Fully Composable and Adequate Verified Compilation with Direct Refinements between Open Modules

LING ZHANG, Shanghai Jiao Tong University, China
YUTING WANG*, Shanghai Jiao Tong University, China
JINHUA WU, Shanghai Jiao Tong University, China
JÉRÉMIE KOENIG, Yale University, USA
ZHONG SHAO, Yale University, USA

Verified compilation of open modules (i.e., modules whose functionality depends on other modules) provides a foundation for end-to-end verification of modular programs ubiquitous in contemporary software. However, despite intensive investigation in this topic for decades, the proposed approaches are still difficult to use in practice as they rely on assumptions about the internal working of compilers which make it difficult for external users to apply the verification results. We propose an approach to verified compositional compilation without such assumptions in the setting of verifying compilation of heterogeneous modules written in first-order languages supporting global memory and pointers. Our approach is based on the memory model of CompCert and a new discovery that a Kripke relation with a notion of memory protection can serve as a uniform and composable semantic interface for the compiler passes. By absorbing the rely-guarantee conditions on memory evolution for all compiler passes into this Kripke Memory Relation and by piggybacking requirements on compiler optimizations onto it, we get compositional correctness theorems for realistic optimizing compilers as refinements that directly relate native semantics of open modules and that are ignorant of intermediate compilation processes. Such direct refinements support all the compositionality and adequacy properties essential for verified compilation of open modules. We have applied this approach to the full compilation chain of CompCert with its Clight source language and demonstrated that our compiler correctness theorem is open to composition and intuitive to use with reduced verification complexity through end-to-end verification of non-trivial heterogeneous modules that may freely invoke each other (e.g., mutually recursively).

CCS Concepts: • Software and its engineering → Formal software verification; Compilers; • Theory of computation → Program verification.

Additional Key Words and Phrases: Verified Compositional Compilation, Direct Refinements, Kripke Relations

ACM Reference Format:

*Corresponding author

Authors’ addresses: Ling Zhang, John Hopcroft Center for Computer Science, School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, China, ling.zhang@sjtu.edu.cn; Yuting Wang, John Hopcroft Center for Computer Science, School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, China, yuting.wang@sjtu.edu.cn; Jinhua Wu, John Hopcroft Center for Computer Science, School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, China, jinhua.wu@sjtu.edu.cn; Jérémie Koenig, Yale University, USA, jeremie.koenig@yale.edu; Zhong Shao, Yale University, USA, zhong.shao@yale.edu.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
© 2024 Copyright held by the owner/author(s).
ACM 2475-1421/2024/1-ART72
https://doi.org/10.1145/3632914

Proc. ACM Program. Lang., Vol. 8, No. POPL, Article 72. Publication date: January 2024.
1 INTRODUCTION

Verified compilation ensures that behaviors of source programs are faithfully transported to target code, a property desirable for end-to-end verification of software whose development involves compilation. As software is usually composed of modules independently developed and compiled, researchers have developed a wide range of techniques for verified compositional compilation or VCC that support modules invoking each other (i.e., open), being written in different languages (i.e., heterogeneous) and transformed by different compilers [Patterson and Ahmed 2019].

We are concerned with VCC for first-order languages with global memory states and support of pointers (e.g., see Gu et al. [2015]; Jiang et al. [2019]; Koenig and Shao [2021]; Song et al. [2020]; Stewart et al. [2015]; Wang et al. [2019]). As it stands now, the proposed approaches are inherently limited at supporting open modules (e.g. libraries) as they either deviate from the native semantics of modules or expose the semantics of intermediate representations for compilation, resulting in correctness theorems that are difficult to work with for external users. In this paper, we investigate an approach that eliminates these limitations while retaining the full benefits of VCC, i.e., obtaining correctness of compiling open modules that is fully composable, adequate, and extensional.

1.1 Full Compositional Adequacy in Verified Compilation

Correctness of compiling open modules is usually described as refinement between semantics of source and target modules. We shall write L (possibly with subscripts) to denote semantics of open modules and write $L_1 \preceq L_2$ to denote that $L_1$ is refined by $L_2$. Therefore, the compilation of any module $M_1$ into $M_2$ is correct iff $[[M_1]] \preceq [[M_2]]$ where $[[M_i]]$ denotes the semantics of $M_i$.

To support the most general form of VCC, it is critical that the established refinements are fully composable, i.e., both horizontally and vertically composable, and adequate for native semantics:

- **Vertical Compositionality:** $L_1 \preceq L_2 \Rightarrow L_2 \preceq L_3 \Rightarrow L_1 \preceq L_3$
- **Horizontal Compositionality:** $L_1 \preceq L'_1 \Rightarrow L_2 \preceq L'_2 \Rightarrow L_1 \oplus L_2 \preceq L'_1 \oplus L'_2$
- **Adequacy for Native Semantics:** $[[M_1 + M_2]] \preceq [[M_1]] \oplus [[M_2]]$

The first property states that refinements are transitive. It is essential for composing proofs for multi-pass compilers. The second property guarantees that refinements are preserved by semantic linking (denoted by $\oplus$). It is essential for composing correctness of compiling open modules (possibly through different compilers). The last one ensures that, given any modules, their semantic linking coincides with their syntactic linking (denoted by $\oplus$). It ensures that linked semantics do not deviate from native semantics and is essential to propagate verified properties to final target programs.

We use the example in Fig. 1 to illustrate the importance of the above properties in VCC where heterogeneous modules are compiled through different compilation chains and linked into a final target module. In this example, a source C module $a.c$ is compiled into an assembly module $a.s$ through a multi-pass optimizing compiler like CompCert: it is first compiled to $a.i_1$ in an intermediate representation (IR) for optimization (e.g., the RTL language of CompCert) and then to $a.i_2$ in another IR for code generation (e.g., the Mach language of CompCert). Finally, it is linked with a library module $b.s$ which is not compiled at all (an extreme case where the compilation chain is empty). The goal is to prove that the semantics of linked target assembly $a.s + b.s$ refines the combined source semantics $[[a.c]] \oplus L_b$ where $L_b$ is the semantic specification of $b.s$, i.e., $[[a.s + b.s]] \preceq [[a.c]] \oplus L_b$. The proof proceeds as follows:

---

Proc. ACM Program. Lang., Vol. 8, No. POPL, Article 72. Publication date: January 2024.
When the interface of every semantics to use the same language interface I CompCertM [Song et al. 2020] and CompCertO [Koenig and Shao 2021].

C compiler) to support VCC, i.e., Compositional CompCert (CompComp) [Stewart et al. 2015], facing them via three well-known extensions of CompCert [Leroy 2023] (the state-of-the-art verified on refinements. We categorize them by their algebraic structures below, and explain the problems transitionality and adequacy. The existing approaches achieve this by imposing algebraic structures 

Note that the interface for a module may not match its native semantics. For example, only allows for interaction with environments through function calls and returns in C. Similarly, interface formulates their interaction with environments. We write L to denote a refinement between two semantics with interfaces a.c and b.s where L only allows for interaction at the assembly level. Note that the interface for a module may not match its native semantics. For example, a.s : C asserts that a.s actually converts assembly level calls/returns to C function calls/returns for interacting with C environments (e.g., extracting arguments from registers and memory to form an argument list for C function calls). In this case, a.s deviates from the native semantics of a.s.

When the interface of [M] is not explicitly given, it is implicitly the native interface of M. We write ≪ I1 ≐ I2 to denote a refinement between two semantics with interfaces I1 and I2. For instance, given ≐ac A ≐ C that relates open semantics at the C and assembly levels, [b.s] ≐ac Lb asserts that [b.s] is the native semantics of b.s and is refined by the C level specification Lb.

For VCC, it is essential that variance of open semantics and refinements does not impede compositionality and adequacy. The existing approaches achieve this by imposing algebraic structures on refinements. We categorize them by their algebraic structures below, and explain the problems facing them via three well-known extensions of CompCert [Leroy 2023] (the state-of-the-art verified C compiler) to support VCC, i.e., Compositional CompCert (CompComp) [Stewart et al. 2015], CompCertM [Song et al. 2020] and CompCertO [Koenig and Shao 2021].

Fig. 2. Refinements in the Existing Approaches to VCC

(a) CompComp

(b) CompCertM

(c) CompCertO

1.2 Problems with the Existing Approaches to Refinements

Despite the simplicity of VCC at an intuitive level, full compositionality and adequacy are surprisingly difficult to prove for any non-trivial multi-pass compiler. First and foremost, the formal definitions must take into account the facts that each intermediate representation has different semantics and each pass may imply a different refinement relation. To facilitate the discussion below, we classify different open semantics by language interfaces (or simply interfaces) which formalize their interaction with environments. We write L : I to denote that L has a language interface I. For instance, [[a.c]] : C denotes that the semantics of a.c has the interface C which only allows for interaction with environments through function calls and returns in C. Similarly, [[a.s]] : A denotes the semantics of a.s where A only allows for interaction at the assembly level. Note that the interface for a module may not match its native semantics. For example, [[a.s]] : C asserts that [[a.s]] actually converts assembly level calls/returns to C function calls/returns for interacting with C environments (e.g., extracting arguments from registers and memory to form an argument list for C function calls). In this case, [[a.s]] deviates from the native semantics of a.s.

For VCC, it is essential that variance of open semantics and refinements does not impede compositionality and adequacy. The existing approaches achieve this by imposing algebraic structures on refinements. We categorize them by their algebraic structures below, and explain the problems facing them via three well-known extensions of CompCert [Leroy 2023] (the state-of-the-art verified C compiler) to support VCC, i.e., Compositional CompCert (CompComp) [Stewart et al. 2015], CompCertM [Song et al. 2020] and CompCertO [Koenig and Shao 2021].

Constant Refinement. An obvious way to account for different semantics in VCC is to force every semantics to use the same language interface I and a constant refinement ≐ I : I → I.
CompCert adopts this “one-type-fits-all” approach by having every language of CompCert to use C function calls/returns for module-level interactions and using a uniform refinement relation \( \preceq_C : C \leftrightarrow C \) known as structured simulation [Stewart et al. 2015]. In this case, vertical and horizontal compositionality is established by proving transitivity of \( \preceq_C \) and symmetry of rely-guarantee conditions of \( \preceq_C \). However, because the C interface is adopted for assembly semantics, adequacy at the target level is lost, making end-to-end compiler correctness not provable as shown in Fig. 2a.

**Sum of Refinements.** A more relaxed approach allows users to choose language interfaces for different IRs from a finite collection \( \{ I_1, \ldots, I_m \} \) and refinements for different passes from a finite set \( \{ \preceq_1, \ldots, \preceq_n \} \) relating these interfaces, i.e., \( \preceq_i : I_1 + \ldots + I_m \bowtie I_1 + \ldots + I_m \). In essence, a constant refinement is split into a sum of refinements s.t. \( L \preceq_1 + \ldots + \preceq_n L' \) holds if \( L \preceq_i L' \) for some \( 1 \leq i \leq n \). Then, every compiler pass can use \( \preceq_1 + \ldots + \preceq_n \) as the uniform refinement relation, which is proven both composable and adequate under certain well-formedness constraints. Fig. 2b depicts such an example where semantics have both C and assembly interfaces (e.g., \( [[a.s]] : \mathcal{A} + \mathcal{C} \)) and the refinement relations \( \preceq_i : \mathcal{A} + \mathcal{C} \leftrightarrow \mathcal{A} + \mathcal{C} (1 \leq i \leq 5) \) are tailored for each pass. This is the approach adopted by CompCertM [Song et al. 2020]. However, the top-level refinement \( \preceq_1 + \ldots + \preceq_n \) is difficult to use by a third party without introducing complicated dependency on intermediate results of compilation. For example, horizontal composition with \( \preceq_1 + \ldots + \preceq_n \) only works for modules self-related by all the refinements \( \preceq_i (1 \leq i \leq n) \). Since \( \preceq_i \)'s are tailored for individual passes, they inevitably depend on the intermediate semantics used in compilation. Such dependency is only exacerbated as new languages, compilers and optimizations are introduced.

**Product of Refinements.** The previous approach effectively “flattens” the refinements for individual compiler passes into an end-to-end refinement. A different approach adopted by CompCertO [Koenig and Shao 2021] is to “concatenate” the refinements for individual passes into a chain of refinements by a product operation \( (_\cdot \cdot _) \) such that \( L \preceq_1 \cdot \preceq_2 L'' \) if \( L \preceq_1 L' \) and \( L' \preceq_2 L'' \) for some \( L' \). Fig. 2c illustrates how it works. Vertical composition is simply the concatenation of refinements. For example, composing refinements for compiling \( a.c \) results in \( [[a.s]] \preceq_3 \cdot \preceq_2 \cdot \preceq_1 [[a.c]] \). Adequacy is trivially guaranteed with native interfaces. However, horizontal composition still depends on the intermediate semantics of compilation because of the concatenation. For example, in Fig. 2c, to horizontally compose with \( [[a.s]] \preceq_3 \cdot \preceq_2 \cdot \preceq_1 [[a.c]] \), it is necessary to show \( L_b \) refines \( [[b.s]] \) via the same product, i.e., to construct intermediate semantics bridging \( \preceq_1, \preceq_2 \) and \( \preceq_3 \).

**Summary.** The existing approaches for VCC either lack adequacy because they force non-native language interfaces on semantics for open modules (e.g., CompCert) or lack compositionality that is truly extensional because they depend on intermediate semantics used in compilation (e.g., CompCertM and CompCertO). Such dependency makes their correctness theorems for compiling open modules (e.g., libraries) difficult to further compose with and incurs a high cost in verification.

### 1.3 Challenges for Direct Refinement of Open Modules

The ideal approach to VCC should produce refinements that directly relate the native semantics of source and target open modules without mentioning any intermediate semantics and support both vertical and horizontal composition. We shall call them *direct refinements of open modules*. For example, a direct refinement between \( a.c \) and \( a.s \) could be \( \preceq_{a.c} : \mathcal{A} \leftrightarrow C \) s.t. \( [[a.s]] \preceq_{a.c} [[a.c]] \). It relates assembly and C without mentioning intermediate semantics, and could be further horizontally composed with \( [[b.s]] \preceq_{a.c} L_b \) and vertically composed by adequacy to get \( [[a.s + b.s]] \preceq_{a.c} [[a.c]] \oplus L_b \). Note that even the top-level refinement is still open to horizontal and vertical composition, making direct refinements effective for supporting VCC for open modules.
The main challenge in getting direct refinements is tied to their “real” vertical composition, i.e., given any direct refinements \( \leq_1 \) and \( \leq_2 \), how to show \( \leq_1 \cdot \leq_2 \) is equivalent to a direct refinement \( \leq_3 \). This is considered very technical and involved (see Hur et al. [2012b]; Neis et al. [2015]; Patterson and Ahmed [2019]; Song et al. [2020]) because of the difficulty in constructing interpolating program states for transitively relating evolving source and target states across external calls of open modules. This problem also manifests in proving transitivity for logical relations where construction of interpolating terms of higher-order types is not in general possible [Ahmed 2006]. In the setting of compiling first-order languages with global memory, all previous work avoids proving real vertical composition of direct refinements. Some produce refinement without adequacy by introducing intrusive changes to semantics to make construction of interpolating states possible. For example, CompComp instruments the semantics of languages with effect annotations to expose internal effects for this purpose. Some essentially restrict vertical composition to closed programs (e.g., CompCertM). Some leave the top-level refinement a combination of refinements that still exposes the intermediate steps of compilation (e.g., CompCertO). Finally, even if the problem of vertical composition was solved, it is not clear if the solution can support realistic optimizing compilers.

1.4 Our Contributions

In this paper, we propose an approach to direct refinements for VCC of imperative programs that addresses all of the above challenges. Our approach is based on the memory model of CompCert which supports first-order states and pointers. We show that in this memory model interpolating states for proving vertical compositionality of refinements can be constructed by exploiting the properties on memory invariants known as memory injections. The solution is based on a new discovery that a Kripke relation with memory protection can serve as a uniform and composable relation for characterizing the evolution of memory states across external calls. With this relation we successfully combined the correctness theorems of CompCert’s passes into a direct refinement between C and assembly modules. We summarize our technical contributions below:

- We prove that \( \text{injp} \)—a Kripke Memory Relation with a notion of memory protection—is both uniform (i.e., memory transformation in every compiler pass respects this relation) and composable (i.e., transitive modulo an equivalence relation). The critical observation making this proof possible is that interpolating memory states can be constructed by exploiting memory protection inherent to memory injections and the functional nature of injections.
- Based on the above observation, we show that a direct refinement from C to assembly can be derived by composing open refinements for all of CompCert’s passes starting from Clight. In particular, we show that compiler passes can use different Kripke relations sufficient for their proofs (which may be weaker than \( \text{injp} \)) and these relations will later be absorbed into \( \text{injp} \) via refinements of open semantics. Furthermore, we show that assumptions for compiler optimizations can be formalized as semantic invariants and, when piggybacked onto \( \text{injp} \), can be transitively composed. Based on these techniques, we upgrade the proofs in CompCertO to get a direct refinement from C to assembly for the full CompCert, including all of its optimization passes. These experiments show that direct refinements can be obtained without fundamental changes to the verification framework of CompCert.
- We demonstrate the simplicity and usefulness of direct refinements by applying it to end-to-end verification of several non-trivial examples with heterogeneous modules that mutually invoke each other. In particular, we observe that C level refinements can be absorbed into the direct refinement of CompCert by transitivity of \( \text{injp} \). Combining direct refinements with full compositionality and adequacy, we derive end-to-end refinements from high-level source specifications to syntactically linked assembly modules in a straightforward manner.
/* client.c */
int result;
void encrypt(int i, void (*p)(int *));
void process(int *r) {
    result = *r;
}
int request(int i) {
    encrypt(i, process);
    return i;
}

/* server.s */
key: long 42
encrypt:
void(*p)(int*));
void process(int *r) {
    // allocate frame
    Pallocframe 24 16 0
    // RSP[8] = i XOR key
    Pmov key RAX
    Pxor RAX RDI
    Pmov RDI 8(RSP)
    // call p( RSP + 8)
    Plea 8(RSP) RDI
    Pcall RSI
    // free frame
    Pfreeframe 24 16 0
}
int request(int i) {
    encrypt(i, process);
    return i;
}

/* server_opt.s */
key is a constant and inlined in code
encrypt:
void(*p)(int*));
void process(int *r) {
    // allocate frame
    Pallocframe 24 16 0
    // RSP[8] = i XOR 42
    Pxori 42 RDI
    Pmov RDI 8(RSP)
    // call p( RSP + 8)
    Plea 8(RSP) RDI
    Pcall RSI
    // free frame
    Pfreeframe 24 16 0
}

(a) Client in C  
(b) Server in Asm  
(c) Optimized Server

Fig. 3. An Example of Encryption Client and Server

The above developments are fully formalized in Coq based on the latest CompCertO which is in turn built on top of CompCert v3.10 (see the data-availability statement at the end of the paper for more details). While the formalisation of our approach is tied to CompCert’s block-based memory model [Leroy et al. 2012], and applied to its particular chain of compilation, we present evidence in §7 that variants of injp could be adapted for alternate memory models for first-order languages, and that it may be extended to support new optimizations. Therefore, this work provides a promising direction for further evolving the techniques for VCC.

1.5 Structure of the Paper

Below we first introduce the key ideas supporting this work in §2. We then introduce necessary background and discuss the technical challenges for building direct refinements in §3. We present our technical contributions in §4, §5 and §6. We discuss the generality and limitations of our approach in §7. We discuss evaluation and related work in §8 and finally conclude in §9.

2 KEY IDEAS

We introduce a running example with heterogeneous modules and callback functions to illustrate the key ideas of our work. This example is representative of mutual dependency between modules that often appears in practice and it shows how free-form invocation between modules can be supported by our approach. As we shall see in §6, our approach also handles more complicated programs with mutually recursive heterogeneity without any problem.

The example is given in Fig. 3. It consists of a client written in C (Fig. 3a) and an encryption server hand-written in x86 assembly by using CompCert’s assembly syntax where instruction names begin with P (Fig. 3b). For now, let us ignore Fig. 3c which illustrates how optimizations work in direct refinements. Users invoke request to initialize an encryption request. It is relayed to the function encrypt in the server with the prototype void encrypt(int i, void (*p)(int*)) which respects a calling convention placing the first and second arguments in registers RDI and RSI, respectively. The main job of the server is to encrypt i (RDI) by XORing it with an encryption key (stored in the global variable key) and invoke the callback function p (RSI). Finally, the client
which have been witnessed in a wide range of verification projects [Gu et al. 2015; Jiang et al. 2019; Koenig and Shao 2021; Song et al. 2020; Stewart et al. 2015; Wang et al. 2019]. Fig. 5a depicts a refinement of open modules discussed in our paper are based on forward simulations between small-step operational semantics (often in the form of labeled transition systems or LTS) which have been witnessed in a wide range of verification projects [Gu et al. 2015; Jiang et al. 2019; Koenig and Shao 2021; Song et al. 2020; Stewart et al. 2015; Wang et al. 2019]. Fig. 5a depicts a refinement of open modules discussed in our paper are based on forward simulations between small-step operational semantics (often in the form of labeled transition systems or LTS) which have been witnessed in a wide range of verification projects [Gu et al. 2015; Jiang et al. 2019; Koenig and Shao 2021; Song et al. 2020; Stewart et al. 2015; Wang et al. 2019].

With the running example, our goal is to verify its end-to-end correctness by exploiting the direct refinement \( \preceq_{ac} : A \iff C \) derived from CompCert’s compilation chain as shown in Fig. 4. The verification proceeds as follows. First, we establish \([\text{client.s}] \preceq_{ac} [\text{client.c}]\) by the correctness of compilation. Then, we prove \([\text{server.s}] \preceq_{ac} L_S\) manually by providing a specification \(L_S\) for the server that respects the direct refinement. At the source level, the combined semantics is further refined to a single top-level specification \(L_{CS}\). Finally, the source and target level refinements are absorbed into the direct refinement by vertical composition and adequacy, resulting in a single direct refinement between the top-level specification and the target program:

\[
[\text{client.s + server.s}] \preceq_{ac} L_{CS}
\]

The refinements of open modules discussed in our paper are based on forward simulations between small-step operational semantics (often in the form of labeled transition systems or LTS) which have been witnessed in a wide range of verification projects [Gu et al. 2015; Jiang et al. 2019; Koenig and Shao 2021; Song et al. 2020; Stewart et al. 2015; Wang et al. 2019].
calling environments). The rely-guarantee conditions are essential for horizontal composition: two refinements $L_1 \sim L_2$ and $L'_1 \sim L'_2$ with complementary rely-guarantee conditions can be composed into a single refinement $L_1 \oplus L_2 \sim L'_1 \oplus L'_2$. However, vertical composition of such refinements is difficult. A naive vertical composition of two refinements (one between $L_1$ and $L_2$ and another between $L_2$ and $L_3$) simply concatenates them together like Fig. 5b, instead of generating a single refinement between $L_1$ and $L_3$ like Fig. 5c.1 This exposes the intermediate semantics (i.e., $L_2$) and imposes serious limitations on VCC as discussed in §1.2. Therefore, to the best of our knowledge, none of the existing approaches fully support the verification outlined in Fig. 4.

To address the above problem, we develop direct refinements with the following distinguishing features: 1) they always relate the semantics of modules at their native interfaces, thereby supporting adequacy; 2) they do not mention the intermediate process of compilation, thereby supporting heterogeneous modules and compilers; 3) they provide direct memory protection for source and target semantics via a Kripke relation, thereby enabling horizontal composition of refinements for heterogeneous modules; 4) most importantly, they are vertically composable. The first three features are manifested in the very definition of direct refinements, which we shall discuss in §2.1 below. We then discuss the vertical composition of direct refinements in §2.2, which relies on the discovery of the uniformity and transitivity of a Kripke relation for memory protection.

### 2.1 Refinement Supporting Adequacy, Heterogeneity and Horizontal Composition

To illustrate the key ideas, we use the top-level direct refinement $\ltimes_{ac}$ in Fig. 4 as an example. In the remaining discussions, we adopt the block-based memory model of CompCert [Leroy et al. 2012] where a memory state consists of a disjoint set of memory blocks. $\ltimes_{ac}$ is a forward simulation that directly relates C and assembly modules with their native language interfaces. By the definition of these interfaces (See §3.1), a C query $q_C = v_f [sg](\overline{\alpha})@m$ is a function call to $v_f$ with signature $sg$, a list of arguments $\overline{\alpha}$ and a memory state $m$; a C reply $r_C = v'@m'$ carries a return value $v'$ and an updated memory state $m'$. An assembly query $q_A = rs@m$ invokes a function with the current register set $rs$ and memory state $m$. An assembly reply $r_A = rs'@m'$ returns from a function with the updated registers $rs'$ and memory $m'$. By definition, $L_2 \ltimes_{ac} L_1$ means that $L_1$ and $L_2$ behave like C and assembly programs at the boundary of modules, respectively. However, there is no restriction on how $L_1$ and $L_2$ are actually implemented internally, which enables source-level specifications with C interfaces like $L_3$ in Fig. 4.

The rely and guarantee conditions imposed by $\ltimes_{ac}$ are symmetric and bundled with the simulation invariants at the boundary of modules. They make assumptions about how C and assembly queries should be related at the call sites and provide conclusions about how the replies should be related after the calls return. Given any matching source and target queries $q_C = v_f [sg](\overline{\alpha})@m_1$ and $q_A = rs@m_2$, it is assumed that

1. The memory states are related by an invariant $j$ known as a memory injection function [Leroy et al. 2012], i.e., memory blocks in $m_1$ are projected into those in $m_2$;
2. The function pointer $v_f$ is related to the program counter register in $rs$;
3. The source arguments $\overline{\alpha}$ are projected either to registers in $rs$ or to outgoing argument slots in the stack frame RSP in $m_2$ according to the C calling convention;
4. The outgoing arguments on the target stack frame are freeable and not in the image of $j$.

The first three requirements ensure that C arguments and memory are related to assembly registers and memory according to CompCert’s C calling convention. The last one ensures outgoing arguments are protected, thereby preserving the invariant of open simulation across external calls.

---

1To simplify the presentation, we often elide the guarantee conditions in figures for simulation.
After the function calls return, the source and target queries \( r_C = res@m'_1 \) and \( r_A = rs'@m'_2 \) must satisfy the following requirements:

1. The updated memory states \( m'_1 \) and \( m'_2 \) are related by an updated memory injection \( j'_1 \);
2. The C-level return value \( res \) is related to the value stored in the register for return value;
3. For any callee-saved register \( r \), \( rs' (r) = rs(r) \);
4. The stack pointer register and program counter are restored.
5. The access to memory during the function call is protected by a *Kripke Memory Relation* \( \text{injp} \) such that the private stack data for other function calls are not modified.

The first two requirements ensure that return values and memories are related according to the calling convention. The following two ensure that registers are correctly restored before returning. The last requirement plays a critical role in rely-guarantee reasoning and enables horizontal composition of direct refinements as we shall see soon.

### 2.1.1 Adequacy and Heterogeneity via Direct Refinement

By definition, \( \prec_{\text{ac}} \) is basically a formalized C calling convention for CompCert with direct relations between C and assembly operational semantics and with invariants for protecting register values and memory states. Adequacy is automatically guaranteed as syntactic linking coincides with semantics linking at the assembly level. That is, given any assembly modules \( a.s \) and \( b.s \), \( [[a.s + b.s]] \prec_{\text{id}} [[a.s]] \oplus [[b.s]] \).

Moreover, \( \prec_{\text{ac}} \) does not mention anything about compilation. It works for any heterogeneous module and compilation chain that meet its requirements, even for hand-written assembly. Take the refinement of \( [[\text{server.s}]] \prec_{\text{ac}} L_5 \) in Fig. 4 as an example. The first few steps of the simulation are depicted in Fig. 6, where \( L_5 \) is an LTS hand-written by us and \( [[\text{server.s}]] \) is derived from the CompCert assembly semantics. Because \( L_5 \) is only required to respect the C interface, we choose a form easy to comprehend where its internal executions are in big steps. Now, suppose the environment calls encrypt with source and target queries initially related by CompCert’s calling convention s.t. \( rs(\text{RD1}) = i \) and \( rs(\text{RS1}) = p \). After the initialization \( I_1 \) and \( I_2 \), the execution enters internal states related by an invariant \( R \). Then, the target execution takes internal steps \( K_2 \) until reaching an external call. This corresponds to executing lines 5-13 in Fig. 3b, which allocates the stack frame RSP, performs encryption by storing \( i \) XOR key at the address RSP+8, and calls back \( p \) with RSP+8. At the source level, these steps correspond to one big-step execution \( K_1 \) which allocates a memory block \( b_0 \), stores \( i \) XOR key at \( b_0 \), and prepares to call \( p \) with \( b_0 \). Therefore, the memory injection in \( R \) maps \( b_0 \) to RSP+8. The source and target execution continue with transitions \( X_1 \) and \( X_2 \) to the external calls to \( p \), return from \( p \) and go on until they return from encrypt.

### 2.1.2 Horizontal Composition via Kripke Memory Relations

The Kripke Memory Relation (KMR) \( \text{injp} \) provides essential protection for private values on the stack, which ensures that simulations between heterogeneous modules can be established and their horizontal composition is feasible.

We illustrate these points via our running example. Assume that the environment calls request in the client with 11 which in turn calls encrypt in the server to get the value 11 XOR 42 = 33.
whose address is passed back to the client by calling `process`. Fig. 7 depicts a snapshot of the memory states and the injection right after `process` is entered (i.e., at line 8 in Fig. 3a), where boxes denote allocated memory blocks, black arrows between blocks denote injections, and red arrows denote pointers. The source semantics allocates one block for each local variable (\(b_i\) for \(i\), \(b_0\) for the encrypted value 33 and \(b_r\) for \(r\)) while the target semantics stores their values in registers or stacks (11 is stored in RDI while 33 on the stack because its address is taken and may be modified by the callee). One stack frame is allocated for each function call which stores private data including pointers to previous frames (\(b_{RSP}\)), return addresses (RA), and callee-saved registers (e.g., RBX).

`injp` is essential for proving simulation for open modules as it guarantees simulation can be re-established after external calls return. Informally, at every external call site, `injp` marks all memory regions outside the footprint (domain and image) of the current injection as private and does not allow the external call to modify those memory regions. From the perspective of server.s, when the snapshot in Fig. 7 is taken, the execution is inside the thick dashed line in Fig. 6 and protected by `injp`. Therefore, all the shaded memory in Fig. 7 are marked as private and protected against the callback to `process`. Indeed, they correspond to either memory values turned into temporary variables (e.g., \(b_i\)) or private stack data (e.g., \(b_{RSP}\), RBX and RA in block \(b_{RSP_1}\)) that should not be touched by `process`. Such protection ensures that when `process` returns, all the private values are still valid, thereby re-establishing the simulation invariant.

The role of `injp` is reversed for the incoming calls from the environment: it guarantees that the entire execution from the initial query to the final reply will not touch any private memory of the environment. Therefore, `injp` is used to impose a reliance on memory protection by external calls and to provide a symmetric guarantee of memory protection for the environment callers. Any simulations with compatible language interfaces that satisfy this rely-guarantee condition can be horizontally composed. For example, we can horizontally compose \([[\text{client.s}] \prec \text{ac} \ [\text{client.c}]]\) and \([[\text{server.s}] \prec \text{ac} \ L_S]\) into \([[\text{client.s}] \oplus [\text{server.s}] \prec \text{ac} \ [\text{client.c}] \oplus L_S]\) in Fig. 4.

### 2.2 Uniform and Transitive KMR for Vertical Composition of Direct Refinements

Direct refinements are only useful if they can be vertically composed, which is critical for composing refinements obtained from individual compiler passes into a single top-level refinement such as \(\prec \text{ac}\) and for further composition with source-level refinements as shown in Fig. 4.

We discuss our approach for addressing this problem by using CompCert and CompCertO as the concrete platforms. It is based on the following two observations. First, `injp` in fact captures the rely-guarantee conditions for memory protection needed by every compiler pass in CompCert. At a high-level, it means that the rely-guarantee conditions as depicted in Fig. 5 can all be replaced...
by injp (modulo the details on language interfaces). Second, injp is transitorily composable, i.e., any vertical pairing of injp can be proved equivalent to a single injp. It means that given two refinements $L_2 \preceq L_1$ and $L_3 \preceq L_2$ as depicted in Fig. 5b, when their rely-guarantee conditions are uniformly represented by injp, they can be merged into the direct refinement $L_3 \preceq L_1$ in Fig. 5c with a single injp as the rely-guarantee condition. We shall present the technical challenges leading to these observations in §3 and elaborate on the observations themselves in §4.

By the above observations, an obvious approach for applying direct refinements to realistic optimizing compilers is to prove open simulation for every compiler pass using injp, and vertically compose those simulations into a single simulation. However, for a non-trivial compiler like CompCert, it means we need to rewire a significant part of its proofs. More importantly, optimization passes in CompCert need additional rely-guarantee conditions as they are based on value analysis. To address the first problem, we start from the refinement proofs with least restrictive KMRs for individual passes in CompCertO [Koenig and Shao 2021], and exploit the properties that these KMRs can eventually be “absorbed” into injp in vertical composition to generate a direct refinement parameterized only by injp. To address the second problem, we propose a notion of semantic invariant that captures the rely-guarantee conditions for value analysis. When piggybacked onto injp, this semantic invariant can be transitively composed along with injp and eventually pushed to the C level. It then becomes a condition for enabling optimizations at the source level, e.g., for supporting the refinement of the optimized server in Fig. 3c. We discuss those solutions in §5.

Finally, we observe that source-level refinements can also be parameterized by injp, which enables end-to-end program verification as depicted in Fig. 4 as we shall discuss in §6.

3 BACKGROUND AND CHALLENGES

3.1 Background

We introduce the memory model, open simulations, and injp which is critical for direct refinements.

3.1.1 Block-based Memory Model. By Leroy et al. [2012], a memory state $m$ (of type mem) consists of a disjoint set of memory blocks. A memory address or pointer $(b, o)$ points to the $o$-th byte in the block $b$ where $b$ has type block and $o$ has type $Z$ (integers). The value at $(b, o)$ is denoted by $m[b, o]$. Values (of type val) are either undefined (Vundef), 32- or 64-bit integers or floats, or pointers of the form Vptr$(b, o)$. For simplicity, we often write $b$ for Vptr$(b, 0)$. The memory operations including allocation, free, read and write are provided and governed by permissions. The permission of a memory cell is ordered from high to low as Freeable $\geq$ Writable $\geq$ Readable $\geq$ NA where Freeable enables all operations, Writable enables all but free, Readable enables only read, and NA enables none. If $p_1 \geq p_2$ then any cell with permission $p_1$ also has permission $p_2$. $\text{perm}(m, p)$ denotes the set of cells with at least permission $p$. For example, $(b, o) \in \text{perm}(m, \text{Readable})$ iff the cell at $(b, o)$ in $m$ is readable. An address with no permission at all is not in the footprint of memory.

Transformations of memory states are captured via partial functions $j: \text{block} \rightarrow [\text{block} \times Z]$ called injection functions, s.t. $j(b) = \emptyset$ if $b$ is removed from memory and $j(b) = [(b', o)]$ if $b$ is shifted (injected) to $(b', o)$ in the target memory. We define mem inj $j = \text{block} \rightarrow [\text{block} \times Z]$. $v_1$ and $v_2$ are related under $j$ (denoted by $v_1 \leftrightarrow_j v_2$) if either $v_1$ is Vundef, or they are both equal scalar values, or pointers shifted according to $j$, i.e., $v_1 = \text{Vptr}(b, o)$, $j(b) = [(b', o')]$ and $v_2 = \text{Vptr}(b', o + o')$.

Given this relation, there is a memory injection between the source memory state $m_1$ and the target state $m_2$ under $j$ (denoted by $m_1 \leftrightarrow_j m_2$) if the following properties are satisfied which ensure preservation of permissions and values under injection:

$$\forall b_1, b_2, o, o', j(b_1) = [(b_2, o')] \Rightarrow (b_1, o) \in \text{perm}(m_1, p) \Rightarrow (b_2, o + o') \in \text{perm}(m_2, p).$$
$$\forall b_1, b_2, o, o', j(b_1) = [(b_2, o')] \Rightarrow (b_1, o) \in \text{perm}(m_1, \text{Readable}) \Rightarrow m_1[b_1, o] \leftrightarrow_j m_2[b_2, o + o'].$$
Memory injections are transitive and necessary for verifying transformations of memory structures (e.g., merging local variables into stack-allocated data and generating a concrete stack frame). For the remaining passes, a simpler relation called memory extension is used instead, which employs an identity injection. Reasoning about permissions under refinements is a major source of complexity.

3.1.2 A Framework for Open Simulations. In CompCertO [Koenig and Shao 2021], a language interface \( A = \langle A^q, A^r \rangle \) is a pair of sets \( A^q \) and \( A^r \) denoting acceptable queries and replies for open modules, respectively. Different interfaces may be used for different languages. The relevant ones for our discussion have been introduced in §2.1 and formally defined as follows. The language interface at the C level is \( C = \langle \text{val} \times \text{sig} \times \text{val}^* \times \text{mem}, \text{val} \times \text{mem} \rangle \) where its queries and replies take the forms \( v_f \langle \text{sg} \rangle (\tilde{o})@m \) and \( v'@m' \), respectively. The language interface at the assembly level is \( A = \langle \text{regset} \times \text{mem}, \text{regset} \times \text{mem} \rangle \) where its queries and replies take the form \( rs@m \).

Open labeled transition systems (LTSs) represent semantics of modules that may accept queries and provide replies at the incoming side and provide queries and accept replies at the outgoing side (i.e., calling external functions). An open LTS \( L : A \to B \) is a tuple \( \langle D, S, I, \to, F, X, Y \rangle \) where \( A (B) \) is the language interface for outgoing (incoming) queries and replies, \( D \subseteq B^q \) a set of initial queries, \( S \) a set of internal states, \( I \subseteq D \times S \times (F \subseteq S \times B^r) \) transition relations for incoming queries (replies), \( X \subseteq S \times A^q \) \((Y \subseteq S \times A^r \times S)\) transitions for outgoing queries (replies), and \( \to \subseteq S \times E^* \times S \) internal transitions emitting events of type \( E \). Note that \( (s, q^O) \) \( \in X \) iff an outgoing query \( q^O \) happens at \( s \); \( (s, e^O, s') \) \( \in Y \) iff after \( e^O \) returns with \( e^O \) the execution continues with an updated state \( s' \).

Kripke relations serve as interfaces of open simulations using LTs. A Kripke relation \( R : W \to \{ S | S \subseteq A \times B \} \) is a family of relations indexed by a Kripke world \( W \); for simplicity, we define \( K_W(A, B) = W \to \{ S | S \subseteq A \times B \} \). A simulation convention relating two language interfaces \( A_1 \) and \( A_2 \) is a tuple \( R = \langle W, R^q : K_W(A^q_1, A^q_2), R^r : K_W(A^r_1, A^r_2) \rangle \) which we write as \( R : A_1 \equiv A_2 \). Simulation conventions serve as interfaces of open simulations by relating source and target language interfaces. For example, a C-level convention \( c : C \equiv C = \langle \text{meminj}, R^q_c, R^r_c \rangle \) relates C queries and replies as follows, where the Kripke world consists of injections and, in a given world \( j \), the values and memory in queries and replies are related by \( j \).

\[
\begin{align*}
(v_f \langle \text{sg} \rangle (\tilde{o})@m, v'_f \langle \text{sg} \rangle (\tilde{o'})@m') & \in R^q_c(j) \iff v_f \to v'_f \land \tilde{o} \to \tilde{o'} \land m \to m'
\end{align*}
\]

\[
\begin{align*}
(v@m, v'@m') & \in R^r_c(j) \iff v \to v' \land m \to m'
\end{align*}
\]

Open forward simulations describe refinement between LTS. To establish an open (forward) simulation between \( L_1 : A_1 \to B_1 \) and \( L_2 : A_2 \to B_2 \), one needs to find two simulation conventions \( R_A : A_1 \equiv A_2 \) and \( R_B : B_1 \equiv B_2 \) that connect queries and replies at the outgoing and incoming sides, and show the internal execution steps and external interactions of open modules are related by an invariant \( R \). This simulation is denoted by \( L_1 \equiv_{R_A \to R_B} L_2 \) and formally defined as follows (for simplicity, we shall write \( L_1 \equiv_{R} L_2 \) to denote \( L_1 \equiv_{R \to R} L_2 \)).

Definition 3.1. Given \( L_1 : A_1 \to B_1, L_2 : A_2 \to B_2, R_A : A_1 \equiv A_2 \) and \( R_B : B_1 \equiv B_2 \), \( L_1 \equiv_{R_A \to R_B} L_2 \) holds if there is some Kripke relation \( R \in K_{W_{b}}(S_1, S_2) \) that satisfies:

1. \( \forall q_1, q_2, (q_1, q_2) \in R^q_{b}(w_b) \Rightarrow (q_1 \in D_1 \Leftrightarrow q_2 \in D_2) \)
2. \( \forall w_b q_1, q_2, s_1, (q_1, q_2) \in R^q_{b}(w_b) \Rightarrow (q_1, s_1) \in I_1 \Rightarrow \exists s_2, (s_1, s_2) \in R(w_b) \land (q_2, s_2) \in I_2. \)
3. \( \forall w_b s_1, s_2, t, (s_1, s_2) \in R(w_b) \Rightarrow s_1 \overset{t}{\rightarrow} s_1' \Rightarrow \exists s'_2, (s'_1, s'_2) \in R(w_b) \land s_2 \overset{t}{\rightarrow} s'_2. \)
4. \( \forall w_b s_1, s_2, q_1, (s_1, s_2) \in R(w_b) \Rightarrow (s_1, q_1) \in X_1 \Rightarrow \exists w_A q_2, (q_1, q_2) \in R^q_{A}(w_A) \land (s_2, q_2) \in X_2. \)
5. \( \forall w_b s_1, s_2, r_1, (s_1, s_2) \in R(w_b) \Rightarrow (s_1, r_1) \in F_1 \Rightarrow \exists r_2, (r_1, r_2) \in R^r_{b}(w_b) \land (s_2, r_2) \in F_2. \)

Proc. ACM Program. Lang., Vol. 8, No. POPL, Article 72. Publication date: January 2024.
Here, property (1) requires initial queries to match; (2) requires initial states to hold under the invariant \( R \); (3) requires internal execution to preserve \( R \); (4) requires \( R \) to be preserved across external calls, and (5) requires final replies to match. According to these properties, a complete forward simulation looks like Fig. 8. From the above definition, it is easy to prove the horizontal and vertical compositionality of open simulations and adequacy for assembly modules, i.e., \( \forall L_1 \ L_2 \ L'_1 \ L'_2 , \ L_1 \leq_R \ L_2 \Rightarrow L'_1 \leq_R \ L'_2 \Rightarrow L_1 \oplus L'_1 \leq_R \ L_2 \oplus L'_2 \) and \( \forall (M_1 \ M_2 : \text{Asm}) , \ [[M_1]] \oplus [[M_2]] \leq_{1\text{d}} [[M_1 + M_2]] \).

The Kripke worlds (e.g., memory injections) may evolve as the execution goes on. Rely-guarantee reasoning about such evolution is essential for horizontal composition of simulations. For illustration, the Kripke worlds at the boundary of modules are displayed in Fig. 8. The evolution of worlds across external calls is governed by an accessibility relation \( \sim_A \) for describing the rely-condition. By assuming \( w_A \sim_A w'_A \), one needs to prove the guarantee condition \( w_B \sim_B w'_B \), i.e., the evolution of worlds in the whole execution respects \( \sim_B \). Simulations with symmetric rely-guarantee conditions can be horizontally composed, even with mutual calls between modules.

Note that the accessibility relation and evolution of Kripke worlds are not explicit in the definition of simulation conventions. Instead, they are implicit by assuming a modality operator \( \diamond \) is always applied to \( R' \) s.t. \( r \in \diamond R'(w) \Leftrightarrow \exists \ w' , w \sim w' \land r \in R'(w') \). We often ignore accessibility and modality when talking purely about simulation conventions in the remaining discussion.

Accessibility relations are mainly for describing evolution of memory states across external calls. For this, simulation conventions are parameterized by Kripke Memory Relations or KMR.

**Definition 3.2.** A Kripke Memory Relation is a tuple \((W, f, \leadsto, R)\) where \( W \) is a set of worlds, \( f : W \rightarrow \text{mem} \ \text{min} \ \text{max} \) a function for extracting injections from worlds, \( \sim \subseteq W \times W \) an accessibility relation between worlds and \( R : \mathcal{K} W (\text{mem, mem}) \) a Kripke relation over memory states that is compatible with the memory operations. We write \( w \sim w' \) for \((w, w') \in \sim \).

We write \( \mathbb{R}_K \) to emphasize that a simulation convention \( \mathbb{R} \) is parameterized by the KMR \( K \), meaning \( \mathbb{R}_K \) shares the same type of worlds with \( K \) and inherits its accessibility relation.

The most interesting KMR is \( \text{injp} \) as it provides protection on memory w.r.t. injections.

**Definition 3.3 (Kripke Relation with Memory Protection).** \( \text{injp} = (W_{\text{injp}}, f_{\text{injp}}, \leadsto_{\text{injp}}, R_{\text{injp}}) \) where \( W_{\text{injp}} = (\text{mem} \times \text{mem} \times \text{mem}) \), \( f_{\text{injp}}(j, \_ , \_ , \_ ) = j \), \((m_1, m_2) \in R_{\text{injp}}(j, m_1, m_2) \Leftrightarrow m_1 \sim_{\text{l}} m_2 \) and \((j, m_1, m_2) \sim_{\text{injp}} (j', m'_1, m'_2) \Leftrightarrow j \in j' \land \text{unmapped}(j) \subseteq \text{unchanged-on}(m_1, m'_1) \land \text{out-of-reach}(j, m_1) \subseteq \text{unchanged-on}(m_2, m'_2), \land \text{mem-acc}(m_1, m'_1) \land \text{mem-acc}(m_2, m'_2) \). Here, \( \text{mem-acc}(m, m') \) denotes monotonicity of memory states such as valid blocks can only increase and read-only data does not change in value. \( \text{unchanged-on}(m, m') \) denotes memory cells whose permissions and values are not changed from \( m \) to \( m' \) and \((b_1, o_1) \in \text{unmapped}(j) \Leftrightarrow j(b_1) = \emptyset \) \((b_2, o_2) \in \text{out-of-reach}(j, m_1) \Leftrightarrow \forall b_1 o'_2 , j(b_1) = [(b_2, o'_2)] \Leftrightarrow (b_1, o_2 - o'_2) \notin \text{perm}(m_1, \text{NA}). \)
By definition, a world \((j, m_1, m_2)\) evolves to \((j', m'_1, m'_2)\) under injp only if \(j'\) is strictly larger than \(j\) and any memory cells in \(m_1\) and \(m_2\) not in the domain (i.e., unmapped by \(j\)) or image of \(j\) (i.e., out-of-reach by \(j\) from \(m_1\)) will be protected, meaning their values and permissions are unchanged from \(m_1\) (\(m_2\)) to \(m'_1\) (\(m'_2\)). An example is shown in Fig. 9 where the shaded regions in \(m_1\) are unmapped by \(j\) and unchanged while those in \(m_2\) are out-of-reach from \(j\) and unchanged. \(m'_1\) and \(m'_2\) may contain newly allocated blocks which are not protected by injp. When injp is used at the outgoing side, it denotes that the simulation relies on knowing that the unmapped and out-of-reach regions at the call side are not modified by external calls. When injp is used at the incoming side, it denotes that the simulation guarantees such regions at initial queries are not modified by the simulation itself.

### 3.2 Challenges for Vertically Composing Open Simulations

As discussed in §2.2, the challenge for constructing direct refinements for multi-pass optimizing compilers lies in their vertical composition. The most basic vertical composition for open simulations is stated below which is easily proved by pairing of individual simulations [Koenig and Shao 2021].

**Theorem 3.4 (V. Comp).** Given \(L_1 : A_1 \rightarrow B_1\), \(L_2 : A_2 \rightarrow B_2\) and \(L_3 : A_3 \rightarrow B_3\), and given \(\mathcal{R}_{12} : A_1 \Leftrightarrow A_2\), \(\mathcal{S}_{12} : B_1 \Leftrightarrow B_2\), \(\mathcal{R}_{23} : A_2 \Leftrightarrow A_3\) and \(\mathcal{S}_{23} : B_2 \Leftrightarrow B_3\),

\[
L_1 \equiv_{\mathcal{R}_{12} \rightarrow \mathcal{S}_{12}} L_2 \equiv_{\mathcal{R}_{23} \rightarrow \mathcal{S}_{23}} L_3 \equiv_{\mathcal{R}_{12} \rightarrow \mathcal{R}_{23} \rightarrow \mathcal{S}_{12} \rightarrow \mathcal{S}_{23}} L_3.
\]

Here, \((\_ \cdot \_)\) is a composed simulation convention s.t. \(\mathcal{R} \cdot \mathcal{S} = (W_R \times W_S, \mathcal{R} \cdot \mathcal{S}, \mathcal{R}' \cdot \mathcal{S}')\) where for any \(q_1\) and \(q_3\), \((q_1, q_3) \in \mathcal{R} \cdot \mathcal{S}(w_R, w_S) \Leftrightarrow \exists q_2, (q_1, q_2) \in \mathcal{R}(w_R) \wedge (q_2, q_3) \in \mathcal{S}(w_S)\) (similarly for \(\mathcal{R}' \cdot \mathcal{S}'\)). Then, given any compiler with \(N\) passes and their refinement relations \(L_i \equiv_{\mathcal{R}_{i-1} \rightarrow \mathcal{S}_{i-1}} L_{i-1}\), \(\ldots\), \(L_N \equiv_{\mathcal{R}_{N+1} \rightarrow \mathcal{S}_{N+1}} L_{N+1}\), we get their concatenation \(L_1 \equiv_{\mathcal{R}_{12} \ldots \mathcal{R}_{N+1} \rightarrow \mathcal{S}_{12} \ldots \mathcal{S}_{N+1}} L_{N+1}\), which exposes internal compilation and weakens compositionality as we have discussed in §1.2.

The above problem may be solved if the composed simulation convention can be refined into a single convention directly relating source and target queries and replies. Given two simulation conventions \(\mathcal{R}, \mathcal{S} : A_1 \leftrightarrow A_2\), \(\mathcal{R}\) is refined by \(\mathcal{S}\) if

\[
\forall w_S q_1, q_2, (q_1, q_2) \in \mathcal{S}(w_S) \Rightarrow \exists w_R, (q_1, q_2) \in \mathcal{R}(w_R) \wedge \\
\forall r_1, r_2, (r_1, r_2) \in \mathcal{R}'(w_R) \Rightarrow (r_1, r_2) \in \mathcal{S}'(w_S)
\]

which we write as \(\mathcal{R} \subseteq \mathcal{S}\). If both \(\mathcal{R} \subseteq \mathcal{S}\) and \(\mathcal{S} \subseteq \mathcal{R}\), then \(\mathcal{R}\) and \(\mathcal{S}\) are equivalent and written as \(\mathcal{R} \equiv \mathcal{S}\). By definition, \(\mathcal{R} \subseteq \mathcal{S}\) indicates any query for \(\mathcal{S}\) can be converted into a query for \(\mathcal{R}\) and any reply resulting from the converted query can be converted back to a reply for \(\mathcal{S}\). By wrapping the incoming side of an open simulation with a more general convention and its outgoing side with a more specialized convention, one gets another valid open simulation [Koenig and Shao 2021]:

**Theorem 3.5.** Given \(L_1 : A_1 \rightarrow B_1\) and \(L_2 : A_2 \rightarrow B_2\), if \(\mathcal{R}'_A \subseteq \mathcal{R}_A : A_1 \leftrightarrow A_2, \mathcal{R}_B \subseteq \mathcal{R}'_B : B_1 \leftrightarrow B_2\) and \(L_1 \equiv_{\mathcal{R}_A \rightarrow \mathcal{R}_B} L_2\), then \(L_1 \equiv_{\mathcal{R}'_A \rightarrow \mathcal{R}'_B} L_2\).

Now, we would like to prove the “real” vertical composition generating direct refinements (simulations). Given any \(L_1 \equiv_{\mathcal{R}_{12} \rightarrow \mathcal{S}_{12}} L_2\) and \(L_2 \equiv_{\mathcal{R}_{23} \rightarrow \mathcal{S}_{23}} L_3\), if we can show the existence of simulation conventions \(\mathcal{R}_{13}\) directly relating source and target semantics s.t. \(\mathcal{R}_{13} \equiv \mathcal{R}_{12} \cdot \mathcal{R}_{23}\), then \(L_1 \equiv_{\mathcal{R}_{13} \rightarrow \mathcal{S}_{13}} L_3\) holds by Theorem 3.4 and Theorem 3.5, which is the desired direct refinement. This composition is illustrated in Fig. 10 where the parts enclosed by dashed boxes represent the concatenation of \(L_1 \equiv_{\mathcal{R}_{12} \rightarrow \mathcal{S}_{12}} L_2\) and \(L_2 \equiv_{\mathcal{R}_{23} \rightarrow \mathcal{S}_{23}} L_3\). The direct queries and replies are split and merged for interaction with parallelly running simulations underlining the direct refinement.
We write $\mathcal{K}$ where $\sqsubseteq$ have many forms since Kripke relations original worlds $w$ memory (i.e., through vertical lines. A dual property we need to prove for the incoming side is shown in Fig. 11b. denote $R_w$ on the outgoing side in Fig. 10. For simplicity, we use $\exists$ depicted in Fig. 11a where black symbols are $\forall$-quantified (conclusions we need to construct). Note that Fig. 11a exactly mirrors the refinement $i$ when it connects memory states s.t. they are accessible from the $2$. Assume $R_i$ is parameterized by KMR $K_i$, showing the existence of $R_{13}$ s.t. $R_{13} \sqsubseteq R_{12} \times R_{23}$ amounts to proving a parallel refinement over the parameterizing KMRs, i.e., there exists $K_{13}$ s.t. $K_{13} \sqsubseteq K_{12} \times K_{23}$ where $K_{12} \times K_{23} = \langle W_{12} \times W_{23}, f_{12} \times f_{23}, \sim_{12} \times \sim_{23}, R_{12} \times R_{23} \rangle$. A more intuitive interpretation is depicted in Fig. 11a where black symbols are $\forall$-quantified (assumptions we know) and red ones are $\exists$-quantified (conclusions we need to construct). Note that Fig. 11a exactly mirrors the refinement on the outgoing side in Fig. 10. For simplicity, we use $w_i$ not only to represent worlds, but also to denote $R_i(w_i)$ (where $R_i$ is the Kripke relation given by KMR $K_i$) when it connects memory states through vertical lines. A dual property we need to prove for the incoming side is shown in Fig. 11b.

In both cases in Fig. 11, we need to construct interpolating states for relating source and target memory (i.e., $m'_2$ in Fig. 11a and $m_2$ in Fig. 11b). The construction of $m'_2$ is especially challenging, for which we need to decompose the evolved world $w'_{13}$ into $w'_{12}$ and $w'_{23}$ s.t. they are accessible from the original worlds $w_{12}$ and $w_{23}$. It is not clear at all how this construction is possible because 1) $m'_2$ may have many forms since Kripke relations are in general non-deterministic and 2) KMRs (e.g., $\inj$) introduce memory protection for external calls which may not hold after the (de-)composition.

Since simulation conventions are parameterized by KMRs, a major obstacle to the real vertical composition of open simulations is to prove KMRs for individual simulations can be composed into a single KMR. For this, one needs to define refinements between KMRs. Given any KMRs $K$ and $L$, $K \sqsubseteq L$ (i.e., $K$ is refined by $L$) holds if the following is true:

$$\forall w_L m_1 m_2, (m_1, m_2) \in R_L(w_L) \Rightarrow \exists w_K, (m_1, m_2) \in R_K(w_K) \land f_L(w_L) \subseteq f_K(w_K) \land$$

$$\forall w'_K m'_1 m'_2, w_K \sim_K w'_K \Rightarrow (m'_1, m'_2) \in R_K(w'_K) \Rightarrow$$

$$\exists w'_L, w_L \sim_L w'_L \land (m'_1, m'_2) \in R_L(w'_L) \land f_K(w'_K) \subseteq f_L(w'_L).$$

We write $K \equiv L$ to denote that $K$ and $L$ are equivalent, i.e., $K \sqsubseteq L$ and $L \sqsubseteq K$.
We show that To overcome the challenge for vertically composing open simulations, we exploit the observation
that \texttt{injp} in fact can be viewed as a most general KMR. Then, the compositionality of KMRs
discussed in §3.2 is reduced to transitivity of \texttt{injp}, i.e., \texttt{injp} \equiv \texttt{injp} \cdot \texttt{injp}.

4 A UNIFORM AND TRANSITIVE KRIPKE MEMORY RELATION

To overcome the challenge for vertically composing open simulations, we exploit the observation
that \texttt{injp} in fact can be viewed as a most general KMR. Then, the compositionality of KMRs
discussed in §3.2 is reduced to transitivity of \texttt{injp}, i.e., \texttt{injp} \equiv \texttt{injp} \cdot \texttt{injp}.

4.1 Uniformity of \texttt{injp}

We show that \texttt{injp} is both a reasonable guarantee condition and a reasonable rely condition for all
the compiler passes in CompCert. It is based on the observation that a notion of private and public
memory can be derived from injections and coincides with the protection provided by \texttt{injp}.

4.1.1 Public and Private Memory via Memory Injections.

Definition 4.1. Given \( m_1 \leftarrow^j_m m_2 \), the public memory regions in \( m_1 \) and \( m_2 \) are defined as follows:

\[
\text{pub-src-mem}(j) = \{ (b, o) \mid j(b) \neq \emptyset \};
\]

\[
\text{pub-tgt-mem}(j, m_1) = \{ (b, o) \mid \exists b' o', j(b') = [(b, o')] \land (b', o - o') \in \text{perm}(m_1, \text{NA}) \}.
\]

By definition, a cell \((b, o)\) is public in the source memory if it is in the domain of \( j \), and \((b, o)\) is
public in the target memory if it is mapped by \( j \) from some valid public source memory. Any memory
not public with respect to \( j \) is private. We can see that private memory corresponds exactly to un-
mapped and out-of-reach memory defined by \texttt{injp}, i.e., for any \( b \) and \( o \), \((b, o) \in \text{pub-src-mem}(j) \equiv (b, o) \notin \text{unmapped}(j) \) and \((b, o) \in \text{pub-tgt-mem}(j, m) \equiv (b, o) \notin \text{out-of-reach}(j, m) \).

With Definition 4.1 and the properties of memory injection (see §3.1.1), we can easily prove
access of pointers in a readable and public source location gets back another public location.

Lemma 4.2. Given \( m_1 \leftarrow^j_m m_2 \),

\[
\forall b_1 o_1, (b_1, o_1) \in \text{pub-src-mem}(j) \Rightarrow (b_1, o_1) \in \text{perm}(m_1, \text{Readable}) \Rightarrow
m_1[b_1, o_1] = \forall \text{ptr}(b'_1, o'_1) \Rightarrow (b'_1, o'_1) \in \text{pub-src-mem}(j).
\]

It implies that readable public memory regions form a “closure” such that the sequences of reads
are bounded inside these regions, as shown in Fig. 12. The horizontal arrows indicates a pointer
value \((b_{i+1}, o_{i+1})\) is read from \((b_i, o_i)\) with possible adjustment with pointer arithmetic. Note that
all memory cells at \((b_i, o_i)s\) and \((b'_i, o'_i)s\) have \texttt{Readable} permission. By Lemma 4.2, \((b_i, o_i)s\) are all
in public regions. By Definition 4.1, the mirroring reads \((b'_i, o'_i)s\) are also in public regions.

Fig. 12. Closure of Public Memory

Fig. 13. Protection of Private Memory by \texttt{injp}

Because of the above difficulties, existing approaches either make substantial changes to semantics
for constructing interpolating states, thereby destroying adequacy [Stewart et al. 2015], or do not
even try to merge Kripke memory relations, but instead leave them as separate entities [Koenig
and Shao 2021; Song et al. 2020]. As a result, direct refinements cannot be achieved.
4.1.2 injp as a Uniform Rely Condition. injp is adequate for preventing external calls from interfering with internal execution for all the compiler passes of CompCert. To illustrate this point, we discuss the effect of injp on two of CompCert’s passes using Fig. 13a as an example where g is an external function. The first pass is SImpLocals which converts local variables whose memory addresses are not taken into temporary ones. As shown in Fig. 13b, x is turned into a temporary variable at the target level which is not visible to g. Therefore, x at the source level becomes private data as its block b_x is unmapped by j, thereby protected by injp and cannot be modified by g. The second pass is Stacking which expands the stack frames with private regions for return addresses, spilled registers, arguments, etc. Continuing with our example, the only public stack data in Fig. 13c is y. All the private data is out-of-reach, thereby protected by injp.

4.1.3 injp as a Uniform Guarantee Condition. For injp to serve as a uniform guarantee condition, it suffices to show the private memory of the environment is protected between initial calls and final replies. During an open forward simulation, all incoming values and memories are related by some initial injection j (e.g., \( \mathcal{S}_1 \leftarrow^j \mathcal{S}_2 \) and \( m_1 \leftarrow^j m_2 \)). In particular, the pointers in them are related by j. Therefore, any sequence of reads starting from pointers stored in the initial queries only inspect public memories in the source and target, as already shown in Fig. 12. The private (i.e., unmapped or out-of-reach) regions of the initial memories are not modified by internal execution. Moreover, because injection functions only grow bigger during execution but never change in value and the outgoing calls have injp as a rely-condition, the initially unmapped (out-of-reach) regions will stay unmapped (out-of-reach) and be protected during external calls. As a result, we conclude that injp is a reasonable guarantee condition for any open simulation.

4.2 Transitivity of injp

The goal is to show the two refinements in Fig. 11 hold when \( K_{ij} = \text{injp} \), i.e., injp \( \equiv \text{injp} \cdot \text{injp} \). As discussed in §3.2, the critical step is to construct interpolating memory states that transitively relate source and target states. The construction is based on two observations: 1) the memory injections deterministically decide the value and permissions of public memory because they encode partial functional transformations on memory states, and 2) any memory not in the domain or range of the partial functions is protected (private) and unchanged throughout external calls. Although the proof is quite involved, the result can be reused for all compiler passes thanks to injp’s uniformity.

4.2.1 injp \( \subseteq injp \cdot injp \). By definition, we need to prove the following lemma:

**Lemma 4.3.** injp \( \subseteq injp \cdot injp \) holds. That is,
\[
\forall j_{12} j_{23} m_1 m_2 m_3, m_1 \leftarrow^j_{12} m_2 \Rightarrow m_2 \leftarrow^j_{23} m_3 \Rightarrow \exists j_{13}, m_1 \leftarrow^j_{13} m_3 \land \\
\exists m'_1 j'_{13}, (j_{13}, m_1, m_3) \sim_{\text{injp}} (j'_{13}, m'_1, m_3) \Rightarrow m'_1 \leftarrow^j_{13} m'_3 \Rightarrow \\
\exists m'_2 j'_{12}, j'_{23}, (j_{12}, m_1, m_2) \sim_{\text{injp}} (j'_{12}, m'_1, m'_2) \land m'_1 \leftarrow^j_{12} m'_2 \\
\land (j_{23}, m_2, m_3) \sim_{\text{injp}} (j'_{23}, m'_2, m'_3) \land m'_2 \leftarrow^j_{23} m'_3.
\]

This lemma conforms to the graphic representation in Fig. 11a. To prove it, an obvious choice is to pick \( j_{13} = j_{23} \cdot j_{12} \). Then, we are left to prove the existence of interpolating state \( m'_2 \) and the

---

\(^2\)In fact, the properties in Definition 3.3 are exactly from CompCert’s assumptions on external calls.
memory and accessibility relations as shown in Fig. 14. By definition, $m_2'$ consists of memory blocks newly allocated with respect to $m_2$ and blocks that already exist in $m_2$. The latter can be further divided into public and private memory regions with respect to injections $j_{12}$ and $j_{23}$. Then, $m_2'$ is constructed following the ideas that 1) the public and newly allocated memory should be projected from the updated source memory $m_1'$ by $j_{12}'$; and 2) the private memory is protected by $\text{injp}$ and should be copied over from $m_2$ to $m_2'$.

We use the concrete example in Fig. 15 to motivate the construction of $m_2'$. Here, the white and green areas correspond to locations in $\text{perm}(_{\text{Readable}})$ (with at least some permission) and in $\text{perm}(_{\text{NA}})$ (with at least readable permission), respectively. Given $m_1 \xrightarrow{j_{12}'} m_2, m_2 \xrightarrow{j_{23}'} m_3$ and $(j_{23} \cdot j_{12}, m_1, m_3) \sim_{\text{injp}} (j_{13}', m_1', m_3')$, we need to define $j_{12}'$ and $j_{23}'$ and build $m_2'$ satisfying $m_1 \xrightarrow{j_{12}'} m_2', m_2' \xrightarrow{j_{23}'} m_3'$. $m_1'$ and $m_3'$ are expansions of $m_1$ and $m_3$ with new blocks and possible modification to the public regions of $m_1$ and $m_3$. Here, $m_1'$ has a new block $b_1'$ and $m_3'$ has two new blocks $b_3'$ and $b_4'$.

We first fix $j_{12}'$, $j_{23}'$ and the shape of blocks in $m_2'$. We begin with $m_2$ and introduce a newly allocated block $b_1'$ whose shape matches $b_1$ in $m_1'$. Then, $j_{12}'$ is obtained by expanding $j_{12}$ with identity mapping from $b_1'$ to $b_1$. Furthermore, $j_{23}'$ is also expanded with a mapping from $b_2'$ to a block in $m_3'$; this mapping is determined by $j_{13}'$.

We then set the values and permissions for memory cells in $m_2'$ so that it satisfies injection and the unchanged-on properties for readable memory regions implied by $(j_{12}, m_1, m_2) \sim_{\text{injp}} (j_{12}', m_1', m_2')$ and $(j_{23}, m_2, m_3) \sim_{\text{injp}} (j_{23}', m_1', m_3')$. The values and permissions for newly allocated blocks are obviously mapped from $m_1'$ by $j_{12}'$. Those for old blocks are fixed as follows. By memory protection provided in $(j_{23} \cdot j_{12}, m_1, m_3) \sim_{\text{injp}} (j_{13}', m_1', m_3')$, the only memory cells in $m_1$ that may have been modified in $m_1'$ are those mapped all the way to $m_3$ by $j_{23} \cdot j_{12}$, while the cells in $m_3$ that may be modified in $m_3'$ must be in the image of $j_{23} \cdot j_{12}$. To match this fact, the only old memory regions in $m_2'$ whose values and permissions may be modified are those both in the image of $j_{12}$ and the domain of $j_{23}$. Those are the public memory with respect to $j_{12}$ and $j_{23}$ and displayed as the gray areas in Fig. 15b. Following idea 1) above, the values and permissions in those regions are projected from $m_1'$ by applying the injection function $j_{12}$. Note that there is an exception: values in read-only public regions are copied over from $m_2$. Following idea 2) above, the remaining old memory regions are private with respect to $j_{12}$ and $j_{23}$ and should have the same values and permissions as in $m_2$.

Note that the accessibility relations $(j_{12}, m_1, m_2) \sim_{\text{injp}} (j_{12}', m_1', m_2')$ and $(j_{23}, m_2, m_3) \sim_{\text{injp}} (j_{23}', m_1', m_3')$ can be derived from $(j_{23} \cdot j_{12}, m_1, m_3) \sim_{\text{injp}} (j_{13}', m_1', m_3')$ because the latter enforces stronger protection than the former. This is due to unmapped and out-of-reach regions getting bigger as memory injections get composed. For example, in Fig. 15, $b_1'$ is mapped by $j_{12}$ but becomes unmapped by $j_{23} \cdot j_{12}$; the image of $b_1'$ in $b_3$ is in reach by $j_{23}$ but becomes out-of-reach by $j_{23} \cdot j_{12}$.
We list the compiler passes and their simulation types in Table 1 (passes on the right follow the self-simulating passes on the left) together with their source and target languages and interfaces (in bold fonts). The passes in black are reused from CompCertO, while those in red are reproved optimizing passes.

### Table 1. Significant Passes of CompCert

<table>
<thead>
<tr>
<th>Languages/Passes</th>
<th>Outgoing $\rightarrow$Incoming</th>
<th>Language/Pass</th>
<th>Outgoing $\rightarrow$Incoming</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clight</td>
<td>$C \rightarrow C$</td>
<td>Constprop</td>
<td>$ro \cdot c_{injp} \rightarrow ro \cdot c_{injp}$</td>
</tr>
<tr>
<td>Self-Sim</td>
<td>$ro \cdot c_{injp} \rightarrow ro \cdot c_{injp}$</td>
<td>CSE</td>
<td>$ro \cdot c_{injp} \rightarrow ro \cdot c_{injp}$</td>
</tr>
<tr>
<td>SimplLocals</td>
<td>$c_{injp} \rightarrow c_{injp}$</td>
<td>Deadcode</td>
<td>$ro \cdot c_{injp} \rightarrow ro \cdot c_{injp}$</td>
</tr>
<tr>
<td>Csharpminor</td>
<td>$C \rightarrow C$</td>
<td>Unusedglob</td>
<td>$c_{injp} \rightarrow c_{injp}$</td>
</tr>
<tr>
<td>Cminorgen</td>
<td>$C \rightarrow C$</td>
<td>Allocation</td>
<td>$wt \cdot c_{ext} \cdot CL \rightarrow wt \cdot c_{ext} \cdot CL$</td>
</tr>
<tr>
<td>Cminor</td>
<td>$c_{injp} \rightarrow c_{injp}$</td>
<td>Mach</td>
<td>$L \rightarrow L$</td>
</tr>
<tr>
<td>Selection</td>
<td>$wt \cdot c_{ext} \rightarrow wt \cdot c_{ext}$</td>
<td>Linear</td>
<td>$L \rightarrow L$</td>
</tr>
<tr>
<td>CminorSel</td>
<td>$c_{ext} \rightarrow c_{ext}$</td>
<td>Stacking</td>
<td>$ltl_{injp} \cdot LM \rightarrow LM \cdot mach_{inj}$</td>
</tr>
<tr>
<td>RTL</td>
<td>$C \rightarrow C$</td>
<td>Machgen</td>
<td>$mach_{ext} \cdot MA \rightarrow mach_{ext} \cdot MA$</td>
</tr>
<tr>
<td>Self-Sim</td>
<td>$c_{injp} \rightarrow c_{injp}$</td>
<td>Asmgen</td>
<td>$A \rightarrow A$</td>
</tr>
<tr>
<td>Tailcall</td>
<td>$c_{ext} \rightarrow c_{ext}$</td>
<td>Asml</td>
<td>$asm_{injp} \rightarrow asm_{injp}$</td>
</tr>
<tr>
<td>Inlining</td>
<td>$c_{injp} \rightarrow c_{injp}$</td>
<td>Asm</td>
<td>$asm_{injp} \rightarrow asm_{injp}$</td>
</tr>
<tr>
<td>Self-Sim</td>
<td>$c_{injp} \rightarrow c_{injp}$</td>
<td>Self-Sim</td>
<td>$asm_{injp} \rightarrow asm_{injp}$</td>
</tr>
</tbody>
</table>

4.2.2 $injp \cdot injp \subseteq injp$. By definition, we need to prove:

**Lemma 4.4.** $injp \cdot injp \subseteq injp$ holds. That is,

\[
\forall j_{13} m_1 m_3, m_1 \xleftarrow{j_{13}} m_3 \Rightarrow \exists j_{12} f_{23} m_2, m_1 \xleftarrow{j_{12}} m_2 \land m_2 \xleftarrow{j_{23}} m_3 \land \\
\forall m_1' m_2' j_{12}' j_{23}' (j_{12}, m_1, m_2) \sim_{injp} (j_{12}', m_1', m_2') \Rightarrow (j_{23}, m_2, m_3) \sim_{injp} (j_{23}', m_2', m_3') \Rightarrow \\
m_1' \xleftarrow{j_{12}'} m_2' \Rightarrow m_2' \xleftarrow{j_{23}'} m_3' \Rightarrow \exists j_{13}' (j_{13}', m_1, m_3) \sim_{injp} (j_{13}', m_1', m_3', m_3') \land m_1' \xleftarrow{j_{13}'} m_1'.
\]

This lemma conforms to Fig. 11b. To prove it, we pick $j_{12}$ to be an partial identity injection ($j_{12}(b) = [b, 0]$ when $j_{12}(b) \not= \emptyset$), $j_{23} = j_{13}$ and $m_2 = m_1$. Then the lemma is reduced to proving the existence of $j'_{13}$ that satisfies ($j_{13}, m_1, m_3) \sim_{injp} (j'_{13}, m'_1, m'_3')$ and $m_1' \xleftarrow{j_{13}'} m_1'$. By picking $j'_{13} = j_{12}' \cdot j_{23}'$, we can easily prove these properties by exploiting the properties of $injp$.

5 DERIVATION OF THE DIRECT REFINEMENT FOR COMPCCERT

In this section, we discuss the proofs and composition of open simulations for the compiler passes of CompCert into the direct refinement $\leq_{dc}$ following the ideas discussed in §2.2. CompCert compiles Clight programs into Asm programs through 19 passes [Leroy 2023], including several optimization passes working on the RTL intermediate language. First, we prove the open simulations for all these passes with appropriate simulation conventions. In particular, we directly reuse the proofs of non-optimizing passes in CompCertO and update the proofs of optimizing passes with semantic invariants. Second, we prove a collection of properties for refining simulation conventions in preparation for vertical composition. Those properties enable absorption of KMRs into $injp$ and composition of semantic invariants. They rely critically on transitivity of $injp$. Finally, we vertically compose the simulations and refine the incoming and outgoing simulation conventions into a single simulation convention $\mathbb{C}$, thereby establishing $\leq_{dc}$ as the top-level refinement $\leq_{ac}$.

### 5.1 Open Simulation of Individual Passes

We list the compiler passes and their simulation types in Table 1 (passes on the right follow the passes on the left) together with their source and target languages and interfaces (in bold fonts). The passes in black are reused from CompCertO, while those in red are reproved optimizing passes. The passes in blue are *self-simulating* passes we inserted; they will be used in §5.3 for refining
composed simulation conventions. Note that we have omitted passes with the identity simulation convention (i.e., simulations of the form $L_1 \leq_{id} L_2$) in Table 1 as they do not affect the proofs.

5.1.1 Simulation Conventions and Semantic Invariants. We first introduce relevant simulation conventions and semantic invariants shown in Table 1. The simulation conventions $c_K : C \equiv C$, $\ltl K : L \equiv L$, $\text{mach}_K : M \equiv M$, and $\text{asm}_K : A \equiv A$ relate the same language interfaces with queries and replies native to the associated intermediate languages. They are parameterized by a KMR $K$ to allow different compiler passes to have different assumptions on memory evolution. Conceptually, this parameterization is unnecessary as we can simply use $\text{injp}$ for every pass due to its uniformity (as discussed in §4.1). Nevertheless, it is useful because the compiler proofs become simpler and more natural with the least restrictive KMRs which may be weaker than $\text{injp}$. CompCertO defines several KMRs weaker than $\text{injp}$: $\text{id}$ is used when memory is unchanged; $\text{ext}$ is used when the source and target memory share the same structure; $\text{inj}$ is a simplified version of $\text{injp}$ without its memory protection. The simulation conventions $\text{CL} : C \equiv L$, $\text{LM} : L \equiv M$ and $\text{MA} : M \equiv A$ capture the calling convention of CompCert: $\text{CL}$ relates C-level queries and replies to those in the $\text{LTL}$ language where the arguments are distributed to abstract stack slots; $\text{LM}$ further relates abstract stack slots with states on an architecture independent machine; $\text{MA}$ relates this state to registers and memory in the assembly language (X86 assembly in our case). As discussed before, some refinements rely on invariants on the source semantics. The semantic invariant $\text{wt}$ enforces that arguments and return values of function calls respect function signatures. $\text{ro}$ is critical for ensuring the correctness of optimizations, which will be discussed next.

5.1.2 Open Simulation of Optimizations. The optimizing passes $\text{Constprop}$, $\text{CSE}$ and $\text{Deadcode}$ perform constant propagation, common subexpression elimination and dead code elimination, respectively. They make use of a static value analysis algorithm for collecting information of variables during the execution. For each function, this algorithm starts with the known initial values of read-only (constant) global variables. It simulates the function execution to analyze the values of global or local variables after executing each instruction. In particular, for global constant variables, their references at any point should have the initial values of constants. For local variables stored on the stack, their references may have initial values or may not if interfered by other function calls. When the analysis encounters a call to another function, it checks whether the address of current stack frame is leaked to the callee directly through arguments or indirectly through pointers in memory. If not, then the stack frame is considered unreachable from its callee. Consequently, the references to local variables on unreachable stack frames after function calls remain to be their initial values. Based on this analysis, the three passes then identify and perform optimizations.

Most of the proofs of closed simulations for those passes can be adapted to open simulation straightforwardly. The only and main difficulty is to prove that information derived from static analysis is consistent with the dynamic memory states in incoming queries and after external calls return. We introduce the semantic invariant $\text{ro}$ and combine it with $\text{injp}$ to ensure this consistency. The above optimization passes all use $\text{ro} \cdot c_{\text{injp}}$ as their simulation conventions (because $\text{RTL}$

Fig. 16. An Example of Constant Propagation

(a) Source Program  
(b) Target Program

const int key = 42;  
void foo(int*);  
int double_key() {  
   int a = key;  
   foo(&key);  
   return a + key;  
} 

int a = 42;  
foo(&key);  
return 84;
conforms to the C interface). The adaptation of optimization proofs for those passes is similar. As an example, we only discuss constant propagation whose correctness theorem is stated as follows:

**Lemma 5.1.** \( ∀(M, M') : \text{RTL}, \text{Constprop}(M) = M' \Rightarrow [\![M]\!] \leq_{\text{ro-injp}} [\![M']\!] \).

Instead of presenting its proof, we illustrate how \( \text{ro} \) and \( \text{injp} \) help establish the open simulation for \( \text{Constprop} \) through a concrete example as depicted in Fig. 16. This example covers optimization for both global constants (e.g., \( \text{key} \)) and local variables (e.g., \( a \)). By static analysis of Fig. 16a, 1) \( \text{key} \) contains 42 at line 4 because \( \text{key} \) is a constant global variable and, 2) both \( \text{a} \) and a contain 42 after the external call to \( \text{foo} \) returns to line 6. Here, the analysis confirms \( \text{key} \) has the value 42 because \( \text{foo} \) (if well-behaved) will not modify a constant global variable. Furthermore, \( \text{a} \) has the value 42 because it resides in the stack frame of \( \text{double_key} \) which is unreachable from \( \text{foo} \) (in fact, \( a \) is the only variable in the frame). As a result, the source program is optimized into Fig. 16b.

We first show that \( \text{ro} \) guarantees the dynamic values of global constants are consistent with static analysis. That is, global variables are correct in incoming memory and are protected during external calls. \( \text{ro} \) is defined as follows:

**Definition 5.2.** \( \text{ro} : C \leftrightarrow C = \langle W_\text{ro}, R^\text{q}_{\text{ro}}, R^\text{c}_{\text{ro}} \rangle \) where \( W_\text{ro} = (\text{symtbl} × \text{mem}) \) and

\[
R^\text{q}_{\text{ro}}(se, m) = \{(v_f[\text{sg}](\overline{v})@m, v_f[\text{sg}](\overline{v})@m) \mid \text{ro-valid}(se, m)\}
\]

\[
R^\text{c}_{\text{ro}}(se, m) = \{(\text{res}@m', \text{res}@m') \mid \text{mem-acc}(m, m')\}
\]

Note that although \( \text{ro} \) takes the form of a simulation convention, it only relates the same queries and replies, i.e., it only enforces invariants on the source side. This kind of simulation conventions are what we called semantic invariants. A symbol table \( se \) (of type \( \text{symtbl} \)) is provided together with memory, so that the semantics can locate the memory blocks and initial values of global definitions. \( \text{ro-valid}(se, m) \) states that the values of global constant variables in the incoming memory \( m \) are the same as their initial values. Therefore, the optimization of key into 42 at line 4 of Fig. 16a is correct. For the call to \( \text{foo} \), \( \text{mem-acc}(m, m') \) ensures that read-only values in memory are unchanged, therefore \( \text{ro-valid} \) is preserved across external calls (i.e., \( \text{ro-valid}(se, m) \Rightarrow \text{mem-acc}(m, m') \Rightarrow \text{ro-valid}(se, m') \)). As a result, replacing key with 42 at line 6 makes sense.

We then show that \( \text{injp} \) guarantees the dynamic values of unreachable local variables are consistent with static analysis. That is, unreachable stack values are unchanged by external calls. This protection is realized by \( \text{injp} \) with shrinking memory injections. Fig. 17 shows the protection of a when calling \( \text{foo} \). Before the external call to \( \text{foo} \), the source blocks \( b_\text{a} \) and \( b_\text{key} \) are mapped to target blocks by the current injection \( j \). The analysis determines that the argument and memory passed to \( \text{foo} \) do not contain any pointer to \( b_\text{a} \). Therefore, we can simply remove \( b_\text{a} \) from \( j \) to get a shrunk yet valid memory injection \( j_i \). Then, \( b_\text{a} \) is protected during the call to \( \text{foo} \). \( b_\text{a} \) is added back to the injection after \( \text{foo} \) returns and the simulation continues.

Finally, \( \text{Unusedglob} \) which removes unused static global variables is verified by assuming that global symbols remain the same throughout the compilation and with a weaker KMR \( \text{inj} \).
5.2 Properties for Refining Simulation Conventions

We present properties necessary for refining the composed simulation conventions in Table 1.

5.2.1 Commutativity of KMRs and Structural Conventions.

**Lemma 5.3.** For Z ∈ {CL, LM, MA} and K ∈ {ext, inj, injp} we have X_K • Z ⊆ Z • Y_K.

This lemma is provided by CompCertO [Koenig and Shao 2021]. X and Y denote the simulation conventions for the source and target languages of Z, respectively (e.g., X = c and Y = 1t1 when Z = CL). If K = injp we get c_injp • CL ⊆ CL • 1t1_injp. This lemma indicates at the outgoing (incoming) side a convention lower (higher) than CL, LM and MA may be lifted over them to a higher position (pushed down to a lower position).

5.2.2 Absorption of KMRs into injp. The lemma below is needed for absorbing KMRs into injp:

**Lemma 5.4.** For any R, (1)R_injp • R_injp ≡ R_injp (2)R_injp ⊑ R_inj • inj (3)R_injp • R_inj • inj ⊑ R_injp (4)R_inj • inj ⊑ R_inj (5)R_ext • R_inj ≡ R_inj (6)R_inj • R_ext ≡ R_inj (7)R_ext • R_ext ≡ R_ext.

The simulation convention R is parameterized over a KMR. Property (1) is a direct consequence of injp • injp = injp, which is critical for merging simulations using injp. The remaining ones either depend on transitivity of injp, or trivially hold as shown by Koenig and Shao [2021].

5.2.3 Composition of Semantic Invariants. Lastly, we also need to handle the two semantic invariants ro and wt. They cannot be absorbed into injp because their assumptions are fundamentally different. Therefore, our goal is to permute them to the top-level and merge any duplicated copies. The following lemmas enable elimination and permutation of wt:

**Lemma 5.5.** For any R_K : C ⇒ C, we have (1)R_K • wt ≡ wt • R_K • wt and (2)R_K • wt ≡ wt • R_K.

ro is more difficult to handle as it does not commute with arbitrary simulation conventions. To eliminate redundant ro, we piggyback ro onto injp and prove the following transitivity property:

**Lemma 5.6.** ro • c_injp ≡ ro • c_injp • ro • c_injp

Its proof is similar to that for c_injp = c_injp • c_injp but with additional reasoning for establishing ro. A graphic presentation of the proof is given in Fig. 18 which mirrors Fig. 11. We focus on the additional reasoning and have omitted the ⊢ injp relations and the worlds for injp in Fig. 18.

![Fig. 18. Transitivity of ro • c_injp](image_url)
that by definition the worlds \((se, m)\) for \(ro\) do not evolve like those for \(injp\). A red circle around a memory state \(m\) indicates the necessity to prove that \(ro\)-valid in \(Rt\) holds for \(m\). The \text{mem-acc} relations over dashed arrows are the properties over replies in \(Rt\) and must also be verified.

The above additional properties are proved based on two observations. First, the properties for queries (i.e., \(ro\)-valid) are propagated in refinement along with copying of memory states. For example, to prove the refinement in Fig. 18a, we are given \(ro\)-valid \((se1, m1)\) and \(ro\)-valid \((se2, m2)\) according to the initial \(Rt\) relations. By choosing \((se1, m1)\) to be the world for the composed \(Rt\), \(ro\)-valid \((se1, m1)\) holds trivially for \(m1\) in the circle. To prove the refinement in Fig. 18b, we need to prove that the interpolating memory state after the initial decomposition satisfies \(Rt\). By choosing \(m1\) to be this state (in the middle circle in Fig. 18b and according to the proof of Lemma 4.4), \(ro\)-valid \((se1, m1)\) follows directly from the initial assumption. Second, the properties for replies (i.e., \text{mem-acc}) have already been encoded into \(\sim injp\) by Definition 3.3. For example, \(m2\) in Fig. 18a is constructed by following exactly Lemma 4.3. Therefore, \text{mem-acc} \((m2, m3)\) trivially holds.

Finally, at the top level, we need \(ro\) and \(wt\) to commute which is straightforward to prove:

**Lemma 5.7.** \(ro \cdot wt \equiv wt \cdot ro\)

### 5.3 Proving the Direct Open Simulation for CompCert

We first insert self-simulations into the compiler passes, as shown in Table 1. This is to supply extra \(Rinj, Rinjp\), and \(ro\) for absorbing \(Rext\) \((Rinj)\) into \(Rinj\) \((Rinjp)\) by properties in Lemma 5.4 and for transitive composition of \(ro\). Self-simulations are obtained by the following lemma:

**Theorem 5.8.** If \(p\) is a program written in Clight or RTL and \(R \in \{ro, cext, cinj, cinjp\}\), or \(p\) is written in Asm and \(R \in \{asmext, asmijn, asmijnjp\}\), then \([[p]] \equiv [[R \Rightarrow P]] \Rightarrow P\) holds.

We unify the conventions at the incoming and outgoing sides. We start with the simulation \(L1 \equiv L2\) which is the transitive composition of compiler passes in Table 1 where

\[
R = ro \cdot cinjp \cdot cinjp \cdot cinjp \cdot wt \cdot cext \cdot cext \cdot cinj \cdot cext \cdot cinjp \cdot cinjp \cdot ro \cdot cinjp
\]

\[
S = ro \cdot cinjp \cdot cinjp \cdot cinjp \cdot wt \cdot cext \cdot cext \cdot cinj \cdot cext \cdot cinjp \cdot ro \cdot cinjp \cdot ro \cdot cinjp
\]

We then find two sequences of refinements \(C \subseteq Rn \subseteq \ldots \subseteq R1 \subseteq R\) and \(S \subseteq S1 \subseteq \ldots \subseteq Sm \subseteq C\), by which and Theorem 3.5 we get the simulation \(L1 \equiv L2\). The direct simulation convention is \(C = ro \cdot wt \cdot CAinjp \cdot asmijnjp\). \(ro\) enables optimizations at \(C\) level while \(wt\) ensures well-typedness. The definition of \(CAinjp\) has already been discussed informally in \(\S.2.1\); its formal definition is given in the technical report [Zhang et al. 2023b]. The last \(asmijnjp\) is irrelevant as assembly code is self-simulating by Theorem 5.8. The final correctness theorem is shown below:

**Theorem 5.9.** *Compilation in CompCert is correct in terms of open simulations,*

\[
\forall (M : \text{Clight}) (M' : \text{Asm}), \text{CompCert}(M) = M' \Rightarrow [[M]] \equiv [[M']] \equiv
\]

We explain how the refinements are carried out at the outgoing side. Refinements at the incoming side are similar. The following is the sequence of refined simulation conventions \(C \subseteq Rn \subseteq \ldots \subseteq R1 \subseteq R\). It begins with \(R\) and ends with \(C\).

\[
(1) \quad ro \cdot c_{injp} \cdot c_{injp} \cdot c_{injp} \cdot wt \cdot c_{ext} \cdot c_{ext} \cdot c_{inj} \cdot c_{ext} \cdot c_{injp} \cdot ro \cdot c_{injp} \cdot ro \cdot c_{injp} \\
\]

\[
(2) \quad ro \cdot c_{injp} \cdot wt \cdot c_{inj} \cdot c_{injp} \cdot ro \cdot c_{injp} \\
\]

\[
(3) \quad ro \cdot c_{injp} \cdot wt \cdot c_{inj} \cdot wt \cdot c_{injp} \cdot ro \cdot c_{injp} \\
\]

Proc. ACM Program. Lang., Vol. 8, No. POPL, Article 72. Publication date: January 2024.
We use server_opt.s instead of server.s to illustrate how optimizations are enabled by ro. The proof for the unoptimized server is similar with only minor adjustments. A formal definition of LTS for L₅ is given below and its transition diagram is given in Fig. 19a.

Definition 6.1. LTS of Lₛ:

\[ Sₕ := \{ \text{Calle } i \ f \ m, \text{Callp } sp \ f \ v \ m, \text{Retp } sp \ m, \text{Rete } m \} ; \]

\[ Iₕ := \{ (\text{Vptr}(b_e, 0)[\text{int} \rightarrow \text{ptr} \rightarrow \text{void}]) (i, v_f) @ m, \text{Calle } i \ f \ m \} ; \]

\[ \rightarrowₕ := \{ (\text{Calle } i \ f \ m, \text{Callp } sp \ f \ v \ m') | (m', sp) = \text{alloc } m \ 0 \ 8 \land \]

\[ m' = m'[sp \leftarrow (i \text{ XOR } m[b_k])] \} \cup \{ (\text{Retp } sp \ m, \text{Rete } m') | m' = \text{free } m \ sp \} ; \]

\[ Xₕ := \{ (\text{Callp } sp \text{Vptr}(b_p, 0) m, \text{Vptr}(b_p, 0)[\text{ptr} \rightarrow \text{void}]) (\text{Vptr}(sp, 0)) @ m \} ; \]

\[ Yₕ := \{ (\text{Callp } sp \ f \ v \ m, \text{res} @ m', \text{Retp } sp \ m') \} ; \]

\[ Fₕ := \{ (\text{Rete } m, \text{Vundef} @ m) \} . \]

The LTS has four internal states as depicted in Fig. 19a. Initialization is encoded in Iₛ. If the incoming query q₁ contains a function pointer Vptr(b_e, 0) which points to encrypt, Lₛ enters Calle i \ f \ m where i and v_f are its arguments. The first internal transition allocates the stack frame sp and stores the result of encryption i XOR m[b_k] in sp where b_k contains key. Then, it enters Callp
We first prove the following source-level refinement where 

\[ \text{well-typed outgoing arguments and return value.} \]

process which is the state before calling 

Therefore, the private stack values of the server are protected. For the unoptimized server, the only 

variables (key) and asserts that it is a constant. The only difference between the specifications for 

server_opt.s and server.s is whether key is a constant in the symbol table. The semantics of assembly module \([\text{server_opt.s]}\) is given by CompCertO whose transition diagram is shown in 

Fig. 19b. All the states, including queries and replies, are composed of register sets and memories. 

Now, we need to prove the following forward simulation. The most important points of the proof 

are how ro enables optimizations and how injp preserves memory across external calls.

**Theorem 6.2.** \( L_S \leq_C [\text{server_opt.s}] \).

At the top level, we expand \( C \) to \( ro \cdot wt \cdot \text{Cinjp} \cdot \text{asm_injp} \) and switch the order of \( ro \) and \( wt \) by Lemma 5.7. By the vertical compositionality (Theorem 3.4), we first establish \( L_S \leq_{wt} L_S \) with the well-typed outgoing arguments and return value. \([\text{server_opt.s}] \leq_{\text{asm_injp}} [\text{server_opt.s}]\) is proved by Theorem 5.8.

We are left with proving \( L_S \leq_{ro \cdot \text{Cinjp}} [\text{server_opt.s}] \). That is, we need to show the simulation diagram in Fig. 20 holds where \( \mathcal{S} = ro \cdot \text{Cinjp} \) (which mirrors Fig. 6). Here, the given assumptions and the conclusions to be proved are represented as black and red arrows, respectively. For the proof, we need an invariant \( R \in \mathcal{K}_{W_{ro \cdot \text{Cinjp}}} (S_0, \text{regset} \times \text{mem}) \). The most important point is that \( ro \) and injp play essential roles in establishing the invariant. First, \( ro\text{-valid} \) is propagated from the initial source query \( q^l_I \) to internal program states. This guarantees that the value of key read from the source memory states is always 42, hence matching the constant in \( \text{Pxor} \) 42 \( \text{RDI} \) in server_opt.s. Second, injp is essential for deriving that memory locations in the target stack frame with offset \( o \) (\( o < 8 \) or \( 16 \leq o \)) are unchanged since they are designated out-of-reach by \( R \). Therefore, the private stack values of the server are protected. For the unoptimized server, the only difference is that we decompose \( L_S \leq_{ro \cdot \text{Cinjp}} [\text{server.s}] \) into \( L_S \leq_{ro} L_S \) which trivially holds and \( L_S \leq_{\text{Cinjp}} [\text{server.s}] \) which can be proved without the help of \( ro \).

### 6.2 End-to-end Correctness Theorem

We first prove the following source-level refinement where \( L_{CS} \) is the top-level specification. Its proof follows the same pattern as Theorem 6.2 but is considerably simpler because the source and target semantics share the same \( C \) interface.

**Lemma 6.3.** \( L_{CS} \leq_{ro \cdot wt \cdot \text{Cinjp}} [\text{client.c}] \oplus L_S \).

We then prove the following simulation, which is immediate from the full compositionality, the adequacy for assembly described in §3.1.2, Theorem 5.9, and Theorem 6.2:

**Lemma 6.4.** \([\text{client.c}] \oplus L_S \leq_C [\text{client.s + server_opt.s}]\).
/* client.c */
#define N 10
int input[N] = {...};
int result[N];
int i;
void encrypt(int i, void (*p)(int *));

void request(int *r) {
    if (i == 0) encrypt(input[i++], request);
    else if (0 < i && i < N) {
        result[i - 1] = *r;
        encrypt(input[i++], request);
    } else result[i - 1] = *r;
}

Fig. 21. Client with Multiple Encryption Requests

For end-to-end direct refinement, we need to absorb Lemma 6.3 into Lemma 6.4. The following theorem is easily derived by applying Lemmas 5.5, 5.6 and 5.7.

**Lemma 6.5.** $C \equiv \mathsf{ro} \cdot \mathsf{wt} \cdot \mathsf{c_{injp}} \cdot C$.

The final end-to-end simulation is immediate by vertically composing Lemma 6.3, Lemma 6.4 and refining the simulation convention using Lemma 6.5.

**Theorem 6.6.** $L_{CS} \leq_{C} [\text{client.s + server_opt.s}]$.

### 6.3 Verification of the Mutually Recursive Client and Server

We introduce a variant of the running example with mutual recursion in Fig. 21. The server remains the same while the client is changed.

request itself is passed as a callback function to encrypt, resulting in recursive calls to encrypt for encrypting and storing an array of values. To perform the same end-to-end verification for this example, we only need to define a new top-level specification $L'_{CS}$ and prove $L'_{CS} \leq_{\mathsf{ro}\cdot\mathsf{wt}\cdot\mathsf{c_{injp}}} [\text{client.c}] \oplus L_S$. Other proofs are either unchanged (e.g., the refinement of the server) or can be derived from Theorem 5.9, full compositionality and adequacy. The complete proofs can be found in the technical report [Zhang et al. 2023b].

### 7 Generality and Limitations of Our Approach

We explain how our approach may be generalized to support other memory models, compilers and optimizations for first-order languages. We also discuss the limitations of our approach.

#### 7.1 Supporting Different Memory Models and Compilers

At a high level, injp is simply a general and transitive relation on evolving functional memory invariants (represented as injections) enhanced with memory protection to guard against modification to private memory by external calls. Many other first-order memory models can be viewed as employing either a richer or a simplified version of injection as memory invariants and equipped with a similar notion of memory protection. For example, the memory model of CompCertS [Besson et al. 2015] extends injections to map symbolic values. The memory refinements in the CH$_2$O memory model [Krebbers 2016] function like injections except that pointer offsets are represented as abstract paths pointing into aggregated data structures. The memory model defined by Kang et al. [2015] explicitly divides a memory state into public and private memory. Its memory invariant is an equivalence relation between public source and target memory which is essentially an identity injection. Therefore, uniform KMRs may be defined for those memory models as variants of injp.

To prove the transitivity of these KMRs, the key is the construction of interpolating memory states after external calls as described in §4.2. This construction is based on the following general ideas: 1) as KMRs are transitively composed, more memory gets protected, 2) the private memory should be identical to the initial memory, and 3) the public memory should be projected from the
updated source memory via memory invariants. As we can see, these ideas are applicable to any memory model with functional memory invariants and a notion of private memory. Therefore, our approach should work for the aforementioned memory models and compilers based on them.

7.2 Supporting Additional Optimization Passes

Given any new optimization pass whose additional rely-guarantee condition can be represented as a semantics invariant $I$, we may piggyback $I$ onto an enriched $\text{injp}$ to achieve direct refinement. To see that, note that any $I$ consists of two parts: a condition for initial queries (e.g., $\text{ro-valid}$ in $\text{ro}$) and a condition for replies (e.g., $\text{mem-acc}$ in $\text{ro}$). By extending $\text{injp}$ to include the latter (just like that Definition 3.3 includes $\text{mem-acc}$), if the enriched $\text{injp}$ is still transitive, then we can easily prove the following proposition which is generalized from Lemma 5.6.

**Proposition 7.1.** For any $I : C \Leftrightarrow C$ and enriched $\text{injp}$, $I \cdot c_{\text{injp}} \equiv I \cdot c_{\text{injp}} \cdot I \cdot c_{\text{injp}}$.

The proof follows exactly the steps for proving Lemma 5.6. It is based on the two observations we made near the end of §5.2.3, i.e., 1) the properties for initial queries of $I$ hold along with copying of memory states, and 2) the properties for replies trivially hold as they are part of the enriched $\text{injp}$.

7.3 Limitations

We discuss the limitations of our approach and possible solutions. First, it does not yet support behavior refinements for whole programs in CompCert [Leroy 2023]. This is a technical limitation and can be solved by reducing open simulations into closed simulations in CompCert. Second, the proof for $\text{Unusedglob}$ assumes that global symbols for removed definitions are preserved as CompCertO’s simulation framework requires the same set of global symbols throughout compilation. We need to weaken this requirement to enable removal of global symbols by compilation. Third, open simulations assume given any input injection $j$, the execution outputs some injection $j'$ related to $j$ by $\text{injp}$. This may not work for memory models with fixed injection functions [Wang et al. 2022]. A possible solution is to enrich $\text{injp}$ to account for this fixed definition. Finally, given a new optimization, if its rely-guarantee condition cannot be described as a semantic invariant $I$ or if $\text{injp}$ enriched with $I$ becomes intransitive, then direct refinements may not be derivable. In this case, we may need stronger restrictions on this optimization for our approach to work.

8 EVALUATION AND RELATED WORK

Our Coq development took about 7 person-months and 18.3k lines of code (LOC) on top of CompCertO. We added 3.7k LOC to prove the transitivity of $\text{injp}$, 3k LOC to verify the compiler passes as discussed in §5.1, 1.2k LOC for composing simulation conventions as described in the rest of §5 and 7.3k LOC for the Client-Server examples. We also ported CompCertM’s example on mutually recursive summation [Song et al. 2020], which adds 3.1k LOC [Zhang et al. 2023b]. For now, the cost of examples is relatively high. However, we observe that a lot of low-level proofs such as pointer arithmetic can be automated by proof scripts, many proofs with predictable patterns can be directly derived from the program structures, and a lot of duplicated lemmas in the examples can be eliminated. We will carry out those exercises in the future which should simplify the proofs significantly. Below we compare our work with other frameworks for VCC and program verification.

8.1 Verified Compositional Compilation for First-Order Languages

In this work, we are concerned with VCC of first-order imperative programs with global memory states and support of pointers. A majority of the work in this setting is based on CompCert. We compare them from the perspectives listed in the first column of Table 2. An answer that is not a simple “Yes” or “No” denotes that special constraints are enforced to support the given feature.
Table 2. Comparison between Work on VCC Based on CompCert

<table>
<thead>
<tr>
<th></th>
<th>CompComp</th>
<th>CompCertM</th>
<th>CompCertO</th>
<th>CompCertX</th>
<th>This Work</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct Refinement</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Vertical Composition</td>
<td>Yes</td>
<td>RUSC</td>
<td>Trivial</td>
<td>CAL</td>
<td>Yes</td>
</tr>
<tr>
<td>Horizontal Composition</td>
<td>Yes</td>
<td>RUSC</td>
<td>Yes</td>
<td>CAL</td>
<td>Yes</td>
</tr>
<tr>
<td>Adequacy</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>End-to-end Verification</td>
<td>No</td>
<td>Yes</td>
<td>Unknown</td>
<td>CAL</td>
<td>Yes</td>
</tr>
<tr>
<td>Free-form Heterogeneity</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Behavior Refinement</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

*Compositional CompCert.* CompCert supports VCC based on interaction semantics which is a specialized version of open semantics with C interfaces [Stewart et al. 2015]. We have already talked about its merits and limitations in §1.2. It is interesting to note that CompCert can also be obtained based on our approach by adopting $c_{injp}$ for every compiler pass and exploiting the transitivity of $c_{injp}$, which does not require the instrumentation of semantics in CompCert.

*CompCertM.* CompCertM supports adequacy and end-to-end verification of mixed C and assembly programs. A distinguishing feature of CompCertM is Refinement Under Self-related Contexts or RUSC [Song et al. 2020]. A RUSC relation is a fixed collection of simulation relations. By exploiting contexts that are self-relating under all of these simulation relations, horizontal and vertical compositionality are achieved. However, refinements based on RUSC relations can be difficult to use as they are not extensional. For example, the complete open refinement relation $\leq_{R_1+...+R_9}$ in CompCertM carries 9 RUSC relations $R_1, \ldots, R_9$ (6 for compiler passes and 3 for source-level verification). To establish the refinement between $a_1.s$ and its specification $L_S$, one needs to prove $L_S$ are self-simulating over all 9 simulation relations. This can quickly get out of hand as more modules and more compiler passes are introduced. By contrast, we only need to prove direct refinement for once and the refinement is open to further horizontal or vertical composition. On the other hand, CompCertM supports behavior refinement of closed programs which we do not yet (See §7.3).

*CompCertO.* Vertical composition is a trivial pairing of simulations in CompCertO, which exposes internal compilation steps. CompCertO tries to alleviate this problem via ad-hoc refinement of simulation conventions. The resulting top-level convention is $\mathbb{C}_{\text{Comp}} = R^* \cdot \text{wt} \cdot \text{CL} \cdot \text{LM} \cdot \text{MA} \cdot \text{asm}_{\text{vainj}}$ where $R = c_{injp} + c_{inj} + c_{ext} + c_{vainj} + c_{vaext}$ is a sum of conventions parameterized over KMRs. In particular, $c_{vaext}$ is an ad-hoc combination of KMR and internal invariants for optimizations. $R^*$ means that $R$ may be repeated for an arbitrary number of times. Since the top-level summation of KMRs is similar to that in CompCertM, we need to go through a reasoning process similar to CompCertM, only more complicated because of the need to reason about internal invariants of optimizations in $c_{vaext}$ and indefinitely repeated combination of all the KMRs by $R^*$. Therefore, it is unknown if the correctness theorem of CompCertO suffices for end-to-end program verification.

*CompCertX.* CompCertX [Gu et al. 2015; Wang et al. 2019] realizes a weaker form of VCC that only allows assembly contexts to invoke C programs, but not the other way around. Therefore, it does not support horizontal composition of modules with mutual recursions. The compositionality and program verification are delegated to Certified Abstraction Layers (CAL) [Gu et al. 2015, 2018]. Furthermore, CompCertX does not support stack-allocated data (e.g., our server example). However, its top-level semantic interface is similar to our interface, albeit not carrying a symmetric rely-guarantee condition. This indicates that our work is a natural evolution of CompCertX.
VCC for Concurrent Programs. VCC for concurrent programs needs to deal with multiple threads and their linking. CASCompCert is an extension of CompComp that supports compositional compilation of concurrency with no (or benign) data races [Jiang et al. 2019]. To make CompComp’s approach to VCC work in a concurrent setting, CASCompCert imposes some restrictions including not supporting stack-allocated data and allowing only nondeterminism in scheduling threads. A recent advancement based on CASCompCert is about verifying concurrent programs [Zha et al. 2022] running on weak memory models using the promising semantics [Kang et al. 2017; Lee et al. 2020]. We believe the ideas in CASCompCert are complementary to this work and can be combined with our approach to achieve VCC for concurrency with cleaner interface and less restrictions.

8.2 Verified Compositional Compilation for Higher-Order Languages

Another class of work on VCC focuses on compilation of higher-order languages. In this setting, the main difficulty comes from complex language features together with higher-order states. A prominent example is the Pilsner compiler [Neis et al. 2015] that compiles a higher-order language into some form of assembly programs. The technique Pilsner adopts is called parametric simulations that evolves from earlier work on reasoning about program equivalence via bisimulation [Hur et al. 2012a]. Another line of work is multi-language semantics [Patterson and Ahmed 2019; Patterson et al. 2017; Perconti and Ahmed 2014; Scherer et al. 2018] where a language combining all source, intermediate and target languages is used to formalize semantics. Compiler correctness is stated as contextual equivalence or logical relations. It seems that our techniques are not directly applicable to those work because relations on higher-order states cannot deterministically fix the interpolating states. A possible solution is to divide the higher-order memory into a first-order and a higher-order part such that the former does not contain pointers to the latter (forming a closure). By encapsulating higher-order programs inside first-order states, we may be able to apply our approach.

The high-level ideas for constructing interpolating states for proving transitivity of injp can also be found in some of the work on program equivalence [Ahmed 2006; Hur et al. 2012b]. To the best of our knowledge, our approach is the first concrete implementation of these ideas that works for a realistic optimizing compiler for imperative languages with non-trivial memory models.

8.3 Frameworks for Compositional Program Verification

Researchers have proposed frameworks for compositional program verification based on novel semantics, refinements and separation logics [Chappe et al. 2023; Gu et al. 2015, 2018; He et al. 2021; Sammler et al. 2023; Song et al. 2023; Xia et al. 2019]. These frameworks aim at broader program verification and may be combined with our approach to generate more flexible end-to-end verification techniques. For example, to support more flexible certified abstraction layers, we may combine our approach with data abstraction in CAL and extend horizontal linking to work with abstraction layers. More details can be found in the technical report [Zhang et al. 2023b].

9 CONCLUSION AND FUTURE WORK

We have proposed an approach to compositional compiler correctness for first-order languages via direct refinements between source and target semantics at their native interfaces, which overcomes the limitations of the existing approaches on compositionality, adequacy and other important criteria for VCC. In the future, we plan to support behavior (trace) refinement for closed programs by reducing our open simulation into the whole-program correctness theorem for the original CompCert. We also plan to combine our work with refinement-based program verification like certified abstraction layers to support more substantial applications. Another research direction is to apply our approach to different memory models and compilers for first-order and higher-order languages, which will better test the limit of our approach and the usefulness of our discoveries.
DATA-AVAILABILITY STATEMENT

The Coq artifact containing the formal developments described in this paper is available on Zenodo [Zhang et al. 2023a]. A technical report of this work is available on arXiv [Zhang et al. 2023b].

ACKNOWLEDGMENTS

We would like to thank our shepherd Yannick Zakowski and the anonymous referees for their helpful feedback which improved this paper significantly. This work is supported in part by the National Natural Science Foundation of China (NSFC) under Grant No. 62002217 and 62372290, and by the Natural Science Foundation of the United States (NSF) under Grant No. 1763399, 2019285, and 2313433. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

REFERENCES


