Types for Precise Thread Interference

Jaeheon Yi Tim Disney
Computer Science Department
University of California, Santa Cruz
Santa Cruz, CA 95064

Stephen N. Freund
Computer Science Department
Williams College
Williamstown, MA 01267

Cormac Flanagan
Computer Science Department
University of California, Santa Cruz
Santa Cruz, CA 95064

Abstract

The potential for unexpected interference between threads makes multithreaded programming notoriously difficult. Programmers use a variety of synchronization idioms such as locks and barriers to restrict where interference may actually occur. Unfortunately, the resulting *actual* interference points are typically never documented and must be manually reconstructed as the first step in any subsequent programming task (code review, refactoring, etc).

This paper proposes explicitly documenting actual interference points in the program source code, and it presents a type and effect system for verifying the correctness of these interference specifications.

Experimental results on a variety of Java benchmarks show that this approach provides a significant improvement over prior systems based on method-level atomicity specifications. In particular, it reduces the number of interference points one must consider from several hundred points per thousand lines of code to roughly 13 per thousand lines of code. Explicit interference points also serve to highlight all known concurrency defects in these benchmarks.

1. Introduction

The widespread adoption of multi-core processors necessitates effective techniques for developing reliable multithreaded software. However, developing and validating multithreaded software is very difficult, in large part because of the potential for nondeterministic interference between concurrent threads. Typically, programmers use a variety of synchronization idioms to restrict where interference may occur. For example, locks, semaphores, or barriers can be used to prevent interference between threads at program points within critical sections.

There is, however, a large gap between *actual interference points* (where interference actually happens in the given program due to its synchronization structure) and *preemptive interference points* (which potentially can occur at any program point under a typical preemptive semantics for thread interleaving). Knowledge of actual interference points is required for almost all reasoning about program behavior, but these interference points are almost never documented in the program source code. Thus every programming task (code review, refactoring, feature extension, etc) must typically begin with the programmer manually reconstructing the actual interference points by analyzing the synchronization structure of the target program.

Manual reconstruction of actual interference points is tedious and error-prone. Moreover, actual interference points are quite sparse in practice, and consequently programmers have a tendency to optimistically assume that the code being analyzed is free of interference and simply perform sequential reasoning about program behavior. This shortcut is sometimes correct, but often results in defects such as race conditions and violations of intended atomicity, determinism, and ordering constraints.

This work was presented at the *International Workshop on Foundations of Object-Oriented Languages (FOOL)*, 2011.

Figure 1: Traveling Salesperson Algorithm

```
1  Object lock;
2  volatile int shortestPathLength;
3
4  compound void searchFrom(Path path) {
5    if (path.length >= ..shortestPathLength)
6      return;
7
8    if (path.isComplete()) {
9      ..synchronized (lock) {
10       if (path.length < shortestPathLength)
11         shortestPathLength = path.length;
12     }
13   } else {
14     for (Path c: path.children())
15       searchFrom#(c);
16   }
17 }
```

The central philosophy behind this paper is that actual interference points should be explicit in the program source code, avoiding the need to reconstruct them before each programming task.

To illustrate the benefits of explicit interference annotations, consider the traveling salesperson algorithm shown in Figure 1. The function `searchFrom` recursively searches through all extensions of a salesperson's `path`, aborting the search whenever the `path`'s length field becomes greater than the length of the shortest complete path found so far (stored in `shortestPathLength`). To exploit multicore processors, multiple instances of `searchFrom` execute concurrently, and the code uses the notation "`..`" to document thread interference in a lightweight manner. For example, the variable `shortestPathLength` is protected by the mutex `lock` for all writes, but racy reads are permitted to maximize performance. Thus, thread interference may occur before the read of `shortestPathLength` on line 5 (because of potential concurrent writes) and before the lock acquire at line 9 (because a concurrent invocation of `searchFrom` may be racing on that lock). We refer to these explicitly annotated points of potential interference as *yield points*, and they document where interleaved actions of concurrent threads may conflict with operations performed by this function.

To facilitate modular reasoning, methods are annotated to indicate their yielding effects. Methods with no internal yield points are declared `atomic`, and methods that may yield are declared `compound`, as illustrated by the declaration of `searchFrom`. Calls to `compound` methods (as on line 15 above) are highlighted with a postfix "`#`" to indicate that the callee contains yield points, and so

invariants over shared state that hold before the call may not hold after the call returns.

This paper presents a type and effect system to verify that explicit interference annotations in a program capture all possible thread interference. The type system is based on Lipton's theory of reduction [33], which characterizes when a sequence of instructions from one thread forms a serializable transaction. Our type system guarantees that the instructions of each well-typed thread consists of a sequence of serializable transactions separated by interference annotations. Consequently, any well-typed program behaves *as if* executing under a *cooperative scheduler*, where context switches happen only at explicitly marked interference points.

This approach provides several benefits. First, manual reconstruction of actual interference points is no longer a required first step of any programming task, since the type system guarantees that all interference points are explicitly documented in the source code. Second, sequential reasoning is applicable to any code fragment that does not include interference annotations. Third, interference annotations provide an explicit reminder to programmers about where their natural tendency to apply sequential reasoning is not applicable, and where they need to account for thread interference. Finally, a preliminary user study has shown that documenting interference points produces a statistically significant improvement in the ability of programmers to identify defects [43].

We describe a prototype implementation called JCC (Java cooperability checker) for the Java programming language. This type checker takes as input a program annotated with interference annotations (including where races may occur) and verifies that the specification holds. Experimental results show that JCC requires annotating only 13 interference points per thousand lines of code. In comparison, our prior type system for method-level atomicity [21] requires the programmer to reason about over 180 interference points per thousand lines of code, mainly because of interference points in methods that cannot be verified as atomic.

Contributions. In summary, the primary contributions of our work include:

- A lightweight and precise notation for specifying interference points.
- A type and effect system for verifying these interference specifications (Section 5).
- An implementation of this type system for the Java programming language (Section 6).
- Experimental results showing that (1) these interference specifications highlight all known concurrency bugs in our benchmarks, and (2) in comparison to prior approaches based on race conditions or atomicity, the type system reduces the number of interference points by an order of magnitude (Section 7).

Previous work explored dynamic cooperability analyses [52], a cooperability type system for a limited imperative calculus [51], and whether cooperability annotations facilitate program understanding [43]. In this paper we extend cooperability to the type system for a large, object-oriented language, implement a type checker for it, and demonstrate that it is effective on a variety of realistic programs.

Comparison to Atomic Non-Interference Specifications. A method or code block is *atomic* if it does not contain any thread interference points [29, 35, 34, 15], and previous studies have shown that as many as 90% of methods are typically atomic [19].

Unfortunately, the notion of atomicity is rather awkward for specifying interference in non-atomic methods such as `searchFrom`. Some particular code blocks in `searchFrom`, such as the synchronized block, can be marked atomic, as shown in Figure 2, but the result is rather verbose and inadequate, since atomic focuses on delimiting blocks where interference does not occur, but still suggests that interference can occur everywhere outside these atomic code blocks, such as on the access to `path` on line 18.

Figure 2: Traveling Salesperson Algorithm with atomic blocks

```
1  Object lock;
2  volatile int shortestPathLength;
3
4  compound void searchFrom(Path path) {
5    atomic {
6      if (path.length >= shortestPathLength)
7        return;
8    }
9
10   if (path.isComplete()) {
11     atomic {
12       synchronized (lock) {
13         if (path.length < shortestPathLength)
14           shortestPathLength = path.length;
15       }
16     }
17   } else {
18     for (Path c: path.children())
19       searchFrom#(c);
20   }
21 }
```

Moreover, the use of atomic blocks requires a *bimodal* reasoning style that combines sequential reasoning inside atomic blocks with preemptive reasoning (pervasive interference) outside atomic blocks. The programmer is responsible for choosing the appropriate reasoning mode for each piece of code according to whether it is inside or outside an atomic block, with obvious room for error.

Comparison to Yield Interference Specifications. An alternative notation for thread interference is the `yield` statements of cooperative multithreading [3, 4, 9], automatic mutual exclusion [30], and some of our earlier work [51, 52]. After annotating a variety of systems with `yield` statements, we concluded that `yield` clarifies *where* interference may occur, but does not address *why* interference may occur. To illustrate this limitation, consider the following revised `searchFrom` implementation that uses `yield` interference specifications:

```
compound void searchFrom(Path path) {
  yield;
  if (path.length >= shortestPathLength)
  ...
}
```

Here, the `yield` suggests that one of the subsequent reads of `path.length` or `shortestPathLength` is interfering or racy, but it is not immediately obvious which one. By comparison, the interference annotation at line 5 in Figure 1 precisely documents that interference occurs on the read of `shortestPathLength`.

2. Documenting Thread Interference

In a well-typed program, each thread should consist of a sequence of serializable transactions that are separated by yields. A programmer must therefore include a yield point right before the first operation of each transaction, using the following annotations.

Yielding Field Accesses. The following syntax denotes a yielding read or write of a racy field `f`, where the yield is performed just before the access to `f`:

```
e..f          // yield before racy read
```

```
e..f = e′     // yield before racy write
```

Yielding Lock Acquires. The following yielding synchronized block performs a yield after the evaluation of the lock expression e, but before the lock acquire. Note that a lock release cannot interfere with any concurrently executing action by another thread and thus never needs to start a new transaction.

```
..synchronized (e) { // yield before acquire
   e′
}
```

Yielding Method Calls. Next, consider an atomic method `m()` whose body performs a complete transaction. In order to sequentially compose two calls to this method, a yield point must occur in between them. For this purpose, we also allow yield annotations on a method call, where the yield occurs after e and e' have been evaluated and right before the call is performed:

```
e..m(e′)    // yield before method call
```

Non-Atomic Method Calls. A method `m()` is non-atomic or `compound` if its body consists of multiple transactions separated by yield points. In that case, we require calls to `m()` to document the potential for yield points inside `m()`, where the postfix annotation "`#`" reminds the programmer that invariants over thread-shared state that hold before the call may not hold after the call, due to the nested yields.

```
e.m#(e′)    // yields inside m
```

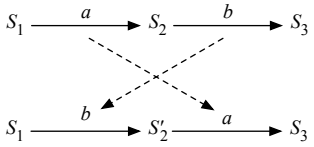
Self-References. As usual, if the target object of a field access or a method call is the self-reference variable `this`, then it can be left implicit as follows:

```
..f         // yield before racy read on this
..f = e′    // yield before racy write to this
..m(e′)     // yield before call on this method
```

3. A Review of Lipton's Theory of Reduction

Our type and effect system reasons about thread interference using Lipton's theory of reduction [33], which describes how certain adjacent operations by different threads in a program trace can be swapped without changing the overall behavior of the trace, and when a sequence of instructions from one thread forms a serializable transaction.

Suppose a trace contains an acquire operation a on a lock `m` that is immediately followed by an operation b of a different thread. That operation b cannot acquire or release `m` (as it is held by the first thread) and so the two operations *commute* — the operations a and b can be swapped without changing the overall behavior or final state of the trace. Thus, acquire operations are *right-movers*.



Similarly, if a lock release operation b by one thread is immediately preceded by an operation a of a different thread, again these two operations can be swapped without changing overall behavior, so each lock release operation is a *left-mover*.

Next, consider an access to a variable `x`. If `x` is race-free, then there are no concurrent accesses to `x` by other threads, so each access commutes with both preceding and following operations of other threads. Put differently, race-free accesses are *both-movers*.

Conversely, if `x` is a racy variable, then an access to `x` cannot in general be swapped with a following (or preceding) operation b, since b in general could be a conflicting access to `x`. Thus, racy accesses are *non-movers*, since they are neither left nor right movers. (To avoid the complexities of Java's relaxed memory model [36], we assume here that all racy variables are `volatile`, but the JCC implementation supports non-volatile racy variables: see Section 6.)

This classification of operations as various kinds of movers then allows us to identify serializable code blocks. In particular, suppose the sequence $\alpha = a_1.\cdots.a_n$ of instructions performed by a particular thread consists of:

1. zero or more right-movers, followed by
2. at most one non-mover, followed by
3. zero or more left-movers.

Any instructions of other threads that are interleaved into α can be commuted out so that α executes serially, without interleaved operations of other threads. In this case, we consider α a *serializable transaction*, or simply a *transaction*.

4. Effects for Cooperability

Our effect system characterize the behavior of each program subexpression using two kinds of effects: *mover effects* and *atomicity effects*.

4.1 Mover Effects

A *mover effect* μ characterizes the behavior of a program expression in terms of how operations of that expression commute with operations of other threads:

$$\mu \;::=\; \mathrm{F} \mid \mathrm{Y} \mid \mathrm{M} \mid \mathrm{R} \mid \mathrm{L} \mid \mathrm{N}$$

F: The mover effect F (for functional) describes expressions whose result does not depend on any mutable state. Hence, re-evaluating a functional expression is guaranteed to produce the same result. (Our type system requires all lock names to be functional to ensure that it does not confuse distinct locks.)

Y: The mover effect Y describes yield operations, denoted as "`..`". These expressions mark transactional boundaries where the current transaction ends and a new one starts.

M: The mover effect M describes both-mover expressions that commute both left and right with concurrent operations by other threads, according to Lipton's theory.

R: The mover effect R describes right-mover expressions.

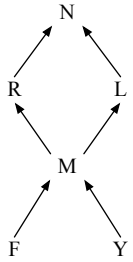
L: The mover effect L describes left-mover expressions.

N: The mover effect N describes non-mover code that may perform a racy access or that may contain right-movers followed by left-movers.

The following tables define the iterative closure (μ^*) and sequential composition ($\mu_1;\mu_2$) of mover effects. These operations are partial (indicated with a "$-$") and may fail if the code between two successive yields would not be serializable. For example, the sequential composition $(\mathrm{L};\mathrm{R})$ is undefined, since their composition is not reducible—code containing a left-mover followed by a right-mover does not form a serializable transaction. Mover effects are ordered by the relation $\sqsubset$ shown via the lattice below.

μ	μ^*
F	F
Y	M
M	M
R	R
L	L
N	—

;	F	Y	M	R	L	N
F	F	Y	M	R	L	N
Y	Y	Y	Y	Y	L	L
M	M	Y	M	R	L	N
R	R	R	R	R	N	N
L	L	Y	L	—	L	—
N	N	R	N	—	N	—



4.2 Atomicity Effects

Each program expression also has an *atomicity effect* τ that summarizes whether the expression performs a yield operation.

$$\tau \quad ::= \quad \text{A} \mid \text{C}$$

Here, A (atomic) means the expression never yields, and C (compound) means the expression may yield. Ordering ($\sqsubset$), iterative closure (τ^*) and sequential composition ($\tau_1;\tau_2$) for atomicity effects are defined by:

$$\begin{array}{rcl} \text{A} & \sqsubset & \text{C} \\ \tau^* & \stackrel{\text{def}}{=} & \tau \\ \tau_1;\tau_2 & \stackrel{\text{def}}{=} & \tau_1 \sqcup \tau_2 \end{array}$$

4.3 Combined Effects

A *combined effect* κ is a pair of a mover effect μ and an atomicity effect τ:

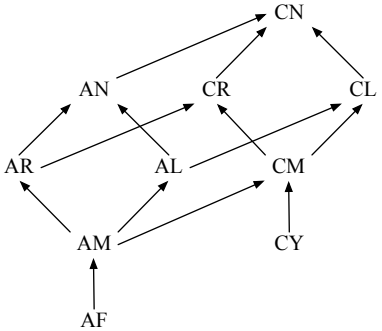
$$\kappa \quad ::= \quad \tau\mu$$

Note that not all combined effects are meaningful; in particular, AY and CF are contradictory: an atomic piece of code may not contain a yield, and code with yields cannot be considered functional.

We define the ordering relation and the join, iterative closure, and sequential composition operations on combined effects in a point-wise manner:

$$\begin{array}{rcll} \tau_1\mu_1 \sqsubseteq \tau_2\mu_2 & \text{iff} & \tau_1 \sqsubseteq \tau_2 \text{ and } \mu_1 \sqsubseteq \mu_2 & \\ \tau_1\mu_1 \sqcup \tau_2\mu_2 & \stackrel{\text{def}}{=} & \tau_3\mu_3 & \text{where } \tau_3 = \tau_1 \sqcup \tau_2 \text{ and } \mu_3 = \mu_1 \sqcup \mu_2 \\ \tau_1\mu_1;\tau_2\mu_2 & \stackrel{\text{def}}{=} & \tau_3\mu_3 & \text{where } \tau_3 = \tau_1;\tau_2 \text{ and } \mu_3 = \mu_1;\mu_2 \\ (\tau\mu)^* & \stackrel{\text{def}}{=} & \tau^*\mu^* & \end{array}$$

The following diagram summarizes the resulting lattice of combined effects·



4.4 Conditional Effects

Based on the previous discussion, the effect of acquiring a lock `m` is AR, since a lock acquire is a right-mover that contains no yield operations. However, if the lock `m` is already held by the current thread, then the re-entrant lock acquire is actually a no-op and could be more precisely characterized as a both-mover AM.

Figure 3: YIELDJAVA **Syntax**

$$\begin{array}{rcl} P \in \mathit{Program} & = & \overline{\mathit{defn}} \\ \mathit{defn} \in \mathit{Definition} & ::= & \texttt{class}\, c\, \texttt{\{}\ \overline{\mathit{field}}\ \overline{\mathit{meth}}\ \texttt{\}} \\ \mathit{field} \in \mathit{Field} & ::= & c\ f \\ \mathit{meth} \in \mathit{Method} & ::= & a\ c\ m\texttt{(}\overline{c\ x}\texttt{)}\ \texttt{\{}\ e\ \texttt{\}} \\ \\ e, \ell \in \mathit{Expr} & ::= & x \mid \texttt{null} \\ & & \mid e_\gamma f \mid e_\gamma f = e \\ & & \mid e_\gamma m\texttt{(}\overline{e}\texttt{)} \mid e_\gamma m\texttt{\#(}\overline{e}\texttt{)} \\ & & \mid \texttt{new}\, c\texttt{(}\overline{e}\texttt{)} \mid e_\gamma\texttt{sync}\ e \mid \texttt{fork}\, e \\ & & \mid \texttt{let}\ x\ \texttt{=}\ e\ \texttt{in}\ e \mid \texttt{if}\ e\ e\ e \mid \texttt{while}\ e\ e \\ \\ \gamma \in \mathit{OptYield} & ::= & \texttt{.} \mid \texttt{..} \\ \\ x, y \in \mathit{Var} & & \\ c, d \in \mathit{ClassName} & & \\ f \in \mathit{FieldName} & = & \mathit{Normal} \cup \mathit{Final} \cup \mathit{Volatile} \\ m \in \mathit{MethodName} & & \\ a \in \mathit{Effect} & & \end{array}$$

We introduce *conditional effects* to capture situations like this where the effect of an operation depends on which locks are held by the current thread. We use ℓ to range over expressions that are functional (F). Such expressions always reliably denote the same lock. An effect a is then either a combined effect κ or an effect conditional on whether a lock ℓ is held:

$$a \quad ::= \quad \kappa \mid \ell\,?\,a_1 : a_2$$

We extend the calculation of iterative closure, sequential composition, and join operations to conditional effects as follows:

$$\begin{array}{rcl} (\ell\,?\,a_1 : a_2)^* & = & \ell\,?\,a_1^* : a_2^* \\ (\ell\,?\,a_1 : a_2);a & = & \ell\,?\,(a_1;a) : (a_2;a) \\ a;(\ell\,?\,a_1 : a_2) & = & \ell\,?\,(a;a_1) : (a;a_2) \\ (\ell\,?\,a_1 : a_2) \sqcup a & = & \ell\,?\,(a_1 \sqcup a) : (a_2 \sqcup a) \\ a \sqcup (\ell\,?\,a_1 : a_2) & = & \ell\,?\,(a \sqcup a_1) : (a \sqcup a_2) \end{array}$$

We also extend the effect ordering to conditional effects. To decide $a_1 \sqsubseteq a_2$, we use an auxiliary relation $\sqsubseteq^h_n$, where h is a set of locks known to be *h*eld by the current thread, and n is a set of locks known *n*ot to be held by the current thread. We define $a_1 \sqsubseteq a_2$ to be $a_1 \sqsubseteq^\emptyset_\emptyset a_2$ and check $a_1 \sqsubseteq^h_n a_2$ recursively as follows:

$$\frac{\kappa_1 \sqsubseteq \kappa_2}{\kappa_1 \sqsubseteq^h_n \kappa_2} \qquad \frac{\ell \notin n \Rightarrow a_1 \sqsubseteq^{h\cup\{l\}}_n a \qquad \ell \notin h \Rightarrow a_2 \sqsubseteq^h_{n\cup\{l\}} a}{\ell\,?\,a_1 : a_2 \sqsubseteq^h_n a} \qquad \frac{\ell \notin n \Rightarrow \kappa \sqsubseteq^{h\cup\{l\}}_n a_1 \qquad \ell \notin h \Rightarrow \kappa \sqsubseteq^h_{n\cup\{l\}} a_2}{\kappa \sqsubseteq^h_n \ell\,?\,a_1 : a_2}$$

A similar notion of ordering was used for conditional atomicities in our previous work on atomicity checkers [21].

5. Type and Effect System

We now formalize our type system for the idealized language YIELDJAVA, a multithreaded subset of Java. This language does not include some Java features, such as primitive types, arrays, inheritance, and interfaces. However, it is sufficient to explore the most salient aspects of reasoning about thread interference. Section 6 describes how our implementation extends this idealized type system to support other Java features.

5.1 Syntax

Figure 3 presents the YIELDJAVA syntax. A program P consists of a sequence of class definitions $\overline{\mathit{defn}}$. Each class definition defn

associates a name with a body containing field and method declarations.

A field declaration includes a class type and a name. Field names are syntactically divided into three categories:

- *Normal* fields are mutable and free of race conditions. They include thread-local fields as well as thread-shared fields that are synchronized, for example, via locks.
- *Final* fields are immutable and thus race-free.
- *Volatile* fields are mutable, and may have concurrent conflicting accesses.

We assume that race freedom for *Normal* fields is verified separately by, for example, a race-free type system [10, 24, 1]. While not permitted in YIELDJAVA, our prototype does support racy accesses to non-`volatile` fields, as described in Section 6.

A method declaration

$$a\ c\ m(\overline{c\ x})\ \texttt{\{}\ e\ \texttt{\}}$$

defines a method m with return type c that takes parameters $\overline{x}$ of type $\overline{c}$. The self-reference variable `this` is implicitly bound to the receiving object in the method body e. The method declaration also includes its effect a, which may include lock expressions referring to any variables in scope, including `this` and $\overline{x}$.

We assume that all programs include a class `Unit`, which has no methods or fields. `Unit` is the type of expressions that produce no meaningful value, such as `while` loops. The language includes the special constant `null`, which has any class type, including `Unit`.

The object allocation expression $\texttt{new}\ c(\overline{e})$ creates a new object of type c and initializes its fields to the values computed from the expression sequence $\overline{e}$. Other YIELDJAVA expressions include field read and update, method calls, variable binding and reference, conditionals, loops, and fork. We also support synchronized blocks $e_1\texttt{.sync}\ e_2$, which are analogous to Java's synchronized statements

$\texttt{synchronized (}e_1\texttt{) \{ }e_2\texttt{ \}}$

Some of these constructs have special forms to document yield points, as summarized by the following table.

Expression Form	Non-Yielding	Yielding
Field Read	$e\texttt{.}f$	$e\texttt{..}f$
Field Write	$e\texttt{.}f = e$	$e\texttt{..}f = e$
Atomic Method Call	$e\texttt{.}m(\overline{e})$	$e\texttt{..}m(\overline{e})$
Non-Atomic Method Call	$e\texttt{.}m\texttt{\#}(\overline{e})$	$e\texttt{..}m\texttt{\#}(\overline{e})$
Lock Synchronization	$e\texttt{.sync}\ e$	$e\texttt{..sync}\ e$

5.2 Type Rules

The YIELDJAVA type system ensures that thread interference is observable only at explicitly annotated yield points. The core of the type system is a set of rules for reasoning about the effect of an expression, as captured by the judgment:

$$P; E \vdash e : c \cdot a$$

Here, e is an expression, c is the type of the expression, and a is an effect describing the behavior of e. The program P is included to provide access to class declarations, and the environment E maps free variables in e to their types:

$$E ::= \epsilon \mid E, c\ x$$

Figure 4 presents the complete set of rules for expressions, as well as additional judgments for reasoning about well-formed environments ($P \vdash E$), types ($P \vdash c$), effects ($P; E \vdash a$), methods ($P; E \vdash meth$), class declarations ($P \vdash defn$), and programs ($P \vdash \text{OK}$). We describe the most important rules defining these judgments:

[EXP VAR] and [EXP NULL] All variables are immutable in YIELDJAVA and thus have cooperability effect AF. That is, a variable access is atomic and yields a constant value. The constant `null` also has effect AF.

[EXP IF] and [EXP WHILE] The effect of a conditional expression is the effect of the *test* expression sequentially composed with the join of the *then* and *else* branches. Similarly, the effect of a while loop $\texttt{while}\ e_1\ e_2$ is the effect of the *test* e_1 composed with the iterative closure $(e_2; e_1)^*$ of the loop body followed by a subsequent evaluation of the test.

[EXP NEW] The object allocation rule first retrieves the definition of the class c from P, and then ensures the arguments $e_{1..n}$ match the types of the fields of c. The effect of the whole expression is the composition of effects of evaluating $e_{1..n}$ composed with the effect AM, reflecting that `new` is not functional (since re-evaluating an object allocation would not return the same object).

[EXP REF] The rule [EXP REF] handles a read $e.f$ of a *Normal* or *Final* field. The rule first checks that e has some type c, and extracts the type d of the field f from P. If f is *Normal* and thus race-free, the effect of accessing the field is AM because the access is guaranteed to commute with steps by other threads. If f is *Final*, the field's value is constant and the effect is AF.

[EXP REF RACE] A racy field read is a non-mover operation, since it may conflict with concurrent accesses by other threads. A racy read $e_\gamma f$ may be annotated with a yield point (if $\gamma =$ "`..`") or not (if $\gamma =$ "`.`"). We use the auxiliary function $[\![\gamma]\!]$ to map γ to the corresponding effect:

$$\begin{array}{rcl} [\![\ ]\!] & : & \mathit{OptYield} \to \mathit{Effect} \\ [\![\texttt{.}]\!] & = & \text{AF} \\ [\![\texttt{..}]\!] & = & \text{CY} \end{array}$$

Thus, if the expression e has effect a, then the non-yielding racy access $e\texttt{.}f$ has effect $(a; \text{AF}; \text{AN})$, whereas the yielding racy access $e\texttt{..}f$ has effect $(a; \text{CY}; \text{AN})$.

[EXP ASSIGN] and [EXP ASSIGN RACE] The rules for field updates are similar to those for field reads, with the additional requirement that *Final* fields cannot be modified.

[EXP SYNC] The type rule for the synchronized statement $\ell_\gamma\texttt{sync}\ e$ first checks that ℓ is a valid lock expression ($P; E \vdash_{\text{lock}} \ell$), meaning that ℓ must have effect AF to guarantee that it always denotes the same lock at run time.

The rule then computes the effect $\mathcal{S}(\ell, \gamma, a)$, where a is the atomicity of e, and γ specifies whether there is a yield point. The function S is defined as follows:

a	$\mathcal{S}(\ell, \gamma, a)$
κ	$\ell\ ?\ \kappa : ([\![\gamma]\!]; \text{AR}; \kappa; \text{AL})$
$\ell\ ?\ a_1 : a_2$	$\mathcal{S}(\ell, \gamma, a_1)$
$\ell'\ ?\ a_1 : a_2$	$\ell'\ ?\ \mathcal{S}(\ell, \gamma, a_1) : \mathcal{S}(\ell, \gamma, a_2)$ if $\ell \neq \ell'$

If the body of the synchronized statement has a basic effect κ and the lock ℓ is already held, then the synchronized statement also has effect κ, since the acquire and release operations are no-ops. Note that in this case the yield operation is ignored, since it is unnecessary.

If the body has effect κ and the lock is not already held, then the synchronized statement has effect $([\![\gamma]\!]; \text{AR}; \kappa; \text{AL})$, since the execution consists of a potential yield point, followed by a right-mover (the acquire), followed by κ (the body), followed by a left-mover (the release).

Figure 4: YIELDJAVA **Type Rules**

$\boxed{P; E \vdash e : c \cdot a}$

[EXP VAR]
$$\frac{P \vdash E \quad E = E_1, c\ x, E_2}{P; E \vdash x : c \cdot \text{AF}}$$

[EXP NULL]
$$\frac{P \vdash E \quad P \vdash c}{P; E \vdash \texttt{null} : c \cdot \text{AF}}$$

[EXP IF]
$$\frac{P; E \vdash e_1 : d \cdot a_1 \quad P; E \vdash e_i : c \cdot a_i \quad \forall i \in 2..3}{P; E \vdash \texttt{if}\ e_1\ e_2\ e_3 : c \cdot (a_1; (a_2 \sqcup a_3))}$$

[EXP WHILE]
$$\frac{P; E \vdash e_1 : c_1 \cdot a_1 \quad P; E \vdash e_2 : c_2 \cdot a_2}{P; E \vdash \texttt{while}\ e_1\ e_2 : \texttt{Unit} \cdot (a_1; (a_2; a_1)^*)}$$

[EXP REF]
$$\frac{P; E \vdash e : c \cdot a \quad \texttt{class}\ c\ \texttt{\{}\ \ldots\ d\ f\ \ldots\ \texttt{\}} \in P \quad f \in \mathit{Normal} \implies a' = \text{AM} \quad f \in \mathit{Final} \implies a' = \text{AF}}{P; E \vdash e.f : d \cdot (a; a')}$$

[EXP ASSIGN]
$$\frac{P; E \vdash e : c \cdot a \quad P; E \vdash e' : d \cdot a' \quad \texttt{class}\ c\ \texttt{\{}\ \ldots\ d\ f\ \ldots\ \texttt{\}} \in P \quad f \in \mathit{Normal}}{P; E \vdash (e.f = e') : d \cdot (a; a'; \text{AM})}$$

[EXP NEW]
$$\frac{\texttt{class}\ c\ \texttt{\{}\ d_i\ x_i\ ^{i \in 1..n}\ \ldots\ \texttt{\}} \in P \quad P; E \vdash e_i : d_i \cdot a_i \quad \forall i \in 1..n}{P; E \vdash \texttt{new}\ c(e_{1..n}) : c \cdot (a_1; \cdots; a_n; \text{AM})}$$

[EXP REF RACE]
$$\frac{P; E \vdash e : c \cdot a \quad \texttt{class}\ c\ \texttt{\{}\ \ldots\ d\ f\ \ldots\ \texttt{\}} \in P \quad f \in \mathit{Volatile}}{P; E \vdash e_\gamma f : d \cdot (a; [\![\gamma]\!]; \text{AN})}$$

[EXP ASSIGN RACE]
$$\frac{P; E \vdash e : c \cdot a \quad P; E \vdash e' : d \cdot a' \quad \texttt{class}\ c\ \texttt{\{}\ \ldots\ d\ f\ \ldots\ \texttt{\}} \in P \quad f \in \mathit{Volatile}}{P; E \vdash (e_\gamma f = e') : d \cdot (a; a'; [\![\gamma]\!]; \text{AN})}$$

[EXP SYNC]
$$\frac{P; E \vdash_{\text{lock}} \ell \quad P; E \vdash e : c \cdot a}{P; E \vdash \ell_\gamma \texttt{sync}\ e : c \cdot \mathcal{S}(\ell, \gamma, a)}$$

[EXP INVOKE ATOMIC]
$$\frac{P; E \vdash e : c \cdot a \quad \texttt{class}\ c\ \texttt{\{}\ \ldots\ \mathit{meth}\ \ldots\ \texttt{\}} \in P \quad \mathit{meth} = a'\ c'\ m(d_i\ x_i\ ^{i \in 1..n})\ \texttt{\{}\ e'\ \texttt{\}} \quad P; E \vdash e_i : d_i \cdot a_i \quad \forall i \in 1..n \quad P; E \vdash a'[\texttt{this} := e, x_i := e_i\ ^{i \in 1..n}] \uparrow a'' \quad a'' \sqsubseteq \text{AN}}{P; E \vdash e_\gamma m(e_{1..n}) : c' \cdot (a; a_1; \cdots; a_n; [\![\gamma]\!]; a'')}$$

[EXP INVOKE COMPOUND]
$$\frac{P; E \vdash e : c \cdot a \quad \texttt{class}\ c\ \texttt{\{}\ \ldots\ \mathit{meth}\ \ldots\ \texttt{\}} \in P \quad \mathit{meth} = a'\ c'\ m(d_i\ x_i\ ^{i \in 1..n})\ \texttt{\{}\ e'\ \texttt{\}} \quad P; E \vdash e_i : d_i \cdot a_i \quad \forall i \in 1..n \quad P; E \vdash a'[\texttt{this} := e, x_i := e_i\ ^{i \in 1..n}] \uparrow a''}{P; E \vdash e_\gamma m\#(e_{1..n}) : c' \cdot (a; a_1; \cdots; a_n; [\![\gamma]\!]; a'')}$$

[EXP FORK]
$$\frac{P; E \vdash e : c \cdot a}{P; E \vdash \texttt{fork}\ e : \texttt{Unit} \cdot \text{AL}}$$

[EXP LET]
$$\frac{P; E \vdash e_1 : c_1 \cdot a_1 \quad P; E, c_1\ x \vdash e_2 : c_2 \cdot a_2 \quad P; E \vdash a_2[x := e_1] \uparrow a_2'}{P; E \vdash \texttt{let}\ x\ \texttt{=}\ e_1\ \texttt{in}\ e_2 : c_2 \cdot (a_1; a_2')}$$

$\boxed{P \vdash E}$

[ENV EMPTY]
$$\frac{}{P \vdash \epsilon}$$

[ENV X]
$$\frac{P \vdash E \quad P \vdash c \quad x \notin \mathit{dom}(E)}{P \vdash (E, c\ x)}$$

$\boxed{P \vdash c}$

[CLASS NAME]
$$\frac{\texttt{class}\ c\ \texttt{\{}\ \ldots\ \texttt{\}} \in P}{P \vdash c}$$

$\boxed{P; E \vdash a}$

[AT BASE]
$$\frac{P \vdash E}{P; E \vdash \kappa}$$

[AT COND]
$$\frac{P; E \vdash_{\text{lock}} \ell \quad P; E \vdash a_i \quad \forall i \in 1..2}{P; E \vdash \ell\,?\,a_1 : a_2}$$

$\boxed{P; E \vdash_{\text{lock}} e}$

[LOCK EXP]
$$\frac{P; E \vdash e : c \cdot \text{AF}}{P; E \vdash_{\text{lock}} e}$$

$\boxed{P; E \vdash a \uparrow a'}$

[LIFT BASE]
$$\frac{P \vdash E}{P; E \vdash \kappa \uparrow \kappa}$$

[LIFT GOOD LOCK]
$$\frac{P; E \vdash_{\text{lock}} \ell \quad P; E \vdash a_i \uparrow a_i' \quad \forall i \in 1..2}{P; E \vdash (\ell\,?\,a_1 : a_2) \uparrow (\ell\,?\,a_1' : a_2')}$$

[LIFT BAD LOCK]
$$\frac{P; E \not\vdash_{\text{lock}} \ell \quad P; E \vdash a_i \uparrow a_i' \quad \forall i \in 1..2}{P; E \vdash (\ell\,?\,a_1 : a_2) \uparrow (a_1' \sqcup a_2')}$$

$\boxed{P; E \vdash \mathit{meth}}$

[METHOD]
$$\frac{P; E, \overline{d\ x} \vdash e : c \cdot a' \quad P; E, \overline{d\ x} \vdash a \quad a' \sqsubseteq a}{P; E \vdash a\ c\ m(\overline{d\ x})\ \texttt{\{}\ e\ \texttt{\}}}$$

$\boxed{P \vdash \mathit{defn}}$

[CLASS]
$$\frac{\mathit{field}_i = d_i\ f_i \quad \forall i \in 1..m \quad P \vdash d_i \quad \forall i \in 1..m \quad P; c\ \texttt{this} \vdash \mathit{meth}_i \quad \forall i \in 1..n}{P \vdash \texttt{class}\ c\ \texttt{\{}\ \mathit{field}_{1..m}\ \mathit{meth}_{1..n}\ \texttt{\}}}$$

$\boxed{P \vdash \text{OK}}$

[PROGRAM]
$$\frac{P = \mathit{defn}_{1..n} \quad P \vdash \mathit{defn}_i \quad \forall i \in 1..n \quad \mathit{ClassesOnce}(P)\ \ \mathit{FieldsOnce}(P)\ \ \mathit{MethodsOnce}(P)}{P \vdash \text{OK}}$$

If the body has conditional effect $\ell\,?\,a_1 : a_2$, where ℓ is the lock being acquired by this synchronized statement, then we ignore a_2 and recursively apply S to a_1, since ℓ is held within the synchronized body.

If the body has an effect that is conditional on some other lock ℓ', then we recursively apply S to both branches.

[EXP LET] This rule for `let` x `=` e_1 `in` e_2 infers effects a_1 and a_2 for e_1 and e_2, respectively. Care must be taken when constructing the effect for this expression because a_2 may refer to the let-bound variable x.

For example, the body of the following `let` expression produces an effect that is conditional on whether the lock `x` is held.

```
let x = e1 in
  x.sync ...
```

Thus, we apply the substitution $[x := e_1]$ to yield a corresponding effect $a_2[x := e_1]$ that does not mention x. However, e_1 may not have effect AF, in which case $a_2[x := e_1]$ may not be a valid effect (because it could contain e_1 as part of a non-constant lock expression). As in our previous work [21], we use the judgment

$$P; E \vdash a_2[x := e_1] \uparrow a_2'$$

to *lift* the effect $a_2[x := e_1]$ to a well-formed effect a_2' that is greater than or equal to $a_2[x := e_1]$.

This "lifting" judgment is defined in Figure 4:

[LIFT BASE] Basic effects are always well-formed and remain unchanged when lifted.

[LIFT LOCK WELL-FORMED] If a conditional effect refers to a well-formed lock, this rule recursively lifts the two component effects.

[LIFT LOCK ILL-FORMED] If a conditional effect refers to an ill-formed lock expression, this rule removes the dependency on this lock by joining together the two recursively-lifted component effects.

[EXP INVOKE ATOMIC] This rule handles calls to atomic methods. The rule first extracts the appropriate method signature from P based on the receiver's type, and it verifies that the actual arguments match the declared parameter types. Finally, the rule computes the cooperability effect of the invocation. The method's specified effect a' may refer to `this` or the parameter names $x_{1..n}$. Therefore, we substitute

- the actual receiver e for `this`, and
- the actual arguments $e_{1..n}$ for the parameters $x_{1..n}$

to produce the effect $a'[\texttt{this} := e, x_i := e_i \ ^{i \in 1..n}]$, and ensure that the resulting effect is valid by lifting it to an effect a'' that is well-formed in the current environment. This effect a'' must be an atomic effect and less than or equal to AN.

Like racy field accesses, a method invocation may be labeled with a yield point. Thus, the overall effect also includes $[\![\gamma]\!]$.

[EXP INVOKE COMPOUND] This rule applies to a method invocation $e_\gamma m\texttt{\#(}e_{1..n}\texttt{)}$ to a non-atomic method. It is similar to [EXP INVOKE ATOMIC], but removes the requirement that the computed effect a'' be atomic.

[EXP FORK] A fork expression `fork` e creates a new thread to evaluate e. Since a fork operation cannot commute past the operations of its child thread, fork operations are left movers and thus never start new transactions.

[METHOD], [CLASS], and [PROG] These rules verify the basic well-formedness requirements of methods, classes, and programs. The [PROG] rule uses the following additional predicates. (See [23] for their precise definition.)

- *ClassesOnce*(P): no class is declared twice in P.
- *FieldsOnce*(P): no field name is declared twice in a class.
- *MethodsOnce*(P): no method name is declared twice in a class.

5.3 Correctness

The extended version of this paper [50] presents a formal semantics for YIELDJAVA. The semantics is defined as a transition relation over run-time states, where a run-time state σ contains a program's class definitions, dynamically allocated objects, and dynamically created threads. This semantics satisfies the standard preservation property, whereby evaluation preserves well-typing of the run-time state. We then define both a *preemptive* and *cooperative* semantics for program behavior. The preemptive semantics interleaves the instructions of the various threads at instruction-level granularity, essentially modeling the behavior of a preemptive scheduler. The cooperative semantics performs context switches between threads only at explicitly marked yield points in the style of cooperative multitasking [3, 4, 9].

The central correctness result for this type system is that well-typed programs behave equivalently under both semantics. That is, if a well-typed program P can reach a final state σ under the preemptive semantics, then it can also reach that final state under the cooperative semantics. Therefore, it is sufficient to reason about the correctness of well-typed programs under the simpler cooperative semantics, since this correctness result also applies to executions under the preemptive semantics (and consequently, to executions on multicore hardware).

6. Implementation

We have developed an implementation called JCC that extends the YIELDJAVA type system to support the Java language.

JCC uses the standard Java field modifiers `final` and `volatile` to classify fields as either *Final* or *Volatile*; all other fields are considered *Normal*. We also introduce one new modifier, `racy`, to capture intentionally racy *Normal* fields. Our implementation assumes that correct field annotations are provided for the input program. Such annotations could be generated using RCC/JAVA [1] or any other analysis technique. For our experiments, we leveraged both that tool, as well as the FASTTRACK [20] race detector, to identify racy fields.

JCC supports annotations on methods to describe their effects. The following three keywords are sufficient to annotate most methods:

- `atomic`: an atomic non-mover method with effect AN.
- `mover`: an atomic both-mover method with effect AM.
- `compound`: a compound non-mover method with effect CN.

These effect annotations appear alongside the standard modifiers for a method, as in:

```
atomic public void m() { ... }
```

They may also be combined to form conditional effects, such as `(this ? mover : compound)`.

To further reduce the burden of annotating methods with cooperability effects, JCC uses carefully chosen defaults when annotations are absent. In essence, it assumes:

- fields are race free and

Access Type	Syntax	Effect
racy read	$e_1..f\#$	a_1; CL
racy write	$e_1..f\# = e_2$	$a_1; a_2$; CL
write-guarded read (lock held)	$e_1.f$	a_1; AM
write-guarded read (lock not held)	$e_1.f$	a_1; AN
write-guarded write (lock held)	$e_1.f = e_2$	$a_1; a_2$; AN
race-free array read	$e_1[e_2]$	$a_1; a_2$; AM
race-free array write	$e_1[e_2] = e_3$	$a_1; a_2; a_3$; AM
racy array read	$e_1[e_2]\#$	$a_1; a_2$; CL
racy array write	$e_1[e_2]\# = e_3$	$a_1; a_2; a_3$; CL

Figure 5. Effects of additional operations (a_i is the effect of e_i).

- all methods are atomic both-movers.

We also permit more expressive (but more verbose) annotations to describe all elements of the conditional effect lattice when the keywords are not sufficient. The following illustrates the full syntax for effect annotations:

```
atomic non-mover void m1() { ... }

this.f ? (atomic functional) : (compound non-mover)
void m2() { ... }
```

As in YIELDJAVA, field accesses and method invocations may be written using "`..`" in place of "`.`" to indicate interference points. Yielding synchronized statements use the syntax

```
..synchronized(e) { ... }
```

The JCC checker verifies that these yield points characterize all possible thread interference. It reports a warning whenever interference may occur at a program point not corresponding to a yield, or if a method's specification is not satisfied by its implementation.

The remainder of this section describes how JCC extends the YIELDJAVA type system to support features of Java programs including subtyping, intentional races, write-guarded data, and arrays[1].

Subtyping and Covariant Cooperability Specifications. One major extension to the presented type system is the support for inheritance and subtyping. We permit the cooperability effect of the method to change covariantly, intuitively requiring, for example, that $b \sqsubseteq a$ in the following class definitions:

```
class C {                    class D extends C {
  a t m() { ... }              b t m() { ... }
}                            }
```

Fields with Races. Although data races should in general be avoided when possible, large programs often have some intentional races, which JCC supports via a `racy` annotation on *Normal* fields. A read from a racy field must be written as $e..f\#$. Here, the double dots as usual indicate a yield point, and the trailing `#` identifies the racy nature of the read (and that the programmer needs to consider the consequences of Java's relaxed memory model [36]). The overall effect of $e..f\#$ is the composition of a yield and a non-mover memory access: $(\text{CY}; \text{AN}) = \text{CL}$. Writes are handled similarly. The rules for computing the effects of these operations, and those discussed below are summarized in Figure 5.

[1] Our implementation does not currently support generic classes due to limitations in the front-end checker upon which JCC is built, but supporting generic types would not be fundamentally problematic.

Write-Guarded Fields. YIELDJAVA also supports write-guarded fields (such as the `shortestPathLength` field in Section 1) for which a protecting lock is held for all writes but not necessarily for reads. For such fields, a read while holding the protecting lock is a both-mover, since there can be no concurrent writes. However, a write with the lock held is still a non-mover, since there may be concurrent reads (that do not hold the lock).

Arrays. The YIELDJAVA checker handles array accesses in a way analogous to *Normal* fields[2]. Racy array accesses must be annotated with "`#`" and are assumed to be yield points.)

7. Experimental Evaluation

To evaluate its effectiveness, we applied JCC to a variety of benchmark programs, including:

- a number of standard library classes from Java 1.4.2_19: namely `Inflater`, `Deflater`, `StringBuffer`, `String`, `PrintWriter`, `Vector`, and `ZipFile`;
- `sparse`, `raytracer`, `sor`, and `moldyn` from the Java Grande suite [31];
- `tsp`, a solver for the traveling salesman problem [47];
- and `elevator`, a real-time discrete event simulator [47].

These programs use a variety of synchronization idioms, and previous work has revealed a number of interesting concurrency bugs in these programs. Thus, they show the ability of our annotations to capture thread interference under various conditions and to highlight unintended, problematic interference. Three of these programs (`raytracer`, `sor`, and `moldyn`) use broken barrier implementations [20]. We discuss those problems below and use versions with corrected barrier code (named `raytracer-fixed`, `sor-fixed`, and `moldyn-fixed`) in our experiments.

All experiments were performed on a 2 GHz dual-core computer with 3 GB memory, using the Java HotSpot 64-bit Server VM, version 1.6.0_24. The JCC checker was able to analyze each of these benchmarks in under 2 seconds.

Figure 6 shows the size of each benchmark program, the time required to manually insert the JCC annotations into each program, and the number of annotations required to enable successful type checking. This count includes all `racy` field annotations, method specifications, and occurrences of "`..`" and "`#`". Even for programs comprising several thousand lines, the annotation burden is quite low. Each program was annotated and checked in about 10 to 30 minutes, and roughly one annotation per 30 lines of code was required. We did have some previous experience using these programs, which facilitated the annotation process, but since we intend JCC to be used during development, we believe our experience reflects the cost incurred by the intended use of our technique.

7.1 Precision of Thread Interference Annotations

Our experiments demonstrate that cooperability annotations serve as clear documentation of where interference may occur, thereby simplifying the complex task of reasoning about program behavior. To quantify this in one specific dimension, we consider the question

> "How many potential interference points must a programmer consider in a program annotated with various forms of non-interference specifications?"

Each specification form provides a particular semantic guarantee about where interference may occur, and we believe that isolating

[2] Note that in Java, it is not possible to indicate that elements of an array are `final` or `volatile`.

Program	Size (lines)	Annot. Time (min.)	Annot. Count	Interference Points					Unintended Yields
				Preemptive	Race	Atomic	AtomRace	Cooperative	
java.util.zip.Inflater	317	9	4	36	12	0	0	0	0
java.util.zip.Deflater	381	7	8	49	13	0	0	0	0
java.lang.StringBuffer	1,276	20	10	210	81	9	2	1	1
java.lang.String	2,307	15	5	230	87	6	2	1	0
java.io.PrintWriter	534	40	109	73	99	130	97	26	9
java.util.Vector	1,019	25	43	185	106	44	24	4	1
java.util.zip.ZipFile	490	30	62	120	105	85	53	30	0
sparse	868	15	19	329	98	48	14	6	0
tsp	706	10	45	445	115	437	80	19	0
elevator	1,447	30	64	454	146	241	60	25	0
raytracer-fixed	1,915	10	50	565	200	105	39	26	2
sor-fixed	958	10	32	249	99	128	24	12	0
moldyn-fixed	1,352	10	39	983	130	657	37	30	0
Total	13,570	231	490	3,928	1,291	1,890	432	180	13

Figure 6. Interference Points and Unintended Yields

interference to as few program points as possible facilitates reasoning about code. The five different specifications (or semantic guarantees) we consider are as follows:

- **Preemptive**: If a program has no synchronization specification, interference must be assumed to possibly occur on any access to a field or any lock acquire, since these operations may conflict with operations of concurrent threads. We exclude operations that do not cause interference, such as accesses to method-local variables, lock releases, method calls, etc. from this count.
- **Race**: If a program's specification distinguishes race-free fields from potentially racy fields, interference may be assumed to occur only on any access to a racy field or any lock acquire.
- **Atomic**: If a program's specification distinguishes atomic methods from non-atomic methods, thread interference may be considered to occur only in non-atomic methods. These interference points include field accesses, lock acquires, and also calls to an atomic method from a non-atomic context.
- **AtomRace**: If a program specification distinguishes both racy fields and atomic methods, interference may be considered to occur in non-atomic methods at racy accesses, lock acquires, and calls to atomic methods.
- **Cooperative**: If a program specification identifies yield points, interference may be considered to occur only at those explicit yield points.

Figure 6 shows the number of interference points in each benchmark under each semantics. Benchmarks in which all methods are atomic, such as `Inflater`, have zero interference points under both atomic and cooperative semantics.

Overall, the results show that information about race conditions and atomic methods provides significant benefits over the Preemptive column. In particular, the number of interference points drops from 3,928 under the Preemptive column (which assumes no non-interference information) to 432 under AtomRace (which essentially characterizes prior techniques based on method-level atomicity and race condition information).

As shown in the "Cooperative" column, our cooperative type and effect system further reduces the number of possible interference points from 432 to 180, essentially because this analysis reasons more precisely about where thread interference may occur.

We sketch two situations illustrating why cooperative reasoning is significantly more precise than AtomRace. First, for the TSP algorithm from Figure 1, AtomRace requires an interference point before each call to the atomic methods `path.isComplete()` and `path.children()` from within a non-atomic method. In contrast, our type system identifies that these two methods are more precisely characterized as both movers (which do not interfere with other threads), and so no yield point is necessary at these calls.

Figure 7: StringBuffer

```
public final class StringBuffer ··· {
  (this?mover:atomic) int length() { ... }
  (this?mover:atomic) void getChars(...) { ... }

  compound
  synchronized StringBuffer append(StringBuffer sb) {
    ...
    int len = sb.length();
    int newcount = count + len;
    if (newcount > value.length)
      expandCapacity(newcount);
    sb..getChars(0, len, value, count);
    count = newcount;
    return this;
  }
}
```

As a second example, consider a non-atomic method that contains two nested synchronized blocks. Under AtomRace, both acquires are an interference point. Under our analysis, both acquires are right-movers and so no yield is required before the second.

The method `StringBuffer.append()` in Figure 7 illustrates this kind of situation. The `append()` method acquires the `this` lock and then calls `sb.length()`, which is atomic (since the lock `sb` is not held). The lock acquire of `this` can move right to just before the `sb.length()` call, so no yield is required at that call. Conversely, a yield *is* required for the call `sb..getChars()`, since it is preceded by actions (such as the call to `sb.length()`) that are not right movers. This yield at `sb..getChars()` highlights that `append()` is `compound`, and may behave erroneously if `sb` is concurrently modified by other threads.

These two examples illustrate why the notions of race conditions and atomic methods are not by themselves sufficient to identify interference points in a precise manner, and the experimental results show that the cooperative type and effect system is significantly more precise in its ability to verify interference points.

Figure 8: RayTracer

```
class RayTracerRunner implements Runnable {
  int id;

  compound public void run() {
    // init
    br.DoBarrier#(id);
    // render

    ..synchronized (scene) {
      for(int i = 0; i < JGFRayTracerBench.nthreads; i++)
        if (id == i)
          JGFRayTracerBench..checksum1 =
            JGFRayTracerBench..checksum1 + checksum;
    }

    br.DoBarrier#(id);
    // cleanup
  }
}
```

Figure 9: Vector

```
interface Collection {
  (this ? mover : atomic) int size();
  (this ? mover : atomic) Object[] toArray(Object a[]);
}
class Vector {
  protected Object elementData[];
  protected int elementCount;

  compound public Vector(Collection c) {
    elementCount = c..size();
    elementData =
      new Object[(int)Math.min( (elementCount*110L)/100,
                                Integer.MAX_VALUE )];
    c..toArray(elementData);
  }
}
```

Figure 10: Original SOR Algorithm

```
class SORRunner implements Runnable {
  double G[][];
  volatile long sync[][];

  compound public void run() {
    ...
      for (int p = 0; p < 2*num_iterations; p++) {
        for (int i = ilow + (p%2); i < iupper; i=i+2) {
          ...
          for (int j=1; j < Nm1; j = j+2){

            G[i][j]# = omega_over_four *
              ( G[i-1][j]# + G[i+1][j]# +
                G[i][j-1]# + G[i][j+1]# ) +
              one_minus_omega * G[i][j]#;
          ...
          }
        }

        sync[id][0]# = sync[id][0]# + 1;
        if (id > 0)
          while (sync[id-1][0]# < sync[id][0]#) ;
        if (id < JGFSORBench.nthreads-1)
          while (sync[id+1][0]# < sync[id][0]#) ;
      }
  }
}
```

7.2 Identifying Defects

The final column in Figure 6 shows the number of *unintended yield point* in each program. These are program points that suffer from thread interference in ways that we determined are unintentional or possibly damaging based on manual code inspection. These unintended yields highlight concurrency bugs, such as atomicity violations and data races [18, 21]. We illustrate how JCC enables the programmer to identify several concurrency bugs in our benchmarks.

RayTracer. The raytracer benchmark uses a barrier `br` to coordinate several rendering threads, as shown in Figure 8. Although the initial barrier code was incorrect, it did not cause other unintended yields. After rendering, each thread acquires the lock `scene` before adding its local `checksum` to the global shared variable `checksum1`. However, each thread creates its own `scene` object, and thus acquiring `scene` fails to ensure mutual exclusion over the updates to `checksum1`. This is made clear by the explicit yields on the reads and writes of `checksum1`. [3]

Vector. A `Vector` constructor (see Figure 9) takes as argument a collection `c` and invokes `c`'s `size` method, allocates an array of that size, and copies elements from `c` into the array. The two yield annotations highlight that `c` may be modified by a concurrent thread in between requesting the size and copying the data, potentially resulting in an incorrectly initialized `Vector` or an `ArrayIndexOutOfBounds` exception. Similar pitfalls for the methods `removeAll(c)` and `retainAll(c)` [21] are also caught by JCC.

SOR. In the `sor` benchmark (see Figure 10), the computation threads synchronize on a barrier implemented as a shared two-dimensional array `sync`. Unfortunately, the barrier is broken, since the `volatile` keyword applies only to the array reference, not the array elements. Thus, the barrier synchronization code at the bottom of the main processing loop may not properly coordinate the threads, leading to races on the data array `G`. This problem is obvious when using JCC because yield annotations must be added in dozens of places, essentially to all accesses of `sync` and `G`.

When the barrier is fixed, we obtain much cleaner code: the yield count decreases from 40 to 12. In particular, the accesses to `G` between barrier calls are free of yields, signifying that between barriers, sequential reasoning is applicable. Figure 6 includes data for `sor-fixed`, the corrected version of the benchmark, which we believe is more representative of multithreaded Java programs.

Moldyn. The `moldyn` benchmark uses a barrier object to synchronize the actions of multiple computation threads. The barrier object maintains an array

```
volatile boolean[] IsDone;
```

to record which threads are currently waiting at the barrier, but the elements of the array are prone to race conditions because again the `volatile` keyword applies only to the array reference and not the array elements. This bug leads to potential races on all data accesses intended to be synchronized by the barrier, and a large number (58) of yield annotations were necessary to document all such cases. Again, we report on the corrected version `moldyn-fixed` in Figure 6.

[3] We also note that if JCC were extended to identify locks used only by a single thread, we could remove the yield on the `synchronized` operation.

8. Related Work

Cooperability. *Cooperative multithreading* is a thread execution model in which context switching between threads may occur only at `yield` statements [3, 4, 9]. That is, cooperative multithreading allows concurrency but disallows parallel execution of threads. In contrast, cooperability guarantees behavior equivalent to cooperative multithreading, but actually allows execution in a preemptive manner, enabling full use of modern multicore hardware.

Automatic mutual exclusion is an execution model that proposes ensuring mutual exclusion by default [30]; `yield` statements demarcate where thread interference is permitted. A critical difference is that these `yield` statements are enforced at run time to provide serializability via transactional memory techniques; in contrast, the JCC checker guarantee serializability statically.

In prior work, we explored a type and effect system for cooperability [51], and dynamic analyses for checking cooperability and inferring yield annotations for legacy programs [52]. Others have explored *task types*, a data-centric approach to obtaining *pervasive atomicity* [32], a notion that is closely related to cooperability.

Race Freedom. A data race occurs when two threads simultaneously access a shared variable without synchronization, and at least one thread writes to that variable. Data races often reflect problems in synchronization, and expose a weak memory model to programmers, compromising software reliability. Data race freedom remedies this issue by guaranteeing behavior equivalent to executing with sequentially consistent [2].

The JCC checker relies on a race analysis to properly annotate racy variables, used as input to the cooperability analysis. There is extensive literature on how to find and fix data races efficiently. Dynamic race detectors may track the happens-before relation [20], implement the lockset algorithm [44], or combine both [38]. Sampling techniques are also used to make race detection more lightweight [8, 16]. Static race detectors may make use of a type system [10, 1], implement a static lockset algorithm [37, 40], or use model checking [41].

Atomicity and Transactional Memory. An atomic block is a lexically-scoped annotation that declares the sequence of instructions in that block to be free of thread interference. Atomicity and transactional memory both focus on ensuring that atomic blocks execute in a serializable manner, thus enjoying freedom from thread interference.

Atomicity is an analysis approach that checks if atomic blocks are serializable. Both static [21, 26, 48] and dynamic tools [19, 49, 22, 17] have been developed to check atomicity. *Transactional memory* enforces serializability of atomic blocks (or transactions) at run time. Both hardware [27, 14] and software [45, 25] implementation techniques have been developed, and the semantics of transactional memory have also been explored in depth [6].

While atomic blocks are clearly beneficial to program reasoning, one must still ascertain whether a piece of code is inside some atomic block to enjoy freedom from thread interference, hence introducing a form of bi-modal reasoning [51]. Furthermore, the regions of code outside atomic blocks are still subject to unconstrained preemptive scheduling, with all the traditional problems such schedulings pose.

Other Properties. *Deterministic parallelism* guarantees that the result of executing multiple threads is invariant across thread schedulings. There are various approaches to obtaining deterministic parallelism: static analyses [7], dynamic analyses [42, 12], as well as run-time enforcement [39, 13].

Linearizability, a popular correctness criterion, guarantees that the concurrent calls to a shared object execute atomically and satisfy a sequential specification [28, 46]. Shape analysis [5] and abstract interpretation [46] have been used to prove linearizability for small programs, while model checking has been used to refute linearizability [11].

9. Summary

Reasoning about the correctness of multithreaded software is notoriously difficulty under the preemptive semantics provided by multiprocessor and multicore architectures. This paper proposes a more modular approach for reasoning about multithreaded software correctness.

Under our approach, software is written using traditional synchronization idioms such as locks, but the programmer also explicitly documents intended sources of thread interference, which we refer to as yield points. The type system of this paper then verifies that these annotations identify all possible situations where interference may occur. Consequently, any well-typed program behaves *as if* it is executing under a cooperative semantics where context switches between threads happen only at yield points.

This cooperative semantics provides a much nicer foundation for subsequent reasoning about program behavior and correctness. In particular, intuitive sequential reasoning is now valid, except at explicitly marked yield points. One interesting avenue of future work would be to incorporate cooperative reasoning into a proof system, such as rely-guarantee reasoning. Another would be to extend cooperability to reason about determinism properties. Finally, the formal system described here could be adapted to other languages, such as C++ or X10.

Acknowledgments. This work was supported by NSF grants CNS-0905650, CCF-1116883 and CCF-1116825.

References

[1] M. Abadi, C. Flanagan, and S. N. Freund. Types for safe locking: Static race detection for Java. *Transactions on Programming Languages and Systems (TOPLAS)*, 28(2):207–255, 2006.

[2] S. V. Adve and K. Gharachorloo. Shared memory consistency models: A tutorial. *IEEE Computer*, 29(12):66–76, 1996.

[3] A. Adya, J. Howell, M. Theimer, W. J. Bolosky, and J. R. Douceur. Cooperative task management without manual stack management. In *Annual Technical Conference*. USENIX Association, 2002.

[4] R. M. Amadio and S. D. Zilio. Resource control for synchronous cooperative threads. In *International Conference on Concurrency Theory (CONCUR)*, 2004.

[5] D. Amit, N. Rinetzky, T. Reps, M. Sagiv, and E. Yahav. Comparison under abstraction for verifying linearizability. In *Computer Aided Verification (CAV)*, 2007.

[6] C. Blundell, E. Christopher, L. Milo, and M. K. Martin. Subtleties of transactional memory atomicity semantics. *Computer Architecture Letters*, 5:2006, 2006.

[7] R. L. Bocchino, Jr., V. S. Adve, D. Dig, S. Adve, S. Heumann, R. Komuravelli, J. Overbey, P. Simmons, H. Sung, and M. Vakilian. A type and effect system for Deterministic Parallel Java. In *Object-Oriented Programming, Systems, Languages, and Applications (OOPSLA)*, 2009.

[8] M. D. Bond, K. E. Coons, and K. S. McKinley. PACER: Proportional detection of data races. In *Conference on Programming Language Design and Implementation (PLDI)*, 2010.

[9] G. Boudol. Fair cooperative multithreading. In *International Conference on Concurrency Theory (CONCUR)*, 2007.

[10] C. Boyapati and M. Rinard. A parameterized type system for race-free Java programs. In *Object-Oriented Programming, Systems, Languages, and Applications (OOPSLA)*, pages 56–69, 2001.

[11] S. Burckhardt, C. Dern, M. Musuvathi, and R. Tan. Line-Up: a complete and automatic linearizability checker. In *Conference on Programming Language Design and Implementation (PLDI)*, 2010.

[12] J. Burnim and K. Sen. Asserting and checking determinism for multithreaded programs. In *International Symposium on Foundations of Software Engineering (FSE)*, 2009.

[13] J. Devietti, B. Lucia, L. Ceze, and M. Oskin. DMP: deterministic shared memory multiprocessing. In *International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*. ACM, 2009.

[14] D. Dice, Y. Lev, M. Moir, and D. Nussbaum. Early experience with a commercial hardware transactional memory implementation. In *International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, 2009.

[15] J. L. Eppinger, L. B. Mummert, and A. Z. Spector. *Camelot and Avalon: A Distributed Transaction Facility*. Morgan Kaufmann, 1991.

[16] J. Erickson, M. Musuvathi, S. Burckhardt, and K. Olynyk. Effective data-race detection for the kernel. In *Operating Systems Design and Implementation (OSDI)*, 2010.

[17] A. Farzan and P. Madhusudan. Monitoring atomicity in concurrent programs. In *Computer Aided Verification (CAV)*, 2008.

[18] C. Flanagan and S. Freund. Adversarial memory for detecting destructive races. In *Conference on Programming Language Design and Implementation (PLDI)*, 2010.

[19] C. Flanagan and S. N. Freund. Atomizer: A dynamic atomicity checker for multithreaded programs. In *Symposium on Principles of Programming Languages (POPL)*, 2004.

[20] C. Flanagan and S. N. Freund. Fasttrack: efficient and precise dynamic race detection. *Commun. ACM*, 53(11):93–101, 2010.

[21] C. Flanagan, S. N. Freund, M. Lifshin, and S. Qadeer. Types for atomicity: Static checking and inference for Java. *Transactions on Programming Languages and Systems (TOPLAS)*, 30(4):1–53, 2008.

[22] C. Flanagan, S. N. Freund, and J. Yi. Velodrome: A sound and complete dynamic atomicity checker for multithreaded programs. In *Conference on Programming Language Design and Implementation (PLDI)*, 2008.

[23] M. Flatt, S. Krishnamurthi, and M. Felleisen. Classes and mixins. In *Symposium on Principles of Programming Languages (POPL)*, pages 171–183, 1998.

[24] D. Grossman. Type-safe multithreading in Cyclone. In *Workshop on Types in Language Design and Implementation (TLDI)*, 2003.

[25] T. Harris and K. Fraser. Language support for lightweight transactions. In *Object-Oriented Programming, Systems, Languages, and Applications (OOPSLA)*, 2003.

[26] J. Hatcliff, Robby, and M. B. Dwyer. Verifying atomicity specifications for concurrent object-oriented software using model-checking. In *International Conference on Verification, Model Checking, and Abstract Interpretation*, 2004.

[27] M. Herlihy and J. E. B. Moss. Transactional memory: architectural support for lock-free data structures. In *International Symposium on Computer Architecture (ISCA)*, 1993.

[28] M. P. Herlihy and J. M. Wing. Linearizability: A correctness condition for concurrent objects. *Transactions on Programming Languages and Systems (TOPLAS)*, 12(3):463–492, 1990.

[29] C. A. R. Hoare. Towards a theory of parallel programming. In *Operating Systems Techniques*, volume 9 of *A.P.I.C. Studies in Data Processing*, pages 61–71, 1972.

[30] M. Isard and A. Birrell. Automatic mutual exclusion. In *Workshop on Hot Topics in Operating Systems (HOTOS)*, 2007.

[31] Java Grande Forum. Java Grande benchmark suite. Available at `http://www.javagrande.org`, 2008.

[32] A. Kulkarni, Y. D. Liu, and S. F. Smith. Task types for pervasive atomicity. In *Object-Oriented Programming, Systems, Languages, and Applications (OOPSLA)*, 2010.

[33] R. J. Lipton. Reduction: A method of proving properties of parallel programs. *Communications of the ACM*, 18(12):717–721, 1975.

[34] B. Liskov, D. Curtis, P. Johnson, and R. Scheifler. Implementation of Argus. In *Proceedings of the Symposium on Operating Systems Principles*, pages 111–122, 1987.

[35] D. B. Lomet. Process structuring, synchronization, and recovery using atomic actions. *Language Design for Reliable Software*, pages 128–137, 1977.

[36] J. Manson, W. Pugh, and S. V. Adve. The Java memory model. In *Symposium on Principles of Programming Languages (POPL)*, 2005.

[37] M. Naik, A. Aiken, and J. Whaley. Effective static race detection for Java. In *Conference on Programming Language Design and Implementation (PLDI)*, 2006.

[38] R. O'Callahan and J.-D. Choi. Hybrid dynamic data race detection. In *Symposium on Principles and Practice of Parallel Programming (PPoPP)*, 2003.

[39] M. Olszewski, J. Ansel, and S. Amarasinghe. Kendo: efficient deterministic multithreading in software. In *International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, 2009.

[40] P. Pratikakis, J. S. Foster, and M. Hicks. Context-sensitive correlation analysis for detecting races. In *Conference on Programming Language Design and Implementation (PLDI)*, 2006.

[41] S. Qadeer and D. Wu. Kiss: keep it simple and sequential. In *Conference on Programming Language Design and Implementation (PLDI)*, 2004.

[42] C. Sadowski, S. N. Freund, and C. Flanagan. SingleTrack: A dynamic determinism checker for multithreaded programs. In *European Symposium on Programming (ESOP)*, 2009.

[43] C. Sadowski and J. Yi. Applying usability studies to correctness conditions: A case study of cooperability. In *Onward! Workshop on Evaluation and Usability of Programming Languages and Tools (PLATEAU)*, 2010.

[44] S. Savage, M. Burrows, G. Nelson, P. Sobalvarro, and T. E. Anderson. Eraser: A dynamic data race detector for multi-threaded programs. *ACM Transactions on Computer Systems (TOCS)*, 15(4), 1997.

[45] N. Shavit and D. Touitou. Software transactional memory. In *ACM SIGACT-SIGOPS Symposium on Principles of Distributed Computing (PODC)*, 1995.

[46] V. Vafeiadis. Automatically proving linearizability. In *Computer Aided Verification (CAV)*, 2010.

[47] C. von Praun and T. Gross. Static conflict analysis for multi-threaded object-oriented programs. In *Conference on Programming Language Design and Implementation (PLDI)*, 2003.

[48] C. von Praun and T. Gross. Static detection of atomicity violations in object-oriented programs. In *Workshop on Formal Techniques for Java-like Programs*, 2003.

[49] L. Wang and S. D. Stoller. Runtime analysis of atomicity for multithreaded programs. *IEEE Transactions on Software Engineering*, 32:93–110, Feb. 2006.

[50] J. Yi, T. Disney, S. N. Freund, and C. Flanagan. Types for precise thread interference. Technical Report UCSC-SOE-11-22, The University of California at Santa Cruz, 2011.

[51] J. Yi and C. Flanagan. Effects for cooperable and serializable threads. In *Workshop on Types in Language Design and Implementation (TLDI)*, 2010.

[52] J. Yi, C. Sadowski, and C. Flanagan. Cooperative reasoning for preemptive execution. In *Symposium on Principles and Practice of Parallel Programming (PPoPP)*, 2011.