true -T 1 10

TUNIQ: Count methods called within a transaction and unique objects
       touched during a transaction.
$ FLEXSTACK=64m OPT_FLEX_SUPPORT_FILES=Support/jutil.jar \
  SPECSUFFIX=TUNIQ \
  CLASSPATH_HOME=.../classpath-0.08-install \
  bin/build-spec-precisec -Dharpoon.synctrans.uniquerwcounters=true \
     -Dharpoon.counters.enabled.synctrans.virgin_{read,write}_array=true \
     -Dharpoon.counters.enabled.synctrans.virgin_{read,write}_array_size=true \
```

```
    -Dharpoon.counters.enabled.synctrans.virgin_{read,write}_object=true \
    -Dharpoon.counters.enabled.synctrans.virgin_{read,write}_object_size=true\
    -Dharpoon.counters.enabled.synctrans.{read,write}_t_{array,object}=true \
    -Dharpoon.counters.enabled.synctrans.{read,write}_nt_{array,object}=true \
    -Dharpoon.counters.enabled.synctrans.trans_call=true \
    -T 6 7
```

These source files were then built and linked against a number of
different configurations of the runtime. The following script was used.

```
-----------------------------------------------------------------------
#!/bin/bash
BENCHMARKS="200CHECK 201COMPRESS 202JESS 205RAYTRACE 209DB 213JAVAC"
BENCHMARKS="$BENCHMARKS 222MPEGAUDIO 227MTRT 228JACK"

function buildone() {
pushd $RT
( ./configure "${CONFIG[@]}" && make clean && make ) || return
popd
mkdir -p $DST || return
tar -C src -xzf src/$SRC.tgz || return
for f in $BENCHMARKS ; do
  make -C src/$SRC/as${f}sa$SUF -j2 CC="$CFLAGS" && \
  mv src/$SRC/as${f}sa$SUF/Java.a ${RT}/${f}sa${SUF}-Java.a && \
  make -C src/$SRC/as${f}sa$SUF clean && \
  make -C ${RT} run${f}sa$SUF && \
  mv ${RT}/${f}sa${SUF}-Java.a ${RT}/run${f}sa$SUF $DST && \
  gzip -9 -f $DST/run${f}sa$SUF
done
/bin/rm -rf src/$SRC/
}

# non-transactional runtime
RT=/home/cananian/Harpoon/Runtime
CFLAGS="gcc-3.4 -I${RT}/include -g -O9"
CONFIG=(--with-precise-c --with-gc=conservative --with-thread-model=heavy \
        --with-classpath=/home/cananian/Harpoon/classpath-0.08-install)

# 0-NO is nontransactional
SRC=NO
DST=0-NO
SUF=
buildone

# transactional runtime
RT=/home/cananian/Harpoon/TRANS/Runtime
CFLAGS="gcc-3.4 -I${RT}/include -g -O9"
CONFIG=(--with-precise-c --with-gc=conservative --with-thread-model=heavy \
        --with-classpath=/home/cananian/Harpoon/classpath-0.08-install \
        --with-transactions --with-object-padding=8)
```

```
# 1-T is pure transactional
SRC=T
DST=1-T
SUF=T
buildone

# 2-NOA only transforms non-array references
SRC=TNOA
DST=2-NOA
SUF=TNOA
CONFIG=(--with-precise-c --with-gc=conservative --with-thread-model=heavy \
        --with-classpath=/home/cananian/Harpoon/classpath-0.08-install \
        --with-transactions --with-object-padding=8 \
        --with-extra-cflags=-DTRANS_NO_ARRAY)
buildone

# 8-NT strips all the transactions, so we can see non-transactional overhead
SRC=NT
DST=8-NT
SUF=NT
CONFIG=(--with-precise-c --with-gc=conservative --with-thread-model=heavy \
        --with-classpath=/home/cananian/Harpoon/classpath-0.08-install \
        --with-transactions --with-object-padding=8 \
        --with-extra-cflags=-DTRANS_NT)
buildone

# 9-NONT strips all the transactions, so we can see non-transactional overhead
# also substitutes plain reads & writes.
SRC=NT
DST=9-NONT
SUF=NT
CFLAGS="gcc-3.4 -I${RT}/include -g -O9 -DDONT_REALLY_DO_TRANSACTIONS"
CONFIG=(--with-precise-c --with-gc=conservative --with-thread-model=heavy \
        --with-classpath=/home/cananian/Harpoon/classpath-0.08-install \
        --with-transactions --with-object-padding=8 \
        --with-extra-cflags="-DDONT_REALLY_DO_TRANSACTIONS -DTRANS_NT")
buildone

# B-NONTSC is similar, but it replaces the stores with LL/SC to see how large
# that effect is.
DST=B-NONTSC
CFLAGS="gcc-3.4 -I${RT}/include -g -O9 -DDONT_REALLY_DO_TRANSACTIONS "
CFLAGS="$CFLAGS -DSTUB_LLSC"
CONFIG=(--with-precise-c --with-gc=conservative --with-thread-model=heavy \
        --with-classpath=/home/cananian/Harpoon/classpath-0.08-install \
        --with-transactions --with-object-padding=8 \
   --with-extra-cflags="-DDONT_REALLY_DO_TRANSACTIONS -DTRANS_NT -DSTUB_LLSC")
buildone
```

```
# 6-NOT is transactional code, but with plain reads & writes substituted
SRC=T
DST=6-NOT
SUF=T
CFLAGS="gcc-3.4 -I${RT}/include -g -O9 -DDONT_REALLY_DO_TRANSACTIONS"
CONFIG=(--with-precise-c --with-gc=conservative --with-thread-model=heavy \
        --with-classpath=/home/cananian/Harpoon/classpath-0.08-install \
        --with-transactions --with-object-padding=8 \
        --with-extra-cflags=-DDONT_REALLY_DO_TRANSACTIONS)
buildone

# E-TUNIQ has code to count unique objects
SRC=TUNIQ
DST=E-TUNIQ
SUF=TUNIQ
CFLAGS="gcc-3.4 -I${RT}/include -g -DDONT_REALLY_DO_TRANSACTIONS"
CONFIG=(--with-precise-c --with-gc=conservative --with-thread-model=heavy \
        --with-classpath=/home/cananian/Harpoon/classpath-0.08-install \
        --with-transactions --with-object-padding=16 \
        --with-extra-cflags=-DDONT_REALLY_DO_TRANSACTIONS)
buildone

# 7-TS is the transactional code w/ some statistics gathering stuff
RT=/home/cananian/Harpoon/TRANS/Runtime
CFLAGS="gcc-3.4 -I${RT}/include -g -O9"
CONFIG=(--with-precise-c --with-gc=conservative --with-thread-model=heavy \
        --with-classpath=/home/cananian/Harpoon/classpath-0.08-install \
        --with-transactions --with-object-padding=8 --with-statistics)
SRC=T
DST=7-TS
SUF=T
buildone

# done!
```

Bibliography

Each citation is followed by a bracketed list of the pages on which it appears. DOI names can be resolved at http://doi.org.

[1] Juan Alemany and Edward W. Felten. Performance issues in non-blocking synchronization on shared-memory multiprocessors. In *Proceedings of the 11th Annual ACM Symposium on Principles of Distributed Computing (PODC '92)*, pages 125–134, Vancouver, British Columbia, Canada, August 1992. ACM Press. ISBN 0-89791-495-3. doi: 10.1145/135419.135446. [on p. 164.]

[2] C. Scott Ananian. The static single information form. Master's thesis, Massachusetts Institute of Technology, Cambridge, MA, September 1999. MIT Technical Report MIT-LCS-TR-801. [on pp. 58 and 64.]

[3] C. Scott Ananian. The FLEX compiler project. http://flex-compiler.csail.mit.edu/, 2007. [on pp. 33 and 57.]

[4] C. Scott Ananian and Martin Rinard. Efficient object-based software transactions. In *Synchronization and Concurrency in Object-Oriented Languages (SCOOL)*, San Diego, CA, October 2005. [on pp. 18 and 58.]

[5] C. Scott Ananian and Martin Rinard. Data size optimizations for Java programs. In *Proceedings of the Joint Conference on Languages, Compilers, and Tools for Embedded Systems (LCTES '03)*, pages 59–68, New York, NY, USA, 2003. ACM Press. ISBN 1-58113-647-1. doi: 10.1145/780732.780741. [on p. 58.]

[6] C. Scott Ananian, Krste Asanović, Bradley C. Kuszmaul, Charles E. Leiserson, and Sean Lie. Unbounded transactional memory. In

Proceedings of the 11th International Symposium on High-Performance Computer Architecture (HPCA-11), pages 316–327, San Francisco, California, February 2005. IEEE Computer Society. doi: 10.1109/HPCA.2005.41. [on pp. 19, 22, 29, 38, 58, 83, 129, and 147.]

[7] C. Scott Ananian, Krste Asanović, Bradley C. Kuszmaul, Charles E. Leiserson, and Sean Lie. Unbounded transactional memory. *IEEE Micro Special Issue: Top Picks from Computer Architecture Conferences*, January/February 2006. doi: 10.1109/MM.2006.26. [on pp. 19, 58, and 129.]

[8] Andrew W. Appel. Simple generational garbage collection and fast allocation. *Software–Practice and Experience*, 19(2):171–183, February 1989. doi: 10.1002/spe.4380190206. [on p. 81.]

[9] Andrew W. Appel and Guy J. Jacobson. The world's fastest Scrabble program. *Communications of the ACM*, 31(5):572–578, May 1988. doi: 10.1145/42411.42420. [on p. 26.]

[10] Henry G. Baker. Shallow binding makes functional arrays fast. *ACM SIGPLAN Notices*, 26(8):145–147, August 1991. ISSN 0362-1340. doi: 10.1145/122598.122614. [on pp. 107, 118, 124, and 125.]

[11] Henry G. Baker, Jr. Shallow binding in Lisp 1.5. *Communications of the ACM*, 21(7):565–569, July 1978. ISSN 0001-0782. doi: 10.1145/359545.359566. [on p. 118.]

[12] Greg Barnes. A method for implementing lock-free shared data structures. In *Proceedings of the 5th Annual ACM Symposium on Parallel Algorithms and Architectures (SPAA)*, pages 261–270. ACM Press, June 1993. ISBN 0-89791-599-2. doi: 10.1145/165231.165265. [on p. 163.]

[13] William S. Beebee and Martin Rinard. An implementation of scoped memory for Real-Time Java. In *Proceedings of Embedded Software: First International Workshop (EMSOFT 2001)*, volume 2211/2001 of *LNCS*, Tahoe City, CA, October 2001. [on p. 58.]

[14] William S. Beebee, Jr. Region-based memory management for Real-Time Java. Master's thesis, MIT Dept. of Electrical Engineering and Computer Science, September 2001. [on p. 58.]

[15] A. Bensoussan, C.T. Clingen, and R.C. Daley. The Multics virtual memory: Concepts and design. *CACM*, 15(5):308–318, May 1972. doi: 10.1145/355602.361306. [on p. 138.]

[16] Andrew Black, Norman Hutchinson, Eric Jul, and Henry Levy. Object structure in the Emerald system. In *Conference Proceedings on Object-Oriented Programming Systems, Languages and Applications (OOPSLA)*, pages 78–86. ACM Press, September 1986. ISBN 0-89791-204-7. doi: 10.1145/28697.28706. [on pp. 29, 168, and 169.]

[17] Colin Blundell, E. Christopher Lewis, and Milo M. K. Martin. Deconstructing transactional semantics: The subtleties of atomicity. In *Fourth Annual Workshop on Duplicating, Deconstructing, and Debunking*, Madison, Wisconsin, June 2005. URL http://www.ece.wisc.edu/~wddd/2005/papers/WDDD05_blundell.pdf. [on pp. 17 and 153.]

[18] Hans Boehm, Alan Demers, and Mark Weiser. A garbage collector for C and C++. http://www.hpl.hp.com/personal/Hans_Boehm/gc/, 1991. [on pp. 59 and 65.]

[19] Gregory Bollella, Benjamin Brosgol, James Gosling, Peter Dibble, Steve Furr, and Mark Turnbull. *The Real-Time Specification for Java*. Addison Wesley Longman, 1st edition, January 2000. ISBN 0201703238. [on p. 24.]

[20] Chandrasekhar Boyapati, Alexandru Salcianu, William Beebee, Jr, and Martin Rinard. Ownership types for safe region-based memory management in Real-Time Java. In *PLDI '03: Proceedings of the ACM SIGPLAN Conference on Programming Language Design and Implementation*, pages 324–337, New York, NY, USA, 2003. ACM Press. ISBN 1-58113-662-5. doi: 10.1145/781131.781168. [on p. 58.]

[21] Tyng-Ruey Chuang. A randomized implementation of multiple functional arrays. In *Proceedings of the 1994 ACM Conference on LISP*

and Functional Programming (LFP), pages 173–184. ACM Press, June 1994. ISBN 0-89791-643-3. doi: 10.1145/182409.156779. [on pp. 107, 117, 118, 124, and 125.]

[22] Cliff Click, Gil Tene, and Michael Wolf. The pauseless gc algorithm. In *Proceedings of the 1st ACM/USENIX international conference on Virtual Execution Environments*, pages 46–56, Chicago, Illinois, June 2005. doi: 10.1145/1064979.1064988. [on p. 149.]

[23] Shimon Cohen. Multi-version structures in Prolog. In *Proceedings of the International Conference on Fifth Generation Computer Systems*, pages 265–274, November 1984. URL http://www.eecs.berkeley.edu/Pubs/TechRpts/1984/CSD-84-178.pdf. [on p. 118.]

[24] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms*. The MIT Press and McGraw-Hill, 2nd edition, 2001. [on p. 28.]

[25] Valentin Dallmeier, Christian Lindig, Andrzej Wasylkowski, and Andreas Zeller. Mining object behavior with ADABU. In *WODA '06: Proceedings of the 2006 International Workshop on Dynamic Systems Analysis*, pages 17–24, New York, NY, USA, 2006. ACM Press. ISBN 1-59593-400-6. doi: 10.1145/1138912.1138918. [on p. 58.]

[26] Dave Dice, Ori Shalev, and Nir Shavir. Transactional locking II. In *International Symposium on Distributed Computing (DISC 2006*, Stockholm, Sweden, September 2006. Swedish Institute of Computer Science. URL http://research.sun.com/scalable/pubs/DISC2006.pdf. [on p. 17.]

[27] *VAX MACRO and Instruction Set Reference Manual*. Digital Equipment Corporation, Maynard, Massachusetts, November 1996. [on p. 165.]

[28] Shahrooz Feizabadi, William Beebee, Jr, Binoy Ravindran, Peng Li, and Martin Rinard. Utility accrual scheduling with Real-Time Java. In *Workshop on Java Technologies for Real-Time and Embedded Systems (JTRES)*, volume 2889/2003 of *LNCS*, pages 550–563, 2003. ISBN 3-540-20494-6. doi: 10.1007/b94345. [on p. 58.]

[29] Cormac Flanagan and Shaz Qadeer. A type and effect system for atomicity. In *Proceedings of the ACM SIGPLAN '98 Conference on Programming Language Design and Implementation*, pages 338–349, Montreal, Quebec, Canada, June 1998. doi: 10.1145/781131.781169. Proceedings published ACM SIGPLAN Notices, Vol. 33, No. 5, May, 1998. [on pp. 30, 170, and 171.]

[30] Catalin A. Francu. Real-time scheduling for Java. Master's thesis, MIT Department of Electrical Engineering and Computer Science, June 2002. [on p. 58.]

[31] Keir Fraser and Tim Harris. Concurrent programming without locks. http://www.cl.cam.ac.uk/research/srg/netos/lock-free/, July 2004. [on p. 17.]

[32] *MPC7450 RISC Microprocessor Family Reference Manual, Rev. 5*. Freescale Semiconductor, Chandler, Arizona, January 2005. MPC7450UM. [on p. 44.]

[33] Jeffery E. F. Friedl. *Mastering Regular Expressions*. O'Reilly & Associates, 2nd edition, July 2002. [on p. 26.]

[34] Ovidiu Gheorghioiu. Statically determining memory consumption of Real-Time Java threads. Master's thesis, MIT Department of Electrical Engineering and Computer Science, June 2002. [on p. 58.]

[35] Ovidiu Gheorghioiu, Alexandru Salcianu, and Martin Rinard. Interprocedural compatibility analysis for static object preallocation. In *POPL '03: Proceedings of the 30th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, pages 273–284, New York, NY, USA, 2003. ACM Press. ISBN 1-58113-628-5. doi: 10.1145/604131.604154. [on p. 58.]

[36] GNU Classpath. The GNU Classpath project. http://classpath.org/, 2004. [on p. 58.]

[37] James Gosling, Bill Joy, Guy Steele, and Gilad Bracha. *The Java Language Specification*. Prentice Hall PTR, 3rd edition, June 2005. [on p. 74.]

[38] Jim Gray. The transaction concept: Virtues and limitations. In *Seventh International Conference of Very Large Data Bases*, pages 144–154, September 1981. [on p. 165.]

[39] Jim Gray and Andreas Reuter. *Transaction Processing: Concepts and Techniques*. Morgan Kaufmann, 1993. [on p. 165.]

[40] Michael Greenwald and David Cheriton. The synergy between non-blocking synchronization and operating system structure. In *Proceedings of the 2nd USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, pages 123–136. ACM Press, October 1996. ISBN 1-880446-82-0. doi: 10.1145/238721.238767. [on pp. 15, 29, and 83.]

[41] Dan Grossman, Jeremy Manson, and William Pugh. What do high-level memory models mean for transactions? In *MSPC '06: Proceedings of the 2006 workshop on Memory system performance and correctness*, pages 62–69, New York, NY, USA, 2006. ACM Press. doi: 10.1145/1178597.1178609. [on p. 34.]

[42] Dirk Grunwald and Soraya Ghiasi. Microarchitectural denial of service: insuring microarchitectural fairness. In *Proceedings of the 35th Annual ACM/IEEE International Symposium on Microarchitecture*, pages 409–418, Istanbul, Turkey, November 2002. doi: 10.1109/MICRO.2002.1176268. [on p. 152.]

[43] Lance Hammond, Vicky Wong, Mike Chen, Brian D. Carlstrom, John D. Davis, Ben Hertzberg, Manohar K. Prabhu, Honggo Wijaya, Christos Kozyrakis, and Kunle Olukotun. Transactional memory coherence and consistency. In *ISCA 31*, pages 102–113, München, Germany, June 2004. ISBN 0-7695-2143-6. doi: 10.1109/ISCA.2004.1310767. [on pp. 83 and 166.]

[44] Tim Harris and Keir Fraser. Language support for lightweight transactions. In *Proceedings of the 18th Annual ACM SIGPLAN Conference on Object-Oriented Programing, Systems, Languages, and Applications (OOPSLA)*, pages 388–402, Anaheim, California, October 2003. ACM Press. ISBN 1-58113-712-5. doi: 10.1145/949305.949340. [on pp. 17, 29, 39, 78, 166, and 167.]

[45] Tim Harris, Simon Marlow, Simon Peyton-Jones, and Maurice Herlihy. Composable memory transactions. In *PPoPP '05: Proceedings of the tenth ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, pages 48–60, New York, NY, USA, 2005. ACM Press. ISBN 1-59593-080-9. doi: 10.1145/1065944.1065952. [on pp. 17 and 23.]

[46] John L. Hennessy and David A. Patterson. *Computer Architecture: A Quantitative Approach*. Morgan Kaufmann, 2nd edition, 1996. ISBN 1-55860-329-8. [on pp. 42, 87, 91, 92, and 109.]

[47] Maurice Herlihy. Wait-free synchronization. *ACM TOPLAS*, 13(1): 124–149, January 1991. ISSN 0164-0925. doi: 10.1145/114005.102808. [on p. 162.]

[48] Maurice Herlihy. A methodology for implementing highly concurrent data objects. *ACM TOPLAS*, 15(5):745–770, November 1993. ISSN 0164-0925. doi: 10.1145/161468.161469. [on pp. 111, 163, and 164.]

[49] Maurice Herlihy and J. Eliot B. Moss. Transactional memory: Architectural support for lock-free data structures. In *Proceedings of the 20th Annual International Conference on Computer Architecture (ISCA). (Also published as ACM SIGARCH Computer Architecture News, Volume 21, Issue 2, May 1993.)*, pages 289–300, San Diego, California, May 1993. ACM Press. ISBN 0-8186-3810-9. doi: 10.1145/165123.165164. [on pp. 15, 16, 29, 36, 39, 123, 132, 137, 165, and 167.]

[50] Maurice Herlihy, Victor Luchangco, Mark Moir, and William N. Scherer, III. Software transactional memory for dynamic-sized data structures. In *Proceedings of the 22nd Annual ACM Symposium on Principles of Distributed Computing (PODC '03)*, pages 92–101, Boston, Massachusetts, July 2003. ACM Press. ISBN 1-58113-708-7. doi: 10.1145/872035.872048. [on pp. 15, 16, 17, and 166.]

[51] Maurice Herlihy, Victor Luchangco, and Mark Moir. A flexible framework for implementing software transactional memory. In *Proceedings of the 21st Annual ACM SIGPLAN Conference on Object-Oriented Programing, Systems, Languages, and Applica-*

tions (OOPSLA), pages 253–262, Portland, Oregon, 2006. ACM Press. ISBN 1-59593-348-4. doi: 10.1145/1167473.1167495. [on p. 17.]

[52] Maurice P. Herlihy. The transactional manifesto: Software engineering and non-blocking synchronization. Invited talk at PLDI, June 2005. http://research.ihost.com/pldi2005/manifesto.pldi.ppt. [on pp. 21 and 23.]

[53] Maurice P. Herlihy. SXM1.1: Software transactional memory package for C#. http://www.cs.brown.edu/people/mph/, May 2005. [on p. 17.]

[54] Maurice P. Herlihy. Impossibility and universality results for wait-free synchronization. In *Proceedings of the 7th Annual ACM Symposium on Principles of Distributed Computing (PODC)*, pages 276–290, Toronto, Ontario, Canada, August 1988. ACM Press. ISBN 0-89791-277-2. doi: 10.1145/62546.62593. [on pp. 15, 161, and 162.]

[55] Maurice P. Herlihy and J. Eliot B. Moss. Transactional support for lock-free data structures. Technical Report 92/07, Digital Cambridge Research Lab, One Kendall Square, Cambridge, MA 02139, December 1992. [on pp. 140 and 165.]

[56] Maurice P. Herlihy, Victor Luchangco, and Mark Moir. Obstruction-free synchronization: Double-ended queues as an example. In *Proceedings of the 23rd IEEE International Conference on Distributed Computing Systems (ICDCS)*, pages 522–529, Providence, Rhode Island, May 2003. doi: 10.1109/ICDCS.2003.1203503. [on pp. 15, 23, and 162.]

[57] Gerard J. Holzmann. *The Spin Model Checker*. Addison-Wesley, 2003. ISBN 0-321-22862-6. [on p. 44.]

[58] *PowerPC Virtual Environment Architecture, Book II, Version 2.02*. IBM, Austin, Texas, January 2005. [on p. 44.]

[59] E. H. Jensen, G. W. Hagensen, and J. M. Broughton. A new approach to exclusive data access in shared memory multiprocessors. Technical Report UCRL-97663, LLNL, Livermore, California, November 1987. [on p. 165.]

[60] Mike Jones. What really happened on Mars? http://research.microsoft.com/~mbj/Mars_Pathfinder/, December 1997. [on p. 15.]

[61] Eric Jul and Bjarne Steensgaard. Implementation of distributed objects in Emerald. In *Proceedings of the International Workshop on Object Orientation in Operating Systems*, pages 130–132, Palo Alto, CA, October 1991. IEEE. doi: 10.1109/IWOOOS.1991.183037. [on pp. 29 and 168.]

[62] Felix S. Klock, II. Architecture independent register allocation. Master's thesis, MIT Dept. of Electrical Engineering and Computer Science, 2001. [on p. 58.]

[63] Tom Knight. An architecture for mostly functional languages. In *LFP*, pages 105–112. ACM Press, 1986. ISBN 0-89791-200-4. doi: 10.1145/319838.319854. [on pp. 15, 132, 137, 140, and 165.]

[64] Anthony LaMarca. A performance evaluation of lock-free synchronization protocols. In *Proceedings of the 13th Annual ACM Symposium on Principles of Distributed Computing (PODC '94)*, pages 130–140, Los Angeles, CA, August 1994. ACM Press. ISBN 0-89791-654-9. doi: 10.1145/197917.197975. [on pp. 83 and 164.]

[65] Leslie Lamport. Concurrent reading and writing. *CACM*, 20(11):806–811, November 1977. ISSN 0001-0782. doi: 10.1145/359863.359878. [on pp. 15 and 161.]

[66] Ruby B. Lee. Precision architecture. *IEEE Computer*, 22(1):78–91, January 1989. doi: 10.1109/2.19825. [on p. 139.]

[67] Thomas Lengauer and Robert Endre Tarjan. A fast algorithm for finding dominators in a flowgraph. *ACM TOPLAS*, 1(1):121–141, 1979. doi: 10.1145/357062.357071. [on p. 64.]

[68] Jeremy Manson and William Pugh. Core semantics of multithreaded Java. In *JGI '01: Proceedings of the 2001 joint ACM-ISCOPE conference on Java Grande*, pages 29–38, New York, NY, USA, 2001. ACM Press. doi: 10.1145/376656.376806. [on pp. 34 and 167.]

[69] Jeremy Manson and William Pugh. Semantics of multithreaded Java. Technical Report CS-TR-4215, Department of Computer Sci-

ence, University of Maryland, College Park, May 2001. URL http://hdl.handle.net/1903/1118. [on pp. 34 and 167.]

[70] Jeremy Manson, William Pugh, and Sarita V. Adve. The Java memory model. In *POPL '05: Proceedings of the 32nd ACM SIGPLAN-SIGACT symposium on Principles of programming languages*, pages 378–391, New York, NY, USA, 2005. ACM Press. doi: 10.1145/1040305.1040336. [on pp. 34 and 167.]

[71] Simon Marlow, Simon L. Peyton Jones, Andrew Moran, and John H. Reppy. Asynchronous exceptions in Haskell. In *PLDI '01*, pages 274–285, Snowbird, Utah, June 2001. ACM Press. ISBN 1-58113-414-2. doi: 10.1145/378795.378858. [on p. 24.]

[72] J. F. Martinez and J. Torrellas. Speculative synchronization: applying thread-level speculation to explicitly parallel applications. In *Proceedings of the 10th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS-X)*, pages 18–29, San Jose, California, October 5–9 2002. ACM Press. ISBN 1-58113-574-2. doi: 10.1145/605397.605400. [on p. 165.]

[73] Henry Massalin and Calton Pu. A lock-free multiprocessor OS kernel. Technical Report CUCS-005-91, Columbia University, New York, NY 10027, June 1991. [on pp. 15, 29, 83, 158, and 162.]

[74] Robert M. Metcalfe and David R. Boggs. Ethernet: Distributed packet switching for local computer networks. *CACM*, 19(7):395–404, July 1976. doi: 10.1145/360248.360253. [on p. 136.]

[75] Maged M. Michael and Michael L. Scott. Simple, fast, and practical non-blocking and blocking concurrent queue algorithms. In *Proceedings of the 15th Annual ACM Symposium on Principles of Distributed Computing (PODC '96)*, pages 267–276, Philadelphia, PA, 23–26 May 1996. ISBN 0-89791-710-3. doi: 10.1145/248052.248106. [on p. 23.]

[76] Sun Microsystems. Java Native Interface. http://java.sun.com/j2se/1.5.0/docs/guide/jni/, 2004. [on pp. 65, 66, and 67.]

[77] Carl D. Offner. Notes on graph algorithms used in optimizing compilers. University of Massachusetts at Boston, 1995. URL http://www.cs.umb.edu/~offner/files/flow_graph.pdf. [on p. 64.]

[78] Chris Okasaki. Purely functional random-access lists. In *Proceedings of the 7th International Conference on Functional Programming Languages and Computer Architecture (FPCA)*, pages 86–95. ACM Press, June 1995. ISBN 0-89791-719-7. doi: 10.1145/224164.224187. [on p. 118.]

[79] Melissa E. O'Neill and F. Warren Burton. A new method for functional arrays. *J. Func. Prog.*, 7(5):487–514, September 1997. doi: 10.1017/S0956796897002852. [on pp. 108, 118, and 125.]

[80] Ravi Rajwar and James R. Goodman. Speculative lock elision: Enabling highly concurrent multithreaded execution. In *Proceedings of the 34th Annual ACM/IEEE International Symposium on Microarchitecture*, pages 294–305, Austin, Texas, December 2001. doi: 10.1109/MICRO.2001.991127. [on pp. 83, 144, and 165.]

[81] Ravi Rajwar and James R. Goodman. Transactional lock-free execution of lock-based programs. In *Proceedings of the 10th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS-X)*, pages 5–17, San Jose, California, October 5–9 2002. ACM Press. ISBN 1-58113-574-2. doi: 10.1145/605397.605399. [on pp. 15, 83, and 165.]

[82] Martin Rinard, Alexandru Sălcianu, and Suhabe Bugrara. A classification system and analysis for aspect-oriented programs. In *SIGSOFT '04/FSE-12: Proceedings of the ACM SIGSOFT 12th International Symposium on Foundations of Software Engineering*, pages 147–158, New York, NY, USA, 2004. ACM Press. ISBN 1-58113-855-5. doi: 10.1145/1029894.1029917. [on p. 58.]

[83] Martin C. Rinard. Implicitly synchronized abstract data types: Data structures for modular parallel programming. *Journal of Programming Languages*, 6:1–35, 1998. URL http://www.cag.lcs.mit.edu/~rinard/paper/jpl98.pdf. [on p. 23.]

[84] Algis Rudys and Dan S. Wallach. Transactional rollback for language-based systems. In *Proceedings of the 2002 International Conference*

on Dependable Systems and Networks (DSN), pages 439–448. IEEE Computer Society, June 2002. ISBN 0-7695-1597-5. doi: 10.1109/DSN.2002.1028929. [on pp. 17 and 166.]

[85] Daniel J. Scales and Kourosh Gharachorloo. Towards transparent and efficient software distributed shared memory. In *Proceedings of the 16th ACM Symposium on Operating Systems Principles (SOSP)*, pages 157–169. ACM Press, October 1997. ISBN 0-89791-916-5. doi: 10.1145/268998.266673. [on p. 43.]

[86] H. Schorr and W. M. Waite. An efficient machine-independent procedure for garbage collection in various list structures. *CACM*, 10(8): 501–506, August 1967. ISSN 0001-0782. doi: 10.1145/363534.363554. [on p. 25.]

[87] Nir Shavit and Dan Touitou. Software transactional memory. In *Proceedings of the 14th Annual ACM Symposium on Principles of Distributed Computing (PODC '95)*, pages 204–213, Ottawa, Ontario, Canada, August 1995. ACM Press. ISBN 0-89791-710-3. doi: 10.1145/224964.224987. [on pp. 15, 17, 29, and 166.]

[88] D.P. Siewiorek, Gordon Bell, and Allen Newell. *Computer Structures: Principles and Examples*. McGraw-Hill, 1982. [on p. 165.]

[89] Janice M. Stone, Harold S. Stone, Philip Heidelberg, and John Turek. Multiple reservations and the oklahoma update. *IEEE Parallel and Distributed Technology*, 1(4):58–71, November 1993. doi: 10.1109/88.260295. [on pp. 15 and 165.]

[90] Alexandru Sălcianu. Pointer analysis and its applications for Java programs. Master's thesis, Massachusetts Institute of Technology, September 2001. [on p. 58.]

[91] Alexandru Sălcianu and Martin Rinard. Pointer and escape analysis for multithreaded programs. In *PPoPP '01: Proceedings of the 8th ACM SIGPLAN Symposium on Principles and Practices of Parallel Programming*, pages 12–23, New York, NY, USA, 2001. ACM Press. ISBN 1-58113-346-4. doi: 10.1145/379539.379553. [on p. 58.]

[92] Alexandru Sălcianu and Martin Rinard. Purity and side effect analysis for Java programs. In *Proceedings of the 6th International Conference on Verification, Model Checking and Abstract Interpretation*, number 3385 in LNCS, pages 199–215, 2005. doi: 10.1007/b105073. [on p. 58.]

[93] Andrew S. Tanenbaum. *Modern Operating Systems*. Prentice Hall, Englewood Cliffs, NJ, 1992. [on p. 14.]

[94] Robert Tarjan. Finding dominators in directed graphs. *SIAM Journal on Computing*, 3(1):62–89, 1974. doi: 10.1137/0203006. [on p. 64.]

[95] Frédéric Vivien and Martin Rinard. Incrementalized pointer and escape analysis. In *PLDI '01: Proceedings of the ACM SIGPLAN Conference on Programming Language Design and Implementation*, pages 35–46, New York, NY, USA, 2001. ACM Press. ISBN 1-58113-414-2. doi: 10.1145/378795.378804. [on p. 58.]

[96] John Whaley. Partial method compilation using dynamic profile information. In *OOPSLA '01: Proceedings of the 16th ACM SIGPLAN Conference on Object-Oriented Programming, Systems, Languages, and Applications*, pages 166–179, New York, NY, USA, 2001. ACM Press. ISBN 1-58113-335-9. doi: 10.1145/504282.504295. [on p. 58.]

[97] John Whaley and Martin Rinard. Compositional pointer and escape analysis for Java programs. In *OOPSLA '99: Proceedings of the 14th ACM SIGPLAN Conference on Object-Oriented Programming, Systems, Languages, and Applications*, pages 187–206, Denver, November 1999. doi: 10.1145/320384.320400. [on pp. 58 and 81.]

[98] Emmett Witchel, Sam Larsen, C. Scott Ananian, and Krste Asanović. Direct addressed caches for reduced power consumption. In *Proceedings of the 34th Annual ACM/IEEE International Symposium on Microarchitecture*, Austin, Texas, December 2001. doi: 10.1109/MICRO.2001.991111. [on pp. 44, 58, and 73.]

[99] Emmett Witchel, Josh Cates, and Krste Asanović. Mondrian memory protection. In *ASPLOS-X: Proceedings of the 10th International Conference on Architectural Support for Programming Languages*

and Operating Systems, pages 304–316, New York, NY, USA, 2002. ACM Press. ISBN 1-58113-574-2. doi: 10.1145/605397.605429. [on p. 58.]

[100] Sharon Zakhour, Scott Hommel, Jacob Royal, Isaac Rabinovitch, Tom Risser, and Mark Hoeber. *The Java Tutorial: A Short Course on the Basics*. Prentice Hall PTR, 4th edition, September 2006. [on p. 77.]

[101] Karen Zee and Martin Rinard. Write barrier removal by static analysis. In *OOPSLA '02: Proceedings of the 17th ACM SIGPLAN Conference on Object-Oriented Programming, Systems, Languages, and Applications*, pages 191–210, New York, NY, USA, 2002. ACM Press. ISBN 1-58113-471-1. doi: 10.1145/582419.582439. [on p. 58.]

[102] Lixin Zhang. UVSIM reference manual. Technical Report UUCS-03-011, University of Utah, March 2003. [on p. 139.]

[103] Yoav Zibin, Alex Potanin, Shay Artzi, Adam Kieżun, and Michael D. Ernst. Object and reference immutability using Java generics. Technical Report MIT-CSAIL-TR-2007-018, MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, March 16, 2007. URL http://hdl.handle.net/1721.1/36850. [on p. 81.]

[104] Craig Zilles and David H. Flint. Challenges to providing performance isolation in transactional memories. In *Fourth Annual Workshop on Duplicating, Deconstructing, and Debunking*, Madison, Wisconsin, June 2005. URL http://www.ece.wisc.edu/~wddd/2005/papers/WDDD05_zilles.pdf. [on pp. 151 and 152.]