SECOMP: Formally Secure Compilation of Compartmentalized C Programs

Jérémy Thibault
Max Planck Institute for Security and Privacy (MPI-SP)
Bochum, Germany
jeremy.thibault@mpi-sp.org

Roberto Blanco
Max Planck Institute for Security and Privacy (MPI-SP)
Bochum, Germany
roberto.blanco@mpi-sp.org

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

Sven Argo
Ruhr University Bochum
Bochum, Germany
sven.argo@rub.de

Arthur Azevedo de Amorim
Rochester Institute of Technology
Rochester, NY, USA
arthur.aa@gmail.com

Aïna Linn Georges
Max Planck Institute for Software Systems (MPI-SWS)
Saarbrücken, Germany
algeorges@mpi-sws.org

Cătălin Hriţcu
Max Planck Institute for Security and Privacy (MPI-SP)
Bochum, Germany
catalin.hritcu@mpi-sp.org

Andrew Tolmach
Portland State University
Portland, OR, USA
tolmach@pdx.edu

Abstract

Undefined behavior in C often causes devastating security vulnerabilities. One practical mitigation is compartmentalization, which allows developers to structure large programs into mutually distrustful compartments with clearly specified privileges and interactions. In this paper we introduce SECOMP, a compiler for compartmentalized C code that comes with machine-checked proofs guaranteeing that the scope of undefined behavior is restricted to the compartments that encounter it and become dynamically compromised. These guarantees are formalized as the preservation of safety properties against adversarial contexts, a secure compilation criterion similar to full abstraction, and this is the first time such a strong criterion is proven for a mainstream programming language. To achieve this we extend the languages of the CompCert verified C compiler with isolated compartments that can only interact via procedure calls and returns, as specified by cross-compartment interfaces. We adapt the passes and optimizations of CompCert as well as their correctness proofs to this compartment-aware setting. We then use compiler correctness as an ingredient in a larger secure compilation proof that involves several proof engineering novelties, needed to scale formally secure compilation up to a C compiler.

CCS Concepts

• **Security and privacy → Logic and verification**; **Formal methods and theory of security**; • **Software and its engineering →** *Compilers*; *Modules / packages*.

CCS '24, October 14–18, 2024, Salt Lake City, UT, USA

ACM ISBN 979-8-4007-0636-3/24/10
https://doi.org/10.1145/3658644.3670288

Keywords

secure compilation; compartmentalization; undefined behavior; dynamic compromise; machine-checked proofs; Coq; CompCert

ACM Reference Format:
Jérémy Thibault, Roberto Blanco, Dongjae Lee, Sven Argo, Arthur Azevedo de Amorim, Aïna Linn Georges, Cătălin Hriţcu, and Andrew Tolmach. 2024. SECOMP: Formally Secure Compilation of Compartmentalized C Programs. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS '24), October 14–18, 2024, Salt Lake City, UT, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3658644.3670288

1 Introduction

Undefined behavior is endemic in the C language: buffer overflows, use after frees, double frees, signed integer overflows, invalid type casts, various concurrency bugs, etc., cause mainstream C compilers to produce code that can behave completely arbitrarily. This leads to devastating security vulnerabilities that are often remotely exploitable, and both Microsoft and Chrome report that around 70% of their high severity security bugs are caused by undefined behavior due to memory safety violations alone [42, 61].

A strong practical mitigation against such vulnerabilities is *compartmentalization* [16, 29, 35], which allows developers to structure large programs into mutually distrustful compartments that have clearly specified privileges and that can only interact via well-defined interfaces. This way, the compromise of some compartments has a limited impact on the security of the whole program. This intuitive increase in security has made compartmentalization and the compartment isolation technologies used to enforce it become widely deployed in practice; e.g., all major web browsers today use both process-level privilege separation [16, 29, 35] to isolate tabs and plugins [53], and software fault isolation (SFI) [44, 60, 66, 71, 73] to sandbox WebAssembly modules [30].

In this paper, we investigate how to provide strong *formal guarantees* for compartmentalized C source code by making the C compiler aware of compartments. We follow Abate *et al.* [4], who argue that a compartment-aware compiler for an unsafe language can restrict the scope of undefined behavior both (a) spatially to just the compartments that encounter it [33], and (b) temporally by still providing protection to each compartment up to the point in time when it encounters undefined behavior. Abate *et al.* formalize this intuition as a variant of a general *secure compilation* criterion called *Robust Safety Preservation (RSP)* [5, 6, 51]. Their RSP variant ensures that any low-level attack against a compiled program's safety properties mounted by compartments dynamically compromised by undefined behavior could also have been mounted at the source level by arbitrary compartments with the same interface and privileges, while staying in the secure fragment of the source semantics, without undefined behavior. This strong guarantee allows source-level security reasoning about compartmentalized programs that have undefined behavior, and thus for which the C standard and the usual C compilers would provide no guarantees whatsoever.

Such strong formal guarantees are, however, notoriously challenging to achieve in practice and to prove mathematically. RSP [5, 6, 51] belongs to the same class of secure compilation criteria as full abstraction [3, 49], for which simple and intuitive but wrong conjectures have sometimes survived for decades [20], and for which careful paper proofs can take hundreds of pages even for very simple languages and compilers [24, 33]. Such proofs are generally so challenging that no compiler for a mainstream programming language that is guaranteed to achieve any such secure compilation criterion has ever been built. Moreover, such secure compilation proofs are at the moment often only done on paper [6, 7, 18, 24, 33, 48–51], even though at the scale of a realistic compiler, paper proofs would be impossible to trust, construct, and maintain. All this stands in stark contrast to *compiler correctness*: CompCert [40]—a realistic C compiler that comes with a machine-checked correctness proof in the Coq proof assistant—has already existed for more than a decade and is used in practice in highly safety-critical applications [38].

In this paper we take an important step towards bridging this gap by devising SECOMP, a formally secure compiler for compartmentalized C code. To do this, we extend the CompCert compiler and its correctness proof to handle isolated compartments that interact only via procedure calls and returns. Although compiler correctness by itself is definitely not enough to prove secure compilation, since it gives up on programs with undefined behavior, we use it as one key ingredient for such a proof. For this we adopt the high-level proof structure proposed by Abate *et al.* [4], who showed how proving their RSP variant can be reduced to showing compiler correctness together with three security-related properties: *back-translation*, *recomposition*, and *blame* (explained in the next section, §2).Proving these properties at scale and achieving formally secure compilation for a compiler for a mainstream programming language were open research challenges, which we solve in this work by bringing the following **novel contributions**:

- We devise the SECOMP compiler for compartmentalized C programs to RISC-V assembly by extending the syntax and semantics of all the 10 languages of CompCert with the abstraction of isolated compartments that can only interact via procedure calls, as specified by cross-compartment interfaces. For CompCert's RISC-V assembly we propose an enforcement-independent characterization of C compartments that relies on a new shadow stack to ensure the well-bracketedness of cross-compartment control flow. We adapt all 19 passes and all optimizations of CompCert to this extension, except cross-compartment inlining and tail-calls, which we disallow.
- In addition to passing scalar values to each other, our compartments can also perform input and output (IO), which was not the case in the very simple languages studied by Abate *et al.* [4]. Our IO model allows pointers to global buffers of scalars to be passed to the system calls implementing IO and also allows these buffers to be changed nondeterministically by these system calls, which goes beyond what was previously possible in CompCert's IO model.
- We extend CompCert's large-scale compiler correctness proof to account for these changes so that we can use it to show secure compilation. Our extension of the correctness proof is elegant and relatively small, even though two of our changes to the semantics of the CompCert languages are substantial: (1) we extend the CompCert memory model with compartments, and (2) we extend the CompCert trace model with events recording cross-compartment calls and returns, as needed for the secure compilation proof.
- We develop a secure compilation proof for SECOMP in Coq, from Clight, the first intermediate language of CompCert and featuring a determinate [25] semantics (as opposed to CompCert C), down to our extension of CompCert's RISC-V assembly. This proof shows the RSP variant of Abate *et al.* [4], capturing the secure compilation of mutually distrustful C compartments that can be dynamically compromised by undefined behavior. We are the first to prove such a strong secure compilation criterion for a mainstream programming language, which makes this a milestone for secure compilation.
- To scale up the secure compilation proofs to SECOMP, we introduce several proof engineering novelties: (1) Because SECOMP uses the memory model of CompCert [41] (extended with compartments), the novel simulation invariants we devise to prove security have to make use of the sophisticated *memory injections* of CompCert [41], which provide a fine-grained characterization of the way memory is transformed during compilation. (2) For back-translation, because system calls may read global buffers, we extend traces with *informative events*, which record *memory deltas*—i.e., changes to global buffers happening during silent steps—and we use those to establish memory injections to prove correctness of the system calls the back-translation generates. (3) For recomposition, we propose a more principled way of proving the required three-way simulation by defining 8 simulation diagrams and providing a general proof that together they imply recomposition. Despite the realistic RISC-V instruction set, we use these diagrams to provide a compact proof of recomposition showing that our RISC-V assembly semantics securely characterizes the compartment abstraction. We discovered that, for recomposition to hold, the stack-spilled call arguments spilled must be protected, so a malicious caller cannot exploit callbacks to covertly change arguments of a previous call.
- The SECOMP secure compilation proofs end at our extension of CompCert's RISC-V assembly, which is the language where CompCert's compiler correctness proofs also end, and whose semantics still maintains CompCert's block-based memory model. As mentioned above, to this language's semantics we added the extra abstraction of isolated compartments, which formally defines *what*

compartment isolation enforcement should do, but which leaves the *how* to lower-level enforcement mechanisms working with a more concrete view of memory as an array of bytes [67, 68] and potentially making use of hardware security features. We additionally show that the compartment isolation abstraction can be enforced at a lower level by designing and prototyping an unverified backend targeting a variant of the CHERI capability machine [69]. For this we extend a recently proposed efficient calling convention enforcing stack safety [28] to our setting of mutually distrustful compartments by introducing capability-protected wrappers to clear registers on calls and returns and to prevent capabilities from being passed between potentially compromised compartments. Various other enforcement mechanisms should be possible though, including SFI [60, 66, 71] and tagged architectures [10, 22], as shown in a simpler setting by Abate *et al.* [4]. At the moment all these lower-level backends are, however, unverified, and extending the secure compilation proofs to cover them is a formidable research challenge that we leave as future work (§11).

Long version. A long version of this paper with additional details is available at http://arxiv.org/abs/2401.16277

Artifact. The SECOMP formally secure compiler is available at https://github.com/secure-compilation/SECOMP and as a permanently archived artifact at https://doi.org/10.5281/zenodo.11007679. SECOMP adds ~38k LoC on top of CompCert, mostly in proofs. In more detail, our extensions to CompCert and its correctness comprise ~5k LoC of specs and ~7k of proofs, an increase of 6.2% and 14.9% respectively. In addition, back-translation involves ~5k LoC of specs and ~6k of proofs; recomposition ~1k LoC of specs and ~8k of proofs; and blame ~1k LoC of specs and ~4k of proofs. Our development is based on CompCert version 3.12 and our experience shows that tracking mainline development is relatively simple, as our changes are orthogonal to the usual CompCert development.

The machine-checked proofs are generally complete and include no further axioms beyond those already existing in CompCert [43], or small adaptions thereof to account for the addition of compartments to the compiler. One exception to this is an axiom assuming that CompCert can successfully compile the results of our back-translation, which would be very tedious to prove in general, but which we have instead thoroughly tested (§7). The other current gap in our Coq formalization is about connecting compiler correctness, back-translation, blame, and recomposition into a single mechanized secure compilation result (Theorem 8.1); instead at the moment the top-level proof and all the steps are complete, but back-translation, recomposition, and blame are still on separate branches that we are currently in the process of merging into the main branch. This is further documented in the `README.md` file.

Outline. We first review the work on which we directly build (§2) and present the key ideas of our work (§3). Then we explain how we extended CompCert and its correctness proof (§4). The following two sections detail the most interesting parts of our secure compilation proof: back-translation (§5) and recomposition (§6). Then we present the assumption that the result of back-translation compiles and how we thoroughly tested it (§7). We put these together into our secure compilation theorem (§8) and then present our lower-level, unverified capability backend (§9). We discuss related work (§10) before concluding with future work (§11). Finally, the appendices include details that we had to cut for space.

2 Background

In this section we briefly review the RSC^{DC}_{MD} secure compilation criterion of Abate *et al.* [4] as well as their high-level proof structure for this criterion, since we make use of both in this paper.

But first, we warm up by reviewing the compiler correctness properties this proof structure makes use of. For this we assume that both the source language (for our security proof this is Clight) and the target language (RISC-V assembly) are given trace-producing semantics. CompCert traces are composed of events recording the calls the whole program makes to system calls performing IO and the results they return. A special event $\mathrm{Undef}(k)$ terminates the trace if undefined behavior is encountered by compartment k (where k is something we added, and which we omit where it is irrelevant). We further extend these traces to cross-compartment calls and returns (Figure 3 in §4). Because the criterion of Abate *et al.* [4] focuses on safety properties [39] we only consider finite prefixes of traces. We write $\mathbf{W_S} \rightsquigarrow m$ when the whole source program $\mathbf{W_S}$ can produce the finite trace prefix m; and analogously $W_T \rightsquigarrow m$ when the whole target program W_T can produce m. The compiler correctness guarantee of CompCert states that if a compiled whole program can produce a trace prefix m (i.e., $\mathbf{W_S}{\downarrow} \rightsquigarrow m$) then the original source program can produce a related trace $m' \preceq m$ (i.e., $\mathbf{W_S} \rightsquigarrow m'$), where the relation $m' \preceq m$ is defined as $m' = m$ when $\mathrm{Undef} \notin m'$, and as $m'_0 \cdot m_1 = m$ when $m' = m'_0 \cdot \mathrm{Undef}(k)$ for some k. Here "$\cdot$" denotes concatenation and m_1 is a completely arbitrary trace suffix that the correctly compiled program is allowed to produce when the source program encounters undefined behavior, which can lead to security vulnerabilities.

Instead of compiling only whole programs though, we assume separate compilation—as proposed by Kang *et al.* [34] and implemented in CompCert since version 2.7—and separately compile a source program $\mathbf{P}$ and a context $\mathbf{C}$, which are intuitively both formed of linked compartments, and which can be linked together to produce a whole program both before compilation using source linking ($\bowtie$), and after compilation using target linking ($\bowtie$). Using these concepts we can define the correctness of a compiler $\downarrow$ like CompCert or our variant SECOMP as follows:

Definition 1 (Backward Compiler Correctness (BCC))**.**

$$\forall \mathbf{C}\ \mathbf{P}\ m.\ (\mathbf{C}{\downarrow} \bowtie \mathbf{P}{\downarrow}) \rightsquigarrow m \Rightarrow \exists m'.\ (\mathbf{C} \bowtie \mathbf{P}) \rightsquigarrow m' \wedge m' \preceq m$$

In CompCert this backward compiler correctness (BCC) definition is proved by forward simulation [40], so one also obtains a forward compiler correctness (FCC) result, that the RSC^{DC}_{MD} proof structure of Abate *et al.* [4] also makes use of. Here, instead of obtaining a related trace prefix they instead assume that the prefix one starts from in the source does not end with undefined behavior:

Definition 2 (Forward Compiler Correctness (FCC) [4])**.**

$$\forall \mathbf{C}\ \mathbf{P}.\ \forall m \not\ni \mathrm{Undef}.\ (\mathbf{C} \bowtie \mathbf{P}) \rightsquigarrow m \Rightarrow (\mathbf{C}{\downarrow} \bowtie \mathbf{P}{\downarrow}) \rightsquigarrow m$$

All variants of C compiler correctness, including the two above, completely give up on the whole program after it encounters undefined behavior. To mitigate this issue, Abate *et al.* [4] propose a secure compilation notion that restricts the scope of undefined behavior to the compartments that encounter it. Such compromised compartments can only influence other compartments via controlled interactions respecting their interfaces and the other abstractions of the source language (e.g., the stack discipline on

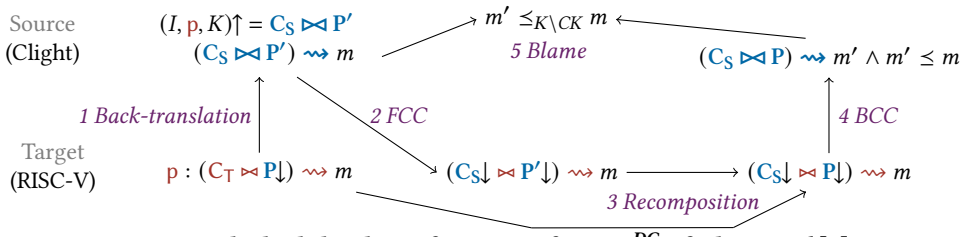


Figure 1: The high-level proof structure for RSC^{DC}_{MD} of Abate *et al.*[4]

calls and returns). Moreover, to model dynamic compromise the scope of undefined behavior is also restricted temporally, by still providing protection to each compartment up to the point in time when it encounters undefined behavior.

Abate *et al.* [4] formalize this intuition as an iterative game in which at each step some (initially empty) set of compartments CK is already compromised and tries to attack the remaining uncompromised compartments $K\setminus CK$, for some set of compartment identifiers K defined in the original compartmentalized program with global interface I, capturing all procedure imports and exports. In each step, the uncompromised compartments are linked together into a source program $\mathbf{P}$ with interface $\lfloor I \rfloor_{K\setminus CK}$, and then $\mathbf{P}$ is compiled and linked with a target context C_T, which puts together the compromised compartments and which has interface $\lfloor I \rfloor_{CK}$. The guarantee obtained at each step in this game is formalized as a property they call RSC^{DC}_{MD}, which is defined then explained below:[1]

Definition 3. A compilation chain satisfies *Robustly Safe Compilation with Dynamic Compromise and Mutual Distrust* (RSC^{DC}_{MD}) if there exists a back-translation function $\uparrow$ that takes interface I, a target execution p producing a trace prefix m, and a compartment identifier k, and generates a source compartment such that

$$\forall K{\subseteq}\mathit{CompIds}.\ \forall I{:}\mathit{Interface}(K).\ \forall CK{\subseteq}K.\ \forall C_T{:}\lfloor I \rfloor_{CK}.\ \forall \mathbf{P}{:}\lfloor I \rfloor_{K\setminus CK}.$$
$$\forall m \not\in \mathrm{Undef}.\forall p : (C_T{\bowtie}\mathbf{P}\downarrow) \rightsquigarrow m.$$
$$\exists C_S{:}\lfloor I \rfloor_{CK}.\ C_S{=}\bowtie_{k\in CK}(I, p, k)\uparrow \wedge \exists m'.(C_S{\bowtie}\mathbf{P}){\rightsquigarrow}m' \wedge m' \preceq_{K\setminus CK} m$$

The premise on the first two lines states that the compound program $C_T \bowtie \mathbf{P}\downarrow$ has an execution p in the target language producing a trace prefix m, which does not end with an undefined behavior event (i.e., for a trace $m \cdot \mathrm{Undef}(k')$ one looks only at the prefix m). The conclusion makes a step towards providing an explanation for m with respect to the source language semantics. For this it calls the back-translation function $\uparrow$ on each of the compromised compartments $k{\in}CK$ and it links together the generated source compartments to obtain a source context C_S with interface $\lfloor I \rfloor_{CK}$. The RSC^{DC}_{MD} property says that the obtained context C_S linked with the original source program $\mathbf{P}$ can produce a trace m' that is related to m by the formula $m' \preceq_{K\setminus CK} m$. This is a variant of the $\preceq$ relation from BCC that ensures that $m' = m$ when $\mathrm{Undef} \notin m'$, and that $m'_0 \cdot m_1 = m$ when $m' = m'_0 \cdot \mathrm{Undef}(k)$ for some *uncompromised* compartment $k{\in}K\setminus CK$. Intuitively either the whole target prefix m can be explained by an execution in the source language, in which case we are done; or the compromised compartments have found a way to use the interface in the source language to trigger an undefined behavior in one of the (so far) uncompromised compartments $k{\in}K\setminus CK$. In this second case, Abate *et al.* [4] will apply RSC^{DC}_{MD} again to an extended set of compromised compartments $CK{\cup}\{k\}$. Because the semantics is determinate [25], with each iterative application of RSC^{DC}_{MD} the execution is "rewound" along the original trace prefix m and longer and longer prefixes of m are explained in the source, until the whole m is explained in terms of the source semantics and a sequence of dynamic compartment compromises.

For proving secure compilation this iterative aspect is less interesting though, and it basically suffices to show RSC^{DC}_{MD} [4]. For this, Abate *et al.* [4] propose the high-level proof structure from Figure 1 that involves compiler correctness (the FCC and BCC properties above) and three additional security-related properties: back-translation, recomposition, and blame. The high-level proof starts by back-translating a global interface I and a target execution p producing a trace prefix m repeatedly to generate each of the compartments $k \in K$ of a whole source program producing m.

Definition 4 (Back-translation)**.** There exists a function $\uparrow$ s.t.

$$\forall K.\ \forall I.\ \forall W_T{:}I.\ \forall CK{\subseteq}K.\ \forall m \not\in \mathrm{Undef}.\forall p : W_T \rightsquigarrow m.$$
$$\exists C_S{:}\lfloor I \rfloor_{CK}.\ \exists \mathbf{P}'{:}\lfloor I \rfloor_{K\setminus CK}.\ C_S{\bowtie}\mathbf{P}'{=}\bowtie_{k\in K}(I, p, k)\uparrow \wedge (C_S{\bowtie}\mathbf{P}'){\rightsquigarrow}m$$

Using the back-translation function $\uparrow$ to generate a *whole* source program $C_S \bowtie \mathbf{P}'$ not only allows the conclusion of Def. 4 to be stated in terms of the usual operational semantics of whole programs ($(C_S \bowtie \mathbf{P}'){\rightsquigarrow}m$), but also allows Abate *et al.* [4] to compile this whole program and make use of FCC in step 2 from Figure 1 to obtain that $(C_S\downarrow \bowtie \mathbf{P}'\downarrow) \rightsquigarrow m$. Then, in step 3 they recompose the compartments from this execution with the ones from original execution $(C_T \bowtie \mathbf{P}\downarrow) \rightsquigarrow m$ to obtain the execution $(C_S\downarrow \bowtie \mathbf{P}\downarrow) \rightsquigarrow m$.

Definition 5 (Recomposition)**.** $\forall K.\ \forall I.\ \forall CK{\subseteq}K.$

$$\forall C_T, C'_T : \lfloor I \rfloor_{CK}.\ \forall P_T, P'_T : \lfloor I \rfloor_{K\setminus CK}.\ \forall m.$$
$$(C_T \bowtie P_T) \rightsquigarrow m \wedge (C'_T \bowtie P'_T) \rightsquigarrow m \Rightarrow (C'_T \bowtie P_T) \rightsquigarrow m$$

Step 4 uses BCC to turn target execution $(C_S\downarrow \bowtie \mathbf{P}\downarrow) \rightsquigarrow m$ back into source execution $(C_S \bowtie \mathbf{P}) \rightsquigarrow m'$, where the relation $m' \preceq m$ accounts for the possibility of undefined behavior in $C_S \bowtie \mathbf{P}$. The context C_S is, however, generated by the back-translation and has no undefined behavior along the trace m, which one shows in step 5 of the proof. So if there is an undefined behavior in m' then this can only be blamed on an as-yet-uncompromised compartment $k \in K\setminus CK$, as required by the conclusion of RSC^{DC}_{MD} from Def. 3.

Definition 6 (Blame)**.** $\forall K.\ \forall I.\ \forall CK{\subseteq}K.\forall C_S{:}\lfloor I \rfloor_{CK}.\ \forall \mathbf{P}, \mathbf{P}'{:}\lfloor I \rfloor_{K\setminus CK}.$

$$\forall m.(C_S \bowtie \mathbf{P}') \rightsquigarrow m \wedge (C_S \bowtie \mathbf{P}) \rightsquigarrow m' \wedge m' \preceq m \Rightarrow m' \preceq_{K\setminus CK} m.$$

[1] First time readers can also skim the remaining technical definitions in this section and focus on the intuitive explanations and the graphical representation from Figure 1. The paper of Abate *et al.* [4] provides a gentler introduction to this proof structure.

3 Key ideas

3.1 Compartment model

Compartmentalization [16, 29, 35] allows developers to structure large programs into mutually distrustful compartments that have limited privileges, are isolated from each other, and can only interact in a controlled way. In this work, we adopt a model that statically partitions C programs into compartments. Every C definition of a procedure or a global variable belongs to a single compartment.

Any block of memory belongs to the compartment that allocated it and compartments do not share memory: each block can only be accessed by the code of the compartment it belongs to. Instead, all interactions between compartments must happen via *cross-compartment calls and returns* that respect the *interfaces* provided by the programmers: each compartment `C` comes with a set of exported procedure declarations (i.e., which procedures it makes available to other compartments), written `C.exports`, and a set of imported procedure declarations (i.e., which procedures it uses from other compartments), written `C.imports`. Compartments must respect these compartment interfaces; otherwise they trigger undefined behavior, whose scope is restricted by our secure compilation property to just the offending compartment. Moreover, in our model compartments can only pass each other scalar values as procedure call arguments and return values. We introduced this last restriction for two reasons: first, a compartment cannot use pointers from another compartment, which defeats the purpose of passing pointers in most cases. More importantly, passing pointers to other compartments would require recording these pointers on the trace, which would significantly complicate back-translation, recomposition, and if also done for pointers to dynamically allocated memory also compiler correctness (see §11).

The compartments also interact with an external environment using *system calls*.[2] These are special, privileged procedures whose semantics is axiomatized in CompCert and that may generate some events. Example of these system calls include volatile memory operations, calls to the heap allocator, or input and output (e.g., reading from the console). The system calls do not belong to a compartment; instead calling them is considered a special kind of internal call that can only change the calling compartment's memory. Following the principle of least privilege, by default compartments do not have access to system calls beyond safe operations like memory allocation and freeing. Instead, the interfaces specify, for each compartment, which system calls they are allowed to use.

3.2 Adding compartments to CompCert

Extension to CompCert's languages. Following the ideas above, we extend all of CompCert's 10 languages, from C to RISC-V assembly, adding syntax describing the compartment breakup and interfaces and semantic checks to ensure all compartments respect these interfaces. As explained above, failing a check triggers undefined behavior for the offending compartment. In particular, we update memory operations to take an additional compartment argument, ensuring compartments cannot access other compartments' memory. Also, at every point where control could pass to another compartment (calls and returns at higher levels, jumps at the lowest level) we add a check that the control transfer respects the interfaces and that compartments only pass each other scalar values. Compilation preserves the program's interfaces and linking two partial programs requires that they have compatible interfaces.

To prove secure compilation (Def. 3 from §2), we also extended the CompCert trace model with two new events: `Event_call` and `Event_return`. These events are generated by cross-compartment calls and returns (or the equivalent jumps in RISC-V assembly), and record enough information on the traces to be able to prove recomposition and back-translation. Since call and return events must not be disturbed by optimization, we disallow cross-compartment tail-call optimization and inlining, as those would substantially change the way compartments interact (e.g., would require merging stack frames belonging to different compartments).

At the RISC-V level, implementing secure compartments required even more care. Without proper protection, adversarial code could make use of the unstructured control-flow inherent to the assembly language to break the compartment abstraction. For instance, an attacker could try to jump to code to which it shouldn't have access. We modified CompCert's RISC-V assembly semantics to prevent this kind of attack, by protecting the compartment abstraction and interfaces. To do so, we observe that calls and returns are only implemented by the compiler using specific instructions—jump-and-link for calls, and indirect jumps for returns—so we forbid all other instructions from changing compartments. Then, when an execution encounters such a jump and attempts to switch to another compartment, we make use of the interfaces and of a newly added shadow stack to decide whether the switch is allowed. If the instruction is a jump-and-link, then the semantics checks whether it is an allowed call according to the interfaces, and then records the return address and the stack pointer on the shadow stack. To decide which indirect jumps to allow to return to a different compartment, we inspect the top of the shadow stack, and make sure that the compartment performing the jump is returning to the right address and has correctly restored the caller's stack pointer. This prevents a malicious compartment from returning using the wrong return address or confusing the caller about its stack. More details about our changes to the semantics are discussed in §4.

Correctness proofs. We updated all of CompCert's 19 passes and the simulation proofs showing FCC (Def. 2) to account for the addition of compartments and the new trace model, which by determinacy implies BCC (Def. 1). The updated proofs mainly rely on the compartment information being correctly preserved by compilation, e.g., procedures do not change compartment, memory blocks that belong to different compartments are not merged, etc.

Adapting CompCert's compiler correctness Coq proof to account for our changes was a substantial amount of work. We wanted to change the proof as little as possible, but since CompCert is a realistic compiler, it was not always obvious from the start how best to do this. Several times, we made design decisions that seemed adequate, but that turned out to be inadequate much later (e.g., choosing at which precise step to insert a given check), when we discovered that they interacted poorly with some particular compilation pass (e.g., intra-compartment inlining or tail-call optimization) or language (e.g., RISC-V assembly). These issues often did not affect the correctness of the compiler, but made the proofs much more difficult,

[2] Our "system calls" correspond in CompCert to the "external" functions that do not get resolved to actual C source code during linking. Such functions are implemented in lower-level, trusted libraries (like libc) and may include operating system calls.

so we had to backtrack and find alternative ways to structure the changes so as to keep the proofs simple.

In the end we found elegant ways of adapting CompCert's compiler correctness Coq proof to account for all the changes above and proved FCC and BCC. Yet, while compiler correctness is an important part of our security proof, it is definitely not sufficient by itself (§1). In the remainder of this section we discuss the other components of the security proof (§2).

3.3 Back-translation from RISC-V to Clight

In our setting, the first step of the proof structure from §2 is to back-translate a finite RISC-V execution prefix into a Clight program that produces the same trace prefix. We show constructively that given a (whole) compartmentalized RISC-V program and a finite trace prefix of that program, there exists a (whole) compartmentalized Clight program with the same interface that can also produce the same trace prefix. Back-translation resembles compilation, but for RSP [5, 6, 51] and variants like the one we consider, the program obtained by back-translation only needs to preserve *one single finite trace prefix*, not every possible execution of the original program.

Based on this observation, prior works [4, 7, 18, 23, 24, 50] use a simple back-translation from a trace prefix to a program. Each procedure of the program consists of a loop over a counter, which records how far the trace has been executed. The body of the loop is a switch over the counter value; the nth case of the switch contains code that will produce the nth event of the trace. Proving such a back-translation correct can usually be done in two steps [4]: first, one proves that all traces generated by a target program satisfy a *well-formedness* condition, and then, that every well-formed trace can be back-translated to a program that produces that same trace. We adapt this back-translation and proof technique to our setting, but to do so, we need to make our events more informative and devise a novel notion of well-formedness of traces made out of these informative events through an intermediate language.

First, the events we use in SECOMP (introduced informally in §3.2 and detailed in Figure 3 in §4) do not contain enough information to directly convert each trace into a Clight statement. In particular, they do not capture all the information necessary to obtain a back-translated program that produces the same events for system calls: if a RISC-V system call produces an event with memory $\mathtt{M}$ and current compartment $\mathtt{C}$, then for the back-translated Clight program to produce the same event with memory $\mathtt{M}'$ and current compartment $\mathtt{C}$, the CompCert-style axiomatization of system calls requires us to prove that $\mathtt{M}$ and $\mathtt{M}'$ are related by some memory injection [41] that is defined at least on the public symbols of $\mathtt{C}$. Put simply, this means that the semantics of system calls is allowed to depend on the content of the calling compartment's global buffers. Since we restrict our semantics to ensure global buffers only contain scalars when calling system calls, this effectively means that $\mathtt{M}$ and $\mathtt{M}'$ must have the same values in $\mathtt{C}$'s global buffers. Yet, the SECOMP events do not include the content of global buffers.

This motivates introducing more *informative events* satisfying two requirements: (1) a RISC-V program always produces a well-formed trace of informative events, and (2) each informative event directly translates into a Clight statement. Using these informative events, we define a novel notion of *well-formed informative trace* and a back-translation from such traces to Clight programs, and we significantly extend the technique of Abate *et al.* [4] to prove the correctness of this back-translation. Informative events augment each system call event (and also cross-compartment call or return events) with a list of the changes to the global buffers since the last informative event. Each of these changes is called a *memory delta*, written δ, and a list of these is written Δ, and is ordered chronologically. Our back-translation uses these deltas to generate Clight code performing the same changes to global buffers before performing the system call.

First, we define a novel notion of well-formedness of a trace of informative events. To do so, we define a new intermediate language[3] with a semantics that characterizes the well-formedness of such traces. Informally, informative traces can be seen as the programs of this language, and the step relation executes these traces by emitting the first informative event of the trace and updating the state according to the event. More precisely, in this language, states s are triples that record the currently executing procedure, a memory, and a cross-compartment call stack; the step relation $s \xrightarrow{\alpha} s'$ relates the states s and s' and produces an informative event α. The rules of the step relation additionally record the conditions necessary for the back-translated code to be proved correct. In particular, if $s \xrightarrow{\alpha} s'$, then applying the deltas Δ from α to the memory $\mathtt{M}$ of s produces the memory $\mathtt{M}'$ of s' (i.e., $\mathtt{mem\_delta\_apply}\ \Delta\ \mathtt{M} = \mathtt{M}'$).

We say that a trace of informative events is *well-formed* when it is produced by the reflexive transitive closure of this step relation. We prove that for any trace prefix m produced by a RISC-V program, there exists a well-formed informative trace $\overline{\alpha}$ that projects to m by removing the additional data recorded in informative events. This allows us to then define a back-translation function on informative traces, and completely forget about the RISC-V semantics when proving the correctness of this back-translation.

This back-translation function operates in the standard way explained above, except that it also uses the memory deltas to generate code that writes the right values to the global buffers before performing a system call. We prove the correctness of the back-translation by induction on the intermediate language execution, using the information provided by the step relation. In order to prove that system calls at the Clight level generate the same events, we maintain an invariant on permissions of global buffers, and show that it can be used together with memory deltas to build a memory injection that allows us to use the axiomatization of system calls from CompCert. We give more details of this in §5.

3.4 Recomposition for RISC-V compartments

Recomposition is essential for the proof structure of §2, as it allows to replace the arbitrary RISC-V context with which we started, with a context that was obtained by back-translation and then compiled. More concretely, starting from two whole target programs $W_1 = C_1 \bowtie P_1$ and $W_2 = C_2 \bowtie P_2$ that can execute to produce the same trace m and that are each split into a *program side* (P_1 respectively P_2) and a *context side* (C_1 respectively C_2), recomposition gives us that the whole program $W_3 = C_1 \bowtie P_2$ produces the same trace m.

The key intuitions behind this kind of proof [4, 6, 23, 24, 33, 48, 50, 51] are as follows: (1) because of determinacy the internal behavior of a compartment only depends on its internal state, and

[3] We see our back-translation function as a compiler from traces to Cminor programs.

on all the information it received from other compartments, or from system calls; (2) our extended trace model is informative enough to capture all information that is exchanged between compartments or obtained from system calls. Because of this, we can relate the execution of W_3 to that of W_1 when executing in a compartment of the context, and to that of W_2 otherwise: the internal state of the compartments of the context part agree in the execution of W_3 and that of W_1, and the same holds for W_3, W_2, and the internal state of the compartments in the program part. This is preserved by silent steps, because they only depend on that same internal state; and because the trace events record every information exchanged, even when switching side, non-silent-step also preserve this information.

Formalizing this idea in order to prove recomposition for our RISC-V semantics extended with compartments is highly complex, as it relies on many low-level details of the RISC-V semantics. For this reason, we propose a generic proof technique that elegantly splits recomposition into several self-contained parts. Our technique introduces eight novel CompCert-like simulation diagrams (described in detail in §6) that provide a structured way to think and reason about recomposition, explicitly separating the definition of invariants, the reasoning about internal steps, and the reasoning about events and cross-compartment communication. Together, the diagrams imply the existence of a novel *three-way recomposition simulation* that itself implies the recomposition theorem (Def. 5).

We define three-way recomposition simulation in the generic setting on labeled transition systems $(S, \rightarrow)$ with *initial* states:

Definition 7 (Three-way recomposition simulation)**.** Given three labeled transition systems $L_1 = (S_1, \rightarrow_1)$, $L_2 = (S_2, \rightarrow_2)$ and $L_3 = (S_3, \rightarrow_3)$, we say there exists a three-way recomposition simulation between L_1, L_2, and L_3 when there exists a relation $\mathcal{R}$ between states of L_1, L_2, and L_3 that satisfies the properties depicted in Figure 2.

For each property from Figure 2, each row represents one of the 3 executions. Arrows represent execution steps, and are annotated with either an event a or silence ε. We denote the reflexive transitive closure of the step relation with $\rightarrow^*$. We use thick, dark purple for the assumptions, and dark green for the conclusions we have to prove (including the existence of states). We denote equality via a double line. We represent the simulation relation as a rectangle around the states it relates. Property (1) states that the simulation relation $\mathcal{R}$ is compatible with initial states. Property (2) states that whenever the first two executions take a step from related states producing the same observable event a, then so can the third one. Properties (3) and (4) state that silent steps in either of the first two executions preserve the simulation relation (and the third execution is allowed to take some silent steps too).

We prove that, given such a three-way simulation between the semantics of programs W_1, W_2, and W_3, then $W_1 \rightsquigarrow m \wedge W_2 \rightsquigarrow m \implies W_3 \rightsquigarrow m$ (which directly implies recomposition). For this it is enough to follow both executions in W_1 and W_2, and to apply the appropriate property, (2), (3), or (4) until all of m is produced.[4]

To define simpler to prove simulation diagrams in §6 we will instantiate the relation $\mathcal{R}$ above with the conjunction of three relations, following ideas of El-Korashy *et al.* [23, 24]: a *strong* relation $\sim$, a *weak* relation $\equiv$, and a *mixed* relation $\mathcal{M}$. Intuitively, the strong relation $\sim$ relates the internal state of the side being executed, the weak relation $\equiv$ relates the internal state of the side not being executed, and the mixed relation $\mathcal{M}$ relates the parts shared between compartments (such as the stack). Given three states s_1, s_2, s_3 that are executing a compartment that's taken from W_1, the strong relation $\equiv$ relates s_1 to s_3, and the weak relation $\sim$ relates s_2 to s_3. Symmetrically, when the current compartment is taken from W_2 then $s_1 \sim s_3$ and $s_2 \equiv s_3$. When switching to a compartment taken from the other side the relations are switched as well.

Compared to prior work [23, 24], a significant challenge in our proofs is that in our realistic setting the three relations above are parameterized by two CompCert *memory injections* [41], one for each of the original executions. Their role is to relate the memory of their respective execution to the memory in the recomposed execution. In particular, these injections do not relate the memory locations of the side that doesn't correspond to their execution, and they are preserved by the simulation steps. Essentially, both weak and strong relations are instantiated so that they relate the memories of the runs according to these injections, and additionally the strong relation also relates the content of the registers. The mixed relation relates the content of the stack and cross-compartment stack, and uses both memory injections.

3.5 Blame for Clight semantics

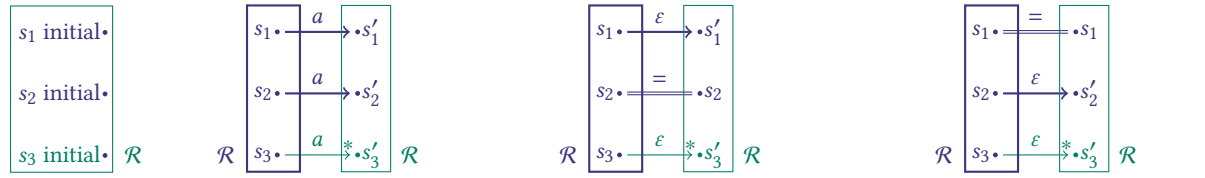
The blame theorem is the final proof step of §2 and shows that the back-translated program $\mathbf{C_S} \bowtie \mathbf{P'}$ is free of undefined behavior along the given trace prefix. From this, it follows that any undefined behavior in $\mathbf{C_S} \bowtie \mathbf{P}$ must come from the original (partial) program $\mathbf{P}$. Blame relates the executions of two whole Clight programs that produce the same trace prefix and share a common set of compartments from $\mathbf{C_S}$—their *context side*—linked with a pair of compatible (same public symbols, imports and exports, etc.), but otherwise arbitrary *program sides* $\mathbf{P}$ and $\mathbf{P'}$, which supply the remaining compartments. Intuitively, because the two executions produce the same trace prefix, the shared context side affects the two program sides of the executions in equivalent ways, and any differences, including undefined behavior, must originate in the different program sides.

Some of the intuitions of our blame proof are similar to the recomposition proof, but carried over to Clight and to a different type of simulation involving three partial program parts arranged in two whole programs. The key challenge of the blame proof lies in the definition of the simulation invariants that relate the two executions. The shared trace prefix forces both executions to run in sync: at any point in time, either the shared context side is driving the two runs *in lockstep*, or each program side is running *independently* from the other until an observable event forces them to re-synchronize. We can apply the ideas outlined above to prove a small number of elementary simulation results for an appropriate relation $\mathcal{R}$, and use these to assemble a full proof of blame:

(1) $\mathcal{R}$ is preserved when *both* whole programs take a *single step*, with both producing the *same event*. (Proved separately starting from the program side and the context side.)
(2) $\mathcal{R}$ is preserved when *one* of the whole programs takes a *silent step* from the *program side* while the other stays put.

On top of these stepwise results, we can build three preservation properties on longer *synchronized executions* of the two programs.

[4] CompCert experts may note that we do not use a notion of decreasing measure or a notion of final states, since we are only concerned with finite execution prefixes.



(1) Initial states **(2) Non-silent step** **(3) Silent step in first execution** **(4) Silent step in second execution**

Figure 2: Three-way simulation properties represented graphically

(1) $\mathcal{R}$ is preserved when *both* whole programs take a sequence of *silent steps*, followed by a pair of *synchronous steps* producing the *same event* on both sides. (Proved separately starting from the program side and the context side.)
(2) $\mathcal{R}$ is preserved after any *shared trace prefix* produced by a pair of executions of the two whole programs.

Finally, synchronized executions allow us to reason about full program runs, which result from the whole program taking steps until it is unable to do so any more; we will simply refer to this as a *whole run*. After a whole program finishes execution, CompCert can reason about its normal or abnormal termination by inspecting the last state of the run. In particular, the semantics defines which states are considered proper *final states*; and all others are considered *stuck states*. A final state corresponds to successful termination, and stuckness corresponds to undefined behavior. The following two key lemmas look at whole runs, the traces produced by those runs, and the side that was in control at the end of the execution (program or context) to blame the program side for any differences.

(1) If $\mathcal{R}$ holds, one whole program runs with a trace m and the context side in control at the end, and the other whole program runs with an extension of m and terminates in a final state, then the first whole program also ended in a final state.
(2) If $\mathcal{R}$ holds, one whole program runs with a trace m, and the other whole program runs with a *strict extension* of the trace, i.e., m followed by a non-empty trace m', then the *program side* of the first program was in control at the end.

The high-level blame proof follows in a relatively straightforward manner from the simulation lemmas for whole runs. The main sources of complexity of the proof pertain to the preservation of the blame invariants in the stepwise simulation lemmas above. Notably, blame requires us to build and maintain asymmetric memory injections that relate the contents of the context sides of a pair of memories and their symbols. However, it cannot completely disregard the program sides of the memories, which have to be compatible, i.e., define the same public symbols, even if they are different. For this reason, the injections must map all public symbols in the whole programs, not just those in the shared context side.

4 Extending CompCert with Compartments

In this section, we detail how we added compartments to CompCert's languages, including RISC-V assembly.

Memory model. We reuse the block-based memory model of CompCert and extend it with compartments. Each memory block belongs to a single compartment that is assigned at allocation and cannot be changed during the execution. Memory operations (reads, writes, frees) are parameterized by the compartment performing the operation, and fail when this compartment does not own the targeted block. This means that compartments cannot share memory or pass data other than by performing calls and returns.

Calls and returns. We extend CompCert's trace model to include two more events capturing compartment transitions (see Figure 3). We write $\overline{x}$ to denote a list of x. The new events are highlighted: `Event_call C C'.f args` captures a cross-compartment call passing `args` from compartment `C` to the procedure `f` of compartment `C'`. Similarly, `Event_return C' C v` represents returning value `v` from compartment `C'` to compartment `C`.

```
a := Event_syscall name v̄ bs v bs
   | Event_vload m ch id ofs v
   | Event_vstore m ch id ofs v
   | Event_annot string v̄
   | Event_call cp cp.id v̄
   | Event_return cp cp v
```

Figure 3: SECOMP events

For all languages but RISC-V, we make use of the existing structure of the semantics to implement these new events, so the only required change is to add the events to the appropriate call and return transitions. We also insert in the semantics dynamic checks to ensure calls and returns conform to the interfaces and do not pass pointers. The dynamic checks act as follows: Internal calls (e.g., a call from `C.f` to `C.g`) are always allowed. System calls are allowed only if the interface allows it. A cross-compartment call from `C.f` to `C'.g` is allowed if (1) `g` $\in$ `C'.exports`; (2) `C'.g` $\in$ `C.imports`; and (3) all of the arguments are scalar values. The dynamic checks for cross-compartment returns are similar: we check that the returned value is a scalar. In all languages before RISC-V control-flow well-bracketedness is ensured by the semantics.

Lastly, we made all registers be caller saved, since on a cross-compartment call we cannot trust the callee compartment to save and restore the caller's registers.[5] We also made the semantics invalidate non-argument registers on cross-compartment calls and non-return registers on cross-compartment returns (by making them undefined values), since recomposition requires all information passed between compartments to be captured by the trace.

Changes to RISC-V. As explained in §3.2, adding compartments to the RISC-V assembly semantics required more extensive changes. For a start we added a boolean flag to the jump-and-link instructions $\texttt{Pjal}_\texttt{s}$ and $\texttt{Pjal}_\texttt{r}$, and to the indirect jump instruction $\texttt{Pj}_\texttt{r}$. When this tag is `true`, the instruction can be used to attempt cross-compartment calls using jump-and-link or returns using indirect jump, but otherwise instructions are not allowed to change the current compartment. Additionally, a change we did to the RISC-V semantics for recomposition to hold is to make stack frames used to pass spilled arguments in cross-compartment calls read-only. This

[5] For simplicity our current implementation does this for all calls, not just cross-compartment ones, but it would be good to improve this aspect in the future.

prevents a malicious compartment to exploit callbacks to modify the spilled arguments it previously passed, for instance by reusing a pointer to a former stack frame that contains such arguments.

Most importantly, our RISC-V semantics makes use of a *cross-compartment shadow stack*, which is a list of records, each containing a return address, a stack pointer, and a procedure signature. Whenever a cross-compartment call occurs, the previous return address and stack pointer, and the callee's signature are pushed on top of the shadow stack. Whenever a cross-compartment return occurs, the semantics checks that the return targets the right address and that the stack pointer has been correctly restored, before executing the return. The procedure signature on the shadow stack includes what register stores the return value, which is used to check that no pointers are passed. This shadow stack is used for abstractly specifying the well-bracketedness [8] of cross-compartment control-flow, and leaves lower-level backends the freedom to enforce this in various ways—for instance, in §9 we describe an enforcement mechanism based on capabilities. Finally, this shadow stack is not used for calls and returns inside a compartment, so backends don't have to do anything special for these.

Buffer-based IO. We also extended the IO model of CompCert from single-character-based to buffer-based IO. Before our change CompCert modeled only very simple IO procedures. System calls and their arguments and return value, which must be scalars or pointers to globals, are recorded as events in the program trace. The behavior of all system calls is described by a single high-level axiomatization that enforces various generic properties, which are sufficient to support the compiler correctness proof. In particular, system calls are required to be *determinate* [25], in the sense that any two calls with the same arguments and results have the same effect on memory, and *receptive*, meaning that any call might return an arbitrary result value of the correct type. While this is adequate to model single-character-based IO procedures like `getchar` and `putchar`, it does not account properly for calls that read or write memory as a side-effect, which are very common in real C code. For example, the system call `read` takes as arguments the number of bytes to read and a buffer address, stores bytes into the buffer, and returns the actual count of bytes that were stored. Two `read` calls might return the same count but store different values in memory (violating determinacy); moreover, the count cannot exceed the requested number of bytes (violating receptivity).

We address these limitations by extending CompCert's system call events to record any bytes loaded from or stored to global memory buffers (the highlighted arguments to `Event_syscall` in Figure 3). We weaken determinacy to allow calls to store different byte values even if they return the same result value, and we weaken receptivity to put procedure-specific constraints on result values and stored bytes. For the latter, we just require that the result and stored bytes might be produced by a call to the procedure with *some* environment and initial memory. To validate our approach, we give detailed models for `read` and `write` system calls, and show that they indeed satisfy the (weakened) properties.

5 Back-Translation Proof Details

We now detail the challenges involved in adapting the back-translation proof of Abate *et al.* [4] to our setting. As explained in §3.3, the main difficulty stems from the fact that when a system call in RISC-V generates an event, we have to prove that it is possible to generate the exact same event in Clight. However, the axiomatization of CompCert's system calls does not allow us to do this easily. Among the axioms CompCert gives us, determinacy, which states that executing the system call in the same memory state yields the same result, is not sufficient. There is indeed little hope of perfectly reproducing the RISC-V memory in Clight.

Instead, CompCert provides another useful axiom: if a system call is executed in some memory $\mathtt{M}_1$ resulting in memory $\mathtt{M}_1'$ while generating a trace, then if executed in some other memory $\mathtt{M}_2$ that $\mathtt{M}_1$ injects into, the same system call with the same arguments results in a memory $\mathtt{M}_2'$ that $\mathtt{M}_1{}'$ injects into, and crucially generates the same trace. Yet using this axiom imposes another condition on the memory injection: it must be defined at least on the public symbols of the environment (of the calling compartment), which include global buffers. This motivates our usage of informative events to record the content of these global buffers inside memory deltas.

Informative events. We consider 3 kinds of informative events:

(1) `Icall f t g` $\overline{\mathtt{v}}$ `sg` Δ represents a cross-compartment call, where `f` is the name of the caller, `t` is the trace event produced by the call, `g` is the name of the callee, $\overline{\mathtt{v}}$ are the arguments, `sg` is the signature of the callee, and Δ are the memory deltas.
(2) `Ireturn f t v` Δ represents a cross-compartment return, where `f` is the current procedure name, `t` is the trace event produced by the return, `v` is the return value, and Δ are memory deltas.
(3) `Isys f t ef` $\overline{\mathtt{v}}$ Δ represents a system call, which is similar to the call case except that it requires the system call descriptor `ef` instead of `g` and `sg`.

In this definition, Δ represents a list of memory deltas. A memory delta δ records a sequence of all operations affecting the contents of memory. For instance, `delta_store ch b o v C` represents a store at location `(b, o)` of value `v` performed by compartment `C`, with memory chunk `ch`. Given an informative event that contains a list of memory deltas Δ, this records relevant operations since the last informative event. Reapplying these allows us to reconstruct the global buffers in Clight. We capture this intuition as part of a novel notion of well-formedness of informative traces using an intermediate language.

Compared to *data-flow events* [23] recording every memory or register operation, our informative events only record operations affecting global buffers. While this means our back-translation doesn't support memory sharing, this also considerably simplifies defining the back-translation and stating and proving the invariants.

Well-formedness of informative events. To characterize the informative traces that can be back-translated, we define a state transition relation that abstracts over the RISC-V semantics. States are triples containing parts of the RISC-V state at the points where events are generated: `state` := `(f, M, fs)`. Intuitively, the first element is the current procedure identifier; the second element is the current memory in the RISC-V execution; and the third element is a simplified view of the stack, seen as a list of procedure identifiers.

The relation $s \xrightarrow{\alpha} s'$ captures the conditions for a state to take a step while producing an informative event, including relations among the global environment `ge` and procedures, memory deltas, and memory updates, as well as the well-bracketedness of cross-compartment calls and returns. It does not include information

irrelevant to the back-translation to Clight though, such as all the low-level details of the RISC-V semantics. For instance, the following (simplified) rule describes the conditions for system calls:

$$\frac{\texttt{mem\_delta\_apply ge C } \Delta\ \texttt{M}_1 = \texttt{M}_2 \qquad \texttt{globals\_scalar ge C}_1\ \texttt{M}_2 \qquad \texttt{wf\_deltas } \delta}{(\texttt{f}_1, \texttt{M}_1, \texttt{k}) \xrightarrow{\texttt{Isys f}_1\ \texttt{(Event\_syscall C ef } \overline{\texttt{v}}\texttt{) ef } \overline{\texttt{v}}\ \Delta} (\texttt{f}_1, \texttt{M}_2, \texttt{k})}$$

This rule abstracts the corresponding rule in the RISC-V semantics: it has the same conditions for the global environment, trace, shadow stack, and global buffers, and adds conditions for updating the memory according to the memory deltas. The rules for calls and returns are similar, except that they also update the stack.

As previewed in §3.3, we say an informative trace is well-formed when it is produced by the reflexive transitive closure of this step relation, starting from a state corresponding to the initial state of RISC-V. We prove that for any prefix m of a trace produced by a RISC-V program, there exists a well-formed informative trace $\overline{\alpha}$, such that $\texttt{proj}(\overline{\alpha}) = m$. Thanks to this result, we can forget about the RISC-V semantics, and simply write a back-translation that takes as an input a well-formed informative trace, and then prove its correctness using the well-formedness.

Back-translation function. Our back-translation function constructs a Clight program by producing procedures from the informative trace. The back-translation converts each informative event to a Clight statement that produces the same event: for instance, a call event is converted to a call instruction. System calls are more interesting: before the system call, the back-translation generates a list of instructions that write to each public variable according to the last memory store recorded for that variable on the trace. The stored values are guaranteed to be scalars by the well-formedness conditions, which is necessary since Clight doesn't allow forging an arbitrary pointer. Eventually, these events are wrapped inside a switch statement and a loop, as described in §3.3.

Correctness of the back-translation. We prove a simulation between the intermediate language describing the well-formedness of the informative trace and the Clight semantics of the program. We prove that given two states, s_i of the intermediate language and s_C of Clight that are related by our simulation invariant, then a transition $s_i \xrightarrow{\alpha} s_i'$ in the intermediate language corresponds to a sequence of transitions $s_C \xrightarrow{t}{}^{*} s_C'$ in Clight, where $t = \texttt{proj}(\alpha)$. Additionally, we prove that s_i' and s_C' are related by our simulation invariant, which does not require the memory of the intermediate language state to inject into the Clight memory; instead, we only maintain the invariant that the global buffers of each compartment are writable. This invariant and the well-formedness of informative events provide enough information for us to construct a memory injection for the global buffers at the point of a system call. Specifically, when encountering a system call, we can use the facts stored in the rule above: $\texttt{mem\_delta\_apply ge C } \Delta\ \texttt{M}_1 = \texttt{M}_2$, $\texttt{wf\_delta } \Delta$, and $\texttt{M}_2$ only contains scalars ($\texttt{globals\_scalar ge C}_1\ \texttt{M}_2$). Using these we show that the back-translation of this informative event writes all the necessary values in the global buffers, and we can construct a memory injection that is only defined on these buffers. We can then apply the CompCert axiom discussed in the 2nd paragraph to prove that the system call succeeds and produces the same event.

Compared to the data-flow back-translation [23], we do not rely on maintaining the invariant that the memories are related by a memory renaming (or a memory injection). Instead, we only have to prove that the part of the memories dedicated to the global buffers are related at specific points in the proof. Establishing this relatedness is also made easier by the fact the global buffers only contain scalars, which allows us to ignore the rest of the memory.

Because we use a finite, `int64` counter to track how many events have been produced, we must limit the length of the (non-informative) trace prefix to 2^{64} events. This limit on prefix lengths is, however, higher than the one coming from the assumption that the compilation of the back-translated program succeeds (see §7).

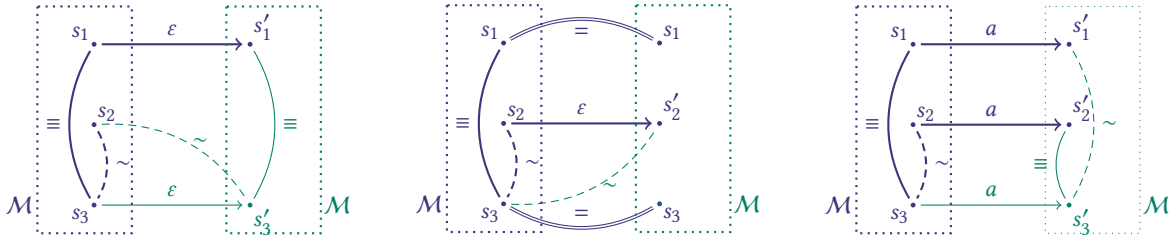
6 Recomposition Proof Details

In §3.4, we described our recomposition proof technique at a high level. Here we detail the proof diagrams that, together, imply the three-way simulation of recomposition. The most important three of these diagrams are depicted in Figure 4. As in §3.4, in each diagram, each row represents one of the 3 executions. Arrows represent execution steps, and are annotated with either an event a or silence ε. We use thick, dark purple for the assumptions, and dark green for the conclusions we have to prove (including existence of states). We depict in dashed line the weak relation, and in plain line the strong relation; and we use dotted rectangles to depict the mixed relation. We denote equality via a double line in Figure 4b.

Figure 4a describes the case where s_1 and s_3 are taking a silent step synchronously and can be read as follows: starting from three states s_1, s_2, and s_3 related by $\mathcal{M}$, such that $s_1 \equiv s_3$ and $s_2 \sim s_3$, and such that s_1 steps silently to s_1', we have to prove that s_3 also steps silently to another state s_3', that the weak and strong relation are reestablished for s_3' (i.e., $s_1' \equiv s_3'$ and $s_2 \sim s_3'$), and that the mixed relation is also reestablished. Similarly, Figure 4b describes the case where s_2 takes a silent step; because it is only weakly related to s_3, we do not require either s_1 or s_3 to take steps as well, but we must still reestablish the relations. Finally, Figure 4c describe the case where both executions produce an event a. Given $s_1 \equiv s_3$ and $s_2 \sim s_3$, related by $\mathcal{M}$, such that $s_1 \xrightarrow{a} s_1'$ and $s_2 \xrightarrow{a} s_2'$, one must prove the existence of s_3' such that $s_3 \xrightarrow{a} s_3'$, i.e., that the three executions advance in lockstep. If the event constitutes a change of control between the two sides (program and context) weak and strong relations are be swapped as illustrated in the diagram; but we also have a similar diagram where the relations are not swapped.

We prove our 8 diagrams together imply the existence of the above three-way simulation, and hence imply the recomposition theorem. To do so, we simply instantiate $\mathcal{R}$ with the conjunction of weak, strong, and mixed relations applied to the appropriate cases.

Applying diagrams to our setting. We now explain how we instantiate the parameters of our proof diagrams. At the RISC-V level, states are the disjoint union of regular states s=($regs$, M, st) and return-states rs=($regs$, M, st, C) where $regs$ is a register set, M is a memory, st a shadow stack, and C a compartment name recording which compartment the execution is returning from. To handle the possibility of having different allocation behavior in each execution, we relate memories and values using two memory injections j_1 and j_2, one for each of the original executions. These memory injections are only defined on their execution's kept compartments,



(a) Silent step in strongly related states (b) Silent step in weakly related states (c) Non-silent step with swapping relations

Figure 4: Recomposition diagrams

and do not describe the other compartments' memory. These memory injections parameterize the relations, are kept as part of the ghost state we maintain, and are updated during the execution. As described previously, the weak and the strong relations relate the memories of the run according to these injections, the strong relation also relates the registers, and the mixed relation relates the stacks, requiring that the stacks represent the same series of calls with the same arguments.

Only the proofs of the 3 diagrams from Figure 4 are interesting, since the remaining 5 follow by exploiting symmetries of the diagrams and of the relations. The first two diagrams require that silent steps don't affect other compartments. This wouldn't hold if stack frames storing spilled arguments were still writable by the caller, since otherwise a compartment could surreptitiously communicate by changing previously passed arguments in an old stack frame. We discovered this issue while trying to complete the proof of these two diagrams and failing, as we couldn't prove that a compartment reading an argument from the stack would get related values according to the memory injection of its side (it could instead get a pointer from the other side that's not in the memory injection). This required us to step back and change the semantics of RISC-V assembly to make old arguments passed on the stack read only, as described in §4. Finally, the last diagram, where an event is produced, is proved by re-establishing the three relations after producing the correct event. This is made possible by the fact that all information that is shared between the two compartments on a call or return appears in the trace event.

While our proof structure is similar to the turn-taking simulation of El-Korashy *et al.* [23], we greatly benefit from the proof diagrams above, as they give us a clear structure to organize the proof. Moreover, because we rely on CompCert's memory injections instead of ad-hoc memory renamings, we are able to reuse much of the machinery that already exists for the compiler correctness proof.

7 Compiling back-translation result

The FCC statement of Abate *et al.* [4] (reproduced in Def. 2) assumed that the compiler is a total function. While this was true of their very simple compiler, it is not the case for realistic compilers like CompCert, so our FCC theorem needs an extra assumption that the compiler can successfully compile $\mathbf{C}$ and $\mathbf{P}$:

Theorem 7.1 (Forward Compiler Correctness (FCC)). $\forall \mathbf{C}\ \mathbf{P}.$
$\forall m \not\in \mathsf{Undef}.\ (\mathbf{C} \bowtie \mathbf{P}) \rightsquigarrow m \wedge \mathbf{C}{\downarrow}\ \textit{and}\ \mathbf{P}{\downarrow}\ \textit{defined} \Rightarrow (\mathbf{C}{\downarrow} \bowtie \mathbf{P}{\downarrow}) \rightsquigarrow m.$

This extra assumption in FCC leads to a new assumption in our RSC^{DC}_{MD} proof, namely that the result of our back-translation can be successfully compiled. But this is not at all easy to prove: CompCert has many sources of partiality and it is not feasible to guarantee in advance that a (well-typed) C program will be successfully compiled. Two passes in CompCert, register allocation and linearization, are not verified but rely on translation validation, which can fail. The data-flow analyzer used in several optimization passes can fail if the analysis doesn't converge after a very large number of steps (e.g. 10^{12}). Several passes (notably Asmgen) use errors to rule out ill-formed code that should never have made it this far, but which is easier to recheck than to prove impossible. Finally, several passes (CminorGen, Inlining, Stacking, etc.) put constraints on the size of the generated stack frames to ensure that offsets within these frames don't overflow a machine word.

Compiling the back-translation is just an artifact of the high-level proof structure (Figure 1), which uses compiler correctness for whole programs to repeatedly move between the source and target languages. Therefore we assume as an axiom that the result of our back-translation, on any trace prefix below a certain length, *can* be successfully compiled. The length bound is needed to account for machine words being finite and other such finite resources.

Assumption 1 (Back-translation successfully compiles)**.**

$$\forall K.\ \forall I.\ \forall \mathbf{W_T}{:}I.\ \forall m \not\in \mathsf{Undef}.\ |m| \leq \mathit{MAX\_TRACE\_LENGTH} \Rightarrow$$
$$\forall \mathrm{p} : \mathbf{W_T} \rightsquigarrow m.\ \forall \mathbf{C}{:}\lfloor I \rfloor_K.\ \mathbf{C} = \bowtie_{k \in K} (I, \mathrm{p}, k)\uparrow \Rightarrow \mathbf{C}{\downarrow}\ \text{is defined}$$

The next paragraphs report how we have systematically tested this assumption for a large number of trace prefixes. In the future one could envision using more compositional compiler correctness results than that of CompCert, recent [36, 54, 57, 72] or upcoming, to potentially overcome the need for this assumption. For now though, we take this assumption as a reasonable cost to pay for a secure compilation proof technique that is the first to scale up to a realistic compiler like CompCert and that only requires an operational semantics for whole programs, which is not compositional, but which simplifies proofs, including CompCert's existing compiler correctness proof that we extended here to isolated compartments.

Property-based testing of Assumption 1. We systematically tested that the Clight programs constructed by our back-translation function can be compiled with CompCert again. Concretely, we experimentally test that $\texttt{ccomp}(\texttt{bt\_fun}(bt, env))$ succeeds for random but *consistent* informative traces bt and environments env (§5). The environment defines a set of compartments, their interfaces, and the available procedures that can be referenced in the traces. We generate random environments by deriving them from random, undirected and connected graphs $\mathcal{G} = (V, E)$. Each vertex $v \in V$ represents a compartment and we associate it with a random, non-empty set of procedures and signatures `v.exports`. Further, for each $(u, v) \in E$ we set `u.imports` to a random, non-empty subset of

v.exports and vice-versa (c.f. §4). The trace is generated consistent with the environment such that (1) each procedure call is allowed (c.f. §4); (2) two calls to the same procedure use the same signature and (3) all values passed as arguments or return values match the signature. For efficiency, we only generate values and in particular memory deltas that are explicitly inspected in the back-translation function and not trivially compiled to skips.

We have been able to successfully compile all Clight programs produced by back-translation for more than 100k pairs of generated environments and traces with up to 880 events and close to 400 events on average. Individual tests with significantly longer traces of more than 150k events also succeeded, but the growing computational costs make it hard to test even longer traces. In total, the traces contained over 16M `Icall` and `Ireturn`, 7M `Isys` and 500M `delta_storev` instances.

8 Top-level RSC_{MD}^{DC} theorem

We follow the general proof diagram from Figure 1 to assemble our previous theorems and obtain the final result that our compilation chain satisfies a variant of RSC_{MD}^{DC} from Def. 3.

Theorem 8.1 (RSC_{MD}^{DC}). *SECOMP satisfies RSC_{MD}^{DC} for all trace prefixes m such that $|m| \leq$ MAX_TRACE_LENGTH.*

The size of the prefixes supported by this theorem is restricted by the need to successfully compile the results of the back-translation (Assumption 1), which we have systematically tested as described in §7. A second disclaimer is that we have not yet finished integrating the Coq proof of this theorem with the Coq proofs of the individual steps, since as mentioned in §1, the proofs of back-translation, recomposition, and blame are complete too, but they were done on separate branches that we are currently in the process of merging.

9 Enforcement using capabilities

To show that the capability abstraction we added to the semantics of CompCert's RISC-V assembly is practically implementable at a lower level, we designed a capability backend for SECOMP. The backend targets an extension of the CHERI RISC-V architecture [69], which provides hardware capabilities; i.e., unforgeable pointers with base and bounds that cannot be circumvented. While various secure calling conventions targeting capabilities have been proposed in recent years [27, 55, 56, 63], our backend is based on the most recent proposal of Georges *et al.* [28], which uses not only the standard capabilities described above, but also CHERI's local capabilities [55], entry, and sealed capabilities. In particular, stack pointers are implemented as local capabilities which can only be stored on the stack or in registers. Hence, a compartment cannot save a capability to some part of the stack for later use, and cannot modify the value of the arguments it passed *a posteriori*, an attack we prevent for languages higher in the compilation chain by making some stack frames read-only (see §4). Additionally, this calling convention is based on two newly proposed kinds of capabilities: uninitialized [27] and directed [28]. In short, uninitialized capabilities prevent reading old values from the stack without excessive clearing [27], and directed capabilities support efficient implementation of stack safety [28]. Our backend targets a lower-level variant of CompCert's RISC-V assembly language with a flat memory model and extended with all these capabilities.

While our calling convention is inspired by Georges *et al.* [28] we had to adapt that design to our setting in two ways: First, because we only enforce compartment isolation, not memory safety, we represent pointers as offsets into a large stack capability or into per-compartment heap capabilities. By not using directed capabilities for stack pointers, we overcome a potential limitation of Georges *et al.*'s [28] calling convention and can store cyclic data structures on the stack. Second, compared to Georges *et al.* [28] we consider a stronger attacker model, in which both the caller and the callee compartments of a call can be compromised. In our model we thus need to always maintain the distinction between the caller and callee compartments and enforce that no capabilities are exchanged between the two. We achieve this by adding privileged wrappers for calls and returns, which ensure that the passed arguments/returns are not capabilities, and which clear all remaining registers.

We built a prototype implementation of this backend in Coq that can already compile simple examples, but that has not yet been thoroughly tested and is not verified. In the short run, one can use property-based testing to get more confidence in its correctness and security. We are also considering the design of a second capability backend inspired by the original work of Watson *et al.* [70] and implemented in CheriBSD [26], which only uses the existing features of CHERI. This second design, however, requires a split stack layout, which is allowed by the C standard and the CompCert memory model, but which changes the RISC-V calling convention. In the long run, formally verifying such backends is a very interesting research challenge, as also discussed in §11.

10 Related work

As explained in §2, we directly build on the work of Abate *et al.* [4], in particular by reusing their RSC_{MD}^{DC} secure compilation criterion and their high-level proof structure. Scaling up these ideas from a very simple compiler for a toy programming language all the way to a verified compiler for the realistic C language was an open research challenge that we overcome in this paper. In our realistic setting, the back-translation, recomposition, and blame steps are more interesting and require several proof engineering novelties and also more sophisticated invariants involving memory injections. Moreover, for the compiler correctness steps, which were assumed by Abate *et al.* [4], we show that with careful design we can extend the massive CompCert proof to compartments with a manageable amount of effort (only around 9% size increase).

The security proofs of Abate *et al.* [4] and also a later variant with pointer passing [23] (also discussed in §11) are both mechanized in Coq. Even if compiler correctness is assumed, these are among the few proofs of secure compilation against adversarial contexts (i.e., for criteria like full abstraction and RSP [6]) that have been mechanized in a proof assistant, with the majority of work in this space being proved only on paper, usually for even simpler languages and compilers [6, 7, 18, 24, 33, 48–51]. One proof that was fully mechanized in Coq is that of Devriese *et al.* [21], who prove modular full abstraction by approximate back-translation for a compiler from the simply typed to the untyped λ-calculus. Jacobs *et al.* [32] prove in Coq the purity of a Haskell-like ST monad stated as full abstraction of a translation from a pure language. Georges *et al.* [28] prove in Coq the security of their calling convention (a variant of which we also use in §9) stated as the full abstraction of

the identity compiler between a secure overlay semantics and the actual semantics of a simple idealized assembly language. Finally, Abate *et al.* [5, §7.1] verify a very simple compiler in Coq illustrating secure compilation when the target language has additional trace events that are not possible in the source.

A more realistic related work is CompCertSFI [14], which builds on previous ideas by Kroll *et al.* [37] to implement portable SFI as a source-to-source transformation in Cminor, an intermediate language of CompCert that comes before optimizations. Pointers are represented as integers and masked in order to offset into a single big array representing all of the sandbox's memory. In addition to masking pointers and using trampolines for functions pointers, CompCertSFI instruments the program to prevent any undefined behavior. This is needed to properly preserve the main security result, showing that all memory accesses stay within the sandbox, down to CompCert's assembly language. An experimental evaluation shows that the overhead of CompCertSFI comes mostly from CompCert itself performing less aggressive optimizations than GCC and Clang. When the proposed SFI transformation is instead used with GCC or Clang the overheads are generally competitive with (P)NaCl [71]. By implementing SFI in an early intermediate language, CompCertSFI can take advantage of all the compiler's optimizations as well as the alignment analysis added by the authors.

Our implementation strategy is different and is not targeted specifically at SFI, but instead at being able to take advantage of hardware features for compartment isolation such as capabilities [13, 69] or the recently proposed support for Hardware Fault Isolation [46]. Another difference is that SECOMP supports an arbitrary number of mutually-distrustful compartments that can interact by calling each other according to clearly specified interfaces. Finally, while one could potentially extend CompCertSFI to achieve a security notion similar to the original RSP [5, 6, 51] (so without mutual distrust), this would require more work, for instance proving compiler correctness with respect to the semantics of source programs, bridging the gap between the memory model used by CompCert and the single memory block model used by CompCertSFI, and devising a verified back-translation between Cminor and higher CompCert languages like C or Clight.

Another verified SFI compiler is vWasm [17], for which the authors proved in F* [59] that Wasm code compiled to x86 can only interact with its host environment via an explicitly provided API. While this security guarantee and interaction model is similar to that of CompCertSFI, the vWasm implementation doesn't take advantage of all standard compiler optimizations, which leads to some performance loss. The security guarantee only talks about the x86 semantics and, as opposed to our work, does not aim at providing source-level security reasoning, even at the Wasm level.

Another realistic work in this space is that of Derakhshan *et al.* [19], who devise a methodology to break up Trusted Execution Environment (TEE) software into concurrently executing C compartments whose security is compositionally verified using semi-automatic tools and which are then correctly compiled using a verified compiler . The formalization of this work is done on paper and the main assumption is that all C compartments are verified, which seems realistic only for small, highly privileged pieces of code, like the TEEs this work considers. Our focus is instead on machine-checked proofs and on compartmentalized C code that can't be formally verified to be even free of undefined behaviors.

Our work targets a variant of RSP, but such preservation of property classes against adversarial contexts is not the only kind of formally secure compilation. Another important kind aims at preserving specific noninterference properties against passive side-channel attackers. For instance, preservation of cryptographic constant time was proved for both the CompCert [11] and Jasmin [12] verified compilers. Another example is guaranteeing that protection against memory probing is preserved by CompCert [15].

Other formal verification work looks at security of low-level enforcement mechanisms, without involving a compiler from a higher-level language. For instance, SFI mechanisms for both x86 [44] and ARM [73] were proved correct in a proof assistant with respect to the semantics of these complex architectures. In these works communication between low-level compartments is done by jumping to a specified set of entry points, while we consider a more structured model that also enforces the correct return discipline. Other work in this space looks at the basic security properties of capability machines, from simpler ones [31, 58] to more realistic ones like CHERI [47] and Arm Morello [13].

11 Future Work

Compiling realistic compartmentalized applications. At the moment we have evaluated SECOMP only on very simple code, for instance C variants of the examples of Abate *et al.* [4]. The main obstacle to compiling more realistic compartmentalized applications is the current inability to communicate non-scalar data. An immediate solution would be to add an IPC-like mechanism for passing the contents of buffers between compartments. A more ambitious solution (discussed next) would be to allow sharing memory by passing capabilities on a machine like CHERI [69]. Another way to make SECOMP more practical would be to extend our interfaces with more fine-grained access control policies for IO [2, 9]. Finally, another interesting direction is connecting with tools for semi-automated compartmentalization of realistic C applications [29].

Pointer passing and memory sharing. As with mainstream compartment isolation mechanisms (e.g., SFI or OS processes), we assume that compartments can only communicate via scalar values, but cannot pass each other pointers to share memory. While secure pointer passing is possible to implement efficiently on a capability machine like CHERI [69] or on the micro-policies tagged architecture [10] and this would allow a more efficient interaction model that is also natural for C programmers, the main challenge one still has to overcome is *proving* secure compilation at scale in the presence of such fine-grained, dynamic memory sharing.

Recent work by El-Korashy *et al.* [23] in a much simpler setting shows that it is indeed possible to prove in Coq the security of an extension of Abate *et al.*'s [4] compiler that allows passing secure pointers (e.g., capabilities) between compartments. With such fine-grained memory sharing, however, proofs become more challenging and the proof technique of El-Korashy *et al.* led to much larger proofs and still has conceptual limitations that one would need to overcome for it to work for CompCert, in particular for supporting memory injections. In fact, even extending CompCert's compiler correctness proof to passing arbitrary pointers seems a challenge, since it would imply a significant change to CompCert's

trace model. In the nearer future we will try to allow more limited forms of memory sharing between compartments, for instance of statically allocated buffers, which could be passed without significantly changing CompCert's trace model.

Building and verifying lower-level backends. Like CompCert's correctness proofs, the SECOMP security proofs currently stop at CompCert's RISC-V assembly language. We extended this language with the abstraction of isolated compartments, which formally defines *what* compartment isolation enforcement should do, but which leaves the *how* to lower-level enforcement mechanisms. Beyond the capability-based backend of §9, various other enforcement mechanisms should be possible, including SFI [60, 66, 71] and tagged architectures [10, 22], as shown in a much simpler setting by Abate *et al.* [4]. Moreover, WebAssembly components [1, 17, 30, 65] could also be a target for such a backend.

At the moment the existing lower-level backends are all unverified. Extending the secure compilation proofs down to cover them is a formidable research challenge that we leave for future work. All existing secure compilation proof techniques in this space [6, 49], including the one we use in the current paper [4], have their origin in proof techniques for full abstraction [49]. But once the memory layout becomes concrete [67, 68] and the code is explicitly stored in memory, we can no longer hide all information about the compartments' code, as would be needed for full abstraction (or in our case for recomposition), so new proof techniques will be needed.

Targeting other hardware architectures. While SECOMP currently targets RISC-V for simplicity, the biggest part of CompCert is architecture independent, and our extension preserves this feature and the only architecture-specific pass is still between Mach and assembly. Also extending the CompCert passes from Mach to x86 or Arm assembly seems feasible, and in particular adding the compartment abstraction to CompCert's semantics for x86 and Arm would be very similar to what we already did for RISC-V. A bigger challenge would be reproving recomposition: while some parts of our proof are generic, such as reducing recomposition to the diagrams from Figure 4, proving these diagrams for huge instruction sets like x86 and Arm would be very tedious. One idea to overcome this challenge would be to assume that an attacker can only run instructions produced by our compiler, and to enforce this in lower-level backends using a combination of W^X memory protection and some amount of control-flow integrity.

Even on these architectures, lower-level backends could implement the compartment abstraction in various ways. On a capability machine like Arm Morello [13], one could still do hardware-supported enforcement using capabilities, as outlined at the end of §9. On a modern x86 one could potentially make use of MPK for implementing gaining efficiency [64]. And if everything else fails, one could still enforce compartment isolation in software using SFI.

From safety to hypersafety. Another interesting direction is extending SECOMP to stronger criteria beyond robust preservation of safety, in particular to hypersafety [6], such as data confidentiality. We expect that SECOMP can be easily adapted to these stronger criteria, by for instance always clearing registers before changing compartments, and also that our RSP proof technique can still apply, by only extending the back-translation step to take finite sets of trace prefixes as input [6, 62]. A very interesting and challenging future work is actually enforcing robust preservation of hypersafety in the lower-level backends with respect to side-channel attacks, including devastating micro-architectural attacks like Spectre [52].

Compositional compiler correctness. Reusing the massive correctness proof of CompCert was definitely worth it for our work, yet the limited compositionality of CompCert [34] lead to a proof technique that relies on an extra assumption (Assumption 1) that realistically we could only test. As mentioned in §7, more compositional compiler correctness results, recent [36, 54, 57, 72] or upcoming, could potentially remove the need for this assumption. Moreover, such compositional compiler correctness results could potentially make our architecture-specific proofs easier, since recomposition could be split into a decomposition step for the target and a composition step for the source [6, 24, 33, 48, 50, 51].

Dynamic compartment creation and dynamic privileges. SECOMP uses a static notion of compartments and static interfaces to restrict their privileges. SECOMP compartments are defined statically by the source program, so are a form of code-based compartmentalization. In the future one could also explore dynamic compartment creation, which would allow for data-based compartmentalization [29], e.g., one compartment per incoming network connection or one compartment per web browser tab or plugin [53]. It would also be interesting to investigate dynamic privileges for compartments, e.g., dynamically sharing memory by passing secure pointers (as discussed above), dynamically changing the compartment interfaces [45], or history-based access control [2, 9].

Acknowledgments. We thank Adrien Durier for participating in early discussions about this work. We are also grateful to the anonymous reviewers at PriSC'23 and CCS'24 for their helpful feedback. This work was in part supported by the European Research Council under ERC Starting Grant SECOMP (715753), by the Deutsche Forschungsgemeinschaft (DFG) as part of the Excellence Strategy of the German Federal and State Governments – EXC 2092 CASA - 390781972, by the National Science Foundation under grants 2048499 and 2314323, and by the European Comission under grant 101070374.

References

[1] The WebAssembly component model.
[2] M. Abadi and C. Fournet. Access control based on execution history. *NDSS*. The Internet Society, 2003.
[3] M. Abadi. Protection in programming-language translations. *Secure Internet Programming*. 1999.
[4] C. Abate, A. Azevedo de Amorim, R. Blanco, A. N. Evans, G. Fachini, C. Hriţcu, T. Laurent, B. C. Pierce, M. Stronati, and A. Tolmach. When good components go bad: Formally secure compilation despite dynamic compromise. *CCS*. 2018.
[5] C. Abate, R. Blanco, Ş. Ciobâcă, A. Durier, D. Garg, C. Hriţcu, M. Patrignani, É. Tanter, and J. Thibault. An extended account of trace-relating compiler correctness and secure compilation, 2021.
[6] C. Abate, R. Blanco, D. Garg, C. Hriţcu, M. Patrignani, and J. Thibault. Journey beyond full abstraction: Exploring robust property preservation for secure compilation. *CSF*, 2019.
[7] P. Agten, R. Strackx, B. Jacobs, and F. Piessens. Secure compilation to modern processors. *CSF*. 2012.
[8] S. N. Anderson, R. Blanco, L. Lampropoulos, B. C. Pierce, and A. Tolmach. Formalizing stack safety as a security property. *CSF*. 2023.
[9] C.-C. Andrici, Ştefan Ciobâcă, C. Hriţcu, G. Martínez, E. Rivas, Éric Tanter, and T. Winterhalter. Securing verified IO programs against unverified code in F*. *Proc. ACM Program. Lang.*, 8(POPL):2226–2259, 2024.
[10] A. Azevedo de Amorim, M. Dénès, N. Giannarakis, C. Hriţcu, B. C. Pierce, A. Spector-Zabusky, and A. Tolmach. Micro-policies: Formally verified, tag-based security monitors. *Oakland S&P*. 2015.
[11] G. Barthe, S. Blazy, B. Grégoire, R. Hutin, V. Laporte, D. Pichardie, and A. Trieu. Formal verification of a constant-time preserving C compiler. *Proc. ACM Program. Lang.*, 4(POPL):7:1–7:30, 2020.

[12] G. Barthe, B. Grégoire, V. Laporte, and S. Priya. Structured leakage and applications to cryptographic constant-time and cost. *CCS*. 2021.

[13] T. Bauereiss, B. Campbell, T. Sewell, A. Armstrong, L. Esswood, I. Stark, G. Barnes, R. N. M. Watson, and P. Sewell. Verified security for the Morello capability-enhanced prototype Arm architecture. *ESOP*. 2022.

[14] F. Besson, S. Blazy, A. Dang, T. Jensen, and P. Wilke. Compiling sandboxes: Formally verified software fault isolation. *ESOP*, 2019.

[15] F. Besson, A. Dang, and T. P. Jensen. Information-flow preservation in compiler optimisations. *2019*. 2019.

[16] A. Bittau, P. Marchenko, M. Handley, and B. Karp. Wedge: Splitting applications into reduced-privilege compartments. *USENIX NSDI*, 2008.

[17] J. Bosamiya, W. S. Lim, and B. Parno. Provably-safe multilingual software sandboxing using WebAssembly. *USENIX Security*. 2022.

[18] M. Busi, J. Noorman, J. V. Bulck, L. Galletta, P. Degano, J. T. Mühlberg, and F. Piessens. Securing interruptible enclaved execution on small microprocessors. *ACM Trans. Program. Lang. Syst.*, 43(3):12:1–12:77, 2021.

[19] F. Derakhshan, Z. Zhang, A. Vasudevan, and L. Jia. Towards end-to-end verified TEEs via verified interface conformance and certified compilers. *CSF*. 2023.

[20] D. Devriese, M. Patrignani, and F. Piessens. Parametricity versus the universal type. *PACMPL*, 2(POPL):38:1–38:23, 2018.

[21] D. Devriese, M. Patrignani, F. Piessens, and S. Keuchel. Modular, fully-abstract compilation by approximate back-translation. *LMCS*, 13(4), 2017.

[22] U. Dhawan, C. Hriţcu, R. Rubin, N. Vasilakis, S. Chiricescu, J. M. Smith, T. F. Knight, Jr., B. C. Pierce, and A. DeHon. Architectural support for software-defined metadata processing. *ASPLOS*. 2015.

[23] A. El-Korashy, R. Blanco, J. Thibault, A. Durier, D. Garg, and C. Hriţcu. SecurePtrs: Proving secure compilation with data-flow back-translation and turn-taking simulation. *CSF*, 2022.

[24] A. El-Korashy, S. Tsampas, M. Patrignani, D. Devriese, D. Garg, and F. Piessens. CapablePtrs: Securely compiling partial programs using the pointers-as-capabilities principle. *CSF*. 2021.

[25] J. Engelfriet. Determinacy implies (observation equivalence = trace equivalence). *TCS*, 36:21–25, 1985.

[26] D. Gao. Compartmentalisation models. Principles of Capability Languages workshop, 2024.

[27] A. L. Georges, A. Guéneau, T. V. Strydonck, A. Timany, A. Trieu, S. Huyghebaert, D. Devriese, and L. Birkedal. Efficient and provable local capability revocation using uninitialized capabilities. *PACMPL*, 5(POPL):1–30, 2021.

[28] A. L. Georges, A. Trieu, and L. Birkedal. Le temps des cerises: efficient temporal stack safety on capability machines using directed capabilities. *PACMPL*, 6(OOPSLA):1–30, 2022.

[29] K. Gudka, R. N. M. Watson, J. Anderson, D. Chisnall, B. Davis, B. Laurie, I. Marinos, P. G. Neumann, and A. Richardson. Clean application compartmentalization with SOAAP. *CCS*. 2015.

[30] A. Haas, A. Rossberg, D. L. Schuff, B. L. Titzer, M. Holman, D. Gohman, L. Wagner, A. Zakai, and J. F. Bastien. Bringing the web up to speed with WebAssembly. *PLDI*, 2017.

[31] S. Huyghebaert, S. Keuchel, C. D. Roover, and D. Devriese. Formalizing, verifying and applying ISA security guarantees as universal contracts. *CCS*. 2023.

[32] K. Jacobs, D. Devriese, and A. Timany. Purity of an ST monad: full abstraction by semantically typed back-translation. *PACMPL*, 6(OOPSLA1):1–27, 2022.

[33] Y. Juglaret, C. Hriţcu, A. Azevedo de Amorim, B. Eng, and B. C. Pierce. Beyond good and evil: Formalizing the security guarantees of compartmentalizing compilation. *CSF*, 2016.

[34] J. Kang, Y. Kim, C.-K. Hur, D. Dreyer, and V. Vafeiadis. Lightweight verification of separate compilation. *POPL*, 2016.

[35] D. Kilpatrick. Privman: A library for partitioning applications. *USENIX FREENIX*. 2003.

[36] J. Koenig and Z. Shao. CompCertO: compiling certified open C components. *PLDI*. 2021.

[37] J. Kroll, G. Stewart, and A. Appel. Portable software fault isolation. *CSF*. 2014.

[38] D. Kästner, U. Wünsche, J. Barrho, M. Schlickling, B. Schommer, M. Schmidt, C. Ferdinand, X. Leroy, and S. Blazy. CompCert: Practical experience on integrating and qualifying a formally verified optimizing compiler. *ERTS*. 2018.

[39] L. Lamport and F. B. Schneider. Formal foundation for specification and verification. In *Distributed Systems: Methods and Tools for Specification, An Advanced Course, April 3-12, 1984 and April 16-25, 1985 Munich*, 1984.

[40] X. Leroy. A formally verified compiler back-end. *JAR*, 43(4):363–446, 2009.

[41] X. Leroy and S. Blazy. Formal verification of a C-like memory model and its uses for verifying program transformations. *JAR*, 41(1):1–31, 2008.

[42] M. Miller. Trends, challenges, and strategic shifts in the software vulnerability mitigation landscape. BlueHat IL, 2019.

[43] D. Monniaux and S. Boulmé. The trusted computing base of the CompCert verified compiler. *ESOP*. 2022.

[44] G. Morrisett, G. Tan, J. Tassarotti, J.-B. Tristan, and E. Gan. RockSalt: Better, faster, stronger SFI for the x86. *PLDI*. 2012.

[45] T. C. Murray, D. Matichuk, M. Brassil, P. Gammie, T. Bourke, S. Seefried, C. Lewis, X. Gao, and G. Klein. seL4: From general purpose to a proof of information flow enforcement. *IEEE S&P*. 2013.

[46] S. Narayan, T. Garfinkel, M. Taram, J. Rudek, D. Moghimi, E. Johnson, C. Fallin, A. Vahldiek-Oberwagner, M. LeMay, R. Sahita, D. M. Tullsen, and D. Stefan. Going beyond the limits of SFI: flexible and secure hardware-assisted in-process isolation with HFI. *ASPLOS*. 2023.

[47] K. Nienhuis, A. Joannou, T. Bauereiss, A. C. J. Fox, M. Roe, B. Campbell, M. Naylor, R. M. Norton, S. W. Moore, P. G. Neumann, I. Stark, R. N. M. Watson, and P. Sewell. Rigorous engineering for hardware security: Formal modelling and proof in the CHERI design and implementation process. *IEEE S&P*. 2020.

[48] M. Patrignani, P. Agten, R. Strackx, B. Jacobs, D. Clarke, and F. Piessens. Secure compilation to protected module architectures. *TOPLAS*, 2015.

[49] M. Patrignani, A. Ahmed, and D. Clarke. Formal approaches to secure compilation: A survey of fully abstract compilation and related work. *ACM Computing Surveys*, 2019.

[50] M. Patrignani and D. Clarke. Fully abstract trace semantics for protected module architectures. *CL*, 42:22–45, 2015.

[51] M. Patrignani and D. Garg. Robustly safe compilation, an efficient form of secure compilation. *ACM Trans. Program. Lang. Syst.*, 43(1), 2021.

[52] M. Patrignani and M. Guarnieri. Exorcising spectres with secure compilers. *CCS*. 2021.

[53] C. Reis and S. D. Gribble. Isolating web programs in modern browser architectures. *EuroSys*. 2009.

[54] M. Sammler, S. Spies, Y. Song, E. D'Osualdo, R. Krebbers, D. Garg, and D. Dreyer. DimSum: A decentralized approach to multi-language semantics and verification. *Proc. ACM Program. Lang.*, 7(POPL):775–805, 2023.

[55] L. Skorstengaard, D. Devriese, and L. Birkedal. Reasoning about a machine with local capabilities: Provably safe stack and return pointer management. *TOPLAS*, 42(1):5:1–5:53, 2020.

[56] L. Skorstengaard, D. Devriese, and L. Birkedal. StkTokens: Enforcing well-bracketed control flow and stack encapsulation using linear capabilities. *JFP*, 31:e9, 2021.

[57] Y. Song, M. Cho, D. Kim, Y. Kim, J. Kang, and C. Hur. CompCertM: CompCert with C-assembly linking and lightweight modular verification. *Proc. ACM Program. Lang.*, 4(POPL):23:1–23:31, 2020.

[58] T. V. Strydonck, A. L. Georges, A. Guéneau, A. Trieu, A. Timany, F. Piessens, L. Birkedal, and D. Devriese. Proving full-system security properties under multiple attacker models on capability machines. *CSF*. 2022.

[59] N. Swamy, C. Hriţcu, C. Keller, A. Rastogi, A. Delignat-Lavaud, S. Forest, K. Bhargavan, C. Fournet, P.-Y. Strub, M. Kohlweiss, J.-K. Zinzindohoue, and S. Zanella-Béguelin. Dependent types and multi-monadic effects in F*. *POPL*. 2016.

[60] G. Tan. Principles and implementation techniques of software-based fault isolation. *FTSEC*, 1(3):137–198, 2017.

[61] The Chromium Project. Memory safety. chromium.org.

[62] J. Thibault and C. Hriţcu. Nanopass back-translation of multiple traces for secure compilation proofs. PriSC, 2021.

[63] S. Tsampas, D. Devriese, and F. Piessens. Temporal safety for stack allocated memory on capability machines. *CSF*. 2019.

[64] A. Vahldiek-Oberwagner, E. Elnikety, N. O. Duarte, M. Sammler, P. Druschel, and D. Garg. ERIM: secure, efficient in-process isolation with protection keys (MPK). *USENIX Security*. 2019.

[65] L. Wagner. What is a WebAssembly component (and why?). WebAssembly Workshop (WAW), 2024.

[66] R. Wahbe, S. Lucco, T. E. Anderson, and S. L. Graham. Efficient software-based fault isolation. *SOSP*, 1993.

[67] Y. Wang, P. Wilke, and Z. Shao. An abstract stack based approach to verified compositional compilation to machine code. *PACMPL*, 3(POPL):62:1–62:30, 2019.

[68] Y. Wang, X. Xu, P. Wilke, and Z. Shao. CompCertELF: verified separate compilation of C programs into ELF object files. *Proc. ACM Program. Lang.*, 4(OOPSLA):197:1–197:28, 2020.

[69] R. N. M. Watson, P. G. Neumann, J. Woodruff, M. Roe, H. Almatary, J. Anderson, J. Baldwin, G. Barnes, D. Chisnall, J. Clarke, B. Davis, L. Eisen, N. W. Filardo, R. Grisenthwaite, A. Joannou, B. Laurie, A. T. Markettos, S. W. Moore, S. J. Murdoch, K. Nienhuis, R. Norton, A. Richardson, P. Rugg, P. Sewell, S. Son, and H. Xia. Capability Hardware Enhanced RISC Instructions: CHERI Instruction-Set Architecture (Version 8). Technical Report UCAM-CL-TR-951, University of Cambridge, Computer Laboratory, 2020.

[70] R. N. M. Watson, J. Woodruff, P. G. Neumann, S. W. Moore, J. Anderson, D. Chisnall, N. H. Dave, B. Davis, K. Gudka, B. Laurie, S. J. Murdoch, R. Norton, M. Roe, S. Son, and M. Vadera. CHERI: A hybrid capability-system architecture for scalable software compartmentalization. *S&P*. 2015.

[71] B. Yee, D. Sehr, G. Dardyk, J. B. Chen, R. Muth, T. Ormandy, S. Okasaka, N. Narula, and N. Fullagar. Native Client: A sandbox for portable, untrusted x86 native code. *CACM*, 53(1):91–99, 2010.

[72] L. Zhang, Y. Wang, J. Wu, J. Koenig, and Z. Shao. Fully composable and adequate verified compilation with direct refinements between open modules. *Proc. ACM Program. Lang.*, 8(POPL):2160–2190, 2024.

[73] L. Zhao, G. Li, B. D. Sutter, and J. Regehr. ARMor: Fully verified software fault isolation. *EMSOFT*. 2011.