Soundness of a Dataflow Analysis for Memory Monitoring

Dara Ly
CEA List
Software Reliability and Security Lab
Gif-sur-Yvette, France
Université d’Orléans
LIFO EA 4022
Orléans, France
firstname.name@cea.fr

Nikolai Kosmatov
Julien Signoles
CEA List
Software Reliability and Security Lab
Gif-sur-Yvette, France
firstname.name@ceas.fr

Frédéric Loulergue
Northern Arizona University
School of Informatics Computing and Cyber Systems
Flagstaff, Arizona, USA
firstname.name@nau.edu

ABSTRACT
An important concern addressed by runtime verification tools for C code is related to detecting memory errors. It requires to monitor some properties of memory locations (e.g., their validity and initialization) along the whole program execution. Static analysis based optimizations have been shown to significantly improve the performances of such tools by reducing the monitoring of irrelevant locations. However, soundness of the verdict of the whole tool strongly depends on the soundness of the underlying static analysis technique. This paper tackles this issue for the dataflow analysis used to optimize the E-ACSL runtime assertion checking tool. We formally define the core dataflow analysis used by E-ACSL and prove its soundness.

KEYWORDS
dataflow analysis, memory monitoring, runtime assertion checking, proof of soundness, formal semantics, E-ACSL tool

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM Reference Format:

1 INTRODUCTION
Context. Memory related errors can provoke serious defects in software [5]. A study based on IBM MVS software [26] reports that about 50% of detected software errors were related to pointers and array accesses. Memory errors account for about 50% of reported security vulnerabilities [27]. This is particularly an issue for a programming language like C that is both the most commonly used for development of critical system software and one of the most poorly equipped with adequate protection mechanisms. The developer remains responsible for correct allocation and deallocation of memory, initialization of variables, pointer dereferencing and manipulation (using casts, offsets, etc.), as well as for the validity of indices in array accesses.

Among the most useful techniques for detecting and locating software errors, runtime assertion checking has become a widely-used programming practice [6]. Runtime checking of memory-related properties can be realized using systematic monitoring of memory operations. However, to do so efficiently is difficult, because of the large number of memory accesses in modern programs.

This paper more specifically considers the problem of memory monitoring of C programs for runtime assertion checking in the context of Frama-C [14], a platform for analysis of C code. Frama-C offers an expressive executable specification language E-ACSL, and comes with a runtime assertion checking plugin, also called E-ACSL [8]. The E-ACSL plugin takes as an input a C program P annotated with an E-ACSL specification, and outputs an instrumented program P’ having the following properties:

- if the execution of P satisfies all the properties expressed in the specification, the functional behavior of P’ is the same as that of the original program P;
- whenever the execution of P violates a property of the specification, the program P’ is aborted and an error is signaled.

This is done by translating the given specification into C code: the generated code implements a monitor verifying at runtime the conformance of the program with regard to the specification. In order to support memory-related E-ACSL annotations for pointers and memory locations (such as being valid, initialized, in a particular block, with a particular offset, etc.), the instrumented program P’ needs to keep track of relevant memory operations previously executed by the program.

Motivation. Previous work demonstrated that the performances of a runtime verification tool can be significantly improved using a preliminary static analysis step [11]. A dedicated dataflow analysis can be used to compute an (over-approximated) set of relevant memory locations that should be monitored. All operations changing the status (i.e. validity, initialization) of these locations should be tracked. All other locations are irrelevant: the monitoring tool does not need to monitor them. For the E-ACSL tool, this technique leads to important performance savings that vary between 60% and 73% [11].

However, this optimization can alter the correctness of the whole tool. Indeed, if the dataflow analysis is not sound, the monitoring of some relevant locations can be missed and their status can be wrongly identified. Since an E-ACSL annotation can check both for a pointer validity or its negation, this obviously can lead both
to false positives (an annotation failure is wrongly reported) and false negatives (a real annotation failure is not reported). For a tool like E-ACSL—whose goal is to check a formal specification and that is in particular used in combination with sound static analysis techniques—in this lack of soundness guarantees is an important issue. Thus, to ensure that runtime assertion checking always provides the correct verdict, the soundness of the underlying dataflow analysis should be formally established.

**Contributions.** The main contributions of this paper include a formal definition of a core dataflow analysis used by the E-ACSL tool and a proof of soundness of this analysis. Another contribution is an operational semantics of the monitor that naturally extends a standard operational semantics. We think that it is particularly suitable when reasoning about memory monitoring tools such as E-ACSL and could be reused in future work.

**Outline.** The paper is organized as follows. Section 2 motivates our work. Section 3 introduces the considered programming and specification languages, in particular their formal semantics. The dataflow analysis is defined in Section 4. Section 5 states and proves the soundness property of the analysis. Section 6 presents related work, and Section 7 gives concluding remarks and future work.

## 2 MOTIVATING EXAMPLE

Consider the toy C function at the left of Fig. 1. It contains a formal assertion guaranteeing that dereferencing pointer \( p \) is safe. This kind of assertions may be automatically generated by Framatic [14].

The E-ACSL tool [23] is able to take this annotated function as input in order to generate the instrumented version at the right of Fig. 1 (where we assume a 32-bit architecture). If executed, this instrumentation nicely stops the execution whenever the assertion is violated, preventing an incorrect memory access. To be able to check at runtime this kind of memory-related annotations, the code generated by E-ACSL records memory operations in the E-ACSL dynamic memory model linked to the instrumented program [28]. For instance, every allocation (resp. deallocation) is recorded through a call to function `store_block` (resp. `delete_block`), while every assignment is recorded through a call to function `full_inšt`.

However, even if the E-ACSL dynamic memory model is heavily optimized, these additional instructions significantly slow down the execution [28]. Fortunately, a complete instrumentation is almost never necessary. Here, for instance, only \( p \) needs to be monitored. Monitoring memory locations not relevant to \( p \) is not useful. In this example, when the annotation is evaluated, \( p \) is equal to the address of \( x \) because of the alias created by the first assignment.

E-ACSL implements an over-approximating dataflow analysis to compute what needs to be monitored. It takes care of aliasing and has already been proved capable to significantly improve the efficiency of the generated code [11]. Here, for instance, all the lines of code in gray may soundly be discarded, but the fact that \( x \) is properly allocated must still be monitored.

The goal of our work is to formalize the core of this analysis in order to prove its soundness. Indeed, if unsound, it would lead to incorrect verdicts by E-ACSL. Here, if the line `store_block(&x, 4)` is omitted, E-ACSL would incorrectly conclude that dereferencing pointer \( p \) is unsafe. Even worse, in other contexts, it could conclude that dereferencing a pointer would be safe while it is not, leading to security weaknesses.

## 3 LANGUAGE DEFINITION

E-ACSL takes as an input a C program annotated with E-ACSL specifications [8]. The whole language is far too complex to be handled formally in this paper. Rather, we present our analysis, its formalization and its proof of correctness on a relevant subset of C annotated with E-ACSL specifications. This subset is large enough to present the main issues tackled in our analysis.

**3.0.1 Syntax.** Figure 2 presents the syntax of our language. Expressions contain arithmetic expressions and pointer accesses (addresses and offset shifts), and are pure (i.e. side-effect free). Since our analysis focuses on memory management, we only detail memory values here.

Basic statements are assignments of an expression to a left-value (which can be either a variable or a pointer access), memory allocations and deallocations, and assertions. Compound statements are sequences, conditionals and loops.

Assertions evaluate terms of a propositional calculus with integer arithmetic and predicates over memory locations. Terms and predicates notably contain five built-in logic functions and predicates (detailed in Fig. 3) usable to express a rich set of memory-related properties [11].

**Execution Memory.** Since our language is a simple C-like language, we use a simplified CompCert memory model [16]. Indeed, the CompCert memory model was developed for the purpose of defining a semantics for C. In this model, the memory is viewed in an abstract manner as a collection of blocks. Physical placement of blocks and their relative position is not modeled. A block contains a range of valid offsets within it and is entirely defined by its bounds at allocation time. Consequently, a valid address is defined by an ordered pair \((b, \delta)\) of a block and an offset. In the following, \( \text{mem} \) denotes the type of memory states, while \( \text{block} \) denotes that of blocks. A memory state associates addresses to values. These values have the type \( \text{val} \) defined as follows:

\[
\text{val} :: = \text{Int}(n) \mid \text{Ptr}(b, \delta) \mid \text{Undef}
\]

Here, \( \text{Undef} \) corresponds to the initial value at any address that was not yet written to.

Four operations may be performed on a given memory state: allocation, deallocation, load and store, respectively defined through the operations `alloca`, `dealloca`, `load` and `store` below.

Assume a memory state \( M \), a block \( b \), three integers \( lo, hi, \delta \) and a value \( c \). Statement `alloca(M, lo, hi)` allocates a new block in \( M \), with lower bound \( lo \) (inclusive) and higher bound \( hi \) (exclusive), and returns the updated memory state with the new block. Statement `dealloc(M, b)` deallocates the block \( b \) in \( M \), returning the updated memory state where \( b \) has bounds \((0, 0)\), which makes it impossible to write to or read from. We model the fact that some operations...
null
The evaluation of expressions draws inspiration from Clight [3], the source language of CompCert, while statements have a small-step semantics akin to that of a While language, as in [19] for instance.

A noticeable trait of our semantics is that it is blocking: any execution ends immediately (cannot continue further) if an assertion is violated. This property is indeed the fundamental property of Frama-C and E-ACSL [7].

The semantics is expressed by means of six forms of evaluation judgments:

- \( M \models \text{lop} t \Rightarrow v \) evaluates a term \( t \) to a value \( v \) in a memory \( M \);
- \( M \models \text{op} e \Rightarrow P \) evaluates a predicate \( P \) to a truth value \( P \) in a memory \( M \);
- \( \langle s, M_1, \overline{M} \rangle \rightarrow \langle M_2, \overline{M}_2 \rangle \) evaluates a statement \( s \) in an execution memory \( M_1 \) and an observation memory \( \overline{M}_1 \), then ends the execution in an execution memory \( M_2 \) and an observation memory \( \overline{M}_2 \);
- \( \langle s_1, M_1, \overline{M}_1 \rangle \rightarrow \langle s_2, M_2, \overline{M}_2 \rangle \) also evaluates a statement \( s_1 \) in an execution memory \( M_1 \) and an observation memory \( \overline{M}_1 \), then continues the execution with the statement \( s_2 \) in an execution memory \( M_2 \) and an observation memory \( \overline{M}_2 \).

We consider only execution of well typed programs (even if the simple type system is omitted here), thus equipped with a typing environment \( \Gamma \). We also associate to a given program an environment \( E \) and initial memory states \( M_0 \) and \( \overline{M}_0 \). The environment \( E \) maps each variable of the program to a block in \( M_0 \), which is otherwise empty.

The evaluation of expressions and left-values is described in Figure 4. Left-values are terms at the left of an assignment, therefore their evaluation yields a memory location to be written to. As mentioned before, evaluating expressions does not alter memory: the expressions are pure. When evaluating a left-value as an expression, the left-value is resolved into a memory location, which is then read in the execution memory with a load. As specified in the C standard [10], reading an uninitialized value is an undefined behavior. Consequently, the evaluation is valid only if the retrieved value is not \texttt{Undefined}.

3.1 Operational semantics

This section introduces an operational semantics for our language. The evaluation of expressions draws inspiration from Clight [3], the source language of CompCert, while statements have a small-step semantics akin to that of a While language, as in [19] for instance.

A noticeable trait of our semantics is that it is blocking: any execution ends immediately (cannot continue further) if an assertion is violated. This property is indeed the fundamental property of Frama-C and E-ACSL [7].

The semantics is expressed by means of six forms of evaluation judgments:

- \( M \models \text{lop} l \Rightarrow (b, \delta) \) evaluates a left-value \( l \) to an address \( (b, \delta) \) in a memory \( M \);
- \( M \models e \Rightarrow v \) evaluates an expression \( e \) to a value \( v \) in a memory \( M \);
- \( M \models t \Rightarrow v \) evaluates a term \( t \) to a value \( v \) in a memory \( M \);
- \( M \models p \Rightarrow P \) evaluates a predicate \( P \) to a truth value \( P \) in a memory \( M \);
- \( \langle s, M_1, \overline{M} \rangle \rightarrow \langle M_2, \overline{M}_2 \rangle \) evaluates a statement \( s \) in an execution memory \( M_1 \) and an observation memory \( \overline{M}_1 \), then ends the execution in an execution memory \( M_2 \) and an observation memory \( \overline{M}_2 \);
- \( \langle s_1, M_1, \overline{M}_1 \rangle \rightarrow \langle s_2, M_2, \overline{M}_2 \rangle \) also evaluates a statement \( s_1 \) in an execution memory \( M_1 \) and an observation memory \( \overline{M}_1 \), then continues the execution with the statement \( s_2 \) in an execution memory \( M_2 \) and an observation memory \( \overline{M}_2 \).

We consider only execution of well typed programs (even if the simple type system is omitted here), thus equipped with a typing environment \( \Gamma \). We also associate to a given program an environment \( E \) and initial memory states \( M_0 \) and \( \overline{M}_0 \). The environment \( E \) maps each variable of the program to a block in \( M_0 \), which is otherwise empty.

The evaluation of expressions and left-values is described in Figure 4. Left-values are terms at the left of an assignment, therefore their evaluation yields a memory location to be written to. As mentioned before, evaluating expressions does not alter memory: the expressions are pure. When evaluating a left-value as an expression, the left-value is resolved into a memory location, which is then read in the execution memory with a load. As specified in the C standard [10], reading an uninitialized value is an undefined behavior. Consequently, the evaluation is valid only if the retrieved value is not \texttt{Undefined}.
The latter leads to a quite imprecise analysis meeting one of our primary concerns to keep the E-ACSL compilation time reasonably fast. Therefore, in the balance between precision and speed of the analysis, we favor speed most of the time. However, even if imprecise, this analysis allows E-ACSL to improve significantly the efficiency of the generated monitor by reducing the instrumentation [11].

The analysis computes for each program point a certain set of variables which defines the zones of memory that will be instrumented for runtime monitoring purposes. At some program point $l$, any block of memory reachable (through dereferencing and offset shifts) from one of the variables computed at $l$ is instrumented. Therefore, although the analysis is formally expressed in terms of variables, it may be useful to consider these variables as representing the set of all memory blocks reachable from them. Instrumenting all these blocks—when maybe only some of them would be strictly necessary—is one of the cases where we trade precision for speed in the analysis.

Reachability is formally defined by viewing memory as a graph where blocks are vertices, and pointers define edges. Following this intuition, we define the relation $\rightarrow_M$ between blocks of a memory $M$ as “given the memory state $M$, $b \rightarrow_M b'$ if there is one value in $b$ that is a pointer to $b'$ with a valid offset”. This definition is
formally expressed as follows:

\[ b \mapsto_M b' \]  
if \( \exists \delta, \delta' \in \mathbb{Z} \) such that

\[ M \models (b, \delta) \]
\[ M \models (b', \delta') \]
\[ \text{load}(M, b, \delta) = \text{Some}(\text{Ptr}(b', \delta')) \]

We define \( \mapsto_M^* \) to be the reflexive, transitive closure of \( \mapsto_M \) and use it to define the set of reachable blocks from a memory value. If \( a \) is a value such that \( M \models a \Rightarrow \text{Ptr}(b_a, \delta_a) \), the set of reachable blocks from \( a \) in \( M \) is defined by:

\[ \mathcal{R}_M(a) = \{ b \in \text{Blocks} \mid b_a \mapsto_M^* b \} \]

### 4.1 Definition

The analysis is classically defined on a labeled Control Flow Graph (CFG) where basic blocks are single statements (skip, assignment, allocation, deallocation, and assertion) and tests (guards of conditionals and loops).

For a given program \( s \) we note \( I(s) \) (resp. \( F(s) \)) the label of the CFG’s initial block (resp. the labels of the set of final blocks). In the example program \( s \) in Figure 8, we have \( I(s) = 1 \) and \( F(s) = \{ 2 \} \).

```
\[
f = 1;  
\text{while } (0 < n) \{  
f = f \ast n;  
n = n - 1;  
\}
\]
```

**Figure 8: CFG of a simple program.**

We define the analysis as a pair \( (\text{live}_{in}, \text{live}_{out}) \) of functions mapping each label of the CFG to a set of program variables. They are assumed to be the least solution of the constraint system presented in Figure 9. As we are solely interested in the soundness of the analysis, we do not study how this solution is computed.

For a given label \( l \), \( \text{live}_{in}(l) \) and \( \text{live}_{out}(l) \) are the set of live variables respectively at the entry and the exit of the block labeled \( l \). (To avoid any risk of confusion between labels and left values, in the remainder of the paper, we denote a left value by \( l.v. \)) Like a liveness analysis, the analysis is **backward**: the information associated to a block is computed from that of its successors in the CFG. Like a liveness analysis, it is **over-approximating**: it combines together information originating from multiple branches using the union operator.

The function \( \text{gen} \) defines how statements generate live variables; while most statements simply transfer the existing liveness information, new liveness information is generated in two cases.

First, predicates generate the base variables of their arguments, as described by the recursive functions \( \rho \) and \( \theta \). The base variable of a memory value \( a \) is the first variable appearing in \( a \)’s syntactical structure. It is, very concretely, the first variable written, in left-to-right order, when writing the source code for \( a \). The base variable is computed by the function \( \text{base} \), inductively defined as following:

```
\[
\text{base}(x) = x  
\text{base}(*a) = \text{base}(a)  
\text{base}(\&x) = x  
\text{base}(\&(*a)) = \text{base}(a)  
\text{base}(a + + e) = \text{base}(a).  
\]
```

Second, in case of an assignment \( l.v = e \) where \( e \) is a pointer, if the left-value \( l.v \) is under monitoring, the assigned pointer \( e \) is generated. The condition "\( l.v \) is under monitoring" is formally expressed by the proposition \( \exists x \in \text{live}_{out}(l), \&x \mapsto_*^\mathcal{A} l.v \). Indeed \( l.v \) is monitored if and only if it is reachable from one of variables in \( \text{live}_{out}(l) \). Since reachability cannot be computed statically, our analysis uses as a parameter a points-to analysis \( \mathcal{A} \) computing which values of the program may point to which others. We use the result of \( \mathcal{A} \) in the form of the relation \( \mapsto_*^\mathcal{A} \), \&x \mapsto_*^\mathcal{A} l.v \) means that there is a path from \&x to \( l.v \) in the points-to graph.

### 4.2 Correctness

Since our analysis aims at determining which blocks must be monitored by \( E\text{-ACSL} \), being correct means that instrumenting only blocks resulting from the analysis must not alter the evaluation of assertions in the output program. Therefore, an informal way to state the correctness could be: "at any program point, altering the observation memory outside of the domain defined by the analysis does not alter the program’s behaviour".

In order to further formalize this idea, we need to define notions for comparing two execution traces of a program. In particular, we want to be able to compare two observation memories, and state that they are the same on a certain set of blocks. Let \( M \) and \( M' \) be observation memories, and \( B \) be a set of blocks. We say that \( M \) and \( M' \) are equivalent on \( B \), and we note \( M \sim_B M' \) if they "have the same values on the blocks in \( B \)"; that is:

\[
\forall b \in B, \left\{ \begin{array}{l}
\forall \delta \in \mathbb{Z}, \quad M \models (b, \delta) \Leftrightarrow M' \models (b, \delta)
\end{array} \right.
\]

\[
\forall \delta \in \mathbb{Z}, \quad \text{bounds}(M, b, \delta) = \text{bounds}(M', b, \delta)
\]

Following Nielson et al. [19] the theorem is composed of two statements: one describes the case of an evaluation step that terminates the program, and the other one the case where it continues thereafter.

**Immediate Termination.**

**Theorem 4.1.** Let \( (s, M_1, M'_1) \rightarrow (M_2, M'_2) \) be an execution step, and let \( M'_1 \) be an observation memory such that \( M'_1 \sim_B M'_1 \), with \( B_1 = \bigcup_{s \in \text{live}_{out}(I(s))} \mathcal{R}_M(s) \).

Then there exists \( M'_2 \) such that \( (s, M_1, M'_1) \rightarrow (M_2, M'_2) \) and \( M'_2 \sim_{B_2} M'_2 \), with \( B_2 = \bigcup_{s \in \text{live}_{out}(I(s))} \mathcal{R}_M(s) \).

**Continued Evaluation.**

**Theorem 4.2.** Let \( (s_1, M_1, M'_1) \rightarrow (s_2, M_2, M'_2) \) be an execution step, and let \( M'_1 \) be an observation memory such that \( M'_1 \sim_B M'_1 \), with \( B_1 = \bigcup_{s \in \text{live}_{out}(I(s_1))} \mathcal{R}_M(s_1) \).

Then there exists \( M'_2 \) such that \( (s_1, M_1, M'_1) \rightarrow (s_2, M_2, M'_2) \) and \( M'_2 \sim_{B_2} M'_2 \), with \( B_2 = \bigcup_{s \in \text{live}_{out}(I(s_2))} \mathcal{R}_M(s_2) \).

The following diagram gives a visual intuition of the correctness property, here in the case of immediate termination:

\[
(s, M_1, M'_1) \rightarrow (M_2, M'_2)
\]

\[
\sim_{B_1} \sim_{B_2}
\]

\[
(s, M_1, M'_1) \rightarrow (M_2, M'_2)
\]
5 PROOF OUTLINE
The proof is by induction on the structure of the evaluation derivation. The most difficult cases are statements that perform memory operations, namely assignment, allocation and deallocation. The other cases are straightforward; the proof relying mainly on the properties of least fixed point of \((\text{live}_{\text{in}}, \text{live}_{\text{out}})\).

We use two kinds of auxiliary lemmas. Some describe properties of the equivalence relation \(\sim\) such as the fact that it is an equivalence relation, how to combine equivalences between the same two memories on multiple domains, or conversely how two memories equivalent on a given domain remain equivalent when the same memory operation is performed on both. Others describe the evolution (or invariance) of reachable blocks sets when memory operations are performed.

5.1 Lemmas.

Lemma 5.1 (Domain Restriction). If \(\mathcal{M} \sim_X \mathcal{M}'\) and \(Y \subseteq X\) then \(\mathcal{M} \sim_Y \mathcal{M}'\).

Lemma 5.2 (Domain Union). If \(\mathcal{M} \sim_X \mathcal{M}'\) and \(\mathcal{M} \sim_Y \mathcal{M}''\) then \(\mathcal{M} \sim_{X \cup Y} \mathcal{M}'\).

Lemma 5.3 (Equivalence Preservation (initialization)). For a given domain \(\mathcal{B}\), two equivalent memories remain equivalent when the same initialize operation is performed on both.

Lemma 5.4 (Reachability from base address). Any block reachable from a given memory location \(a\) is also reachable from its base address:

\[
\forall a \in \text{MemVal}, \forall \mathcal{M}, \mathcal{R}_\mathcal{M}(a) \subseteq \mathcal{R}_\mathcal{M}(x) \quad \text{base}(a)
\]

Lemma 5.5 (Writing outside of a reachable set does not modify it). Let \(x\) be a variable. For a given memory state \(M_1\), if \(\text{store}(M_1, b, \delta, v) = \text{Some}(M_2)\) with \(b \notin \mathcal{R}_\mathcal{M}_1(x)\), then \(\mathcal{R}_\mathcal{M}_1(x) = \mathcal{R}_\mathcal{M}_1(\&x)\).

Lemma 5.6 (Writing an integer does not modify reachable set). Suppose that \(M_1\) and \(M_2\) are two memory states in the trace of a well-typed program, such that \(\text{store}(M_1, b, \delta, \text{Int}(n)) = \text{Some}(M_2)\). Then for any variable \(x\), the set of reachable blocks from \(x\) is the same in \(M_1\) and \(M_2\):

\[
\forall x \in E, \mathcal{R}_{M_1}(\&x) = \mathcal{R}_{M_2}(\&x).
\]

Lemma 5.7 (Maximum extension of a reachable set). If \(M_1\) and \(M_2\) are two memory states such that \(\text{store}(M_1, b, \delta, \text{Ptr}(b_1, \delta_1)) = \text{Some}(M_2)\), then the following inclusion is verified: \(\mathcal{R}_{M_1}(a) \subseteq \mathcal{R}_{M_1}(a) \cup \{b^\prime | b^\prime \mapsto_{M_1} b^\prime\}\).

5.2 Case of Assignments.
We detail the main case of the induction: that of assignments. Given an execution step \(\langle [lv = e^\prime], M_1, \mathcal{M}_1\rangle \rightarrow (M_2, \mathcal{M}_2)\) and an observation memory \(\mathcal{M}_1\) such that \(\mathcal{M}_1 \sim_{\mathcal{B}_1} \mathcal{M}_1\), we want to show the existence of \(\mathcal{M}_2\) such that:

- \(\langle [lv = e^\prime], M_1, \mathcal{M}_1\rangle \rightarrow (M_2, \mathcal{M}_2)\) is a valid execution step.
- \(\mathcal{M}_2 \sim_{\mathcal{B}_2} \mathcal{M}_2\).

\(\mathcal{B}_1\) and \(\mathcal{B}_2\) are defined as:

\[
\mathcal{B}_1 = \bigcup_{v \in \text{live}_{\text{in}}(l)} \mathcal{R}_{M_1}(\&x), \quad \mathcal{B}_2 = \bigcup_{v \in \text{live}_{\text{out}}(l)} \mathcal{R}_{M_1}(\&x).
\]

The execution step has the following form:

\[
M_1 \stackrel{e}{\rightarrow} v \quad \text{store}(M_1, b_1, \delta_1, v) = \text{Some}(M_2) \quad M_1 \stackrel{lv}{\rightarrow} \quad \text{initialize}(M_1, b_1, \delta_1) \quad \mathcal{M}_1 \rightarrow \mathcal{M}_2
\]

We define \(\mathcal{M}_2 = \text{initialize}(\mathcal{M}_1, b_1, \delta_1)\), and prove that \(\mathcal{M}_2 \sim_{\mathcal{B}_2} \mathcal{M}_2\). We distinguish between the case where the block written to is monitored \(b_1 \in \mathcal{B}_2\) and the case where it is not.

In both cases we make use of the following relation:

\[
\mathcal{M}_2 \sim_{\{b\} \circ \mathcal{M}_1} \sim_{\mathcal{B}_1} \mathcal{M}_1 \sim_{\{b\} \circ \mathcal{M}_2},
\]

in which the first and last equivalences are derived from Lemma 5.3 and the middle one is an hypothesis of our theorem.

5.2.1 case \(b_1 \notin \mathcal{B}_2\). Let us prove that \(\mathcal{B}_2 \subseteq \mathcal{B}_1\). Since \(\forall l, \text{live}_{\text{in}}(l) \supseteq \text{live}_{\text{out}}(l)\), it suffices to show that \(\forall x \in \text{live}_{\text{out}}(l), \mathcal{R}_{M_1}(\&x) = \mathcal{R}_{M_2}(\&x)\). Since for any of these sets we have \(b_1 \notin \mathcal{R}_{M_2}(\&x)\) we can apply Lemma 5.5 to conclude.
We can now rewrite the assumption \( b_1 \not\in B_2 \) as \( B_2 \subseteq \{ b_1 \}^c \) so that all domains involved in equation 1 are superset of \( B_2 \). The conclusion follows from Lemma 5.1.

5.2.2 case \( b_1 \in B_2 \). Using Lemma 5.2, we consider \( B_2 \) as the (disjoint) union of \( B_2 \setminus \{ b_1 \} \) and \( \{ b_1 \} \), and we show the equivalence on each of these subdomains. Here we have to consider the type of the assigned expression.

\[ \text{subcase int.} \] If the expression has the type int, we can use Lemma 5.6 to prove that \( B_2 \subseteq B_1 \), using the same method as previously. Applying Lemma 5.1 to 3 then yields \( M_2 \sim_{B_2 \setminus \{ b_1 \}} M_2' \).

Besides, since \( b_1 \in B_2 \subseteq B_1 \) and \( M_1 \sim B_1 \), \( M_2 \) have more specifically \( M_1 \sim \{ b_1 \} M_1' \). By Lemma 5.3 we can conclude that \( M_2 \sim_{\{ b_1 \}} M_2' \).

\[ \text{subcase } \tau^\ast. \] If the expression is a pointer \( v = \text{Ptr}(b_2, \delta_e) \) for some \( (b_2, \delta_e) \), we can use Lemma 5.7 to approximate the evolution of reachable sets: \( R_{M_1}(a) \subseteq R_{M_1}(a) \cup R_{M_1}(e) \). We consider the union of these inclusions:

\[ B_2 \subseteq \bigcup_{x \in \text{live}_{out}(l)} R_{M_1}(\text{in } x) \cup R_{M_1}(\text{out } x). \]

Let us prove that terms of the right side are both subsets of \( B_1 \). For the first one, using the definition of \( B_2 \) and the inclusion \( \text{live}_{\text{out}}(l) \subseteq \text{live}_{\text{in}}(l) \) we can write

\[ \bigcup_{x \in \text{live}_{\text{out}}(l)} R_{M_1}(\text{in } x) \subseteq \bigcup_{x \in \text{live}_{\text{out}}(l)} R_{M_1}(\text{out } x) = B_1. \]

For the second one, we use Lemma 5.4 in conjunction with the fact that \( \text{base}(e) \in \text{live}_{out}(l) \). This corresponds to the first case in the definition of the generation function \( \text{gen}(\{ l v = e \}) \), which is defined by the condition: \( \exists x \in \text{live}_{out}(l), \text{lv} \mapsto \text{A}^* l v \). This condition is necessarily verified here: since \( b_1 \in B_2 \), by definition of \( B_2 \) there is some \( x \in \text{live}_{out}(l) \) such that \( b_1 \in R_{M_1}(\text{in } x) \). Let \( b_x \) be the block allocated for \( x \). By definition of \( R_{M_1}(l) \), \( b_x \mapsto \text{A}^* b_1 \). Finally, applying the correctness property of the points-to analysis to this relation yields the expected result: \( x \mapsto \text{A}^* \text{lv} \) with \( x \in \text{live}_{out}(l) \).

Now we can deduce from equation 2 that \( B_2 \subseteq B_1 \), and conclude with the same arguments as for the subcase int.

6 RELATED WORK

The E-ACSL plugin aims at performing runtime assertion checking for C code, specified in a rich specification language E-ACSL. Being a part of the Frama-C framework, it can be combined with other analyses for quite unique usages. For example, runtime assertion checking can be used to assist the user in case of proof failures during deductive verification [21].

E-ACSL can also be used as a dynamic bug finder, see for example [25, 29] for a recent comparison of such sanitizers. These tools implement various techniques, in particular concerning the storage of metadata necessary for the analysis. There are roughly three main techniques.

A shadow memory is in linear correspondence with the program memory. AddressSanitizer [22] stores 1 shadow byte per aligned 8-byte sequences in the application. Shadow memory does not incur a large runtime overhead but its adaptation can be quite tricky when richer metadata is necessary, as it is the case for E-ACSL [28].

Another technique used to attach additional metadata to each pointer is the so-called fat pointers technique [1, 20]. There is significant overhead in this case, and instrumented code cannot directly be linked with non instrumented code because of differences in memory layout.

Finally, metadata can be stored in a separate data structure [17]. In a previous version of E-ACSL, metadata was stored in a Patricia trie [15]. This technique allows the instrumented code to be linked with non instrumented code and to store richer metadata but execution time overhead may be significant.

Mehlich [17] mentions that “the current implementation of CheckPointer could be improved by using static analysis”. Jakobsson et al. [11] improved E-ACSL by a static analysis (that we formalize in this paper) in order to avoid monitoring memory locations that do not need to be monitored. They combine shadow memory (used for memory locations that only require validity or initialization checks) with a Patricia trie (for locations requiring more complex checks). Other sanitizers rely on static analysis to improve the performances of runtime checking [18, 24, 30]. The work that is the closest to ours in intent and using a proved analysis is CCured [18], but the techniques are different. CCured is based on fat pointers, their static analysis is a type system, and only memory safety is checked. The correctness of the analysis is ensured by a theorem stating that the execution of an instrumented program can either terminate correctly or stop because the instrumentation has detected an error, ruling out an incorrect termination because of an invalid memory access. The proof assumes that there is no dynamic memory allocation in the program.

Our ultimate goal is to be able to extract a correct E-ACSL-like plugin from a Coq development. This goal necessitates an implementation and verification in Coq of the presented static analysis extended to CompCert C. Our formalization follows the CompCert memory model [16]. There are several works on the verification of static analyses with a proof assistant. We only briefly present some papers related to the Coq proof assistant. The CompCert compiler itself contains static analysis for the sake of optimization [2], including a generic implementation of the Kildall algorithm [13]. This is however related to the RTL language of CompCert. Like Verasco [12] the presented memory analysis will rather be conducted at the level of the Cminor language. Unlike Verasco, we plan to directly implement the analysis as a dataflow analysis rather than as an abstract interpretation. [9] is a generic formalization of a solver that can be used for static analysis. However the algorithms are formalized as relations and cannot be extracted to executable code. [4] is an extractable dataflow analysis in Coq in the context of Java.

7 CONCLUSION AND PERSPECTIVES

We have formalized the dataflow analysis underlying a major optimization of the E-ACSL tool for runtime assertion checking and memory debugging. This optimization was shown to bring 60% to 73% performance savings [11]. The analysis is proved correct with regard to an operational semantics which reflects the use of a second memory store to monitor the program memory state.
Soundness of a Dataflow Analysis for Memory Monitoring

HILT 2017, November 5-6, 2018, Boston, MA, USA

We believe that making the memory monitoring explicit in the language semantics (via the so-called observation memory) significantly eases formal reasoning about optimizations performed at this level, and can be beneficial in other contexts. We plan to further explore this notion in our ongoing effort towards a mechanized formalization of E-ACSL. This future work can be carried out in three main directions.

First, the core language presented here can be extended to be more representative of the real E-ACSL language, making the formalization wider. Features such as structures and function calls strongly influence the memory layout of program memory space, so this would most probably need some adaptation.

Second, this work is still mostly pen-and-paper. We could make the formalization deeper by porting it to the Coq proof assistant.

Finally, in this work we have assumed that E-ACSL gives the described semantics to an annotated program by translating annotations into C code, but this translation is not formalized yet. Since this translation is the core E-ACSL, verifying it has the highest priority in the global formalization effort of E-ACSL.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their helpful comments and feedbacks. This work was partly supported by project VESSEDIA, which has received funding from the European Union’s Horizon 2020 Research and Innovation Program under grant agreement No 731453. The work of the first author was partially funded by a Ph.D. grant of the French Ministry of Defence.

REFERENCES

A APPENDIX

This appendix contains a detailed soundness proof of our dataflow analysis. Section B recaps the formal statement of the theorem and presents the global structure of the proof. Section C lists the lemmas used along the proof; they are stated informally, and only the main idea of their proof is given. Please note that they numbering differs from the one used in the article body. The remaining sections detail all the cases of the induction:

- memory operations, which represent the main difficulty of this proof, are tackled in Section D
- Section E includes the others cases of execution in which the program immediately terminates
- Section F handles control-flow related statements.

B THEOREM STATEMENT AND PROOF STRUCTURE

Immediate Termination. Let \( s, M_1, M_1' \) be an execution step, and let \( M_1 \) be an observation memory such that \( M_1 \sim b_1 M_1' \), with \( b_1 = \bigcup_{x \in \text{live}_{\text{ref}}(f(x))} R_{M_1}(\&x) \).

Then there exists \( M_2' \) such that \( s, M_2, M_2' \) be an execution step, and let \( M_2 \) be an observation memory such that \( M_2 \sim b_2 M_2' \), with \( b_2 = \bigcup_{x \in \text{live}_{\text{ref}}(f(x))} R_{M_2}(\&x) \).

Continued Evaluation. Let \( s_1, M_1, M_1' \) be an execution step, and let \( M_1 \) be an observation memory such that \( M_1 \sim b_1 M_1' \), with \( b_1 = \bigcup_{x \in \text{live}_{\text{ref}}(f(x))} R_{M_1}(\&x) \).

Then there exists \( M_2' \) such that \( s_2, M_2, M_2' \) be an execution step, and let \( M_2 \) be an observation memory such that \( M_2 \sim b_2 M_2' \), with \( b_2 = \bigcup_{x \in \text{live}_{\text{ref}}(f(x))} R_{M_2}(\&x) \).

B.1 Structure of the Proof

The proof is by induction on the structure of the evaluation derivation. The most difficult cases are statements that perform memory operations, namely assignment, allocation and deallocation. The other cases are straightforward.

C INTERMEDIATE LEMMAS

C.1 Properties of \( \sim_B \)

Lemma C.1 (Equivalence Relation). \( \forall X, \sim_{X} \) is an equivalence relation on observation memories.

Lemma C.2 (Domain Restriction). If \( M_1 \sim X M_2 \) and \( Y \subseteq X \) then \( M_1 \sim Y M_2 \).

Lemma C.3 (Domain Union). If \( M_1 \sim X M_2 \) and \( M_1 \sim Y M_2 \) then \( M_1 \sim X \cup Y M_2 \).

C.2 Properties of \( \&_{\text{base}}(\cdot) \)

Variable Address. The base address of any memory location is the address of a given variable: \( \forall a \exists x, \text{base}(a) = \&x \).

Consequently, if we consider a trace of a given program, a base address always evaluates to the same block ; that is, the block pointed to by the base address does not depend on the point of the trace where the evaluation takes place.

Lemma C.7 (Reachability from Base Address). Any block reachable from a given memory location \( a \) is also reachable from its base address:

\[ \forall a \in \text{Memval}, \forall M, R_{M}(a) \subseteq R_{M}(\&\text{base}(a)). \]

Proof: by induction on the structure of \( a \).

C.3 Evolution of \( R_M(\&x) \) when writing in \( \text{M} \)

The following lemmas describe the behavior of reachable blocks sets under memory updates. They are best understood by viewing the memory as a directed graph with blocks as vertices and pointers defining edges. From this point of view, the fact that successors of a block are always in the same reachable set as this block itself makes the proof of these lemmas straightforward.

Lemma C.8 (Writing outside of a reachable set does not modify it). Let \( x \) be a variable. For a given memory state \( M_1 \), if \( \text{store}(M_1, b, \delta, \nu) = \text{Some}(M_2) \) with \( b \notin R_{M_1}(\&x) \), then

\[ R_{M_1}(\&x) = R_{M_2}(\&x). \]

Lemma C.9 (Writing an integer does not modify reachable set). Suppose that \( M_1 \) and \( M_2 \) are two memory states in the trace of a well-typed program, such that \( \text{store}(M_1, b, \delta, \text{Int}(n)) = \text{Some}(M_2) \).

Then for any variable \( x \), the set of reachable blocks from \( x \) is the same in \( M_1 \) and \( M_2 \):\n
\[ \forall x \in E, R_{M_1}(\&x) = R_{M_2}(\&x). \]

Lemma C.10 (Maximum extension of a reachable set). If \( M_1 \) and \( M_2 \) are two memory states such that

\[ \text{store}(M_1, b, \delta, \text{Ptr}(b, \delta)) = \text{Some}(M_2), \]

then the following inclusion is verified:

\[ R_{M_1}(a) \subseteq R_{M_2}(a) \cup \{ b' \mid b_{\nu} \mapsto^{*}_{M_1} b' \}. \]

D MEMORY OPERATIONS

Let \( s, M_1, M_1' \rightarrow (M_2, M_2') \) be an execution step, and let \( M_1' \) be an observation memory such that \( M_1 \sim b_1 M_1' \), with \( b_1 = \bigcup_{x \in \text{live}_{\text{ref}}(f(x))} R_{M_1}(\&x) \).

We want to show that there exists \( M_2' \) such that \( s, M_1, M_1' \rightarrow (M_2, M_2') \) and \( M_2 \sim b_2 M_2' \), with \( b_2 = \bigcup_{x \in \text{live}_{\text{ref}}(f(x))} R_{M_2}(\&x) \).

D.1 Assignment

The execution step has the form:
We define \( \overline{M_2} = \text{initialization}(M_2, b_1, \delta_1) \) and prove that \( \overline{M_2} \sim \overline{B_2} \overline{M_2} \). We distinguish the case where the block written to is monitored \( (b_1 \in B_2) \) and the case where it is not.

In each of these cases we make use of the following assumption:

\[
\overline{M_2} \sim (b_1) \overline{M_1} \sim \overline{B_1} \overline{M_1} \sim (b_1) \overline{M_2} \tag{3}
\]

in which the first and last equivalences are derived from Lemma C.4 and the middle one is an hypothesis of our theorem.

D.1.1 case \( b_1 \notin B_2 \). Let us prove that \( B_2 \subseteq B_1 \). Since \( l, \text{liveout}(l) \supseteq \text{liveout}(l) \), it suffices to show that \( \forall x \in \text{liveout}(l), R_{M_1}(\langle kx \rangle) = R_{M_1}(\langle kx \rangle) \). Since for any of these sets we have \( b_1 \notin R_{M_1}(\langle kx \rangle) \) we can apply Lemma C.8 to conclude.

We can now rewrite the assumption \( b_1 \notin B_2 \) as \( B_2 \subseteq B_1 \) so that the all domains involved in 3 are supersets of \( B_2 \). The conclusion follows from Lemma C.2.

D.1.2 case \( b_1 \in B_2 \). Using Lemma C.3, we consider \( B_2 \) as the (disjoint union of \( B_2 \setminus \{ b_1 \} \) and \( \{ b_1 \} \), and we show the equivalence on each of these subdomains. Here we have to consider the type of the assigned expression.

subcase \( \text{int} \). If the expression has the type \( \text{int} \), we can use Lemma C.9 to prove that \( B_2 \subseteq B_1 \), using the same method as previously. Applying Lemma C.2 to 3 then yields \( \overline{M_2} \sim \overline{B_2} \{ b_1 \} \overline{M_2} \).

Besides, since \( b_1 \in B_2 \subseteq B_1 \) and \( M_1 \sim \overline{B_1} \overline{M_1} \) we have more specifically \( \overline{M_1} \sim \overline{B_1} \overline{M_1} \). By Lemma C.4 we can conclude that \( \overline{M_2} \sim (b_1) \overline{M_2} \).

subcase \( \text{ptr} \). If the expression is a pointer \( v = \text{ptr}(b_2, \delta_2) \) for some \( (b_2, \delta_2) \), we can use Lemma C.10 to approximate the evolution of reachable blocks sets: \( R_{M_1}(a) \subseteq R_{M_1}(a) \cup R_{M_1}(e) \). We consider the union of these inclusions:

\[
B_2 \subseteq \bigcup_{x \in \text{liveout}(l)} R_{M_1}(\langle kx \rangle) \cup R_{M_1}(e) \tag{4}
\]

Let us prove that terms of the right side are both subsets of \( B_1 \).

For the first one, using the definition of \( B_1 \) and the inclusion \( \text{liveout}(l) \subseteq \text{livein}(l) \) we can write

\[
\bigcup_{x \in \text{liveout}(l)} R_{M_1}(\langle kx \rangle) \subseteq \bigcup_{x \in \text{livein}(l)} R_{M_1}(\langle kx \rangle) = B_1.
\]

For the second one, we use Lemma C.7 in conjunction with the fact that \( \text{base}(e) \in \text{livein}(l) \). This corresponds to the first case in the definition of the generation function \( gen([lv \leftarrow e]) \), which is defined by the condition: \( \exists x \in \text{liveout}(l), kx \to^{x} lv \). This condition is necessarily verified here: since \( b_1 \in B_2 \), by definition of \( B_2 \) there is some \( x \in \text{liveout}(l) \) such that \( b_1 \in R_{M_1}(\langle kx \rangle) \). Let \( b_2 \) be the block allocated for \( x \). By definition of \( R_{M_1}(e) \), \( b_2 \to^{x} b_2 \). Finally, applying the correctness property of the points-to analysis to this relation yields the expected result: \( x \to^{x} lv \) with \( x \in \text{liveout}(l) \).

Now we can deduce from 4 that \( B_2 \subseteq B_1 \), and conclude with the same arguments as for the subcase int.

### D.2 Allocation

D.2.1 Equivalence on intermediate state. Since \( M_1 \) is simply \( M_1 \) with a new block and no pointer to this block, reachable sets are the same in both memories. Consequently,

\[
\forall x \in \text{livein}(l), R_{M_1}(\langle kx \rangle) = R_{M_1}(\langle kx \rangle)
\]

and thus \( B_1 = B_1 \).

Since \( M_1 \sim B_1 \overline{M_1} \) and the same allocation is performed on both memories, by Lemma C.5 the equivalence is preserved: \( M_1 \sim \overline{B_1} \overline{M_1} \).\( B_1 \) and \( B_1 \) being equal, the conclusion follows.

D.2.2 Equivalence on final state. This setting is very similar to the case of a pointer assignment (which is only natural since a pointer assignment is indeed performed after the allocation of a new block).

We want to prove \( M_2 \sim B_2 \overline{M_2} \), assuming the equivalence on the intermediate state: \( M_1 \sim \overline{B_1} \overline{M_1} \), \( b \) being a freshly allocated block in both \( M_1 \) and \( M_1 \), we have \( M_1 \sim (b) \overline{M_1} \), which means that we can extend the equivalence domain by using Lemma C.3:

\[
M_1 \sim B_1 \cup (b) \overline{M_1}.
\]

Case \( b_1 \notin B_1 \). Then by Lemma C.8, \( \forall x \in \text{livein}(l), R_{M_1}(\langle kx \rangle) \) is unaffected by the write at \( b_1 \), so \( B_1 = B_2 \). We can thus use Lemma C.4 to transfer the equivalence to the next program state: \( M_2 \sim B_2 \cup (b) \overline{M_2} \). To conclude, we use the domain equality \( B_1 = B_2 \) and restrict the domain back to \( B_2 \) by Lemma C.2.

Case \( b_1 \in B_1 \). By Lemma C.10, \( \forall x \in \text{livein}(l), R_{M_1}(\langle kx \rangle) \subseteq R_{M_1}(\langle kx \rangle) \cup \{ b \} \). Considering the union on all \( x \in \text{livein}(l) \) we get: \( B_2 \subseteq B_1 \cup \{ b \} \). Conclusion follows by Lemma C.4.
D.3 Deallocation

\[
M_1 \models_v a \Rightarrow (b, 0)
\]
\[
\text{free}(M_1, b) = \text{Some}(M_2) \quad \text{delete\_block}(M_1, b) = M_2
\]

\[\langle \text{free}(a); \mid_1, M_1 \rangle \rightarrow (M_2, M_2)\]

\(M_2\) is obtained by deleting one block from \(M_1\) so the connectivity between blocks in \(M_2\) is necessarily lower than in \(M_1\). More formally, for any variable \(x, R_{M_1}(x) \subseteq R_{M_2}(x)\). Since \(\text{live}_{in}(I) \supseteq \text{live}_{out}(I)\) by definition of the transfer function, this implies \(B_1 \supseteq B_2\). Therefore, by domain restriction (Lemma C.2), \(M_1 \sim B_1 M_1'\). Conclusion follows from Lemma C.6.

E OTHER CASES OF IMMEDIATE TERMINATION

E.1 Skip

\[
\text{eval-skip}\quad \langle \text{skip}, M_1, \overline{M_1} \rangle \rightarrow (M_1, \overline{M_1})
\]

The memory state and the equivalence domain are the same before and after the statement execution, so the equivalence is trivially preserved.

E.2 End of a loop

\[
\text{eval-whileEnd}\quad M \models_v e \Rightarrow \text{Int}(0)
\]
\[
\text{eval-whileCont}\quad \langle \text{while e} \rangle s, M, \overline{M} \rightarrow (M, \overline{M})
\]

By definition of the functions \(\text{live}_{in}, \text{live}_{out}\), \(\text{live}_{in}(I) \supseteq \text{live}_{out}(I)\). Since the memory is unchanged, this implies \(B_1 \supseteq B_2\). Conclude with Lemma C.2.

E.3 Assertion

\[
\text{eval-assert}\quad M \models_v p \Rightarrow \top
\]
\[
\langle \ast \rangle \overline{e} \text{ assert } \langle p \rangle; * / M, \overline{M} \rightarrow (M, \overline{M})
\]

\(\text{live}_{in}(I) \supseteq \text{live}_{out}(I) = \text{live}_{in}(I(s))\) by definition of the analysis. Since the memory is unchanged, this implies \(B_1 \supseteq B_2\) and we conclude with Lemma C.2.

F CASES OF CONTINUED EXECUTION

Let \(<s_1, M_1, \overline{M_1}>\) and \(<s_2, M_2, \overline{M_2}>\) be an execution step, and let \(\overline{M_1}'\) be an observation memory such that \(\overline{M_1} \sim B_1 \overline{M_1}'\), with \(B_1 = \cup_{x \in \text{live}_{in}(I(s_1))} R_{M_1}(x)\).

We want to show that there exists \(\overline{M_2}'\) such that \(<s_1, M_1, \overline{M_1}>\) \(\sim (s_2, M_2, \overline{M_2}')\), with \(B_2 = \cup_{x \in \text{live}_{in}(I(s_2))} R_{M_2}(x)\).

The case addressed in this section is that of execution steps which do not terminate immediately; these cases correspond to the statements defining the program’s control flow. Since these statements do not modify any of the memory stores, the reachable block sets are the same before and after the statement execution.

Therefore, in order to prove some domain inclusion \(B_1 \supseteq B_2\), it suffices to prove the corresponding live variable sets inclusion \(\text{live}_{out}(I(s_1)) \supseteq \text{live}_{out}(I(s_2))\). In all the following cases, this inclusion is a direct consequence of the relation between the dataflow analysis result (\(\text{live}_{in}, \text{live}_{out}\)) and the CFG structure, or in other words the fact that \(\text{live}_{in}, \text{live}_{out}\) is the least fixed point of the system defining it.

Dara Ly, Nikolai Kosmatov, Julien Signoles, and Frédéric Loulergue

F.1 Sequence

There are two subcases: either the first statement is a single, atomic statement (assignment, allocation…), or it is a composed statement (loop, sequence, conditional…).

F.1.1 Subcase eval-seqEnd.

\[
\text{eval-seqEnd}\quad \langle s_1, M_1, \overline{M_1} \rangle \rightarrow (s_2, M_2, \overline{M_2})
\]
\[
\langle s_1 s_2, M_1, M_1 \rangle \rightarrow (s_2, M_2, \overline{M_2})
\]

Since the memory is unchanged, this implies \(B_1 \supseteq B_2\) and we can restrict the equivalence to \(B_2\) by Lemma C.2: \(M_2 \sim B_2 M_2'\). We conclude by replacing \(M_2\) with \(M_2'\) in the evaluation derivation:

\[
\langle s_1, M_1, \overline{M_1} \rangle \rightarrow (M_2, \overline{M_2})
\]
\[
\langle s_1 s_2, M_1, M_1 \rangle \rightarrow (s_2, M_2, \overline{M_2})
\]

F.1.2 Subcase eval-seqCont.

\[
\text{eval-seqCont}\quad \langle s_1, M_1, \overline{M_1} \rangle \rightarrow (s_2, M_2, M_2)
\]
\[
\langle s_1 s_2, M_1, M_1 \rangle \rightarrow (s_2, M_2, M_2)
\]

The proof is the same as in the previous case, substituting \(\text{live}_{in}(I(s_1))\) for \(\text{live}_{out}(I(s_1))\).

F.2 Conditional Branching

\[
\text{eval-ifTrue}\quad M \models_v e \Rightarrow \text{Int}(n) \quad n \neq 0
\]
\[
\langle \text{if } e \rangle \text{ then } s_1 \text{ else } s_f, M, \overline{M} \rightarrow (s_t, M, \overline{M})
\]
\[
M \models_v e \Rightarrow \text{Int}(0)
\]
\[
\langle \text{if } e \rangle \text{ then } s_1 \text{ else } s_f, M, \overline{M} \rightarrow (s_t, M, \overline{M})
\]

Here \(M_1, M_1', M_2, \overline{M_1}\) are all equal (to \(\overline{M}\)). Therefore it suffices to show that \(B_1 \supseteq B_2\), that is:

\[
\bigcup_{x \in \text{live}_{in}(l)} R_{M_1}(x) = \bigcup_{x \in \text{live}_{in}(I(s))} R_{M_2}(x)
\]

where \(s = s_t\) or \(s_f\), depending on the case. This follows directly from:

\(\text{live}_{in}(l) \supseteq \text{live}_{in}(I(s)) \cup \text{live}_{in}(I(s_f))\)

(definition of the transfer function).

F.3 Loop

\[
\text{eval-whileCont}\quad M \models_v e \Rightarrow \text{Int}(n) \quad n \neq 0
\]
\[
\langle \text{while } e \rangle s, M, \overline{M} \rightarrow (s \text{ while } e, s, M, \overline{M})
\]

Since there no memory operation is performed, is suffices to show the inclusion: \(\text{live}_{in}(l) \supseteq \text{live}_{in}(I(s \text{ while } e, s))\). Applying the definition of \(I\) to the right hand side, we have \(I(s \text{ while } e, s) = I(s, s)\), so that we are left to prove:

\(\text{live}_{in}(l) \supseteq \text{live}_{out}(I(s))\)

which results from the definition of \(\text{live}_{in}, \text{live}_{out}\).