Sequential Reasoning for Optimizing Compilers under Weak Memory Concurrency

Minki Cho∗
Seoul National University
Korea
minki.cho@sf.snu.ac.kr

Sung-Hwan Lee∗
Seoul National University
Korea
sunghwan.lee@sf.snu.ac.kr

Dongjae Lee
Seoul National University
Korea
dongjae.lee@sf.snu.ac.kr

Chung-Kil Hur
Seoul National University
Korea
gil.hur@sf.snu.ac.kr

Ori Lahav
Tel Aviv University
Israel
orilahav@tau.ac.il

Abstract
We formally show that sequential reasoning is adequate and sufficient for establishing soundness of various compiler optimizations under weakly consistent shared-memory concurrency. Concretely, we introduce a sequential model and show that behavioral refinement in that model entails contextual refinement in the Promising Semantics model, extended with non-atomic accesses for non-racy code. This is the first work to achieve such result for a full-fledged model with a variety of C11-style concurrency features. Central to our model is the lifting of the common data-race-freedom assumption, which allows us to validate irrelevant load introduction, a transformation that is commonly performed by compilers. As a proof of concept, we develop an optimizer for a toy concurrent language, and certify it (in Coq) while relying solely on the sequential model. We believe that the proposed approach provides useful means for compiler developers and validators, as well as a solid foundation for the development of certified optimizing compilers for weakly consistent shared-memory concurrency.

CCS Concepts:
• Theory of computation → Concurrency; Operational semantics;
• Software and its engineering → Semantics; Compilers.

Keywords: Relaxed Memory Concurrency; Operational Semantics; Compiler Optimizations

∗The first two authors contributed equally to this work.
The idea is to rely solely on sequential reasoning, and our main contribution is a novel sequential (i.e., single-threaded) semantics that can be safely used for analyzing thread-local optimizations under a full-fledged weak memory model.

The proposed approach goes hand in hand with the fact that compiler writers’ intuition for thread-local optimizations stems from inspecting sequential code, since, intuitively speaking, non-racy code behaves just like sequential code. In fact, validating optimizations that are correct in sequential programs has been one of the main goals in weak memory models design. Our results provide a formal justification of this intuition, and give grounds for development, verification, and testing of optimizations based on a sequential model.

**Example 1.1.** As a concrete simple example, consider an optimization pass that avoids unnecessary reads by locally applying a simplified “store-to-load forwarding” (SLF) as captured by the following pattern:

\[
\begin{align*}
x^{\text{na}} & := v; b := x^{\text{na}} \Rightarrow x^{\text{na}} := v; b := v
\end{align*}
\]

where \(x\) is a shared variable, the \(\text{na}\) superscript denotes non-atomic access to memory, \(v\) is an arbitrary value, and \(b\) is a thread local register. In sequential programs this transformation is clearly sound. We aim to rely on sequential reasoning for justifying this transformation under weak memory. \(\triangle\)

While the idea of using sequential semantics to assist reasoning on concurrent programs is not new, our results provide two important advantages. First, while previous work [9, 10, 17] studied a simple concurrency model based on locks or atomic blocks, the current paper is the first to realize this idea for a rich weak memory model with a wide spectrum of concurrency features, including atomic accesses of several kinds. In particular, we demonstrate that the proposed sequential semantics is sufficiently expressive to validate certain intricate optimizations of non-atomic accesses across atoms, which are performed by mainstream compilers (as we observed on armv8-a clang 11.0.1 and x86-64 GCC 11.2).

**Example 1.2.** Continuing Example 1.1, a more interesting SLF pass eliminates reads across other instructions:

\[
\begin{align*}
x^{\text{na}} & := v; a := x^{\text{na}} \Rightarrow x^{\text{na}} := v; a := v
\end{align*}
\]

What patterns of synchronization accesses (composed of C/C++ atomics) may be included in \(\alpha\) (besides the fact that it should not contain writes to \(x\)) has been a source of confusion before [6, §2.2]. As we show below, the model proposed in this paper allows one to analyze the soundness of this pass relying solely on a sequential model. \(\triangle\)

Second, in contrast to prior work [9, 10, 17], we are not targeting a concurrency model based on the “catch-fire” mechanism, which triggers undefined behavior (UB) for data races like in C/C++11 [2]. The practical significance of this choice is that (irrelevant) load introduction is a sound program transformation in our model. In contrast, this transformation for non-atomics can never be generally sound in a catch-fire model, since it may introduce data races in the target program that do not exist in the source. Allowing load introduction is necessary to support optimizations based on speculation, which are commonly performed by compilers (clang, in particular), e.g., as a part of loop invariant code motion, loop unswitching, load-widening or when loading a vector while only a subset of elements is needed.\(^2\) (In fact, the “freeze” instruction recently introduced in LLVM is a tool to support branching on a possibly undefined value, which is often a result of load introduction [22].)

**Example 1.3.** Consider an optimization pass performing loop invariant code motion (LICM) following the pattern:

\[
\begin{align*}
\text{while } B \text{ do } \{ \alpha ; a := x^{\text{na}} ; \beta \} & \rightarrow c := x^{\text{na}} ; \text{while } B \text{ do } \{ \alpha ; a := c ; \beta \}
\end{align*}
\]

In the case that the loop never executes (when \(B\) is false), a possibly racy (irrelevant) load of \(x\) is introduced. Thus, this transformation is unsound in catch-fire models. In contrast, we aim to validate such transformations, and again use sequential reasoning for their formal justification (with appropriate restrictions on \(\alpha\) and \(\beta\); see §4). \(\triangle\)

To demonstrate that sequential reasoning is adequate for validating soundness of optimizations under a weak memory model, we (formally) establish the adequacy of sequential reasoning for verifying optimizations w.r.t. PS2.1 [8]. The latter is a recent version of the “promising semantics”, a well-studied model [19, 23, 37, 38] addressing the infamous “out-of-thin-air” problem that admits efficient mapping schemes to modern architectures, as well as several critical programming guarantees. Since this model does not include non-atomic accesses, we extend it with such accesses. In this extension, inspired by LLVM [28], to allow load introduction, which is notoriously hard to support in a relaxed memory model [16], racy non-atomic reads retrieve “undefined” values that can be later “frozen” into any non-deterministic value [22] (rather than invoking UB as in C/C++11).

All in all, the contributions of this paper are:

1) We develop a sequential model, called SEQ, instrumenting a standard sequential memory with additional mechanisms to reason about program transformations under a weak memory model (§2). The sequential model abstracts away complicated interference of other threads, by, in particular, tracking permissions to perform non-atomic accesses on certain locations which are non-deterministically gained and dropped with acquire/release atomic accesses. We present two notions of behavioral refinement in SEQ that ensure refinement under weak memory with arbitrary concurrent context: a simple one (§2) that suffices for the vast majority of optimizations (including all those involving solely

\(^2\)See https://llvm.org/docs/Passes.html [Accessed Nov-21].
non-atomics), and a more refined one (§3) needed for certain transformations involving a non-atomic write and a release/relaxed atomic access. We note that typical programmers need not know about the sequential model SEQ, which is only needed for compiler developers.

2) We develop a certified optimizer for a small C-like concurrent language (§4). The optimizer performs four passes of thread-local optimizations: store-to-load forwarding, load-to-load forwarding, dead (overwritten) store elimination, and loop invariant code motion, based on a standard fixpoint analysis in an abstract semantics representing properties of the program executions in SEQ. The optimizer is implemented and certified in Coq. Importantly, its correctness is proved relying solely on simulation in SEQ, without any reference to the underlying (much more complex) promising model. Thus, we view this optimizer as a proof of concept demonstrating the applicability of our approach to verify optimization passes, possibly as an extension of CompCert.

3) We extend PS2.1 with non-atomics, with UB for write-write races and undefined value for write-read races (§5). All results for PS2.1 in [S] are ported to PSna, the extended model. We believe that PSna can be useful as a model for an intermediate representation (IR) language, such as LLVM. In turn, defensive programmers may rely on one of the data-race freedom (DRF) guarantees of the model (see [8]), each of which ensures certain stronger and simpler semantics provided that certain races are avoided.

4) We prove an adequacy theorem allowing one to derive the correctness of optimizations in PSna from (simple or refined) behavioral refinement in SEQ (§6).

Our results are fully mechanized in Coq, building on top of the existing formalization of PS2.1. Our development is available in the artifact accompanying this paper.

2 The Sequential Permission Machine

We introduce a sequential model, which we denote by SEQ, and present a notion of behavioral refinement between sequential programs that we will show to be adequate for reasoning about optimizations of concurrent programs under weak memory consistency: if a target program behaviorally refines a source program, then the source program can be replaced by the target program under any concurrent context assuming weak memory semantics (specifically, PSna).

To understand the intuitions behind SEQ it is important to keep in mind the optimizations we aim to validate. First, since non-atoms are not meant for synchronization, all optimizations allowed in sequential code, including load introduction, should be validated on code involving solely non-atoms. The important exception here is unused store introduction, which is sound in sequential code (although existing compilers avoid this transformation, possibly due to security reasons—we would not want to expose secrets in memory), but trivially unsound in concurrent code, as an unused store of one thread may be read by others.

In contrast, we do not aim to allow optimizations on atomic accesses and fences. Understanding optimizations on synchronization code (via atomics) via sequential reasoning is unnatural, and, even if possible, it will significantly complicate our sequential model. Also, since atomic accesses are relatively rare in concurrent programs and often confined in libraries that are manually optimized by experts, the possible performance gain is rather limited. Although these optimizations were extensively studied (especially for C/C++11 [11, 39]), to the best of our knowledge, existing compilers do not perform such optimizations.

Finally, we also aim to allow optimizations of non-atomics across atomics. As mentioned in the introduction, these are performed by mainstream compilers and have been a source of confusion before. We will also validate reorderings of relaxed accesses and non-atomics, as well as roach-motel reorderings (one-sided reordering of release/acquire and non-atomics) [34], which are not performed by current mainstream compilers but are naturally supported in SEQ.

Concurrency constructs. We assume that shared memory locations are divided into atomic locations (Locat) and non-atomic locations (Locna), and there is no mixing of atomic and non-atomic accesses to the same location. To simplify the presentation in the paper, we only present a fragment of the model consisting of non-atomics and release/acquire and relaxed reads and writes. Our Coq development includes more features: atomic read-modify-writes (RMWs), release sequences, fences (including sequentially consistent fences), strong relaxed accesses (which do not allow “load buffering” behaviors), and system calls. This covers all C/C++11 features as in [21], except for sequentially consistent accesses which are not supported by the promising semantics.

Program representation in the paper. To keep the presentation abstract, rather than introducing a concrete programming language syntax, we assume that the programming language is represented as a labeled transition system (LTS), with transitions labeled with the action that is performed. Below we use \( \sigma \) to denote the program state, which stores the rest of the program to run and the current local register file. Transitions take one of the following forms:

- \( \sigma \rightarrow \sigma' \): silent transitions that do not communicate with the memory (e.g., conditionals and local assignments).
- \( \sigma \xrightarrow{\text{choose}(v)} \sigma' \): transitions resolving a non-deterministic choice (Remark 3 in §6 explains why we need to expose these transitions).

The problem in supporting such mixing is the lower step of PSna that modifies values of outstanding promises. We describe in [1, Appendix E], why the lower operation is needed for validating certain compiler transformations and the challenge it creates for adequacy of sequential reasoning.
error state, we do not formally specify, is meant for intuitive understanding of our examples. In our Coq formulation, programs are represented using interaction trees [45], which provide a convenient formalism supporting this operation.

Values. We assume a parametric set Val of values. To support racy non-atomic reads, Val should contain a distinguished element, called "undefined value" and denoted by undef. In the refinement notions below, we allow the target program to return any value if the source returns undef. For this matter, a partial order ⊆ on Val is defined by: v ⊆ v' ⇔ v = v' ∨ v' = undef. This order is lifted to (partial) functions to Val pointwise.

Remark 1. We follow LLVM assuming that branching on undef invokes UB. A "freeze" instruction can be used to nondeterministically choose a defined value for undef, which is captured by a choose(v) transition in the LTS.

States of SEQ. In addition to the current program state σ, each state S = ⟨σ, P, F, M⟩ of SEQ keeps track of:

- Permission set: a set P ⊆ Locna of non-atomic locations that may be safely accessed. Intuitively, if x /∈ P, then the access to x is racy.
- Written (non-atomic) locations set: a set F ⊆ Locna of non-atomic locations that were written to by the thread. We track this set in the states (and later use in the definition of behavior refinement) in order to ensure that non-atomic write introduction is unsound in SEQ.
- Memory: a function M : Loc → Val assigning a value to every non-atomic location. Since we are not aiming to support optimizations on atoms, there is no need to keep the values of the atomic locations in the memory. (In the refinement notions below, we require that the sequence of atomic accesses generated by the source and the one generated by the target match.)

Transitions of SEQ. The transitions are given in Fig. 1. Each transition S → S' dictates its preconditions (which always require a corresponding program step), the way the different components of the state are updated, and possibly the label e recorded in the trace when the transition is invoked. The latter is essential for the definition of refinement, which imposes conditions relating the traces (i.e., sequences of transition labels) of SEQ generated by the source program to those generated by the target program.

Concretely, silent and choice/relaxed transitions have no additional preconditions and, except for the program component, do not modify the state. The choice/relaxed transitions are recorded in the transition label as they need to match in traces of the source and of the target.

Non-atomic values are handled differently depending on the permission set: when the program performs a non-atomic read from location x, it loads from the memory if x is in the permission set (na-read); or loads undef otherwise (racy-na-read). In turn, when the program performs a non-atomic write to x, it writes to memory and adds x to the set of written locations if x is in the permission set (na-write), or invokes UB (by setting the program state to ⊥) otherwise.
We use standard notations for traces: which, intuitively speaking, accounts for any possible inter-development but elided here to simplify the presentation.

Acquire and release accesses are used for synchronization in the underlying concurrency model. Although they provide more fine grained control than locks, it is helpful to understand an acquire read as a lock acquisition, and a release write as a lock release. In SEQ, these steps non-deterministically update the permission set and the memory, which, intuitively speaking, accounts for any possible interaction with the concurrent environment.

Concretely, ACQ-READ non-deterministically (i) gains permissions for some set of locations (intuitively, these permissions are acquired from other threads), and (ii) gets new values (recorded in a partial function $V : \text{Loc} \rightarrow \text{Val}$) for the locations in this set. Dually, REL-WRITE non-deterministically loses permissions for some set of locations (intuitively, they are released to other threads). The REL-WRITE transition also resets the written locations set $F$. Thus, $F$ tracks written non-atomic locations since the last release, which is needed in order to ensure that (possibly racy) writes cannot be introduced after a release, even if the location was written to (by the source) before the release (see Example 2.10).

In addition, ACQ-READ and REL-WRITE record in the trace (i.e., on their transition labels) the permission set before and after the transition, the written locations set, and the current memory (its updated part in ACQ-READ and "(potentially) released" memory in REL-WRITE). All these are needed for having sufficiently expressive traces that allow us to define an adequate refinement notion.

**Behavioral refinement.** We first define what constitutes a behavior in SEQ.

**Definition 2.1.** A behavior (in SEQ) is a pair of the form $\langle tr, r \rangle$, where $tr$ is a finite sequence of transition labels, and $r$ is either $\text{trm}(v, F, M)$ denoting normal termination returning $v$ with written flags set $F$ and memory $M$, $\text{prt}(F)$ denoting a partial (ongoing) execution with current written flags set $F$, or $\bot$ denoting erroneous termination. We write $S \models \langle tr, r \rangle$ to mean that a state $S$ generates the behavior $\langle tr, r \rangle$, which is inductively defined as follows:

$$
\begin{align*}
   r &= \begin{cases} 
   \text{trm}(v, F, M) & \sigma = \text{return}(v) \\
   \bot & \sigma = \bot \\
   \text{prt}(F) & \text{otherwise}
   \end{cases} \\
   \langle \sigma, P, F, M \rangle \models \langle e, r \rangle
\end{align*}
$$

We use standard notations for traces: $\epsilon$ for the empty trace, $tr_1 \cdot tr_2$ for appending traces, and $\alpha \in tr$ for occurrence of a

---

5 Lock acquisition requires an acquire RMW, which is included in our Coq development but elided here to simplify the presentation.

**Example 2.2.** For a program state $\sigma$ that corresponds to $x^{1x} := 1; y^{\text{na}} := 2; \text{return}(3)$, the state $S = (\sigma, P, 0, M)$ with $y \in P$ has the following behaviors in SEQ: $(e, \text{prt}(0))$, $(\langle W^{1x}(x, 1), \text{prt}(0) \rangle)$, $(\langle W^{1x}(x, 1), \text{prt}(0) \rangle)$, and the terminating behavior $(\langle W^{1x}(x, 1), \text{trm}(3, \{y\}, M[y \mapsto 2]) \rangle)$. If $y \notin P$, then $(\langle W^{1x}(x, 1), \bot \rangle)$ is the only terminating behavior. △

Next, we present the (first) notion of behavioral refinement between programs in SEQ. As standard, one may start by requiring that every behavior $\langle tr_{\text{tgt}}, r_{\text{tgt}} \rangle$ of the target program is also a behavior of the source program. However, to support various optimizations, naive inclusion does not suffice. Instead, we allow the source to generate a matching behavior $\langle tr_{\text{src}}, r_{\text{src}} \rangle$ that is "less commited" than $\langle tr_{\text{tgt}}, r_{\text{tgt}} \rangle$ (denoted $\langle tr_{\text{tgt}}, r_{\text{tgt}} \rangle \sqsubseteq \langle tr_{\text{src}}, r_{\text{src}} \rangle$); the source may return $\text{def}$ when the target returns a normal value $(e_{\text{tgt}} \subseteq v_{\text{src}})$, end with "less defined" memory $(M_{\text{tgt}} \subseteq M_{\text{src}})$, and write to more non-atomic locations $(F_{\text{tgt}} \subseteq F_{\text{src}})$. The same holds along the trace: for values recorded in atomic writes, for memories recorded in release writes, and for written locations sets recorded in acquire and release accesses. Finally, UB by the source allows any continuation by the target. All these are formally captured by the next definition.

**Definition 2.3.** The relation $\sqsubseteq$ on transition labels, traces, and behaviors is given by:

1. Transition labels:

$$
\begin{align*}
   \epsilon & \sqsubseteq \epsilon \\
   v_{\text{tgt}} & \sqsubseteq v_{\text{src}} \\
   F_{\text{tgt}} & \sqsubseteq F_{\text{src}} \\
   V_{\text{tgt}} & \sqsubseteq V_{\text{src}}
\end{align*}
$$

2. Traces:

$$
\begin{align*}
   e_{\text{tgt}}^1 \cdot \ldots \cdot e_{\text{tgt}}^n & \sqsubseteq e_{\text{src}}^1 \cdot \ldots \cdot e_{\text{src}}^k \\
   \forall k, e_{\text{tgt}}^k & \sqsubseteq e_{\text{src}}^k
\end{align*}
$$

3. Behaviors:

$$
\begin{align*}
   \langle tr_{\text{tgt}}, tr_{\text{src}}, r_{\text{src}} \rangle & \sqsubseteq \langle tr_{\text{src}}, tr_{\text{src}}, r_{\text{src}} \rangle \\
   \langle tr_{\text{tgt}}, tr_{\text{src}}, r_{\text{src}} \rangle & \sqsubseteq \langle tr_{\text{src}}, tr_{\text{src}}, r_{\text{src}} \rangle
\end{align*}
$$

**Definition 2.4.** We write $S_{\text{src}} \sqsubseteq S_{\text{src}}$ if $S_{\text{src}} \downarrow \langle tr_{\text{tgt}}, r_{\text{tgt}} \rangle$ implies that $S_{\text{src}} \downarrow \langle tr_{\text{src}}, r_{\text{src}} \rangle$ for some behavior $\langle tr_{\text{src}}, r_{\text{src}} \rangle$ such that $\langle tr_{\text{tgt}}, r_{\text{tgt}} \rangle \sqsubseteq \langle tr_{\text{src}}, r_{\text{src}} \rangle$. A program state $\sigma_{\text{tgt}}$ behaviorally refines a program state $\sigma_{\text{src}}$, denoted by $\sigma_{\text{tgt}} \sqsubseteq \sigma_{\text{src}}$, if $\sigma_{\text{tgt}} \sqsubseteq \sigma_{\text{src}}$, or $\sigma_{\text{tgt}} \sqsubseteq \sigma_{\text{src}}$ for every $P, F, M$. We write $\text{prog}_{\text{src}} \sim \text{prog}_{\text{tgt}}$ for two code snippets $\text{prog}_{\text{src}}$ and $\text{prog}_{\text{tgt}}$, we mean that for any (sequential) context $C$, the state $\sigma_{\text{tgt}}$ that corresponds to $C[\text{prog}_{\text{tgt}}]$ (with some

---
initial register file) behaviorally refines the state $\sigma_{\text{src}}$ that runs $C[prog_{\text{src}}]$ (with the same initial register file). We write $prog_{\text{src}} \not\leadsto prog_{\text{tgt}}$ for the negation of $prog_{\text{src}} \leadsto prog_{\text{tgt}}$.

**Remark 2.** Our results that are formalized in Coq also address reasoning about program transformations in SEQ. In particular, they allow one to lift local refinement properties (such as the ones listed in the examples below) to any sequential context $C$. Concretely, we define a simulation relation between SEQ states and prove that it admits certain congruence properties. Then, by establishing simulation between $prog_{\text{src}}$ and $prog_{\text{tgt}}$, we can derive $prog_{\text{src}} \leadsto prog_{\text{tgt}}$ as defined above. Since these techniques are fairly standard in compiler verification, and our main focus is to reduce reasoning about concurrent code to reasoning about sequential code, we omit these details from the paper. In the example below, we intend to give the right intuitions, rather than precise refinement arguments.

**Example 2.5 (Reordering of non-atomics).** Non-atomic accesses to different locations can be freely reordered in SEQ, e.g., $a := x^{na}; y^{na} := v \not\leadsto y^{na} := v; a := x^{na}$ where $x \neq y$. Consider a general context $C$, and let $\sigma_{\text{src}}$ and $\sigma_{\text{tgt}}$ be the program states that correspond to $C[a := x^{na}; y^{na} := v]$ and $C[y^{na} := v; a := x^{na}]$, respectively, with the same initial register file. Suppose that $(\sigma_{\text{src}}, P, F, M) \not\models (\tau_{\text{tgt}}, P_{\text{tgt}})$. Then, by executing two steps in the source (a read from $x$ followed by a write to $y$) at the same time the target executes its write to $y$, one can show that $(\sigma_{\text{src}}, P, F, M) \not\models (\tau_{\text{tgt}}, P_{\text{tgt}})$. In particular, after the target performs the read, the source and the target reach the same program state.

On the other hand, reordering of non-atoms to the same location is disallowed, e.g., the reordering of a load followed by a store $a := x^{na}; x^{na} := 1 \not\leadsto x^{na} := 1; a := x^{na}$. Indeed, for the context $C = \{\text{return}(a)\}$, we have $(\sigma_{\text{tgt}}, P, F, M) \not\models (\epsilon, \text{trm}(1, F \cup \{x\}, M[x \mapsto 1]))$ starting from a state with $x \in P$ and $M(x) = 2$. However, the only terminating behavior that is generated by the source program from this state is $(\epsilon, \text{trm}(2, F \cup \{x\}, M[x \mapsto 1]))$. △

**Example 2.6 (Eliminations of non-atomics).** Behavioral refinement holds for the following pairs, in which a non-atomic access is eliminated:

(i) $x^{na} := v; x^{na} := v' \leadsto x^{na} := v'$
(ii) $x^{na} := v; a := x^{na} \not\leadsto x^{na} := v; a := v$
(iii) $a := x^{na}; b := x^{na} \not\leadsto a := x^{na}; b := a$
(iv) $a := x^{na}; x^{na} := a \not\leadsto a := x^{na}$

Note that for the read-before-write elimination ((iv) above), the written locations set in the final state of the source may be a strict superset of the one of the target ($F_{\text{tgt}} \subset F_{\text{src}}$), which is allowed in the definition above. Conceptually, the optimized program may avoid writes to some locations that are performed by the source.

In contrast, the introduction of a write after a read is unsound due to the conditions on the written locations set $F$. For example:

$$a := x^{na}; \text{if } a \neq v \text{ then } x^{na} := v \not\leadsto a := x^{na}; x^{na} := v$$

In this example, starting from $F = \emptyset$ and permission to access $x$, the target ends its execution with $F_{\text{tgt}} = \{x\}$, while the source has $F_{\text{src}} = \emptyset$.

In turn, the other introductions of non-atomics obtained as converses of (i)-(iii) above provide additional instances of refinements in SEQ. This intuitively corresponds to the fact that non-atomics are not permitted to perform synchronization. △

**Example 2.7 (Reordering across loops).** Reordering a non-atomic write before a possibly infinite local computation is unsound, as it introduces a write if the loop never terminates:

$$\text{while } (...) \text{ do } ...; x^{na} := v \not\leadsto x^{na} := v; \text{while } (...) \text{ do } ...$$

In SEQ, when starting without permission on $x (x \notin F)$, the target program generates the behavior $(\epsilon, \bot)$, but the source may not be able to generate a matching behavior if the loop is not terminating.

A variant of this example demonstrates why we have to match partial traces with the condition that $F_{\text{tgt}} \subseteq F_{\text{src}}$:

$$a := x^{na}; \quad a := x^{na};$$

if $a \neq 1$ then $x^{na} := 1 \not\leadsto$ if $a \neq 1$ then $x^{na} := 1;$

while $(...) \text{ do } ...; x^{na} := 2 \quad x^{na} := 2; \text{while } (...) \text{ do } ...$$

If the target invokes UB and generates $(\epsilon, \bot)$, then it must be the case that we started without permission on $x$, and the source may invoke UB and generate $(\epsilon, \bot)$ as well. Thus, in order to obtain a behavior of the target that is not matched by a behavior of the source, in case the loop is non-terminating, we must consider behaviors before termination $(\epsilon, \text{prf}(F))$. Indeed, when starting with permission on $x$ and $M(x) = 1$, the target generates $(\epsilon, \text{prf}(\{x\}))$, but, if the loop does not terminate, the only behavior of the source is $(\epsilon, \emptyset)$.

In contrast, reads may be reordered with possibly non-terminating local computation:

$$\text{while } (...) \text{ do } ...; a := x^{na} \leadsto a := x^{na}; \text{while } (...) \text{ do } ...$$

Indeed, in partial traces only the written locations set $F$ has to match, and this set is the same in executions of the two programs. △

**Example 2.8 (Unused load elimination and introduction).** The transformations that eliminate/introduce an unused load $a := x^{na} \leadsto \text{skip}$ and $\text{skip} \leadsto a := x^{na}$ trivially correspond to behavioral refinements in SEQ. For the latter, we need that a non-atomic read without permission does not invoke UB. We note that in the paper presentation, we have to assume that a does not occur in the context. Nevertheless, our Coq formalization that uses interaction trees can easily express a computation that reads a value and does not return the result to its continuation. This computation is interchangeable with a no-op under any context. △
Example 2.9 (Reordering of atomics and non-atomics). Reorder-
ing of atomic and non-atomic accesses follows the “roach-motel” principle. The following are forbidden:

(i) \[ a := x^{\text{acq}}; y^{\text{na}} := v ; y^{\text{na}} := v ; a := x^{\text{acq}} \]

(ii) \[ y^{\text{na}} := v ; x^{\text{rel}} := v \rightarrow x^{\text{rel}} := v ; y^{\text{na}} := v' \]

(iii) \[ a := x^{\text{acq}} ; b := y^{\text{na}} \rightarrow b := y^{\text{na}} ; a := x^{\text{acq}} \]

(iv) \[ a := y^{\text{na}} ; x^{\text{rel}} := v \rightarrow x^{\text{rel}} := v ; a := y^{\text{na}} \]

In (i), starting without permission on \( y \) (\( y \not\in P \)), the target invokes UB, thus generating the behavior \((\epsilon, \bot)\). The source, however, has to perform the acquire read before invoking UB, thus generating terminating behaviors of the form \((R^{\text{acq}}(\_ , \_))\) only.

In (ii), if the release write to \( x \) loses the permission on \( y \) (transferring from a state with \( y \in P \) to a state with \( y \not\in P \)), then the target program invokes UB, generating a behavior of the form \((\_ , \bot)\). However, since we start with \( y \in P \), the source does not invoke UB, and cannot generate any behavior of the form \((\_ , \bot)\).

In (iii), for the context \( C = \_ ; \text{return}(b) \), if a permission on \( y \) is gained by the acquire read from \( x \) and we start with \( y \not\in P \) (and \( M(y) \neq \text{undef} \)), the target generates a behavior of the form \((R^{\text{acq}}(x, \_ , P \cup \{y\}, \_ \_ \_ , 0, \_ , \_ , \_ , \_ ))\), whereas the source cannot perform racy read on \( y \).

In (iv), for the context \( C = \_ ; \text{return}(a) \), if we start with \( y \in P \), and this permission is lost by the release write, the target generates a behavior of the form \((\_ , \_ , \_ \_ \_ , 0, \_ , \_ , \_ , \_ , \_ , \_ ))\), whereas the source cannot perform racy read on \( y \) at all.

Next, the following converses of the above are validated:

(i') \[ y^{\text{na}} := v ; a := x^{\text{acq}} \rightarrow a := x^{\text{acq}} ; y^{\text{na}} := v \]

(ii') \[ b := y^{\text{na}} ; a := x^{\text{acq}} \rightarrow a := x^{\text{acq}} ; b := y^{\text{na}} \]

(iii') \[ x^{\text{rel}} := v ; a := y^{\text{na}} \rightarrow a := y^{\text{na}} ; x^{\text{rel}} := v \]

For (i') we use the fact that acquire transitions of the target can be matched by acquire transitions of the source annotated with \( F^{\text{rel}} \subseteq F^{\text{acq}} \) (since we may have \( y \in F^{\text{rel}} \) but not \( y \in F^{\text{acq}} \) when performing the acquire), as well as the fact that the source may invoke UB earlier than the target (in particular, \((R^{\text{acq}}(\_ , \_)) \subseteq (\_ , \_))\). In turn, (iii') and (iv') demonstrate the need in allowing the source to return \text{undef} when the target returns a defined value. This is needed, for instance, if the acquire read in (iii') gains permission on \( y \).

Finally, despite being a valid roach-motel reordering, the converse of (ii) is disallowed by the current behavior refinement. It is supported by the more refined notion in § 3.△

Example 2.10. Stores cannot be introduced even if they already occur before a release write:

\[ x^{\text{na}} := v ; y^{\text{rel}} := 1 \rightarrow x^{\text{na}} := v ; y^{\text{rel}} := 1 ; x^{\text{na}} := v \]

Intuitively, if a write is protected by a lock, another one should not be introduced after the lock is released. Formally, since release writes reset the written locations set, the target’s terminating behavior has \( x \in F \), while the source ends with \( F = \emptyset \). In contrast, the transformation is validated with \( r^{\text{rlx}} \) instead of \( r^{\text{rel}} \) above. △

Example 2.11 (Store-to-load forwarding across atomics). Reads can be eliminated after writes across atomics:

\[ x^{\text{na}} := v ; a := x^{\text{acq}} ; b := x^{\text{na}} \rightarrow x^{\text{na}} := v ; a := x^{\text{acq}} ; y^{\text{rel}} := v' \]

where \( a \in \{a := y^{\text{rlx}} , y^{\text{rlx}} := v' \rightarrow a := y^{\text{acq}} , y^{\text{rel}} := v' \} \). If we start with \( x \not\in P \), the source raises UB, and generates \((\epsilon, \bot)\), which matches any target behavior. Otherwise, the relevant write step of the source sets \( M(x) = v \), and \( M(x) \) is not altered by \( a \). In particular, it is important here that an acquire read can only modify values of locations that gained permission. Thus, either \( v \) is read by \( b := x^{\text{na}} \), or, if \( \alpha \) corresponds to a release write that transferred the permission on \( x \), the source will read \text{undef}, and we have \( v \subseteq \text{undef} \). In any case, the source matches every behavior of the target. △

Example 2.12. Reads cannot be eliminated after writes across release-acquire pairs:

\[ x^{\text{na}} := v ; y^{\text{rel}} := v' ; a := x^{\text{acq}} ; b := x^{\text{na}} \rightarrow \]

\[ x^{\text{na}} := v ; y^{\text{rel}} := v' ; a := x^{\text{acq}} ; b := v \]

Intuitively, another thread may safely access \( x \) between the release and the acquire and change its value. To see this in SEQ, consider the execution of the target program when permission to \( x \) is lost by the release write, and regained by the acquire read. The updated portion of the memory \( V \), including a new (non-deterministic) value for \( x \), is recorded on the acquire transition in the trace. To match the behavior of the target, the source program has to have the same updated memory in its acquire transition, and when \( V(x) \neq v \), the source will not be able to later read \( v \) from \( x \). This example demonstrates the need in updating the values in memory (for locations that gained permission) in acquire steps. △

3 Advanced Behavior Refinement

As we show in §6, the above notion of behavioral refinement in SEQ is adequate for reasoning about optimizations in the promising semantics. As shown above, it is also precise enough to verify a variety of optimizations. However, optimizations including both an atomic access and a non-atomic write are beyond its power: although they are meant to be sound (and they are sound in the promising semantics), the above notion invalidates them. In this section, we discuss this issue that stems from two different reasons, and then present a more refined notion of behavioral refinement (implied by the simple one above) that addresses this challenge. We note that, since our result in §6 provides contextual refinement, one may mix and match—prove most optimization passes using the simple notion in §2, and use the one of this section for several more involved program transformations.

Late UB. A simple example of a sound optimization that is invalidated by the above notion is the following:

\[ a := y^{\text{rlx}} ; y^{\text{na}} := v \rightarrow y^{\text{na}} := v ; a := x^{\text{rlx}} \]
Indeed, the reasoning in Example 2.9 (i) that shows why an acquire read followed by a non-atomic write cannot be reordered applies as is in this case as well: starting without permission on \( y (y \not\in P) \), the target program invokes UB, thus generating the behavior \( (\epsilon, \perp) \). The source, however, has to perform the relaxed read before invoking UB, thus generating terminating behaviors \( (tr_{src}, \perp) \) with \( tr_{src} \) consisting of a \( R^{l_1x} \) label. Intuitively, however, this should not matter since the source program anyway invokes UB, in which case the target’s behavior is immaterial. Thus, we would like to allow to match any behavior \( (tr_{tgt}, r_{tgt}) \) of the target program by any UB behavior \( (tr_{src}, \perp) \) of the source. Nevertheless, for two reasons, this solution requires extra care.

First, it essentially allows reordering of any access with an operation that invokes UB, e.g., \( a := 1/0 \leadsto a := 1/0 \land a \). As the next example shows, in concurrent settings this reordering must be invalidated if \( a \) contains an acquire read.

**Example 3.1.** Consider the following optimizations:

\[
\begin{align*}
  &a := x^{r_1x}; &a := x^{r_1x}; &y^{l_1x} := 1; \\
  &\text{if } a = 1 \text{ then } &\text{if } a = 1 \text{ then } &a := x^{l_1x}; \\
  &a := x^{acq}; \leadsto b := 1/0; \leadsto \ldots \leadsto &a := x^{acq}; \\
  &b := 1/0 &b := 1/0; &b := 1/0; \\
  &\text{else } y^{l_1x} := 1 &\text{else } y^{l_1x} := 1 &a := x^{acq}
\end{align*}
\]

First, we perform the (unsound) reordering of an acquire read with UB-invoking operation. Then, a sequence of standard optimizations (start with \( b := 1/0 \leadsto y^{l_1x} := 1; b := 1/0 \), then hoist \( y^{l_1x} := 1 \) from both branches of the conditional and reorder it with \( a := x^{r_1x} \)) lead to the program on the right. Now, if the concurrent context consists of another thread with the code: \( c := y^{r_1x}; x^{rel} := c \), then UB is possible for the target, but not for the source.

Thus, to keep invalidating the reordering of an acquire operation followed by UB, we require that there are not any acquire accesses in the suffix of the source trace \( tr_{src} \) (in the source’s path towards UB) that does not match the target’s trace \( tr_{tgt} \). For instance, this invalidates the transformations \( a := x^{acq}; b := 1/0 \leadsto b := 1/0; a := x^{acq} \) and \( a := x^{acq}; y^{na} := v \leadsto y^{na} := v; a := x^{acq} \); if we start without permission on \( y \), the target’s behavior \( (\epsilon, \perp) \) does not match any source behavior.

Second, it is crucial to make sure that for executing this suffix, the source does not make assumptions on the environment. To see this, consider the following example:

\[
\begin{align*}
  &a := x^{r_1x}; &a := x^{r_1x}; &y^{l_1x} := 1; \\
  &\text{if } a = 1 \text{ then } &\text{if } a = 1 \text{ then } &a := x^{l_1x}; \\
  &b := 1/0 \leadsto b := 1/0; a := x^{r_1x}; &b := 1/0; a := x^{r_1x}; &\text{while (...) do {}} \leadsto \text{while (...) do {}}
\end{align*}
\]

Without additional restrictions, since we allow any behavior of the target to be matched with \( R^{l_1x}(x, 1), \perp \) of the source, this transformation, which is clearly unsound (even in sequential programs), will be validated by SEQ. Intuitively speaking, what went wrong here is that the source matches the UB of the target by reading 1 from \( x \), whereas a concurrent environment may not provide this option.

To address this issue, when we match the UB of the target by a suffix of source trace that leads to UB we need to make sure that the source avoids making assumptions on the concurrent environment. Technically, we achieve this by assuming that read values of atomic reads, permission gains and losses, and memory updates \( V \) are dictated by an oracle, which intuitively represents a possible concurrent environment. We then require that behavioral refinement holds for any oracle (which has to satisfy certain progress and monotonicity conditions that any environment satisfies). In particular, in the example above, the source has to match the target’s UB also for an oracle that forces the source to read \( x \neq 1 \), in which case the source cannot invoke UB. In contrast, in the earlier example for the need in late UB \((a := x^{r_1x}; y^{na} := v \leadsto y^{na} := v; a := x^{r_1x})\), if we start without permission on \( y \), the source invokes UB for any oracle as its write is invalid if \( a \) contains an acquire read.

**Writes across release.** Roach motel reordering of a release write followed by a non-atomic write pose an additional challenge to sequential reasoning:

\[
x^{rel} := v; y^{na} := v' \leadsto y^{na} := v'; x^{rel} := v
\]

Even if we modify behavioral refinement as discussed above, some behaviors of the target are not matched by the source. Concretely, starting with permission on \( y \), the target’s label \( w^{rel}(x, v, P, P', F_{tgt}, V_{tgt}) \) must have \( y \in F_{tgt} \) and \( V_{tgt}(y) = v' \), whereas the source is confined by the initial state (which may have \( y \not\in F_{src} \) or \( V_{src}(y) \neq v' \)). Intuitively, this should not be a problem: the source is going to write to \( y \) after the release, and other threads can observe that write.

To solve this, we need to allow the source to generate release labels with different written locations set and memory compared to the target’s labels, provided that later on the source will write to the non-atomic locations that were different. Technically, we achieve this by parameterizing behavioral refinement with a “commitment set” \( R \), which is a set of non-atomic locations that the source program must write to before it terminates or executes an acquire read. (Fulfilling commitments after an acquire read corresponds to the disallowed reordering of writes after an acquire read.) Initially, the commitment set is empty. Then, we modify (remove fulfilled commitments and add new ones) this set with every release transition. In the end of the execution and with every acquire transition, we verify that all commitments were fulfilled. Finally, the non-terminating behaviors \( \langle tr, prt(F) \rangle \) should allow the source program to continue its execution and fulfill the outstanding commitments.

**Advanced behavior refinement.** The above solutions are formalized as follows. First, when checking for refinement between a source program and a target program in SEQ, we use an oracle to represent the environment of the thread. To pass only relevant information to the oracle, we use the following notation for stripping transition labels.
(extended pointwise to traces):
\[ |e| = \begin{cases} R^{\text{acq}}(x, v, P, P', F, V) & \text{for } e = R^{\text{acq}}(x, v, P, P', F, V) \\ W^{\text{rel}}(x, v, P, P') & \text{for } e = W^{\text{rel}}(x, v, P, P', F, V) \\ e & \text{otherwise} \end{cases} \]

Definition 3.2. An oracle \( \Omega \) is an LTS over stripped transition labels such that the following hold:
- Progress: In every state \( w \) of \( \Omega \) and for every \( x \in \text{Loc}^s \), \( v \in \text{Val} \), and \( P \subseteq \text{Loc}^a \), transitions choose(\( _\_ \)), \( R^\text{lr}(x, \_ \_ ) \), \( W^\text{lr}(x, v) \), \( R^{\text{acq}}(x, \_ \_ , P, \_ \_ \_ , \_ \_ \_ ) \), and \( W^{\text{rel}}(x, v, P, \_ \_ \_ , \_ \_ \_ ) \) are enabled for some (valid) values of "\_".
- Monotonicity: If \( w \xrightarrow{\_ \_ \_ } w' \) and \( e \subseteq e' \), then \( w \xrightarrow{\_ \_ \_ } w' \).

The progress condition allows the source to continue its execution and fulfill its commitments after the target has terminated. Monotonicity is required to allow the refinement of \( \text{undef} \) in the source by any defined value. We say that a trace \( tr \) is allowed by an oracle \( \Omega \), denoted by \( tr \in \Omega(\_ \_ \_ ) \), if \( [tr] \) is a trace of \( \Omega \) (i.e., a sequence of symbols that \( \Omega \) can execute, starting from its initial state).

Next, the notion of a behavioral refinement up to a commitment set is formulated in Fig. 2. It modifies the one in Def. 2.3 by allowing the source to invoke UB later than the target (in beh-failure), while tracking and checking the commitment set \( R \). Each time a release write \( W^{\text{rel}}(x, v, P, P', F, V) \) is added to the target’s trace, we compare it to the matching one by the source \( W^{\text{rel}}(x, v, P, P', F, V) \), and set the new commitment set \( R' \) to consist of the locations \( y \) that: (i) were written to by the target but not by the source (\( y \in F_{tgt} \setminus F_{src} \)); (ii) have a value in the target memory that does not redefine the value of the source (\( V_{tgt}(y) \not\subseteq V_{src}(y) \)); or (iii) were included in the previous commitment set and not written yet by the source (\( y \in R \setminus F_{src} \)). Upon termination or acquire read, in addition to \( F_{tgt} \subseteq F_{src} \), we require that all outstanding commitments were fulfilled by the source (\( R \subseteq F_{src} \)). Finally, refinement of non-terminating behaviors beh-partial allow the source to take more steps (but not acquire reads) for fulfilling its commitments. Since the written locations set is reset with every release write, to see what locations the source has written to, we add all \( F \) sets in the release operations in the source’s trace to those in the final \( F \) set.

With the above definition, the more refined behavioral refinement notion is stated as follows:

Definition 3.3. A program state \( \sigma_{tgt} \) weakly behaviorally refines a program state \( \sigma_{src} \), denoted by \( \sigma_{tgt} \subseteq_w \sigma_{src} \), if for every oracle \( \Omega \), if \( \langle \sigma_{tgt}, P, F, M, V \rangle \not\subseteq \langle \sigma_{src}, P, F, M, V \rangle \) then \( \langle \sigma_{src}, P, F, M, V \rangle \not\subseteq \langle \sigma_{src}, P, F, M, V \rangle \) and \( \langle \sigma_{src}, P, F, M, V \rangle \not\subseteq \langle \sigma_{src}, P, F, M, V \rangle \).

Example 3.5 (Overwritten store elimination across atomics). Consider the elimination of a write after another write to the same location across an atomic access:
\[ x^{na} := v ; \alpha ; \sigma_{src} \Rightarrow \sigma_{src} \]
where \( \alpha \in \{ b := y^{\text{lr}}, y^{\text{lr}} := v_y, b := y^{\text{acq}}, y^{\text{acq}} := v_y \} \). The three cases except for \( \alpha = y^{\text{acq}} := v_y \) are easily validated by the simple behavioral refinement in SEQ (here it is needed that the source may have larger \( F \) sets).

The case that \( \alpha \) is a release write should be also considered sound,\(^6\) since, roughly speaking, other threads that can observe \( x^{na} := v \) can always also observe \( x^{na} := v' \) instead. (In particular, this optimization is sound in the promising semantics.) Nevertheless, the simple refinement notion in §2 invalidates this optimization (\( \sigma_{tgt} \not\subseteq_w \sigma_{src} \)): starting with permission on \( x \), the memory recorded in release writes in the source is confined to have \( M(x) = v \), while the target has the value of the initial memory. In turn, we do have \( \sigma_{tgt} \equiv_w \sigma_{src} \). In particular, consider the empty context, and let \( \text{rel}(P, F, M, u) \equiv W^{\text{rel}}(y, v, P, P', F, M(x \mapsto u)) \). If we start with permission on \( x \) and do not release it, then for \( r = \text{trm}(\text{unit}, (x, M[x \mapsto v'])) \), \( \langle \text{rel}(\{x\}, \{x\}, v, r) \rangle \not\subseteq \langle \text{rel}(\{x\}, \{x\}, 0, M(x)) \rangle \) follows from \( \langle \epsilon, r \rangle \subseteq \langle \epsilon, r \rangle \). If we start with permission on \( x \) and release it, then \( \langle \text{rel}(\{x\}, \{0\}, v, r) \rangle \not\subseteq \langle \text{rel}(\{x\}, \{0\}, M(x)) \rangle \) follows from \( \langle \epsilon, \bot \rangle \subseteq \langle \epsilon, \bot \rangle \). If we start without permission on \( x \), then, using beh-failure, we have:
\[ \langle \text{rel}(\{0\}, \{0\}, v, \bot) \rangle \not\subseteq \langle \epsilon, \bot \rangle. \]
\[ \triangle \]

\(^6\)Currently, it is not performed by mainstream compilers (checked for armv8-a clang 11.0.1 and x86-64 GCC 11.2).
We implemented in Coq a verified optimizer that optimizes where weak memory consistency! represented by the following abstract tokens: acquire pair after the most recent write. This information is memory value to forward and a flag for detecting a release-assigns two kinds of information to each shared variable: a later loads, possibly across atomic operations, but not across in the SLF pass, which forwards values written by stores to in [1, Appendix D]. Figure 3 depicts the analysis performed and optimization in detail. The other passes are described contains no reads from x indicates that x contains no writes to x and no release write has been executed after the write; and x ⊢ o(v) transitions to •(v) for a release write; and •(v) transitions to T for an acquire read. To show termination, we have proved that the analysis reaches a fixpoint in at most three iterations when analyzing a loop.

Given the analysis result at each program point, SLF transforms a read a := xna into a register assignment a := v if the token to x is •(v) or o(v) at that program point. Intuitively, having x ⊢ •(v) or x ⊢ o(v) means that no release-acquire pair has been executed since v was written to x, thus the memory value of x is still v even if the thread has lost the permission to x. The transformation is sound since the thread will read v or undef from x depending on whether the permission to x has been lost or not. Formally, a reachable SEQ state (σ, P, F, M) is related to the analysis result at the relevant program point as follows:

\[
\forall x, \begin{cases}
    x \in P \land v \subseteq M(x) & \text{if } x \nLeftarrow o(v) \\
    x \in P \Rightarrow v \subseteq M(x) & \text{if } x \nLeftarrow •(v)
\end{cases}
\]

Figure 4 describes how the analysis and optimization work for a concrete program example.

The verification of the passes is executed by (i) establish the soundness of each analysis; and (ii) from the soundness, derive a simulation in SEQ between the source and target codes. The simulation relation in SEQ (given in [1, Appendix A]) ensures advanced behavioral refinement as defined in §3.5. This verification strategy follows the standard approach of CompCert, and, importantly, the optimizer is fully verified in Coq relying solely on sequential reasoning.

### 5 Non-atomics in the Promising Semantics

We present the extension of PS\textsuperscript{2.1} with non-atomic accesses, which we denote by PS\textsuperscript{na}. At the core of this extension is an operational race detection, so UB is invoked on write-write races and undef is read on read-write races. Unlike in SEQ (§2), we allow the mixing of atomic and non-atomic accesses to the same location (so we assume one set Loc of locations), which means that a race may involve only one non-atomic

\footnotetext{[1]}{In fact, this is the same simulation used in the adequacy proof (see §6), so the optimizer correctness argument is independent of behavioral refinement in SEQ—it goes directly from simulation in SEQ to simulation in PS\textsuperscript{na}.}
access. As for SEQ, we only present a simplified fragment of the full model omitting fences, RMWs, release sequences, and system calls, which are all covered by our Coq formalization.

Roughly, in the promising model the memory is a set of timestamped messages capturing all previous writes. Each thread maintains a view, pointing to the latest message the thread has observed for each location, used to restrict future reads/writes of the thread. Promises are used to allow read-write reordering—a thread may promise to perform a write in the future and add a message to memory before the write is being executed. To avoid "causality cycles" (a.k.a. thin-air behaviors) every step has to be accompanied by certification: by running alone the thread has to be able to fulfill its outstanding promises.

Figure 5 presents the thread configuration steps and the machine steps. Next, we discuss the new parts, highlighted in the figure. We refer the reader to [8, 19, 23] for explanations of the atomics fragment of PS

Normal non-atomic accesses are handled by read and write transitions. A non-atomic read (read) from x behaves exactly as a relaxed one: reads from a message with timestamp t that is greater than or equal to the thread’s view of x, and updates the view to include t. In turn, non-atomic writes (write) require a non-trivial extension. When a thread executes a non-atomic write to a location x, it may add multiple arbitrary messages with the bottom view (denoted by ⊥, a view smaller than any other view) to x before adding a message with the appropriate value. In addition, some of the messages preceding the final one may be valueless non-atomic messages of the form $u = x \otimes \top \in \text{NAMsg}$, which we introduce for detecting races. Writing multiple messages in one non-atomic write allows the splitting of non-atomic writes which is needed in order to allow certain program transformations (see [1, Appendix B]).

Two loads from x are optimized to register assignments. To illustrate the analysis, the code is annotated with abstract tokens to x. The first instruction $x^{\text{na}} := 42$ induces UB if there is no permission on x. Therefore, the permission on x is guaranteed, with the memory value 42 at x (which is represented by $x \mapsto \circ(42)$). Since x is already permissioned, its value is not updated by the $l := y^{\text{acq}}$, thereby maintaining the abstract state of x. Upon a conditional, we keep analyzing each branch separately, and then join the results. On the then branch, $a := x^{\text{na}}$ will load 42 from the memory as the abstract state indicates. For the next instruction, $y^{\text{rel}} := 1$, the abstract state of x transitions to $x \mapsto \circ(42)$ as the permission on x can be dropped by the release write, while the memory value at x is maintained. Finally, the branch is merged and the analysis results are joined (following the partial order on the abstract tokens). The effect of the last instruction, $b := x^{\text{na}}$, depends on the permission on x. If there is no permission on x, undef is read, which can be replaced by 42 by definition. In turn, the abstract state of x tells us 42 must be loaded if there is a permission on x. From the above analysis, we conclude that the two loads can be replaced with register assignments.

Figure 4. An example optimization by SLF including the underlying analysis in SEQ.

Racy accesses are naturally defined: a non-atomic access to x is racy if the thread is unaware of some message with location x ($V(x) < m.t$), and an atomic access to x is racy if the thread is unaware of some non-atomic message with location x. Using race-helper, a thread reads undef when performing a racy read (racy-read), and invokes UB on a racy write (racy-write).

For supporting the compiler transformation that replaces an undef by a non-undef value, we note that the promise lowering step in PS

Machine steps. A machine state, which consists of the different thread states ($T$) and a main memory (M), can take a step by one of the threads taking a sequence of steps (machine: normal, possibly invoking UB (machine: failure)). Normal steps (machine: normal) require “certification” the thread that passes control to the scheduler has to show that by running alone it can fulfill all its promises.

Example 5.1. The following demonstrates how promises and the racy read step work:

$$a := x^{\text{na}}; \quad \text{//undef} \quad y^{\text{rel}} \ := \ 1$$

Here, the left thread may promise $y = 1$, since by running alone, it is able to execute the read from x and fulfill its promise. Then, the right thread reads 1 from y and writes 1 to x. (Other messages may be also added to x before the x = 1 message.) Now, the non-atomic read from x of the left thread is racy since there is a message of x with timestamp larger than the thread’s view of x. Thus, the thread reads undef from x and fulfills the promise $y = 1$.

Results. We ported to PS

We assume that $u \cdot \text{view} = \bot$ for $u \in \text{NAMsg}$.
We write $m = (x, v, V) \in Msg$ if for either $src = \perp$ or $src \neq \perp$, one of the following holds: (i) $e_1 = e_2$ and $e_1 = e_2$; (ii) $e_1 = R^a(x, v_1)$ and $e_2 = R^a(x, v_2)$, and $v_1 = v_2$; or (iii) $e_1 = \text{choose}(v_1)$ and $e_2 = \text{choose}(v_2)$, and $v_1 \neq v_2$.

Theorem 6.2 (Adequacy). If $a_{\text{tgt}} \models w \models a_{\text{src}}$ (Def. 3.3) and $a_{\text{src}}$ is deterministic, then $a_{\text{tgt}} \models |a_1| \cdots |a_n| \models a_{\text{tgt}} \models a_{\text{src}} |a_1| \cdots |a_n$ (Def. 5.3) for any programs $a_1, \ldots, a_n$.

To prove this theorem, we first show that $a_{\text{tgt}} \models w \models a_{\text{src}}$ implies the existence of a simulation relation between the source and target in SEQ (detailed in [1, Appendix A]). Then, we show that a simulation in SEQ implies the existence of a simulation in $PS^{na}$. For this purpose, we lift steps in SEQ to thread steps in $PS^{na}$. This raises three significant challenges.
First, there is a large gap between SEQ’s simple states and the complex states of PS\textsuperscript{na}. Second, in PS\textsuperscript{na}, we should consider interference by other threads at every point, whereas in SEQ, memory states are changed only in release/acquire steps. Third, we need to show how promise steps of the target in PS\textsuperscript{na} are simulated by the source and establish a PS\textsuperscript{na} certification execution for every step of the source.

The key idea for the first point is that even though PS\textsuperscript{na} has complex states, not all its complexity affects non-atomic steps. In fact, a memory in SEQ can be seen as an approximation of a state in PS\textsuperscript{na} capturing only the part related to non-atomic steps. The value of a location \(x\) in SEQ corresponds to the value of the message pointed by the thread view on \(x\) in PS\textsuperscript{na}, and a permission on \(x\) in SEQ means that there is no racy message with the thread in PS\textsuperscript{na}. Since non-atomic and relaxed accesses do not change the thread view on other locations, states in SEQ are not changed after non-atomic and relaxed accesses. In turn, an acquire read in PS\textsuperscript{na} may increase the thread view, which corresponds to the modified values and gained permissions in acquire steps of SEQ.

For the second point, we need a novel insight on the promising semantics: in a machine step, it suffices to have promise steps followed by non-promise steps ending with a release write (or thread termination). This implies that racy messages of other threads are added only when a release write is executed, which corresponds to SEQ losing permissions only on a release write.

For the third point, we construct the certification steps of the source execution from those of the target. The challenge here is the two cases where the target thread fulfills its promise while the source cannot: (i) when there is no source step corresponding to a write step of the target fulfilling a promise; and (ii) when the written message by the source has a different value than the target’s. This challenge is addressed by the commitment set of the advanced refinement, which ensures that, in both cases, the source thread should be able to write to the problematic locations in the future, thereby allowing the source to establish its certification.

Remark 3. Theorem 6.2 does not hold without the determinism premise (see [1, Appendix C] for an example). This stems from a drawback of the promising semantics (rather than due to SEQ) that we encountered while developing SEQ. Concretely, the promising semantics disallows the reordering of an internal non-deterministic choice followed by a release write. By exposing non-deterministic choices (via \texttt{choose(\_)} labels), we invalidate these reorderings in SEQ and obtain adequacy for deterministic programs. (Nevertheless, the reordering of non-deterministic choices and non-atomic accesses is fully allowed by SEQ.) We leave to future work to improve the promising semantics to allow this reordering, which will allow one to remove the \texttt{choose(\_)} labels from SEQ.

7 Conclusion and Related Work

We developed a sequential model, SEQ, for reasoning about compiler optimizations in a rich weak memory model (concretely, PS\textsuperscript{na}, an extension of PS2.1 with non-atomic accesses), and demonstrated its applicability for compiler verification. This provides the first formal result establishing the adequacy of sequential reasoning for a full-fledged weak memory model without relying on catch-fire semantics for races, accompanied by the first non-trivial certified optimization algorithms for weak memory concurrency. While the ideas and intuitions behind the sequential reasoning are general, adequacy is specifically proved for PS\textsuperscript{na}. Nevertheless, we believe that SEQ can be adapted for reasoning about optimizations in other weak memory models.

Having a sequential model for compiler optimizations paves the way for future work, which has seemed to be out of reach when dealing with complicated concurrency models. This includes the extension of CompCert to weak memory concurrency (in fact, our sequential model is not far from compilers’ model of C, and the simple refinement in §2 may well suffice), as well as of automatic tools like Alive2 [29] for SMT-based translation validation.

Our results have two main limitations. First, SEQ requires the memory layouts of the source and target to be identical, which rules out certain compiler transformations that are performed by CompCert and its extensions mentioned below (although register promotion is supported by PS\textsuperscript{na}). To the best of our knowledge, these are the only thread-local optimizations on non-atomic that compilers actually perform that are unsound under SEQ. Second, our refinement notion is not termination preserving (which requires fairness assumptions, possibly following [20]). Addressing these issues is left to future work.

Next, we discuss the relation to previous work.

Sequential reasoning. The closest to our work is the work by Cuellar et al. [10] (see also [3, 9]) who develop a concurrency semantics, called “concurrent permission machine” (CPM), for CompCert that allows sequential reasoning on program optimizations. Their model has catch-fire semantics, using locks to avoid races. They also present a version of concurrent separation logic that can be used to show that a given program is race-free. While our use of permissions is inspired by these works, our results go beyond lock-based programs, and demonstrate the applicability of sequential reasoning for a significantly more involved model: (i) we handle C11-like atomic access and fences of different modes (from which locks can be implemented); (ii) the model of [9, 10] treats lock/unlock as unknown functional calls, thus forbids optimizations across locks (since they are not performed by CompCert) in contrast to our model that allows optimizations across atomics; (iii) we validate load introduction which is unsound in CPM (in fact, we found out that distinguishing read-only and write permissions, as done in
[9, 10], does not suffice when write-read races are not UB, and developed the idea of written locations set (F) instead; and (iv) the target model in [9, 10] is x86-TSO, which is much simpler than the promising model studied here. All these aspects pose significant challenges in the design of the sequential model and its adequacy proof.

Certified compilation of concurrent programs. Jiang et al. [17] presented CASCompCert, an extension of CompCert deriving certified compilation of concurrent programs from the correctness of sequential compilation, which, in particular, preserves termination. The main difference from our work is that CASCompCert targets DRF programs under sequential consistency (SC), and assumes that racy code (e.g., for the implementation of locks) is confined in manually written assembly assuming x86-TSO and has race-free SC abstractions. As [9, 10], CASCompCert does not support optimizations across locked regions, reorderings of non-atomic and atomic events, and load introduction.

Another extension of CompCert, called thread-safe CompCertX, was presented by Gu et al. [13] in the context of the certified concurrent abstraction layers framework (CCAL). They assume SC as the underlying model, and do not support optimizations on shared non-atomics, which are ubiquitous in concurrent programming.

Earlier work extended CompCert to concurrency [40, 43, 44], for the case that both the source and the target programs have x86-TSO semantics [32] using direct TSO reasoning for the relevant optimization passes. In our terms, this assumes that all accesses are atomic with semantics stronger than release/acquire, rendering various optimizations on non-racy code unsound. Indeed, these optimizations are not performed in the optimization passes of CompCertTSO.

Verification of compiler transformations. Many papers study the correctness of compiler optimizations under certain weak memory models. In particular, Burckhardt et al. [4] develop a denotational approach for compiler optimizations based on the rewrites performed by the target architecture; Ševčík [41] investigates optimizations under a general catch-fire model using locks and synchronization (a.k.a. volatile) accesses; and Vafeiadis et al. [39] provide an extensive study (in Coq) of program transformations in the C/C++11 model [2]. The approach of [39] requires understanding of the C/C++11 model and reasoning about all possible contexts. Another important difference is that the claims in [39] are on the trace-level (represented by execution graphs) leaving implicit the connection to programs.

Based on [39], testing methods and tools for checking the correctness of compiler optimizations were developed [5, 31] and applied on randomly generated programs. Roughly speaking, these validators match (full program) source and target executions and check that the matching adheres to the set of allowed transformations.

Dodds et al. [11] developed a technique and a tool for verifying transformations in the fragment of C11 consisting of release/acquire atomics, non-atomics, and SC fences. They presented a denotational framework for establishing contextual refinement and provided a push-button tool (which does not support non-atomics) using the Alloy model checker.

Program-logics-based approaches. Recently, Gäher et al. [12] developed a separation logic (based on Iris [18]) for contextual refinement in a catch-fire model with SC atomics, allowing, in particular, optimizations involving both atomics and non-atomics. Their refinement preserves termination under fairness assumptions, and allows certain optimizations that modify the memory layout mappings. Interestingly, they considered sequential reasoning as a limitation of previous work, but, as we show, such reasoning does not have to identify atomic accesses with external function calls, and is, thus, capable of reasoning about a variety of optimizations.

Earlier work developed a rely-guarantee relational framework, which also provides means for establishing soundness of program transformations in the presence of assumptions about the environment [26, 27]. It assumes SC as the underlying model, and requires rely-guarantee reasoning for encoding, e.g., data-race-freedom, versus sequential reasoning that ensures refinement under any context as we present.

Compilation scheme correctness. A compilation correctness proof is not only about optimizations, and should also include the correctness of the “mapping schemes” to different architectures. In particular, the aforementioned works, including [9, 10, 17], include the correctness of mappings targeting the x86-TSO architecture. Additional proofs of the correctness of mapping schemes between more complex models appear in [30, 36]. For the promising semantics, mapping correctness was established in Coq [23] for multiple architectures (and the proof trivially generalizes to the extension with non-atomics) via IMM [37]. The latter provides an intermediate model between the programming language models and the various multicore architectures, which can be adapted to accommodate revised models on both sides.

Acknowledgments
We thank the anonymous PLDI reviewers for their helpful feedback. Chung-Kil Hur is the corresponding author. Minki Cho, Sung-Hwan Lee, Dongjae Lee, and Chung-Kil Hur were supported by Samsung Research Funding Center of Samsung Electronics under Project Number SRFC-IT2102-03. Ori Lahav was supported by the Israel Science Foundation (grant number 1566/18) and by the Alon Young Faculty Fellowship. This research was supported in part by the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement no. 851811).
References


