Certified Concurrent Abstraction Layers

Ronghui Gu  
Yale University, USA

Zhong Shao  
Yale University, USA

Jieung Kim  
Yale University, USA

Xiongnan (Newman) Wu  
Yale University, USA

Jérémie Koenig  
Yale University, USA

Vilhelm Sjöberg  
Yale University, USA

Hao Chen  
Yale University, USA

David Costanzo  
Yale University, USA

Tahina Ramananandro  
Microsoft Research, USA

Abstract

Concurrent abstraction layers are ubiquitous in modern computer systems because of the pervasiveness of multithreaded programming and multicore hardware. Abstraction layers are used to hide the implementation details (e.g., fine-grained synchronization) and reduce the complex dependencies among components at different levels of abstraction. Despite their obvious importance, concurrent abstraction layers have not been treated formally. This severely limits the applicability of layer-based techniques and makes it difficult to scale verification across multiple concurrent layers.

In this paper, we present CCAL—a fully mechanized programming toolkit developed under the CertiKOS project—for specifying, composing, compiling, and linking certified concurrent abstraction layers. CCAL consists of three technical novelties: a new game-theoretical, strategy-based compositional semantic model for concurrency (and its associated program verifiers), a set of formal linking theorems for composing multithreaded and multicore concurrent layers, and a new CompCertX compiler that supports certified thread-safe compilation and linking. The CCAL toolkit is implemented in Coq and supports layered concurrent programming in both C and assembly. It has been successfully applied to build a fully certified concurrent OS kernel with fine-grained locking.

1 Introduction

Abstraction layers (e.g., circuits, ISA, device drivers, OS kernels, and hypervisors) are widely used in modern computer systems to help reduce the complex interdependencies among components at different levels of abstraction [3, 49]. An abstraction layer defines an interface that hides the implementation details of its underlying software or hardware components. Client programs built on top of each layer are understood solely based on the interface, independent of the layer implementation.

As multicore hardware and multithreaded programming become more pervasive, many of these abstraction layers also become concurrent in nature. Their interfaces not only hide the concrete data representations and algorithmic details, but also create an illusion of atomicity for all of their methods: each method call is viewed as if it completes in a single step, even though its implementation contains complex interleavings with operations done by other threads. Herlihy et al. [19, 20] advocated using layers of these atomic objects to construct large-scale concurrent software systems.

Figure 1 presents a few common concurrent layer objects in a modern multicore runtime. Here we use the light gray color to stand for thread-local (or CPU-local) objects, blue (also with round dots in their top-right corner) for objects shared between CPU cores, green for objects exported and shared between threads, and orange for threads themselves. Above the hardware layers, we must first build an efficient and starvation-free spinlock implementation [37]. With spinlocks, we can implement shared objects for sleep and pending thread queues, which are then used to implement the thread schedulers, and the primitives yield, sleep, and wakeup. On
and developed under the CertiKOS project [16] for building certified concurrent abstraction layers. As shown in Fig. 2, CCAL consists of a novel compositional semantic model for concurrency, a collection of C and assembly program verifiers, a library for building layered refinement proofs, a thread-safe verified C compiler based on CompCertX [15], and a set of certified linking tools for composing multiprocessed or multicore layers.

We define a certified concurrent abstraction layer as a triple \((L_1[A], M, L_2[A])\) plus a mechanized proof object showing that the layer implementation \(M\), running on behalf of a thread set \(A\) over the interface \(L_1\), indeed faithfully implements the desirable interface \(L_2\) above. Our compositional semantics model is based upon ideas from game semantics [39]. It enables local reasoning such that the implementation can be first verified over a single thread \(t\) by building \((L_1[\{t\}], M, L_2[\{t\}])\) without worrying too much about the concurrency and the guarantees can then be propagated to the whole concurrent machine by parallel compositions.

Following Gu et al. [15], certified concurrent layers enforce termination-sensitive contextual correctness property. In the concurrent setting, this means that every certified concurrent object satisfies not only a safety property (e.g., linearity) [10, 20] but also a progress property (e.g., starvation-freedom) [34].

The CCAL toolkit has already been used in multiple large-scale verification projects under CertiKOS: Gu et al. [16] have successfully used CCAL to build the world’s first fully certified concurrent OS kernel; Kim et al. [24] used CCAL to verify the safety and liveness of a complex MCS lock implementation [37]. Neither of these two papers [16, 24] explained the internals of CCAL and how and why it can work so effectively.

This paper, rather than focusing on the applications of CCAL, gives an in-depth exploration of the CCAL toolkit itself and how it can be used for building various certified concurrent objects. Over Gu et al. [16], this paper presents the following three technical contributions:

- We introduce a new compositional semantic model for shared-memory concurrent abstract machines and prove a general parallel layer composition rule. We show how our new framework is used to specify, verify, and compose various concurrent objects at different levels of abstraction (see Fig. 1).

- We show how to apply standard simulation techniques [15, 27] to verify the safety and liveness of concurrent objects in a unified setting. Because our environment context specifies not just the environment’s past events but also future events, we can readily impose temporal invariants such as fairness requirements (for schedulers) or definite actions [31] (for releasing locks). This allows us to give full specifications for lock primitives and support vertical
composition of starvation-free atomic objects, none of which have ever been possible before \cite{31}.

- We have also developed a new thread-safe version of the CompCertX compiler \cite{15} that can compile certified concurrent C layers into assembly layers. To support certified multithreaded linking, we have developed a new extended algebraic memory model (for CompCertX) whereby stack frames allocated for each thread are combined to form a single coherent CompCert-style memory.

**Scope and Paper Outline.** While the notion of certified concurrent layer can potentially be applied to a more general setting \cite{35, 50}, in this paper, we focus on shared-memory concurrent program modules as described in Anderson and Dahlin \cite{2} and Herlihy and Shavit \cite{19}, which are sufficient to verify layers as shown in Fig. 1. Section 7 discusses related work and puts our work in broader perspective. Both the CCAL toolkit and all our assembly (or C) machines assume strong sequential consistency for shared primitives. Adding support for relaxed memory models is left as future work.

## 2 Overview

The key challenge of verifying concurrent systems is how to untangle the complexities of module dependencies and interleaving, and then verify different parts independently and locally at the layers they belong to. To address this issue, we introduce a layer-based approach to formally specify, certify, and compose these (concurrent) layers.

In this section, to illustrate our layered techniques, we will walk through a small example (see Fig. 3) that uses a lock to protect a critical section. In this example, the client program \(P\) has two threads running on two different CPUs; each thread makes one call to the primitive \(\text{foo}\) provided by the concurrent layer interface \(L_2\). The interface \(L_2\) is implemented by the concurrent module \(M_2\), which in turn is built on top of the interface \(L_1\). The method \(\text{foo}\) calls two primitives \(\text{f}\) and \(\text{g}\) in a critical section protected by a lock. The lock is implemented over the interface \(L_0\) using the ticket lock algorithm \cite{37} in module \(M_1\). The lock maintains two integer variables \(n\) (the “now serving” ticket number) and \(t\) (i.e., next ticket number). The lock acquire method \(\text{acq}\) simply increments the “now serving” ticket number and \(t\). These primitives are provided by \(L_0\) and implemented using x86 atomic instructions. \(L_0\) also provides the primitives \(\text{f}\) and \(\text{g}\) that are later passed on to \(L_1\), as well as a no-op primitive \(\text{hold}\) called by \(\text{acq}\) to announce that the lock has been taken.

![Figure 2. System architecture of the CCAL programming toolkit.](image)

![Figure 3. Certified concurrent layers involving ticket locks.](image)
**Certified Abstraction Layers.** Gu et al. [15] defines a certified sequential abstraction layer as a predicate \( L' \rightarrow_R M : L \) plus a mechanized proof object for the predicate, showing that the layer implementation \( M \), built on top of the interface \( L' \) (which we call the underlay interface), indeed faithfully implements the desirable interface \( L \) above (which we call the overlay interface) via a simulation relation \( R \).

Here, the implementation \( M \) is a program module written in assembly (or C). A layer interface \( L \) consists of a set of abstract states and primitives. An abstract layer machine based on \( L \) is just the base assembly (or C) machine extended with abstract states and primitives defined in \( L \). The implements relation \( (\subseteq_R) \) is formally defined as a forward simulation [28, 36, 38, 44] with the (simulation) relation \( R \).

A certified layer enforces a contextual correctness property: a correct layer is like a "certified compiler," converting any safe client program \( P \) running on top of \( L \) into one that has the same behavior but runs on top of \( L' \) (i.e., by "compiling" abstract primitives in \( L \) into their implementation in \( M \)). If we use \( \llbracket l \rrbracket_L \) to denote the behavior of the layer machine based on \( L \), the correctness property of \( L' \rightarrow_R M : L \) is written formally as \( \forall P. \llbracket P \oplus M \rrbracket_L \subseteq_R \llbracket P \rrbracket_L \) where \( \oplus \) denotes a linking operator over programs \( P \) and \( M \).

**Certified Concurrent Layers.** To support concurrency, each layer interface \( L \) is parameterized with a "focused" thread set \( A \) (where \( A \subseteq D \) and \( D \) is the domain of all thread/CPU IDs). The layer machine based on a concurrent layer interface \( L[A] \) specifies the execution of threads in \( A \) (with threads outside \( A \) considered as the environment). For the example in Fig. 3, the domain \( D \) is \{1, 2\}. If we treat \{1\} as the focused thread set, the environment contains thread 2. For readability, we often abbreviate \( L[\{i\}] \) as \( L[i] \) where \( i \in D \).

A concurrent layer interface extends its sequential counterpart with a set of abstract shared primitives and a global log \( L \). Unlike calls to thread-local primitives which are not observable by other threads, each shared primitive call (together with its arguments) is recorded as an observable event appended to the end of the global log. For example, \( \text{FAI}_t \) (see Fig. 3) called from thread \( i \) takes a log \( I \) to a log \( "I*\text{FAI}_t" \) with the symbol "\*" means "cons-ing" an event to the log.

To define the semantics of a concurrent program \( P \) in a generic way, we develop a novel compositional (operational) model based upon ideas from game semantics [39]. Each run of \( P \) over \( L[D] \) is viewed as playing a game involving members of \( D \) (plus a scheduler): each participant \( i \in D \) contributes its play by appending events into the global log \( L \); its strategy \( \varphi_i \) is a deterministic partial function from the current log \( I \) to its next move \( \varphi_i(I) \) whenever the last event in \( I \) transfers control back to \( i \).

For example, suppose thread \( i \) only invokes \( \text{FAI}_t \), its strategy \( \varphi_i \) can be represented as an automaton:

![Automaton Diagram]

Figure 4. Environment contexts and parallel layer composition.

Suppose the global log is equal to \( L \) when the control is transferred to \( i \) (denoted as \( "i \)\)). Thread \( i \) first generates the event \( i.\text{FAI}_t \) (denoted as \( "i.\text{FAI}_t" \)) and returns the ticket number \( t \) (denoted as \( "t" \)) calculated from \( L \). It then becomes idle (denoted as a reflexive edge labeled with \( "?i', \epsilon'" \)) and will not produce any more events. The ticket number \( t \) is calculated by a function that counts the fetch-and-increment events in \( L \). Such functions that reconstruct the current shared state from the log are called replay functions.

The scheduler (denoted as \( \varphi_0 \)) acts as a judge of the game. At each round, it picks one thread to make a move (and generate events) using its strategy. The behavior of the whole layer machine (denoted as \( \llbracket L[D] \rrbracket \)) is then just the set of logs generated by playing the game under all possible schedulers.

When focusing on a subset of threads \( A \), the semantics (or execution) of the (concurrent) layer machine based on an interface \( L[A] \) is defined over its set of valid environment contexts. Each environment context (denoted as \( E \)) provides a strategy for its "environment," i.e., the union of the strategies by the scheduler plus those participants not in \( A \).

For example, Fig. 4 shows a system with two threads \( (t_1 \) and \( t_2 \)) and a scheduler. On the left, it shows one execution of method \( \text{foo} \) over the layer machine \( L'[\{t_1\}] \) under a specific environment context \( E' \). Here, \( E'_i \) is the union of the strategy \( \varphi' \) for the scheduler and \( \varphi' \) for thread \( t_2 \). In the middle, it shows the execution of \( \text{foo} \) (invoked by \( t_2 \)) over \( L'[\{t_2\}] \) under the environment context \( E'_2 \). On the right, it shows the interleaved execution of two invocations to \( \text{foo} \) over \( L'[\{t_1, t_2\}] \) where the environment context \( E' \) is just the scheduler strategy \( \varphi'_0 \).

Given an environment context \( E \) which also contains a specific scheduler strategy, the execution of \( P \) over \( L[A] \) is deterministic; the concurrent machine will run \( P \) when the control is transferred to any member of \( A \), but will ask \( E \) for the next move when the control is transferred to the environment.

To enforce the safety of environmental moves, each layer interface also specifies its set of valid environment contexts. This validity corresponds to a generalized version of the "rely" (or "assume") condition in rely-guarantee-based reasoning [7, 8, 11, 52, 58]. Each layer interface can also provide its own "guarantee" condition. These conditions are simply expressed as invariants over the global log.
Local Layer Interface. Consider the case where the focused thread set is a singleton \( i \). Since the environmental executions (including the interleavings) are all encapsulated into the environment context, \( L[i] \) is actually a sequential-like (or local) interface parameterized over \( E \). Before each move of a client program \( P \) over this local interface, the layer machine first repeatedly asks \( E \) for environmental events until the control is transferred to \( i \). It then makes the move based on received events. Thus, the semantics of running \( P \) over \( L[i] \) (denoted as \( (P)_{L[i]} \)) can also be viewed as a strategy.

The correctness property asserting that a concurrent module on top of a local layer interface indeed satisfies its specification (i.e., a more abstract strategy) is defined by the strategy simulation via a simulation relation \( R \) for logs.

Definition 2.1. (\( \leq_R \)) We say a strategy \( \phi \) is simulated by another strategy \( \phi' \) with a simulation relation \( R \) and write “\( \phi \leq_R \phi' \)”, if, and only if, for any two related (by \( R \)) environmental event sequences and any two related initial logs, we have that for any \( \log \) \( l \) produced by \( \phi \), there must exist a \( \log \) \( l' \) that can be produced by \( \phi' \) such that \( l \) and \( l' \) also satisfy \( R \).

Consider the \( \text{acq} \) method of the ticklock module \( M_1 \) running over \( L_0[i] \) (see Fig. 3). Its specification can be represented as the following strategy \( \phi'_{\text{acq}}[i] \):

\[
\begin{align*}
\tau_E, !i.\text{acq},_t, & \downarrow t \\
\tau_E, !i.\text{get},_n, & \downarrow t \\
\tau_E, !_i.\text{hold}, & \downarrow t
\end{align*}
\]

We write \( \tau_E \) for querying \( E \). We can prove that the simulation “\( (\text{acq})_{L_0[i]} \leq_{\text{id}} \phi'_{\text{acq}}[i] \)" holds for the identical relation: for any equal \( E \) and equal initial state, if \( \phi'_{\text{acq}}[i] \) takes one step, \( \text{acq} \) can take one (or more) steps to generate the same event and the resulting states are still equal. This correctness property is also used to define certified concurrent layers:

\[
L_0[i] \rightarrow_{\text{id}} \text{acq} : \phi'_{\text{acq}}[i] := (\text{acq})_{L_0[i]} \leq_{\text{id}} \phi'_{\text{acq}}[i]
\]

Let “\( M_1 := \text{acq} \otimes_{\text{rel}} \)" and “\( L'_{1[i]} := \phi'_{\text{acq}}[i] \otimes_{\text{rel}} \phi'_{\text{rel}}[i] \)". By showing that the lock release satisfies its specification (i.e., “\( L_0[i] \rightarrow_{\text{rel}} \phi'_{\text{rel}}[i] \)"), and by the horizontal composition rule (see Sec. 3.3), we have:

\[
L_0[i] \rightarrow_{\text{id}} M_1 : L'_{1[i]} := (M_1)_{L_0[i]} \leq_{L[i]} L'_{1[i]}
\]

The notations are extended to a set of strategies, meaning that each strategy of \( L'_{1[i]} \) simulates one of the \( (M_1)_{L_0[i]} \).

Higher-level Strategies. Although the specifications above (e.g., \( \phi'_{\text{acq}}[i] \)) are abstract (i.e., language independent), low-level implementation details and interleavings within the module are still exposed. For example, \( \phi'_{\text{acq}}[i] \) reveals the loop that repeatedly interacts with the environment to check the serving ticket number. To simplify the verification of components using locks, we have to refine the strategies of \( L'_{1[i]} \) to a higher-level interface \( L[i] \) that is atomic:

\[
L_0[i] +_{R_1} M_1 : L_1[i] \oplus L_1[i] +_{R_2} M_2 : L_2[i]
\]

vertical composition

\[
L_0[i] +_{R_1 \circ R_2} M_1 \circ M_2 : L_2[i]
\]

thread-safe composition

\[
\epsilon_{i} \in \{1, 2\}
\]

\[
L_0[i] +_{R_1 \circ R_2} \text{CompCertX}(M_1 \circ M_2) : L_2[i]
\]

parallel composition

\[
\forall P, P \oplus \text{CompCertX}(M_1 \circ M_2) \leq_{R_1 \circ R_2} \text{soundness theorem}
\]

Figure 5. Layer verification of the ticket lock example using CCAL.

We write \( L_1[i] := \phi'_{\text{acq}}[i] \) to denote queries \( E \) and produces a single event \( i.\text{acq} \). It then enters a so-called critical state (marked as gray) to prevent losing the control until the lock is released. Thus, there is no need to ask \( E \) in critical state.

To prove the strategy simulation between \( L'_{1[i]} \) and \( L_1[i] \), we have to prove “relies” (i.e., validity) conditions \( R \) over the environment context of \( L'_{1[i]} \):

- \( \emptyset, L'_{1[i]} : R'_{hs} \): the scheduler strategy \( \phi'_{hs} \) must be \( \text{fair} \).
- \( \emptyset, L'_{1[i]} : R'_{j} \) (\( j \neq i \)): lock-related events generated by \( \phi'_{j} \) must follow \( \phi'_{\text{acq}}[j] \) and \( \phi'_{\text{rel}}[j] \), and the held locks will eventually be released.

These conditions ensure that the loop (waiting for the ticket to be served) in \( \phi'_{\text{acq}}[i] \) terminates. Also, they can be used to prove that each run of \( L'_{1[i]} \) is captured by \( L_1[i] \). For example, if the scheduler strategy \( \phi'_{hs} \) schedules as “\( 1, 2, 2, 1, 1, 2, 1, 2, 2, 1, 2, 1, 2, 2, 1, 2 \)" running \( P \) (see Fig. 3) over \( L'_{1[D]} \) generates the log:

\[
L'_{1[D]} := (1.\text{acq}) \oplus (2.\text{acq}) \oplus (2.\text{get}) \oplus (1.\text{get}) \oplus (1.\text{hold}) \oplus (2.\text{get}) * (1.\text{f}) \oplus (2.\text{get}) \oplus (1.\text{g}) \oplus (1.\text{inc}) \oplus (2.\text{get}) \oplus (2.\text{hold})
\]

This interleaving can be captured by a higher-level scheduler \( \phi_{hs} \) producing “\( 1, 2, 2 \)" (recall that thread 1 is in the critical state while holding the lock), and the generated log at \( L_1[D] \) is:

\[
L_{1[D]} := (1.\text{acq}) \oplus (1.\text{f}) \oplus (1.\text{g}) \oplus (1.\text{rel}) \oplus (2.\text{acq})
\]

Although logs (and events) at these two layers are different, the order of lock acquiring and the resulting shared state (calculated from logs by replay functions) are exactly the same. By defining the relation \( R_1 \) over logs as mapping events \( i.\text{acq} \) to \( i.\text{hold}, i.\text{rel} \) to \( i.\text{inc}_n \) and other lock-related events to empty ones, we can prove:

\[
L'_{1[i]} \leq_{R_i} L_1[i]
\]
Then by the predicate (2.1) and the weakening rule (i.e., the Wx rule in Fig. 9), we have that:

$$L_0[i] \vdash_{id} R_1 = R_1, M_1 : L_1[i]$$  \hspace{1cm} (2.2)

Similarly, for the foo method (i.e., $M_z$ in Fig. 3), we can also introduce a low-level strategy $\varphi'_{foo[i]}$ as the first step:

$$L'_2[i] := \varphi'_{foo[i]} : \varphi'_{foo[i]} : \varphi''_{foo[i]}$$

Then we prove that a high-level atomic interface $\varphi_{foo}$:

$$L_2[i] := \varphi_{foo[i]} : \varphi_{foo[i]} : \varphi''_{foo[i]}$$

simulates (with some $R_2$) $\varphi'_{foo}$, which in turn simulates $\varphi_{foo}$:

$$L''_2[i] \leq R, L_2[i] \quad L_1[i] \vdash_{id} M_2 : L'_2[i]$$  \hspace{1cm} (2.3)

Based on $L''_2[i]$, we can derive the “guarantee” condition $G$ of the thread $t$ saying that held locks are always released within three steps, which is consistent but more concrete than the rely condition $R$ defined above.

**Parallel Layer Composition.** We say that two layer interfaces $L[t_1]$ and $L[t_2]$ are compatible if the guarantee $G$ of each interface implies the other interface’s rely conditions $R$. The new compositional model allows us to prove a general parallel layer composition rule: if $L[t_1]$ is compatible with $L[t_2]$, $L[t_1]$ is compatible with $L[t_2]$, and “$L[t] \vdash R M : L[t]$” holds for every $t \in \{t_1, t_2\}$, then we have “$L[t_1, t_2] \vdash R M : L[t_1, t_2]$.”

Figure 4 shows how to compose certified local layers (one for $t_1$ and another for $t_2$) to build a certified layer for the entire machine (with $t_1$ and $t_2$ both focused).

**Thread-Safe CompCertX and Layer Linking.** Since the local interface is sequential-like, we can adapt the CompCertX compiler [15] to be thread-safe by merging the stack frames of threads on the same CPU into a single stack. In this way, certified C layers can be compiled into certified assembly layers. We can then apply the horizontal, vertical, and the new parallel layer composition rules (see Sec. 3.3) to construct the certified concurrent layer for the entire system (see Fig. 5). Finally, from “$L'[D] \vdash R M : L[D]$,” the soundness theorem enforces a strong contextual refinement property saying that, for any client program $P$, we have that for any log $l$ in the behavior $[P \oplus M]_{L[D]}$, there must exist a log $l'$ in the behavior $[P]_{L[D]}$ such that $l$ and $l'$ satisfy $R$.

**Theorem 2.2 (Soundness).**

$$L'[D] \vdash R M : L[D] \Rightarrow \forall P, [P \oplus M]_{L'[D]} \sqsubseteq_R [P]_{L[D]}$$

### 3 Concurrent Layer Interface and Calculus

In this section we instantiate our compositional model for the x86 multicore hardware and explain the concurrent layer interface and the layer calculus in more detail.

```plaintext
Function $\sigma_{pull}'$ (s: State) (b: Loc) :=

match s.a.status b with
| free => ret s(1: s.c.pull(b):s.l){a.status.b: own s.c}
| _ => none (* get stuck*)
end.
```

**Figure 6.** Pseudocode of the pull specification of $M_{x86}$ in Coq.

### 3.1 Multiprocessor Machine Model

The multiprocessor machine model $M_{x86}$ is defined by the machine state, the transition relation, and the memory model.

**Machine State.** As shown in Fig. 7, the state of $M_{x86}$ is denoted as a tuple “$s := (c, f_p, m, a, l)$” where the components are the current CPU ID $c$, all CPUs’ private states $f_p$ (i.e., a partial map from CPU ID to private state $p$), a shared memory state $m$, an abstract state $a$, and a global event log $l$. The private state $p$ consists of CPU-private memory $m$ (invisible to other CPUs) and a register set $r$. The shared memory state $m$ is shared among all CPUs. Each location $b$ in both local and shared memories contains a memory value $v$. The abstract state $a$ is generally used in our layered approach to summarize in-memory data structures from lower layers. It is not just a ghost state, because it affects program execution when making primitive calls. The global log $l$ is a list of observable events, recording all shared operations that affect more than one CPU. Events generated by different CPUs are interleaved in the log, following the actual chronological order of events.

**Transition Relation.** The machine $M_{x86}$ has two types of transitions that are arbitrarily and nondeterministically interleaved: program transitions and hardware scheduling.

**Program transitions** are one of three possible types: instruction executions, private primitive calls, and shared primitive calls. The first two types are “silent”, in that they do not generate events. Shared primitives, on the other hand, provide the only means for accessing and appending events to the global log. The transitions for instructions only change $p$, $pm$, and $m$, and are defined as standard operational semantics for C or x86-assembly, similar to (and in fact based on) the operational semantics used in CompCert [28]. Primitive calls are specific to our style of verification: they directly specify the semantics of function $f$ from underlying layers as a relation $\sigma_f$ defined in Coq. This relation specifies how the state is updated after $f$ is called with the given arguments and what the return value is.

**Hardware scheduling** transitions change the current CPU ID $c$ to some ID $c'$ (recorded as a scheduling event) and can be arbitrarily interleaved with program transitions. In other words, at any step, $M_{x86}$ can take either a program transition staying on the current CPU, or a hardware scheduling to another CPU. The behavior of a client program $P$ over this multicore machine (denoted as $[P]_{M_{x86}}$) is a set of global logs generated by executing $P$ via these two kinds of transitions.
We now zoom in on the execution of a subset of CPUs A. We introduce a "push/pull" memory model,

Memory Model. We introduce a “push/pull” memory model for the shared memory m (the private memory is separately handled in ρ), which encapsulates the shared memory operations into push/pull events and can detect data races.

In this model, each shared memory location b is associated with an ownership status in the abstract state a, which can only be manipulated by two shared primitives called pull and push. The pull operation modifies the ownership from "free" to "owned by c", after which shared memory accesses can be performed by CPU c. The push operation frees the ownership and records its memory updates in the log. Figure 6 shows the specification σ_pull, where "r {i : v}" means updating the record r at field i with value v.

If a program tries to pull a not-free location, or tries to access or push to a location not owned by the current CPU, a data race may occur and the machine gets stuck. One goal of concurrent program verification is to show that a program is data-race free; in our setting, we accomplish this by showing that the program does not get stuck.

3.2 Concurrent Layer Interface

We now zoom in on the execution of a subset of CPUs A, introducing the concurrent layer interface L[A] defined as a tuple (L, R, G). The machine based on this concurrent interface is "open" in the sense that it is capable of capturing a subset of the CPUs and then be composed with any acceptable execution of the rest of CPUs. The domain of the private state map f_p is also this captured (or focused) subset. The interface L[A] equips this open machine with a collection of primitives that are defined in L and can be invoked at this level, the rely condition R that specifies a set of acceptable environment contexts, and the guarantee condition G that the log l should hold. The instruction transitions are defined as before, but all hardware scheduling is replaced by queries to the environment context.

Environment Context. E is a partial function from a CPU ID to its strategy φ. A strategy is an automata that generates events in response to given logs. When focusing on a CPU set A, all the observable behaviors of the hardware scheduling and the program transitions of other CPUs can be specified as a union of strategies (i.e., E). Thus, whenever there is a potential interleaving, the machine can query E about the events from other CPUs (and the scheduler).

These environmental events cannot influence the behaviors of instructions and private primitive calls. This also applies to shared memory read/write because the push/pull memory model encapsulates other CPUs’ effects over the shared memory into push/pull events. Thus, during the execution of instructions and private primitives, it is unnecessary to query E, and the appended environmental events will be received by the next so-called query point, that is, the point just before executing shared primitives.

To be more specific, at each query point, the machine repeatedly queries E. Each query takes the current log l as the argument and returns an event (i.e., “E(c’, l’)”) from a CPU c’ not in A. That event is then appended to l, and this querying continues until there is a hardware transition event back to A (assuming the hardware scheduler is fair). In the following, we write “E[A, l]” to mean this entire process of extending l with multiple events from other CPUs.

Rely and Guarantee Conditions. The R and G of the layer interface specify the validity of the environment context and the invariant of the log (containing the locally-generated events). After each step of threads in A over interface L[A], the resulting log l must satisfy the guarantee condition L[A], G, i.e., l ∈ L[A], G(c) if c is the current CPU ID indicated by l. To prove that guarantee conditions always hold, we not only need to validate the events generated locally but also need to rely on the validity of the environment context. The rely condition L[A], R specifies a set of valid environment contexts, which take valid input logs and return a valid list of events.

CPU-Local Layer Interface. When focusing on a single CPU c, L[c] is called a CPU-local layer interface. Its machine state is (ρ, m, a, l), where ρ is the private state of the CPU c and m is just a local copy of the shared memory.

This m can only be accessed locally by c. The primitives push/pull of L[c] "deliver" the effects of shared memory operations. Figure 8 shows their specifications, which depend on a replay function E_shared to reconstruct the shared memory value v for some location b and check the well-formedness (i.e., no data race occurs) of the resulting log.

Since σ_pull is parameterized with E, it can also be viewed as the following special strategy with private state updates:
Fixpoint \( R_{\text{shared}} \) \((l :: \text{Loc}) (b :: \text{Loc}) (c :: \text{Id}) \) :=
match \( l \) with
| nil \=> ret (vundef, free) (* initial value *)
| e :: l' \=> (* l' • a *)
do r \<- \( R_{\text{shared}} \) l' b; (* Haskell syntax sugar*)
end.

end.

| _ => None (* get stuck *)
| (v, free), e :: l
end.

Figure 8. Pseudocode of push/pull specifications of \( L[c] \) in Coq.

The layer machine enters the critical state after calling pull by holding the ownership of a shared location. It exits the critical state by invoking push to free the ownership.

3.3 Concurrent Layer Calculus

To build and compose concurrent layers “\( L[A] + M : L'[A] \)” we introduce a layer calculus shown in Fig. 9. We borrow the notations from Gu et al. [15]: “\( \Box \)” stands for an empty program module. “\( \Theta \)” computes the union of two modules (or two layers’ primitive collections), and “\( (i \leftrightarrow \cdot) \)” is a singleton map with a pointer or location \( i \) as its domain.

Composition Rules. The vertical composition rule (Vcomp) allows us to verify the modules \( M \) and \( N \) (where \( N \) may depend on \( M \)) in two separate steps, while the horizontal composition rule (Hcomp) enables local reasoning for independent modules \( M \) and \( N \) belonging to the same level. These two composition rules can only compose layers over the same CPU set.

Layers on different CPUs can be composed by the parallel composition rule (Pcomp) if simulation relations are the same, and both overlay and underlay interfaces are compatible. This compatibility is denoted as “\( \text{compat}(L[A], L[B], L[A \cup B]) \)” It says that each guarantee condition of \( L[A] \) implies the corresponding rely condition of \( L[B] \) and vice versa. The composed interface \( L[A \cup B] \) merges the primitives of two layers and is equipped with stronger guarantees and weaker rely conditions. The machine based on this composed layer interface only queries \( \mathcal{E} \) about the events not from \( A \cup B \).

Multicore Linking Theorem. By composing all the CPUs in the machine (denoted as the set \( D \)), the resulting layer interface does not depend on any environmental events except those from the hardware scheduler. We construct such a layer interface \( L_{\text{x86}}[D] \) using the primitives provided by the hardware \( M_{\text{x86}} \). We can then prove a contextual refinement from \( M_{\text{x86}} \) to \( L_{\text{x86}}[D] \) by picking a suitable hardware scheduler of \( L_{\text{x86}}[D] \) for every interleaving (or log) of \( M_{\text{x86}} \).

Theorem 3.1 (Multicore Linking).
\[
\forall P, P \parallel_{\text{x86}} \preceq R \parallel L_{\text{x86}}[D]
\]
This theorem ensures that all code verification over \( L_{\text{x86}}[D] \) can be propagated down to the x86 multicore hardware \( M_{\text{x86}} \).

Building Leaf Certified Layers. As the unit of certified concurrent layers, leaf layers can be built by applying the Fun rule, which requires to prove the strategy simulation. Two most common patterns, fun-lift and log-lift, for this proof have already been shown in Sec. 2. The fun-lift pattern abstracts a concrete implementation into a low-level strategy without changing the potential interleaving. In this pattern, language dependent details (e.g., silent moves changing temporal variables) are hidden and data representation details (e.g., memory values carried by push events) are replaced with abstract state values.

The log-lift pattern always involves the events merging and the interleavings shuffling to form an atomic interface.

4 Building Certified Multicore Layers

In this section, we start to show how to apply our techniques to verify shared objects in the CCAL toolkit. All layers are built upon the CPU-local layer interface \( L_{\text{x86}}[c] \).

4.1 Spinlocks

Spinlocks (e.g., the ticket lock algorithm described in Sec. 2) are one of the most basic synchronization methods for multicore machines; they are used as building blocks for shared objects and more sophisticated synchronizations.

A spinlock enforces mutual exclusion by restricting CPU access to a memory location \( b \). Therefore, lock operations can be viewed as “safe” versions of push/pull primitives. For example, when the lock acquire for \( b \) succeeds, the corresponding shared memory is guaranteed to be “free”, meaning that it is safe to pull the contents to the local copy at this point (line 4 in Fig. 10). We now show how to build layers for the spinlock in Fig. 10, which uses a ticket lock algorithm. Note that query points are denoted as “\( \Box \)” in pseudocode.

Bottom Interface \( L_{\text{x86}}[c] \). We begin with the CPU-local interface \( L_{\text{x86}}[c] \) extended with shared primitives \( \text{FAI}_t \), \( \text{get}_n \), and \( \text{inc}_n \). These primitives directly manipulate the lock state \( t \) (next ticket) and \( n \) (now serving ticket) via x86 atomic instructions. The lock state can be calculated by a replay function \( R_{\text{ticket}} \) counting \( t \), \( \text{FAI}_t \) and \( \text{inc}_n \) events.

Fun-Lift to \( L_{\text{lock-low}}[c] \). We have shown how to establish the strategy simulation for this low-level interface \( L_{\text{lock-low}}[c] \)


\[
\begin{align*}
&\frac{[\kappa]|L[c]| \leq_R \sigma}{L[c] \vdash_{id} i \mapsto \kappa \mapsto \sigma} \\
&\frac{\text{EMPTY}}{L[A] \vdash_{id} \top : L[A]} \\
&\frac{L[A] \vdash_{R} M : L_{1}[A]}{L'[A], L = L_{1}[A], \mathcal{L} \circ \oplus @ \mathcal{L}, \mathcal{L}} \\
&\frac{L'[A], R = L_{1}[A], \mathcal{R}}{L_{2}[A], \mathcal{R}} \\
&\frac{L'[A], \mathcal{G} = L_{1}[A], \mathcal{G} = L_{2}[A], \mathcal{G}}{L_{2}[A] \vdash_{S} N : L_{3}[A]} \\
&\frac{L_{2}[A] \vdash_{S} N : L_{3}[A]}{L_{1}[A] \vdash_{R \circ S} M \circ N : L_{3}[A]} \\
&\frac{L[A] \vdash_{R} M \circ N : L_{3}[A]}{L'[A] \vdash_{R \circ S} M : L_{3}[A]} \\
&\frac{L[A] \vdash_{R} M : L_{1}[A]}{L_{1}[A] \vdash_{R \circ S} M : L_{2}[A]} \\
&\frac{L_{1}[A] \vdash_{R \circ S} M : L_{2}[A]}{L_{2}[A] \vdash_{S} N : L_{3}[A]} \\
&\frac{L_{2}[A] \vdash_{S} N : L_{3}[A]}{L_{1}[A] \vdash_{R \circ S} M \circ N : L_{3}[A]} \\
&\frac{L_{2}[A] \vdash_{S} N : L_{3}[A]}{L_{1}[A] \vdash_{R \circ S} M \circ N : L_{3}[A]}
\end{align*}
\]

**Figure 10.** The fine-grained layer calculus in the concurrent setting.

1. **void acq (uint b) {**
   2.     \textbf{uint} myt = mFAIL_{t}(b);
   3.     \textbf{push(b);}
   4.     \textbf{while (!get_n(b) = myt)}{**
   5.         \textbf{>inc_n(b);}
   6.         \textbf{pull(b); // acts as hold}()**
   7.     }**

**Figure 10.** Pseudocode of ticket lock using push/pull.

(i.e., \(L'_{1}[c]\), see Sec. 2). Note that \(\{\text{acq}\}_{\text{lock,low}}[c]\) contains extra silent moves (e.g., assigning myt, line 2 in Fig. 10) compared with \(\varphi_{\text{acq}}[c]\). The simulation relation \(R_{\text{lock}}\) not only states the equality between logs but also maps the lock state in the memory to the ones calculated by \(R_{\text{ticket}}\). Here we must also handle potential integer overflows for \(t\) and \(n\). We can prove that, as long as the total number of CPUs (i.e., \#CPU) in the machine is less than \(2^{32}\) (determined by uint), the mutual exclusion property will not be violated even with overflows.

**Log-Lift to \(L_{\text{lock}}[c]\).** We then lift the \(\text{acq}\) and \(\text{rel}\) primitives to an atomic interface, meaning that each invocation produces exactly one event in the log (see Sec. 2). These atomic lock interfaces (or strategies) are similar to \(\text{push/pull}\) specifications, except that the former ones are safe (i.e., will not get stuck). This safety property can be proved using rely conditions \(L_{\text{lock}}[c]\), \(R\) saying that, for any CPU \(c' \neq c\), its \(c'.\text{acq}\) event must be followed by a sequence of its own events (generated in the critical state) ending with \(c'.\text{rel}\). The distance between \(c'.\text{acq}\) and \(c'.\text{rel}\) in the log is less than some number \(n\).

By enforcing the fairness of the scheduler in rely conditions, saying that any CPU can be scheduled within \(m\) steps, we can show the liveness property (i.e., starvation-freedom): the while-loop in \(\text{acq}\) terminates in \(\lceil n \times m \times \#\text{CPU} \rceil\) steps.

### 4.2 Shared Queue Object

Shared queues are widely used in concurrent programs, e.g., as the list of threads in a scheduler, etc. In previous work [31], due to the lack of layering support, the verification of any shared object required inlining the lock implementation and duplicating the lock-related proofs. In the following, we illustrate how to utilize concurrent abstraction layers to verify a shared queue module using fine-grained locks.

**Fun-Lift to \(L_{q}[c]\).** The shared queues are implemented as doubly linked lists, and are protected by spinlocks. For example, the dequeue (\(\text{deQ}\)) operation first acquires the spinlock associated with queue \(i\), then performs the actual dequeue operation in the critical state, and finally releases the lock. Instead of directly verifying \(\text{deQ}\) in one shot, we first introduce an intermediate function \(\text{deQ}_{-t}\), which contains code that performs the dequeue operation over a local copy, under the assumption that the corresponding lock is held. Since no environmental queries are needed in the critical state, building concurrent layers for \(\text{deQ}_{-t}\) is similar to building a sequential layer [15]: we first introduce the abstract states \(a.tcbp\) and \(a.tdpq\), which stand for the thread control block (i.e., tcb) array and the thread queue array. The abstract \(tdqp\) is a partial map from the \textit{queue index} to an \textit{abstract queue}, which is represented as a list of tcb indices. Then we can show that \(\text{deQ}_{-t}\) meets its specification \(\sigma_{\text{deQ}_{-t}}\):

1. **function \(\sigma_{\text{deQ}_{-t}} : E (\text{State}) (i : \text{Loc}) :**
2.     \textbf{do } r \leftarrow E_{\text{shared}} s . l i . s . c ; (* \text{replay ownership} *)
3.     \textbf{match } r \textbf{ with }
4.     \textbf{ | } (_) , \textbf{ own } s . c \textbf{ => } (* \text{if the lock of queue } i \text{ is held} *)
5.     \textbf{ | } TD :: q \textbf{ => } ret (s (s . a . tdpq . i . q), TD)
6.     \textbf{ | } - \textbf{ => } ret (s, -1) (* return -1 for empty queue *)
7.     \textbf{end}
8.     \textbf{| } \textbf{ => } None (* get stuck *)
9.     \textbf{end}.

**Fun- and Log-Lift to \(L_{q\_\text{high}}[c]\).** Finally, we have to show that the \(\text{deQ}\) function that wraps \(\text{deQ}_{-t}\) with lock primitives
5 Building Certified Multithreaded Layers

Multithreaded programs have to deal with interleavings triggered by not only the hardware scheduler but also the explicit invocation of thread scheduling primitives. In this section, we introduce the certified layers dealing with scheduling primitives, a new concept of thread-local layer interfaces equipped with compositional rules, and a thread-safe version of CompCertX.

5.1 Certified Layers for Scheduling Primitives

Based on the shared thread queues provided by the multicore toolkit (see Sec. 4.2), we introduce a new layer interface $L_{\text{btd}}[c]$ that supports multithreading. At this layer interface, the transitions between threads are done using scheduling primitives, implemented in a mix of C and assembly.

In our multithreaded setting, each CPU $c$ has a private ready queue $rdq$ and a shared pending queue $pendq$ containing the threads woken up by other CPUs. A thread yield sends the first pending thread from $pendq$ to $rdq$ and then switches to the next ready thread. There are also many shared sleeping queues $slpq$. When a sleeping thread is woken up, it will be directly appended to the ready queue if the thread belongs to the currently-running CPU. Otherwise, it will be appended to the pending queue of the CPU it belongs to.

Thread switching is implemented by the context switch function $cswitch$, which saves the current thread’s kernel context (i.e., $ra$, $ebp$, $ebx$, $esi$, $edi$, esp), and loads the context of the target thread. This $cswitch$ (invoked by $yield$ and $sleep$) can only be implemented at the assembly level, as it does not satisfy the C calling convention. A scheduling primitive like $yield$ first queries $E$ to update the log, appends its own event, and then invokes $cswitch$ to transfer the control.

This layer interface introduces three new events $c.yield$, $c.sleep(i, lk)$ (sleep on queue $i$ while holding the lock $lk$), and $c.wakeup(i)$ (wakeup the queue $i$). These events record the thread switches, which can be used to track the currently-running thread by a replay function $R_{\text{sched}}$.

5.2 Multithreaded Layer Interface

The CPU-local interface $L_{\text{btd}}[c]$ captures the execution of the whole thread set of CPU $c$ and does not support thread-local verification. Ideally, we would like to formally reason about each thread separately and later compose the proofs together obtaining a global property. Thus, we introduce a new layer interface that is compositional and only focuses on a subset of thread running on CPU $c$.

Let $T_a$ denote the whole thread set running over CPU $c$. Based upon $L[c]$, we construct a multithreaded layer interface $"L[c][T_a]:= (L[c].L[c].R \cup R_{T_a}, L[c].G[|T_a]|)"$ which is parameterized over a focused thread set $T_a \subseteq T_c$. Besides $T_q$, strategies of other threads running on $c$ form a thread context $E'$. Rely conditions of this multithreaded layer interface extend $L[c].R$ with a valid set of $E'$ (denoted as $"R_{T_a}"$) and guarantee conditions replace $L[c].G(c)$ with the invariants held by threads in $T_a$ (denoted as $"L[c].G[|T_a]|"$). Since our machine model does not allow preemption, $E'$ will only be queried during the execution of scheduling primitives, which have two kinds of behaviors depending on whether the target thread is focused or not.

Consider the above execution with $T_a = \{0, 1\}$. Whenever an execution switches (by $yield$ or $sleep$) to a thread outside of $T_a$ (i.e., the yellow $yield$ above), it takes environmental steps (i.e., inside the red box), repeatedly appending the events returned by the environment context $E$ and the thread context $E'$ to the log until a $y$yield event indicates that the control has switched back to a focused thread. Whenever an execution switches to a focused one (i.e., the blue $yield$ above), it will perform the context switch without asking $E/E'$ and its behavior is identical to the one of $L_{\text{btd}}[c]$.

Composing Multithreaded Layers. Multithreaded layer interfaces with disjoint focused thread sets can also be composed in parallel (using an extended Pcomp rule) if the guarantee condition implies the rely condition for every thread. The resulting focused thread set is the union of the composed ones, and some environmental steps are "replaced by" the local steps of the other thread set. For example, if we compose $T_a$ in the above example with thread 2, the previously yellow $yield$ of thread 0 will then switch to a focused thread.
void acq_q(uint l) {  /* */  ql_busy[l] = get_tid();  /* */}
if (ql_busy[l] != -1) {  /* */  void rel_q (uint l) {  /* */  ql_busy[l] = get_tid();  /* */  rel(ql_loc(l));  /* */  }
} else {  /* */  sleep(l);  /* */  void acq_q(uint l) {  /* */  }
}  /* */

Figure 11. Pseudocode of queuing lock.

Here, the event list $l_t$ generated by $E$ and $E'$ has been divided into two parts: "$i_{ta} \bullet c$.yield" (generated by thread 2) and $l_{tb}$ (consisting of events from threads outside $\{0,1,2\}$).

**Multithreaded Linking.** When the whole $T_c$ is focused, all scheduling primitives fall into the second case and never switch to unfocused ones. Thus, its scheduling behaviors are equal to the ones of $L_{h}{\text{d}}[c]$. By introducing a multithreaded layer interface $L_{h}{\text{d}}[c][T_c]$ that contains all the primitives of $L_{h}{\text{d}}[c]$, we can prove the following theorem:

**Theorem 5.1 (Multithreaded Linking).**

$$L_{h}{\text{d}}[c] \leq_{\text{id}} L_{h}{\text{d}}[c][T_c]$$

This theorem guarantees that, once the multithreaded machine based on $L_{h}{\text{d}}[c][T_c]$ captures the whole thread set, the properties of threads running on top can be propagated down to the layer with concrete scheduling implementations.

5.3 Thread-Local Layer Interface

If a multithreaded interface $I[c][T]$ focus only on a single thread $t \in T_c$, yield and sleep primitives always switch to an unfocused thread and then repeatedly query $E$ and $E'$ until yielding back to $t$.

We can prove that this yielding back procedure in our system always terminates. This proof relies on the fact that the software scheduler is fair and every running thread gives up the CPU within a finite number of steps. We call $I[c][T]$ a "thread-local" layer interface because scheduling primitives always end up switching back to the same thread; they do not modify the kernel context (i.e., ra, ebp, ebx, esi, edi, esp) and effectively act as a "no-op", except that the shared log gets updated. Thus, these scheduling primitives indeed satisfy C calling conventions.

5.4 Queuing Lock

Based upon thread-local layer interfaces, we build additional synchronization toolkits, such as a queuing lock (see Fig. 11). With queuing locks, waiting threads are put to sleep to avoid busy spinning. Reasoning about this locking algorithm is particularly challenging since its C implementation utilizes both spinlocks and low-level scheduler primitives (i.e., sleep and wakeup). This verification task can be decomposed into a bunch of layers above $L_{h}{\text{d}}[c][T]$ using CCAL.

The correctness property of a queuing lock consists of two parts: mutual exclusion and starvation freedom. The lock implementation (Fig. 11) is mutually exclusive because the busy value of the lock (ql_busy) is always equal to the lock holder’s thread ID. This busy value is set either by the lock requester when the lock is free (line 6 of Fig. 11) or by the previous lock holder when releasing the lock (line 12). With the atomic interface of the spinlock, the starvation-freedom proof of queuing lock is mainly about the termination of the sleep primitive call (line 4). By showing that all the lock holders will eventually release the lock, we prove that all the sleeping threads will be added to the pending queue or ready queue within a finite number of steps. Thus, sleep will terminate thanks to the fair software scheduler. Note that all these properties proved at the C level can be propagated down to the assembly level using the thread-safe CompCertX.

5.5 Thread-Safe Compilation and Linking

In this section, we show how to adapt Gu et al.’s CompCertX verified separate compiler [15, §6] to handle programs that call scheduling primitives. Section 5.3 shows how thread-local layer interfaces allow us to give C style specifications to scheduling primitives (yield and sleep) which are partly implemented in assembly. Thus, code of each thread can be verified at the C level over $L_{h}{\text{d}}[c][T]$ and individual threads can then be composed into programs on $L_{h}{\text{d}}[c]$ by Thm. 5.1. However, it is still challenging to show that the compiled programs at the assembly level are also compatible with this parallel composition because of a small snag which we glossed over until now: stack frames. In the CompCert memory model [29], whenever a function is called, a fresh memory block has to be allocated in the memory for its stack frame. This means that, on top of the thread-local layer $L_{h}{\text{d}}[c][T]$, a function called within a thread will allocate its stack frame into the thread-private memory state, and conversely, a thread is never aware of any newer memory blocks allocated by other threads. In comparison, on top of the CPU-local layer $L_{h}{\text{d}}[c]$, all stack frames have to be allocated in the CPU-local memory (i.e., thread-shared memory) regardless of which thread they belong to; thus, in the thread composition proof, we need to account for all such stack frames.

Our solution is based on a special memory extension [29, §5.2] that only removes the access permissions of some memory blocks. To enable the thread composition, we extended the semantics of yield and sleep on the thread-local layer $L_{h}{\text{d}}[c][T]$. Besides generating a c.yield/c.sleep event, such a scheduling primitive also allocates empty memory blocks as "placeholders" for other threads’ new stack frames during this yield/sleep. These empty blocks are the ones without...
any access permissions. We write “nb(m)” to denote the total number of blocks in m, and write “liftnb(m, n)” as the memory extended from m by allocating n new empty blocks.

With the extended semantics for scheduling primitives, we can prove that a ternary relation “m1 ⊗ m2 = m” holds between the private memory states m1, m2 of two disjoint thread sets and the thread-shared memory state m after the parallel composition. This relation among memory states is called the “algebraic memory model”, which is defined by the axioms shown in Fig. 12.

Rule Nn states that the block number of the composed memory m is equal to “max(nb(m1), nb(m2)).” Rule Comm says that the parallel memory composition is commutative. Rule Ld and St state that the behaviors of memory load and store (over m1 or m2) are preserved by the composed memory m. It is because that every non-shared memory block of m1 either does not exist in m2 or corresponds to an empty block in m2, and vice versa.

All the remaining rules in Fig. 12 share the condition “nb(m1) ≤ nb(m2).” This condition indicates that thread 2 is “more-recently scheduled/running” because only running thread can allocate memory blocks. Thus, memory allocations on m2 can be preserved by the composed memory m (see Rule Alloc). In addition, if thread 2 is still the next scheduled thread and there are n new stack frames allocated by threads other than {1, 2}, we can then simply allocate n empty blocks in m2, which will be preserved by m (see Rule Lift-R). If thread 1 is the next thread to run, after allocating n new empty blocks to m1, the composed memory m only need to allocate the blocks that have not been captured by m2 (see Rule Lift-L).

Based on the parallel composition for two memory states, we can use Rule Lift-R and Lift-L to generalize to N threads by saying that m is a composition of the private memory states “m1, . . . , mN” of N threads (on a single processor) if, and only if, there exists a memory state m’ such that m’ is a composition of “m1, . . . , mN-1” and mN ⊗ m’ = m holds.

6 Evaluation and Experience

We have implemented the CCAL toolkit (see Fig. 2) in the Coq proof assistant. Table 1 presents the number of lines (in Coq) for each component in Fig. 2. The auxiliary library contains the common tactics and lemmas for 64 bit integers, lists, maps, integer arithmetic, etc.

Case Studies. To evaluate the framework itself, we have implemented, specified, and verified various concurrent programs in the framework. Table 2 presents some of the statistics with respect to the implemented components. As for lock implementations, their source code contains not only the code of the associated functions, but also the data structures and their initialization. In addition to the top-level interface, the specification contains all the specifications used in the intermediate layers. For both the ticket and MCS locks, the simulation proof column includes the proof of starvation freedom (about 1,500 lines) in addition to the correctness proof. The gap between the underlying C implementation and the high-level specification of the locks also contributes to the large proof size for these components. For example, intermediate specification of the ticket lock uses an unbounded integer for the ticket field, while the implementation uses a binary integer which wraps back to zero. Similarly, the queue is represented as a logical list in the specification, while it is implemented as a doubly linked list.

### Table 2. Statistics for implemented components.

<table>
<thead>
<tr>
<th>Component</th>
<th>LOC</th>
<th>Asm verifier</th>
<th>Simulation library</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ticket lock</td>
<td>74</td>
<td>615</td>
<td>1,080</td>
</tr>
<tr>
<td>MCS lock</td>
<td>287</td>
<td>1,569</td>
<td>2,299</td>
</tr>
<tr>
<td>Local queue</td>
<td>377</td>
<td>554</td>
<td>748</td>
</tr>
<tr>
<td>Shared queue</td>
<td>20</td>
<td>107</td>
<td>190</td>
</tr>
<tr>
<td>Scheduler</td>
<td>62</td>
<td>153</td>
<td>166</td>
</tr>
<tr>
<td>Queuing lock</td>
<td>112</td>
<td>255</td>
<td>992</td>
</tr>
</tbody>
</table>

### Table 1. Lines of proofs in Coq for the toolkit.

<table>
<thead>
<tr>
<th>Component</th>
<th>LOC</th>
<th>Component</th>
<th>LOC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Auxiliary library</td>
<td>6,200</td>
<td>Multilayer linking</td>
<td>17,000</td>
</tr>
<tr>
<td>C verifier</td>
<td>2,200</td>
<td>Multithread linking</td>
<td>10,000</td>
</tr>
<tr>
<td>Asm verifier</td>
<td>800</td>
<td>Multico core linking</td>
<td>7,000</td>
</tr>
<tr>
<td>Simulation library</td>
<td>1,800</td>
<td>Thread-safe CompCertX</td>
<td>7,500</td>
</tr>
</tbody>
</table>

Figure 12. Algebraic memory model
Our development is compositional. Both ticket and MCS locks share the same high-level atomic specifications (or strategies) shown in Sec. 2. Thus the lock implementations can be freely interchanged without affecting any proof in the higher-level modules using locks. When implementing the shared queue library, we also reuse the implementation and proof of the local (or sequential) queue library: to implement the atomic queue object, we simply wrap the local queue operations with lock acquire and release statements. As shown in Table 2, using verified lock modules to build atomic objects such as shared queues is relatively simple and does not require many lines of code.

Following the same philosophy, Gu et al. [16] has further extended our work with paging-based dynamically allocated virtual memory, device drivers with in-kernel interrupts, a synchronous inter-process communication (IPC) protocol using the queuing lock, a shared-memory IPC protocol with a shared page, and Intel hardware virtualization support; our CCAL toolkit was used to produce the world’s first fully certified concurrent OS kernel with fine-grained locking.

**Performance Evaluation.** We have measured the performance of the ticket lock on an Intel 4-Core i7-2600S (2.8GHz) processor with 16GB memory. Initially, the ticket lock implementation incurred a latency of 87 CPU cycles in the single core case. After a short investigation, we found that we forgot to remove some function calls to “logical primitives” used for manipulating ghost abstract states. After we removed these extra null calls, the latency dropped down to only 35 CPU cycles. Gu et al. [16] also presented performance evaluations of their OS kernel built using CCAL.

**Limitations.** Our concurrent machine models assume strong sequential consistency (SC) for atomic primitives. Previous work [53] demonstrated that race-free programs on a TSO model do indeed behave as if executing on a sequentially consistent machine. Since safe programs on our push/pull model are race-free, we believe extending our work from SC to TSO is promising. In our future work, we will formalize and integrate this proof in Coq. Furthermore, the current event-based contextual refinement proofs still require quite a bit of manual proof. We are working on developing more automation tactics to further cut down the proof effort. In addition to this general toolkit that can support a broad range of concurrent programs, we also plan to provide more aggressive automation for commonly-used concurrent programming patterns, either through additional tactic libraries or using specific program logics targeting such patterns.

7 Related Work and Conclusions

**Certified Abstraction Layers.** Gu et al. [15] presented the first formal account of certified abstraction layers and showed how to apply layer-based techniques to build certified system software. The layer-based approach differs from Hoare-style program verification [4, 21, 41, 47] in several significant ways. First, it uses the termination-sensitive forward simulation techniques [27, 36] and proves a stronger contextual correctness property rather than simple partial or total correctness properties (as done for Hoare logics). Second, the overlay interface of a certified layer object completely removes the internal concrete memory block (for the object) and replaces it with an abstract state suitable for reasoning; this abstract state differs from auxiliary or ghost states (in Hoare logic) because it is actually used to define the semantics of the overlay abstract machine and the corresponding contextual refinement property. Third, as we move up the abstraction hierarchy by composing more layers, each layer interface provides a new programming language that gets closer to the specification language—it can call primitives at higher abstraction levels while still supporting general-purpose programming in C and assembly.

Our CCAL toolkit follows the same layer-based methodologies. Each time we introduce a new concrete concurrent object implementation, we replace it with an abstract atomic object in its overlay interface. All shared abstract states are represented as a single global log, so the semantics of each atomic method call would need to replay the entire global log to find out the return value. This seemingly “inefficient” way of treating shared atomic objects is actually great for compositional specification. Indeed, it allows us to apply game-semantic ideas and define a general semantics that supports parallel layer composition.

**Abstraction for Concurrent Objects.** Herlihy and Wing [20] introduced linearizability as a key technique for building abstraction over concurrent objects. Developing concurrent software using a stack of shared atomic objects has since become the best practice in the system community [2, 19]. Linearizability is quite difficult to reason about, and it is not until 20 years later that Filipovic et al. [10] showed that linearizability is actually equivalent to a termination-insensitive version of the contextual refinement property. Gotsman and Yang [14] showed that such equivalence also holds for concurrent languages with ownership transfers [43]. Liang et al. [31, 34] showed that linearizability plus various progress properties [19] for concurrent objects is equivalent to various termination-sensitive versions of the contextual refinement property. These results convinced us that we should prove termination-sensitive (contextual) simulation when building certified concurrent layers as well.

**RGSim and LiLi.** Building contextual refinement proofs for concurrent programs (and program transformations) is challenging. Liang et al. [31–33] developed the Rely-Guarantee-based Simulation (RGSim) that can support both parallel composition and contextual refinement of concurrent objects. Our contextual simulation proofs between two concurrent layers can be viewed as an instance of RGSim if we
extend RGSim with auxiliary states such as environment contexts and shared logs. This extension, of course, is the main innovation of our new compositional layered model. Also, all existing RGSim systems are limited to reasoning about atomic objects at one layer; their client program context cannot be the method body of another concurrent object, so they cannot support the same general vertical layer composition as our work does.

**Treatment of Parallel Composition.** Most concurrent languages (including those used by RGSim) use a parallel composition command \((C_1 \parallel C_2)\) to create and terminate new threads. In contrast, we provide thread spawn and join primitives, and assign every new thread a unique ID (e.g., \(t\), which must be a member of the full thread-ID domain set \(D\)). Parallel layer composition in our work is always done over the whole program \(P\) and over all members of \(D\). This allows us to reason about the current thread’s behaviors over the environment’s full strategies (i.e., both past and future events). Even if a thread \(t\) is never created, the semantics for running \(P\) over \(L[t]\) is still well defined since it will simply always query its environment context to construct a global log.

**Program Logics for Shared-Memory Concurrency.** A large body of new program logics [5, 7, 8, 13, 18, 22, 23, 30, 40, 43, 45, 46, 51, 52, 56–59] have been developed to support modular verification of shared-memory concurrent programs. Most of these follow Hoare-style logics so they do not prove the same strong contextual simulation properties as RGSim and our layered framework do. Very few of them (e.g., [46]) can reason about progress properties. Nevertheless, many of these logics support advanced language features such as higher-order functions and sophisticated non-blocking synchronization, both of which will be useful for verifying specific concurrent objects within our layered framework. Our use of a global log is similar to the use of compositional subjective history traces [32]; the main difference is again that our environment context can talk about both past and future events but a history trace can only specify past events.

Both CIVL [18] and FCSL [51] attempt to build proofs of concurrent programs in a “layered” way, but their notions of layers are different from ours in three different ways: (1) they do not provide formal foundational contextual refinement proofs of linearizability as shown by Filipovic et al. [10] and Liang et al. [34]; (2) they do not address the liveness properties; (3) they have not been connected to any verified compilers.

**Compositional CompCert.** Stewart et al. [54] developed a new compositional extension of the original CompCert compiler [27] with the goal of providing thread-safe compilation of concurrent Clight programs. Their interaction semantics also treats all calls to synchronization primitives as external calls. Their compiler does not support a layered ClightX language as our CompCertX does, so they cannot be used to build concurrent layers as shown in Fig. 1.

**Game Semantics.** Even though we have used game-semantic concepts (e.g., strategies) to describe our compositional semantics, our concurrent machine and the layer simulation is still defined using traditional small-step semantics. This is in contrast to several past efforts [1, 12, 42, 48] of modeling concurrency in the game semantics community which use games to define the semantics of a complete language. Modeling higher-order sequential features as games is great for proving full abstraction, but it is still unclear how it would affect large-scale verification as done in the certified software community. We believe there are great potential synergies between the two communities and hope our work will promote such interaction.

**OS Kernel Verification.** There has been a large body of recent work on OS kernel verification including seL4 [25, 26], Verve [60], and Ironclad [17]. None of these works have addressed the issues on concurrency with fine-grained locking. Very recently, Xu et al. [59] developed a new verification framework based on RGSim and Feng et al.’s program logic [9] for reasoning about interrupts; they have successfully verified many key modules (in C) in the \(\mu\)C/OS-II kernel, though so far, they have not proved any progress properties.

**Conclusions.** Abstraction layers are key techniques used in building large-scale concurrent software and hardware. In this paper, we have presented CCAI—a novel programming toolkit developed under the CertiKOS project for building certified concurrent abstraction layers. We have developed a new compositional model for concurrency, program verifiers for concurrent C and assembly, certified linking tools, and a thread-safe verified C compiler. We believe these are critical technologies for developing large-scale certified system infrastructures in the future.

**Acknowledgments**

We would like to thank our shepherd Grigore Rosu and anonymous referees for helpful feedbacks that improved this paper significantly. This research is based on work supported in part by NSF grants 1521523 and 1715154 and DARPA grants FA8750-12-2-0293, FA8750-16-2-0274, and FA8750-15-C-0082. Tahina Ramananandro’s work was completed while he was employed at Reservoir Labs, Inc. Hao Chen’s work is also supported in part by China Scholarship Council. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA or the U.S. Government.
References


