On-Stack Replacement, Distilled

Daniele Cono D’Elia
delia@dis.uniroma1.it
Sapienza University of Rome
Italy

Camil Demetrescu
demetres@dis.uniroma1.it
Sapienza University of Rome
Italy

Abstract

On-stack replacement (OSR) is essential technology for adaptive optimization, allowing changes to code actively executing in a managed runtime. The engineering aspects of OSR are well-known among VM architects, with several implementations available to date. However, OSR is yet to be explored as a general means to transfer execution between related program versions, which can pave the road to unprecedented applications that stretch beyond VMs. We aim at filling this gap with a constructive and provably correct OSR framework, allowing a class of general-purpose transformation functions to yield a special-purpose replacement. We describe and evaluate an implementation of our technique in LLVM. As a novel application of OSR, we present a feasibility study on debugging of optimized code, showing how our techniques can be used to fix variables holding incorrect values at breakpoints due to optimizations.

CCS Concepts → Software and its engineering → Compilers; Correctness; Just-in-time compilers; Software testing and debugging;

Keywords → Dynamic compilers, deoptimization, debugging.

ACM Reference Format:

1 Introduction

On-stack replacement (OSR) is a mechanism employed by modern runtimes to support on-the-fly transitions between variants of a currently executing function. OSR was first prototyped in the SELF VM [24] and has since become a fundamental element of multi-tier architectures in advanced runtimes. OSR is commonly used to transfer the execution between a base version of a function, which either is being interpreted or a baseline compiler has produced code for it, to a more optimized version, usually generated by a dynamic compiler based on the run-time behavior. As speculative optimizations are essential for high performance, deoptimization may become necessary: it can be supported either by reconstructing a frame for the interpreter or the baseline compiled code, which both have a fixed stack layout, or by performing OSR into dynamically generated code as in Jikes RVM [15]. Following the classification proposed in [29], we use OSR to refer to both optimizing and deoptimizing transitions.

A multi-tier approach increases the implementation and maintenance costs for a VM. Compilers have to encode the metadata and glue code needed to get the program state to a correct resumption point. For instance, unoptimized frame reconstruction normally requires the creation of scope descriptors accounting for program location and live variables at each deoptimization point. Deoptimization entry points may also be placed manually in the base tier as in V8 [45]. Also, restrictions may be put on optimizations to preserve the feasibility of OSR at specific points.

As a key motivation for our work, we argue that generalizing OSR to support transitions at arbitrary program points in the face of common classes of code transformations may not only be of interest for a VM implementor, but more importantly pave the way to unprecedented interesting use cases for OSR. For instance, one wishes to collect information about a program crash in an optimized production environment: OSR could be used to revert the state so that a core dump would reflect the expected behavior at the source level, reporting correct values for variables that are live at the crash point. As a second example, a program can be obfuscated to prevent security attacks via dynamic diversity by randomly diverting execution between different program versions at arbitrary execution points. Furthermore, OSR could be employed to switch between instrumented and uninstrumented code in tools that perform IR instrumentation such as dynamic memory safety checkers.

Contributions and Overview. In this paper we contribute to the theory and practice of OSR with a framework that explores soundness and flexibility of OSR transitions between different versions of a program in the face of common code transformations. The techniques presented here can be implemented in a general-purpose compiler with a modest
effort, and may provide useful building blocks to compiler architects. Our contributions are three-fold:

- We distill the essence of OSR to an abstract program morphing problem: code transformations can be made OSR-aware in isolation, and then composed in arbitrary ways. For a class of common optimizations, we devise an algorithm to generate the state realignment code possibly required for a transition to take place. A key observation is that for several classes of code transformations it is enough to reconstruct just the values of live variables rather than the entire state.
- We show that light changes are needed to accommodate our techniques in a general-purpose compiler. We discuss an implementation of our ideas in LLVM, and assess it over code drawn from popular benchmarks.
- We present a feasibility study of a novel OSR application in optimized-code debugging. While optimizations can cause variables to hold inconsistent values at breakpoints, our techniques are able to reconstruct most values expected at the source level.

Section 2 introduces formal preliminaries. Section 3 and 4 describe our OSR theoretical framework. An implementation is discussed in Section 5 and evaluated in Section 6. The feasibility study is presented in Section 7. Section 8 addresses related work. Section 9 presents directions for future work.

2 Language Framework

In this section we introduce some basic definitions used in our representation of programs and code transformations.

2.1 Program Syntax and Semantics

Our discussion is based on a minimal imperative language whose syntax is reported in Figure 1.

**Definition 2.1 (Program).** A program is a sequence of instructions of the form:

\[ P = \langle I_1, I_2, \ldots, I_n \rangle \in \text{Prog} = \bigcup_{i=2}^{\infty} \text{Instr}_{i} \]

where:

- \( I_i \in \text{Instr} \) is the \( i \)-th instruction of the program, indexed by program point \( i \in [1, n] \)
- \( I_1 = \text{in} \) is the initial instruction
- \( \forall i \in [2, n-1] : I_i \neq \text{in} \) • • • \( I_i \neq \text{out} \) • • •
- \( I_n = \text{out} \) is the final instruction

Instruction in must appear at the beginning of a program and specifies the variables that must be defined prior to entering the program. Similarly, out occurs at the end and specifies the variables that are returned as output. We indicate by \( e[x] \) that \( x \) is a variable of the expression \( e \in \text{Expr} \), and by \( |P| = n \) the number of instructions in \( P = \langle I_1, I_2, \ldots, I_n \rangle \).

In order to provide a formal semantics, we need to introduce definitions for program state and memory store:

\[
\begin{align*}
\text{Instr} &::= \quad \text{Var} ::= \text{Expr} \\
& \quad | \text{if} ( \text{Expr} ) \text{ goto } \text{Num} \\
& \quad | \text{goto } \text{Num} \\
& \quad | \text{skip} \text{ abort} \\
& \quad | \text{in } \text{Var} \cdots \text{Var} \text{ out } \text{Var} \cdots \text{Var}
\end{align*}
\]

\[
\begin{align*}
\text{Expr} &::= \text{Num} | \text{Var} | \text{Expr} + \text{Expr} | \ldots \\
\text{Var} &::= x \mid y \mid z \mid \ldots \\
\text{Num} &::= \ldots \mid -2 \mid -1 \mid 0 \mid 1 \mid 2 \mid \ldots
\end{align*}
\]

**Figure 1. Program Syntax**

\[ I_1 = x:=e \land (s, e) \downarrow v \]
\[ (s, l) \Rightarrow (s, m) \]
\[ I_2 = \text{goto } m \]
\[ (s, l) \Rightarrow (s, m) \]
\[ I_3 = \text{skip} \]
\[ (s, l) \Rightarrow (s, l + 1) \]
\[ I_4 = \text{if } (e) \text{ goto } m \land (s, e) \downarrow 0 \]
\[ (s, l) \Rightarrow (s, l + 1) \]
\[ I_5 = \text{if } (e) \text{ goto } m \land (s, e) \downarrow 0 \land v \neq 0 \]
\[ (s, l) \Rightarrow (s, m) \]
\[ I_6 = \text{in } x \land y \cdots \land (s(x) \neq \bot \land \sigma(y) \neq \bot \cdots \land (s(1), 1) \Rightarrow (s(2), 2) \]
\[ I_7 = \text{out } x \land y \cdots \land (s(x) \neq \bot \land \sigma(y) \neq \bot \cdots \land (s(n), n) \Rightarrow (s(x), \cdot \cdot \cdot, n + 1) \]

**Figure 2. Big-step semantics for the language of Figure 1.**

Transition relation \( \Rightarrow_{\Sigma} \text{State } \times \text{State} \) is defined using metavariables \( x, y \in \text{Var}, e \in \text{Expr} \), and \( m \in \text{Num} \). For a transition to apply, we assume that \( I_1 \) is defined, i.e., \( l \in [1, n] \).

**Definition 2.2 (Memory Store).** A memory store is a total function \( \sigma : \text{Var} \rightarrow \mathbb{Z} \cup \{ \bot \} \) that associates integer values to defined variables, and \( \bot \) to undefined variables. We denote by \( \Sigma \) the set of all possible memory stores.

By \( \sigma[x \leftarrow v] \) we denote the same memory store function as \( \sigma \), except that \( x \) takes value \( v \). Furthermore, for any \( A \subseteq \text{Var} \), \( \sigma|_A \) denotes \( \sigma \) restricted to the variables in \( A \), i.e., \( \sigma|_A(x) = \sigma(x) \) if \( x \in A \) and \( \sigma|_A(x) = \bot \) if \( x \notin A \).

**Definition 2.3 (Program State).** The state of a program \( P = \langle I_1, I_2, \ldots, I_n \rangle \) is described by a pair \( (s, l) \), where \( s \) is a memory store and \( l \in [1, n] \) is the program point of the next instruction to be executed. We denote by \( \text{State} = \Sigma \times \mathbb{N} \) the set of all possible program states.

In Figure 2 we provide a big-step semantics for the language using the transition relation \( \Rightarrow_{\Sigma} \subseteq \text{State } \times \text{State} \) that specifies how a single instruction of a program \( P \) affects its state. Our description relies on the relation \( [I_1] \subseteq (\Sigma \times \text{Expr}) \times \mathbb{Z} \) to describe how expressions are evaluated in a given memory store.

**Definition 2.4 (Program Semantic Function).** The semantic function \( [P] : \Sigma \rightarrow \Sigma \) of a program \( P \) as:

\[ \forall \sigma \in \Sigma : [P](\sigma) = \sigma' \iff (s, 1) \Rightarrow_{\Sigma}^* (s', |P| + 1) \]

where \( \Rightarrow_{\Sigma}^* \) is the transitive closure of \( \Rightarrow_{\Sigma} \).

Note that a program has undefined semantics if its execution on a given store does not reach the final out instruction. This
accounts for infinite loops, abort instructions, exceptions, and ill-defined programs or input stores. We define the notion of program semantic equivalence as follows:

**Definition 2.5 (Program Equivalence).** Two programs \( p_1 \) and \( p_2 \) are semantically equivalent iff \( [p_1] = [p_2] \).

Finally, we provide a definition of trace of a transition system that will be useful to reason about feasible executions:

**Definition 2.6 (Traces).** For a transition system \((S, R \subseteq S^2)\) a trace starting at \( s \in S \) is a sequence \( \tau = \langle s_0, s_1, \ldots, s_i, \ldots \rangle \) such that \( s_0 = s \) and \( \forall i \geq 0 : s_i \in \tau \wedge s_i R s_{i+1} \iff s_{i+1} \in \tau \).

We denote by \( T_{R,s} \) the system of all traces starting from \( s \).

For the deterministic transition relation \( \Rightarrow_r \) of our language, the system of traces \( T_{\Rightarrow_r,(\sigma,1)} \) of the execution transition system \((\text{Store}, \Rightarrow_r)\) contains a single trace starting at state \((\sigma, 1)\) for any initial store \( \sigma \). We denote such trace by \( T_{(\sigma, 1)} \).

### 2.2 Program Properties and Transformations

In this section we present a formalism based on computation tree logic (CTL) that is amenable to automated reasoning [9]. The same formalism can be used to reason about program properties and to express program transformations.

**Program Properties.** A property can be expressed using a Boolean formula with free meta-variables that combine facts and formula \( \psi \) holds; \( \forall l \in [1, n] \) in a program \( p = \langle l_1, \ldots, l_n \rangle \), with meta-variables \( e, e' \in \text{Expr}, x, y \in \text{Var}, \) and \( l, m \in \text{Num} \).

holds iff on all backward paths \( (\hat{A}) \) starting at all the predecessors \( (\vec{A}) \) of \( l, x \) has been defined somewhere, and there is at least one forward path from \( l \) that eventually reads \( x \).

This allows us to define the set of live variables at any given point of a program, which will play a central role in this paper.

**Definition 2.7 (Live Variables).** The set of live variables of a program \( p \) at point \( l \) is defined as:

\[
\text{live}(p, l) \triangleq \{ x \in \text{Var} : l \models \text{live}\{x\}\}
\]

Later on we will also use two global predicates: \( \text{freevar}(x, e) \), which holds iff \( x \) is a free variable of expression \( e \), and \( \text{conlit}(c) \), which holds if expression \( c \) is a constant literal.

**Program Transformations.** To describe transformations, we use rewrite rules with side conditions drawn from CTL formulas in a similar manner to [26, 28]. We consider generalized rules that transform multiple instructions at once.

**Definition 2.8 (Rewrite Rule).** A rule \( T \) has the form:

\[
T = m_1 : I_1 \Rightarrow I_1 \cdots m_k : I_r \Rightarrow I_r \quad \text{if } \phi
\]

where \( \forall k \in [1, r], m_k \) is a meta-variable that denotes a program point, \( I_1 \) and \( I_r \) are program instruction patterns that can contain meta-variables, and \( \phi \) is a side condition that states whether the rule can be applied to the input program.

The following example encodes a trivial peephole optimization based on a weak form of operator strength reduction [10], using three meta-variables \( m, x, \) and \( y \):

\[
m : y := 2 * x \Rightarrow y := x \quad \text{if } \text{true}
\]

Rules can be applied to concrete programs by a transformation engine based on model checking. When a substitution \( \theta \) can bind free meta-variables with program objects so that \( \theta(\phi) \) is satisfiable in \( p \) and \( \theta(\hat{I}_k) = I_{\theta(m_k)} \in p \) for some \( k \in [1, r] \), then \( I_\theta(m_k) \) is replaced with \( \theta(\hat{I}_k) = I_{\theta(m_k)} \in p' \):

**Definition 2.9 (Rule Semantics).** Let \( T \) be a rewrite rule as in Definition 2.8. The transformation function \([T] : \text{Prog} \rightarrow \text{Prog}\) is defined as follows:
We say that the mapping is semantics-preserving if for any program \( p \) it holds \( \llbracket p \rrbracket = \llbracket p' \rrbracket \), where \( p' = \llbracket T \rrbracket (p) \).

## 3 On-Stack Replacement Framework

OSR embodiments have to ensure that, whenever control is transferred from a currently running function version to another one, execution can transparently continue without altering the intended program semantics. In the adaptive optimization practice, optimizing OSR transitions typically happen at places where state realignment is simple, i.e., at a method or loop entry. The placement of deoptimizing OSR points is determined by the runtime: it can emit them for all instructions that might deoptimize [37], or group them and resume execution from the last instruction in the deoptimized code that causes outside-visible effects [49].

A framework for reasoning on the feasibility and soundness of state transfers could not only be valuable in generalized OSR applications such as those outlined in Section 1, but provide novel ideas to VM architects as well. Previous work has shown that performing OSR at more points can enable additional optimization opportunities [13]. Similar considerations can be made for deoptimization in multi-tier VMs: for instance, current strategies may hamper optimizations in the deoptimization target code, such as those that make new values alive across deoptimization points [45].

We propose a model based on OSR mappings with compensation code. Code transformations become OSR-aware in isolation, and can then be combined in arbitrary ways.

### 3.1 OSR Mappings

An OSR mapping encodes the machinery required to transfer execution from a point \( l \) in a program \( p \) to a point \( l' \) in a program \( p' \), relying on a compensation code \( c \) to fix the memory store so that execution can safely continue in \( p' \). Ideally, this should be done for any realizable state \( s \) of \( p \), i.e., \( s \) is reachable by \( p \) from some initial store \( \delta \), to a realizable state \( s' \) of \( p' \). For each point \( l \) where the mapping is defined, \( c \) computes the expected values for live variables at the OSR landing point \( l' \) in \( p' \) from the live store at \( l \) in \( p \). We formalize these concepts in the following definition:

**Definition 3.1 (OSR Mapping).** For any \( p, p' \in \text{Prog} \), an OSR mapping from \( p \) to \( p' \) is a (possibly partial) function \( M_{pp'} : \llbracket 1, \| p \| \] \times \text{Prog} \) such that:

\[
\forall \delta \in \Sigma, \forall s = (\sigma, l) \in \tau_{\delta}: l \in \text{dom}(M_{pp'}),
\exists \delta' \in \Sigma, \exists s' = (\sigma', l') \in \tau_{\delta'}: M_{pp'}(l, c) = (l', c) \land \llbracket c \rrbracket(\sigma)_{\text{live}(p, l)} = \sigma'_{\text{live}(p', l')}
\]

We say that the mapping is strict if \( \delta' = \delta \).

### Theorem 3.2.

Given any program \( p \in \text{Prog} \) and any initial store \( \delta \in \Sigma \), it holds that for any state \( (\sigma, l) \in \tau_{\delta} \):

\[
(\sigma, l) \Rightarrow (\sigma', \| p \|) \iff (\sigma_{\text{live}(p, l)}, l) \Rightarrow_{p} (\sigma'', \| p \|)
\]

where \( \sigma'_{\text{live}(p', l')} = \sigma''_{\text{live}(p', l')} \) specifies the content of the output variables of the program.

**Proof (Sketch).** To prove the claim, it is sufficient to reason inductively on the number of transitions required to reach the final instruction at program point \( \| p \| \) using the following argument. For any state \( (\sigma_1, l_1) \in \tau_{\delta} \), it holds:

\[
(\sigma_1, l_1) \Rightarrow (\sigma_2, l_2) \iff (\sigma_1_{\text{live}(p, l_1)}, l_1) \Rightarrow_{p} (\sigma_2', l_2)
\]

where \( \sigma_2'_{\text{live}(p, l_2)} = \sigma_1_{\text{live}(p, l_1)} \), that is \( \sigma_2' \) and \( \sigma_2' \) have the same live variables at \( l_2 \). This follows by noticing that, if \( l_1 \) kills/creates a live variable when executed on \( (\sigma_1, l_1) \), then it does the same when executed on \( (\sigma_1_{\text{live}(p, l_1)}, l_1) \). Also, if an expression is evaluated by \( l_1 \), then all variables it contains must be live, otherwise we would be contradicting the definition of liveness. Hence, they are defined with the same value in both \( \sigma_1 \) and \( \sigma_1_{\text{live}(p, l_1)} \), leading to the same effect of \( l_1 \). Therefore, the resulting stores \( \sigma_2 \) and \( \sigma_2' \) coincide on the set of live variables of \( p \) at \( l_2 \).

The scenario of Definition 3.1 is illustrated in Figure 4(a). Note that we use traces to describe feasible execution states for the two programs.

If \( p' \) has been derived from \( p \) via a code transformation \( T \), we say that \( T \) can be made OSR-aware if we can build a forward mapping \( M_{pp'} \) and a backward mapping \( M_{p'p} \). When a mapping is partial, there are points where OSR is not supported.

To simplify the discussion, we have used a weak notion of store equality restricted to live variables, leaving load/store instructions for discussion in Section 5.3.

The correctness of fixing only live variables in an OSR mapping stems from the following claim, which states that replacing at any time a store with another one where we would have obtained by running \( p' \) instead of \( p \).
3.2 OSR Mapping Composition

A relevant property of our framework is that OSR mappings can be composed. This allows us to make transformation OSR-aware in isolation and apply them in sequence, as their compensation code is amenable to program composition.

Definition 3.3 (Program composition). We say that two programs \( p = \langle l_1, \ldots, l_n \rangle \) and \( p' = \langle l'_1, \ldots, l'_n' \rangle \) are composable if \( l_n = \text{out} x_1, \ldots, x_k \) and \( l'_1 = \text{in} x'_1, \ldots, x'_k \), with \( \{x'_1, \ldots, x'_k\} \subseteq \{x_1, \ldots, x_k\} \). For any pair of composable programs \( p, p' \), we define \( p \circ p' = \langle l_1, \ldots, l_{n-1}, \tilde{l}_2, \ldots, \tilde{l}_n' \rangle \), where \( \forall i \in [1, n'] \), \( \tilde{l}_i \) is obtained from \( l'_i \) by relocating each goto target \( m \) with \( m + n - 2 \).

The semantic function \( \llbracket p \circ p' \rrbracket (p) \) is defined as \( \llbracket p' \rrbracket (\llbracket p \rrbracket (p)) \) for any input store \( p \). Mapping composition is as follows:

Theorem 3.4 (Mapping Composition). Let \( p, p', p'' \in \text{Prog} \), let \( M_{p,p'} \) and \( M_{p',p''} \) be OSR mappings as in Definition 3.1, and let \( M_{p,p'} \circ M_{p',p''} \) be a composition of mappings defined as follows:

\[
\forall i \in \text{dom}(M_{p,p'}) : M_{p,p'}(l_i) = (l_i', c) \wedge l_i' \in \text{dom}(M_{p',p''}) : M_{p',p''}(l_i') = (l''_i, c') \implies (M_{p,p'} \circ M_{p',p''})(l_i) = (l''_i, c \circ c')
\]

Then \( M_{p,p'} \circ M_{p',p''} \) is an OSR mapping from \( p \) to \( p'' \).

The theorem allows us to flexibly combine transformation rules, provided that an OSR mapping between the original and modified programs can be produced for each rule.

3.3 Discussion

Our framework aims at supporting classic and also speculative optimizations, making them OSR-aware in isolation. Observe that while strict mappings are a natural fit for semantics-preserving transformations, Definition 3.1 is rather general, as a non-strict mapping may capture transitions to a program \( p' \) that is not semantically equivalent to \( p \). For instance, \( p' \) may contain speculatively optimized code, or just some optimized fragments of \( p \) in as tracing JIT [3, 17, 18]. Execution in \( p' \) can then be invalidated by firing an OSR back to \( p \) or to some other recovery program.

Minor modifications are required to include function calls and memory accesses, which we support already in the implementation. However, one limitation is that the current formalization would not allow an executing function to trigger an OSR for its caller: we plan to address further replacement in future work. Also, we wish to add support for optimizations such as software pipelining [26] where the OSR landing location depends on the run-time values of variables.

4 Automating OSR Mapping Generation

In this section we describe how to automatically construct OSR mappings between program versions generated using compiler transformations \( T \) that satisfy a natural property that we call live-variable equivalence.

\[
\begin{align*}
\text{Constant propagation (CP)} \\
m : x := e[v] \implies x := e[c] \\
\text{if conlit(c) \wedge m \models \Box_{\ell}(\text{def}(v) U \text{stmt}(v := c))} \\
\text{Dead code elimination (DCE)} \\
m : x := e \implies \text{skip} \\
\text{if } m \models \Box_{\ell}(\text{use}(x)) \\
\text{Code hoisting (Hoist)} \\
p : \text{skip} \implies x := e \\
q : x := e \implies \text{skip} \\
\text{if } p \models \Box_{\ell}((\text{def}(v) \lor \text{point}(q)) \wedge \\
q \models \Box_{\ell}((\text{def}(v) \lor \text{point}(q)) \lor \text{trans}(e) U \text{point}(p))
\end{align*}
\]

Figure 5. Rewrite rules for LVE common transformations.

4.1 Live-Variable Equivalent Transformations

Intuitively, live-variable equivalence of a transformation \( T \) states that the live variables a program \( p \) and its transformed version \( p' = [T](p) \) have in common at any corresponding state hold the same values. By corresponding states we mean those that would be reached at any time by running simultaneously the two programs from the same initial store. To characterize this property more precisely and prove that a transformation is live-variable equivalent, we need to introduce some formal machinery based on bisimulation.

Definition 4.1 (Program Bisimulation). A relation \( R \subseteq \text{State} \times \text{State} \) is a bisimulation relation between two programs \( p \) and \( p' \) if for any input store \( \sigma \in \Sigma \) it holds:

\[
s \in \tau_{\sigma} \wedge s' \in \tau'_{\sigma} \wedge s \models s' \implies \begin{align*}
1) & s \Rightarrow p s_1 \implies s' \Rightarrow p' s'_1 \wedge s_1 \Rightarrow p s'_1 \\
2) & s \Rightarrow p' s'_1 \implies s \Rightarrow p s_1 \wedge s_1 \Rightarrow p s'_1
\end{align*}
\]

Our notion of bisimulation between programs \( p \) and \( p' \) requires that \( R \) be a bisimulation between transition systems \( (\tau_{\sigma}, \Rightarrow_p) \) and \( (\tau'_{\sigma}, \Rightarrow_{p'}) \) for any store \( \sigma \in \Sigma \). This implies that for any \( \sigma, \tau_{\sigma} \) is finite if \( \tau_{\sigma} \) is finite; also, if they are finite, then they have the same length. This assumption can be made without loss of generality, as equal length of traces can be enforced by padding programs with \text{skip} statements.

As a second ingredient, we need to characterize relations between states that tolerate differences, as those that we would get by comparing the corresponding states of two different, but semantically equivalent, program versions:

Definition 4.2 (Partial State Equivalence). For any function \( A : \mathbb{N} \rightarrow 2^{\mathbb{V}ar} \), the partial state equivalence relation \( R_A \subseteq \text{State} \times \text{State} \) is defined as:

\[
R_A \triangleq \{(s, s') : s = (\sigma, l) \wedge s' = (\sigma', l) \wedge \sigma \mid A(l) = \sigma' \mid A(l)\}
\]

Using this notion, we say that two programs \( p \) and \( p' \) are live-variable bisimilar if for any initial store they hold at any time the same values for variables that are live in both programs:

Definition 4.3 (Live-Variable Bisimilar Programs). \( p \) and \( p' \) are live-variable bisimilar (LVB) if \( R_A \) is a bisimulation relation between them, where \( A = l \mapsto l \in (p, l) \cap l (p', l) \).
We now discuss how to enhance an existing LVE transformation.

**Theorem 4.5.** Transformations CP, DCE, and Hoist of Figure 5 are live-variable equivalent.

The argument for the proof follows the bisimulation relations that take as input the $\Delta M$ for an input program $p$. For CP, $R$ is the identity relation, while for DCE and Hoist it is piecewise-defined on the indexes of the traces. Notice that Hoist expects $\text{id}$, which is simply the transform of $\text{id}$, $\text{id}$, and $\text{id}$.

**Algorithm 1:** Value reconstruction for LVE programs.

<table>
<thead>
<tr>
<th>Procedure</th>
<th>Description</th>
</tr>
</thead>
</table>
| `reconstruct(x, p, l, p', l', l''`)` | Takes a variable $x$, the OSR origin and destination points $l$ and $l'$ in programs $p$ and $p'$, respectively, and an additional point $l''$ in $p'$. It builds a compensation code fragment $c$ that assigns $x$ with the value it would have had at $l''$ just before reaching $l'$, had execution been carried on in $p'$ instead of $p$. This scenario is illustrated in Figure 4(b). The workflow of OSR_trans is as follows. To construct an OSR mapping $M_{\text{def}}$ from $p$ to $p'$, it attempts to build a compensation code $c$ for each ordered pair $(l, l')$ of program points in the $\Delta_{\text{def}}$ function returned by `apply(p, T)`. $c$ is a program that takes as input the $x_1, \ldots, x_k$ live variables for the OSR source location $l$ in $p$, and assigns the correct values to the $x_1', \ldots, x_k'$ live variables for execution to resume at $l'$ in $p'$.

When dealing with an LVE transformation, reconstruct needs only to be invoked for variables that are live at the OSR destination but not at the source. The LVB hypothesis for $p$ and $p'$ entails that for any pair of corresponding locations, variables that are live in both programs hold the same value. We exploit this property in the formulation of `reconstruct` in Algorithm 1 for LVE transformations.

**Algorithm description.** reconstruct is called with $l'_{\text{AT}} = l'$ by OSR_trans. The algorithm first checks whether there is a unique reaching definition for $x$ at some point $l'_{\text{DEF}}$ in $p'$ reaching $l'_{\text{AT}}$. In the presence of multiple reaching definitions, the algorithm gives up. If $x$ is live both at the origin $l$ and at the destination $l'$, and the definition of $x$ at $l'_{\text{DEF}}$ that reaches $l'_{\text{AT}}$ is also a unique reaching definition for $x$ at $l'$ (line 4), then $x$ would have assumed at $l'_{\text{AT}}$ the same value available at $l'$. For the LVB hypothesis, if $x$ is already available at the origin no value reconstruction is needed (return at line 4).

If $x$ is not available at $l$, reconstruct iterates over all the constituents of the assignment $x := e$ computed at $l'_{\text{DEF}}$, i.e., the variables that occur in $e$, and recursively builds code to compute the values that they would have assumed at $l'_{\text{DEF}}$ just before reaching $l'$ if execution had been carried on in $p'$. Once the recursively generated code has been added to $c$, the assignment $x := e$ is appended to it. Note that we mark program points to avoid work repetition.

Construction of the $M_{\text{def}}$ mapping is analogous up to index renaming, thus details are not reported. The LVB hypothesis is a sufficient condition for the correctness of our approach:

**Theorem 4.6.** For any program $p$ and LVE transformation $T$, if `apply(p, T)` $\triangleq (p', \Lambda', \Delta')$ where $\Lambda' : [1, |p|] \rightarrow [1, |p'|]$ is the identity mapping between program points, then OSR_trans($p, T$) yields strict forward and backward OSR mappings $M_{\text{def}}$ and $M_{\text{def}}$, respectively, for programs $p$ and $p'$.

**Discussion.** Theorem 3.4 provides us with the flexibility to compose transformations generated from the LVE formulation given in Algorithm 1 for reconstruct with mappings from other transformations. Hence, it would be interesting to explore variants of the algorithm tailored to other classes of optimizations (e.g., vectorization-based ones) as well.

---

1 Predicate `ud(x, p, l_0, l_k)` captures the existence in a program $p$ of a unique definition, located at $l_0$, for variable $x$ that reaches location $l_k$. 
Algorithm 1 currently gives up when analyzing a variable with multiple reaching definitions. Compilers using a static single assignment (SSA) representation [12] model such a variable using a $\phi$ function. Gating functions [36] might offer a solution to us, as they capture the control conditions that determine which incoming definition for a $\phi$ function will provide the value. Compensation code could thus evaluate them against the current state and materialize a value accordingly. We plan to explore this direction as future work.

5 Implementation

In this section we first describe the steps required for implementors to accommodate our ideas in a general-purpose compiler, then we present details for an embodiment in LLVM.

5.1 Mapping Program Points and Variables

In our abstract model, we implicitly map variables by the same name, and unrealistically assume that corresponding states for LVB programs are located at the same points.

In practice, OSR embodiments and deoptimization handlers typically use scope descriptors to track program points and local variables [45]. When such metadata are not available, one can instrument optimizations at places where IR manipulations happen. We argue that for LVE transformations, it is enough to track the following primitive actions:

1. $\text{add}(\text{inst}, \text{loc})$: insert a new instruction at $\text{loc}$
2. $\text{delete}(\text{loc})$: delete instruction at $\text{loc}$
3. $\text{hoist}(\text{loc}, \text{newLoc})$: hoist instr. from $\text{loc}$ to $\text{newLoc}$
4. $\text{sink}(\text{loc}, \text{newLoc})$: sink instr. from $\text{loc}$ to $\text{newLoc}$
5. $\text{replace}(\text{oldOp}, \text{newOp}, [\text{inst}])$: uses of an operand in $\text{inst}$ or in the entire code are replaced with another

This comes at a little cost in some environments: for instance, the profiling technique described in [51] relies on similar actions (e.g., node elimination and movement, value node replacement) that are exposed by the Graal compiler as IR graph manipulations. Others such as LLVM require instead manual placement of hooks inside the transformations, regardless of where one wants to place an OSR point. Once the tracking is done, no further changes to LVE optimizations are required to support OSR.

5.2 Reconstructing Missing Variables

Once optimizations have been applied and IR mappings created by apply, compensation code can be automatically built using reconstruct. When an SSA representation is used, implementing Algorithm 1 is easier, as the reaching definition for a variable is unique at any point it dominates.

In general, a code optimizer might decide to keep a variable alive to support deoptimization at some location, provided that the optimized version can still execute faster. In a similar spirit, we devise two variants of reconstruct: a live version using only live variables at the OSR source, and an avail version that keeps alive already-computed values that the algorithm cannot reconstruct otherwise. Keeping a variable $x$ alive should not be a concern in terms of register pressure increase, as registers are occupied for the minimum necessary: when $x$ is assigned to a register, a compiler spills it when about to be clobbered, to only reload it later when an OSR is about to happen; when assigned to a stack location instead, $x$ is loaded to a physical register only at OSR time.

5.3 Supporting Memory Manipulation

Compilers rely on load and store operations to transfer values between memory and registers. A simple sufficient condition for correctness of transitions is that store instructions are executed at the same program point in all versions. Indeed, when two program versions assign to a variable with a load from the same address, and the variable is live at some same program point in both versions, then the value read from memory has to be the same in both versions. Our implementation preserves the store invariant above while allowing instructions that do not access memory to be hoisted above or sunk below a store instruction. Notice that common LLVM optimizations such as loop hoisting and code sinking deal with store instructions in a similar manner.

A possible extension for scenarios where the above assumption might be too restrictive is as follows. Suppose that a store is sunk during optimization. For each CFG location between the original location and the insertion point: (a) in an optimizing OSR, no compensation code is required, as the store has been executed already, and re-executing it at the insertion point will be harmless; (b) in a deoptimizing OSR, we realign the memory state by executing the sunk store, which has not been reached yet in the optimized version.

5.4 LLVM Implementation

We conclude the section by reporting implementation details specific to LLVM. For the reader not familiar with it, optimizations are expressed as passes that manipulate IR in SSA form and are shared by front-ends for a variety of languages. Architecture-specific optimizations are performed instead in the back-end. For simplicity, front-ends are allowed to place variables on the stack: the mem2reg pass will then promote stack references to registers in SSA form.

Compared to our abstract model, IR conditionals and assignment instructions uniquely determine program locations, while virtual registers correspond to variables in the store.

Making Passes OSR-Aware. Our implementation of apply takes as input a function and a sequence of optimizations, clones the function, optimizes the clone, and eventually constructs a mapping for program points and local variables between the two versions by processing the history of applied actions (Section 5.1). IR mapping construction is implemented in $\sim$300 C++ LOC.
We make OSR-aware a number of standard optimizations, including aggressive dead code elimination (ADCE), constant propagation (CP), common subexpression elimination (CSE), loop-invariant code motion (LICM), sparse conditional constant propagation (SCCP), and code sinking (Sink). We also augment the natural loop canonicalization (LC) and the loop-closed SSA form-construction (LCSSA) utility passes required by LICM. Optimizations performed in the back-end such as instruction scheduling and register allocation do not require instrumentation, as we operate at IR level.

Table 1 reports figures for our edits to the original code, along with the number of primitive actions we track in each pass. Changes involve updating a CodeMapper object (which tracks IR updates) by instrumenting points where manipulations happen. The last column describes changes to utility actions (Section 5.1) often mimic common utilities for LLVM IR manipulation, thus are simple to place once one has an understanding of the high-level behavior of a pass.

reconstruct can take advantage of peculiarities of LLVM IR, too. For instance, we identify and reconstruct ϕ-nodes that always evaluate to the same value, such as those artificially inserted when constructing the LCSSA form. We also capture implicit aliasing information from replace actions.

**OSR Transitions.** LLVM is used as a dynamic compiler in several projects. The OSRKit library [13] provides the machinery to devise an OSR transition between two program points, requiring the VM’s front-end to provide a mapping between virtual registers in the two versions and the state realignment code possibly required for the transition. [13] tackles the engineering aspects for supporting OSR in LLVM, presenting a case study in which glue code is hand-written to support type specialization via dynamic inlining in MATLAB. We hopefully make a step forward by showing for LVE transformations how to automatically generate glue code (and IR mappings) and compose them on top of OSRKit.

OSRKit can insert an OSR point in a function f at a location l to a variant f′, guarding it with a user-provided condition. The transition is modeled as a call that transfers the live state to a continuation function f_{to} that can be generated ahead of time or on the fly. f_{to} will execute any compensation code c placed in its entry block before jumping to the resumption point l′. As OSRKit works at IR level, we provide it with a mapping between live IR registers at the OSR point, and a sequence c of instructions that can materialize values that are live in f′ but not in f. As OSR metadata, we only maintain the CodeMapper for the IR objects of the two code versions, as glue code is generated on demand by reconstruct.

Note that as f_{to} is a specialization of f′ having l′ as unique entry point, it can typically execute faster than the original f′ [15]. In our case, deleting unreachable blocks yields more compact code, possibly improving register allocation, too. Also, f′_{to} in principle can be further optimized, including c.

Due to its placement, c does not affect optimization decisions nor the steady-state overhead from the presence of OSR points, which is minimal for OSRKit [13]. c is executed only once and is typically small in practice (Table 3). OSR guards are transparent to our framework, as we reason about uninstrumented code variants. To extend the liveness range of a virtual register for avail, it is sufficient to add it as argument for the call to f_{to}. An optional OSR edge probability can be used to refine native code generation.

**6 Evaluation**

A preliminary investigation on classic benchmarks suggests that our techniques may enable bidirectional OSR almost everywhere in the face of common code optimizations.

**6.1 Benchmarks and Environment**

We integrate our techniques in the TinyVM artifact\(^2\) from the OSRKit paper [13] for IR optimization, JIT compilation, and benchmarking. Construction and composition of OSR

---

\(^2\)Additions are available at https://github.com/dcdelia/tinyvm.
mappings takes place when a sequence of passes is applied to a function to obtain an optimized variant of it.

We evaluate our technique on the SPEC CPU2006 [21] and the Phoronix PTS [38] benchmarking suites, reporting data for a subset of their C/C++ benchmarks. We profile the hottest method in each benchmark and when it accounts for at least 5% of the execution time, we generate LLVM IR for an f_base version using clang with the sole mem2reg pass enabled. We then produce an f_opt version by applying the optimizations discussed in Section 5.4.

To validate our implementation, we compile and run the code we generate for a sample of all feasible OSR source-destination pairs. We use a machine equipped with an Intel Core i7-3632QM, Ubuntu 14.10 64-bit and LLVM 3.6.2.

6.2 Results

In Table 2 we report aggregate figures for IR manipulations performed by the optimizations. While the f_opt version is typically shorter than its f_base counterpart, it might have a larger number of ϕ-nodes: most extra nodes are commonly generated during the LCSSA-form construction and optimized away in the back-end. SCCP can eliminate a large number of unreachable blocks from ffmpeg, while CSE performs the majority of deletions in the other benchmarks.

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>f_base</th>
<th>f_base</th>
<th>f_opt</th>
<th>f_opt</th>
<th>add</th>
<th>del</th>
<th>live</th>
<th>avail</th>
<th>sink</th>
<th>live</th>
<th>avail</th>
<th>C = ()</th>
<th>live</th>
<th>avail</th>
</tr>
</thead>
<tbody>
<tr>
<td>bzip2</td>
<td>657</td>
<td>32</td>
<td>596</td>
<td>44</td>
<td>16</td>
<td>77</td>
<td>12</td>
<td>3</td>
<td>73</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>h264ref</td>
<td>671</td>
<td>28</td>
<td>576</td>
<td>36</td>
<td>9</td>
<td>105</td>
<td>4</td>
<td>21</td>
<td>102</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hmrmer</td>
<td>568</td>
<td>6</td>
<td>383</td>
<td>8</td>
<td>2</td>
<td>187</td>
<td>3</td>
<td>1</td>
<td>187</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>namd</td>
<td>1737</td>
<td>159</td>
<td>1636</td>
<td>224</td>
<td>68</td>
<td>169</td>
<td>36</td>
<td>73</td>
<td>162</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>perbench</td>
<td>5574</td>
<td>305</td>
<td>5001</td>
<td>355</td>
<td>86</td>
<td>667</td>
<td>96</td>
<td>28</td>
<td>627</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sjeng</td>
<td>1940</td>
<td>93</td>
<td>1540</td>
<td>105</td>
<td>13</td>
<td>413</td>
<td>20</td>
<td>34</td>
<td>413</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>soplex</td>
<td>195</td>
<td>2</td>
<td>154</td>
<td>2</td>
<td>0</td>
<td>41</td>
<td>2</td>
<td>4</td>
<td>41</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bullet</td>
<td>587</td>
<td>24</td>
<td>553</td>
<td>42</td>
<td>26</td>
<td>60</td>
<td>37</td>
<td>3</td>
<td>52</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ddraw</td>
<td>590</td>
<td>37</td>
<td>545</td>
<td>49</td>
<td>13</td>
<td>58</td>
<td>25</td>
<td>6</td>
<td>58</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ffmpeg</td>
<td>618</td>
<td>34</td>
<td>462</td>
<td>40</td>
<td>11</td>
<td>168</td>
<td>9</td>
<td>17</td>
<td>103</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fluorstones</td>
<td>288</td>
<td>29</td>
<td>284</td>
<td>39</td>
<td>14</td>
<td>20</td>
<td>3</td>
<td>0</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>vp8</td>
<td>334</td>
<td>41</td>
<td>299</td>
<td>60</td>
<td>19</td>
<td>54</td>
<td>17</td>
<td>34</td>
<td>54</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

OSR points grows almost to 100% for all benchmarks. The ability to identify ϕ-nodes yielding a constant value is crucial to support OSR at ~20% of the program points in bully1.

In Table 3 we report average and peak size of the compensation code c generated by live and avail for feasible OSR points. Notice that averages are calculated on different sets of program points, i.e., avail extends the set from live. The assignment step of Algorithm 1 (line 8) generates an average number of instructions typically smaller than 20, with the notable exception of perlbench. Its hottest function highly benefits from CSE: no less than 583 out of its 667 deleted instructions (~10% of the size of f_base) are removed by it. Local CSE may reduce the size of the OSR entry block, too. However, we do not optimize c by default as the impact of compensation code on performance is hardly noticeable.

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>K (avail)</th>
<th>f_base</th>
<th>f_opt</th>
<th>c (=)</th>
<th>live</th>
<th>avail</th>
<th>C = ()</th>
<th>live</th>
<th>avail</th>
</tr>
</thead>
<tbody>
<tr>
<td>bzip2</td>
<td>657</td>
<td>32</td>
<td>596</td>
<td>44</td>
<td>16</td>
<td>77</td>
<td>12</td>
<td>3</td>
<td>73</td>
</tr>
<tr>
<td>h264ref</td>
<td>671</td>
<td>28</td>
<td>576</td>
<td>36</td>
<td>9</td>
<td>105</td>
<td>4</td>
<td>21</td>
<td>102</td>
</tr>
<tr>
<td>hmrmer</td>
<td>568</td>
<td>6</td>
<td>383</td>
<td>8</td>
<td>2</td>
<td>187</td>
<td>3</td>
<td>1</td>
<td>187</td>
</tr>
<tr>
<td>namd</td>
<td>1737</td>
<td>159</td>
<td>1636</td>
<td>224</td>
<td>68</td>
<td>169</td>
<td>36</td>
<td>73</td>
<td>162</td>
</tr>
<tr>
<td>perbench</td>
<td>5574</td>
<td>305</td>
<td>5001</td>
<td>355</td>
<td>86</td>
<td>667</td>
<td>96</td>
<td>28</td>
<td>627</td>
</tr>
<tr>
<td>sjeng</td>
<td>1940</td>
<td>93</td>
<td>1540</td>
<td>105</td>
<td>13</td>
<td>413</td>
<td>20</td>
<td>34</td>
<td>413</td>
</tr>
<tr>
<td>soplex</td>
<td>195</td>
<td>2</td>
<td>154</td>
<td>2</td>
<td>0</td>
<td>41</td>
<td>2</td>
<td>4</td>
<td>41</td>
</tr>
<tr>
<td>bullet</td>
<td>587</td>
<td>24</td>
<td>553</td>
<td>42</td>
<td>26</td>
<td>60</td>
<td>37</td>
<td>3</td>
<td>52</td>
</tr>
<tr>
<td>ddraw</td>
<td>590</td>
<td>37</td>
<td>545</td>
<td>49</td>
<td>13</td>
<td>58</td>
<td>25</td>
<td>6</td>
<td>58</td>
</tr>
<tr>
<td>ffmpeg</td>
<td>618</td>
<td>34</td>
<td>462</td>
<td>40</td>
<td>11</td>
<td>168</td>
<td>9</td>
<td>17</td>
<td>103</td>
</tr>
<tr>
<td>fluorstones</td>
<td>288</td>
<td>29</td>
<td>284</td>
<td>39</td>
<td>14</td>
<td>20</td>
<td>3</td>
<td>0</td>
<td>16</td>
</tr>
<tr>
<td>vp8</td>
<td>334</td>
<td>41</td>
<td>299</td>
<td>60</td>
<td>19</td>
<td>54</td>
<td>17</td>
<td>34</td>
<td>54</td>
</tr>
</tbody>
</table>
Table 3. Average and peak size $|c|$ of the compensation code generated by the two versions of reconstruct. $|K_{\text{avail}}|$ is the size of the set of variables to be kept artificially alive for OSR at program points represented by the top bars in Figure 7 and 8.

| Benchmark | $|c| \leftarrow \text{live}$ | $|c| \leftarrow \text{avail}$ | $|K_{\text{avail}}|$ | $|c| \leftarrow \text{live}$ | $|c| \leftarrow \text{avail}$ | $|K_{\text{avail}}|$ |
|---|---|---|---|---|---|---|
| bzip2 | Avg 4.3 Max 14 | Avg 4.73 Max 13 | Avg 3.6 Max 8 | Avg 1.55 Max 4 | Avg 1.77 Max 4 | Avg 1.47 Max 4 |
| bzip2 | Avg 2.9 Max 5 | Avg 3.37 Max 5 | Avg 1.02 Max 2 | Avg 4.46 Max 9 | Avg 2.82 Max 9 | Avg 1.45 Max 7 |
| h264ref | Avg 16.11 Max 23 | Avg 16.63 Max 24 | Avg 4.02 Max 7 | Avg 1 Max 1 | Avg 1 Max 1 | Avg 1.02 Max 2 |
| v8 | Avg 18.61 Max 28 | Avg 17.82 Max 28 | Avg 3.38 Max 6 | Avg 1.5 Max 2 | Avg 5.93 Max 15 | Avg 4.74 Max 18 |
| perlbench | Avg 46.12 Max 57 | Avg 45.82 Max 57 | Avg 1.24 Max 12 | Avg 4.09 Max 12 | Avg 4.22 Max 12 | Avg 1.37 Max 11 |
| sjeng | Avg 9.72 Max 21 | Avg 18.52 Max 32 | Avg 4.2 Max 12 | Avg 1.29 Max 2 | Avg 1.67 Max 11 | Avg 4.09 Max 14 |
| soplex | Avg 5.02 Max 7 | Avg 4.38 Max 7 | Avg 2.34 Max 4 | Avg 3.3 Max 3 | Avg 3.3 Max 3 | Avg 1.0 Max 1 |
| bullet | Avg 16.69 Max 46 | Avg 15.93 Max 46 | Avg 6.15 Max 17 | Avg 1 Max 1 | Avg 1.26 Max 3 | Avg 1.14 Max 2 |
| dcraw | Avg 7.6 Max 15 | Avg 7.32 Max 15 | Avg 1.97 Max 7 | Avg 1.68 Max 2 | Avg 3.84 Max 6 | Avg 4.06 Max 8 |
| ffmpeg | Avg 5.05 Max 8 | Avg 4.03 Max 8 | Avg 1.85 Max 3 | Avg 1.94 Max 5 | Avg 1.95 Max 6 | Avg 1.08 Max 4 |
| fhourstones | Avg 4.5 Max 6 | Avg 4.98 Max 6 | Avg 1.7 Max 2 | Avg 0 Max 0 | Avg 1.12 Max 4 | Avg 1.42 Max 4 |
| vp8 | Avg 10.51 Max 16 | Avg 10.13 Max 17 | Avg 2.35 Max 6 | Avg 5.74 Max 13 | Avg 5.51 Max 13 | Avg 1.18 Max 5 |
| Avg | Avg 12.26 Max 20.50 | Avg 12.81 Max 21.50 | Avg 2.62 Max 7.17 | Avg 2.30 Max 4.58 | Avg 2.87 Max 7.33 | Avg 2.00 Max 6.67 |

In Table 3 we report statistics on the size of the compensation code generated across feasible OSR points and the number of variables to be kept alive by $avail$. Compared to the optimizing OSR scenario, the size of $c$ is much smaller, suggesting that shorter portions of execution need to be reconstructed in a deoptimizing OSR. Although we report a 0 size for fhourstones, some realignment is still needed: our code contains minor optimizations not discussed here that can detect when there is a live alias for a variable $x$ that can be used in its place throughout the IR.

7 Symbolic Debugging of Optimized Code

In this section we present a feasibility study of how our algorithms for compensation code generation can provide useful novel building blocks for optimized-code debuggers. On prominent C benchmarks, reconstruct is able to recover the expected source-level values for the vast majority of scalar user variables that might not be reported correctly by a debugger due to the effects of classic compiler optimizations.

7.1 Background

A source-level (or symbolic) debugger allows a programmer to monitor an executing program at the source-language level. Interactive mechanisms are typically provided to the user to halt/resume the execution at breakpoints, and to inspect the state of the program in terms of its source language.

The importance of the design and use of these tools was already clear in the ‘60s [14]. Optimizations are desirable in a production environment, and bugs can surface when they are enabled: a debuggable translation may hide them, or differences in timing behavior may trigger race conditions. Also, optimization may be mandatory due to memory limitations, efficiency reasons, or other platform-specific constraints [1].

Hennessy [20] describes the classic conflict between the use of optimizations and the ability to debug a program symbolically. A debugger provides the user with the illusion that the source program is executing one statement at a time. Optimizations preserve semantic equivalence, but can alter the structure and the intermediate results of the program.

Two problems surface when trying to symbolically debug optimized code [2, 25]: i) the code location problem, as the debugger must determine the position of a breakpoint in the optimized code, and ii) the data location problem, as users expect values of variables at a breakpoint to be consistent with the source code, but instructions may have been deleted or reordered, or values been “lost” due to register allocation.

When attempting to debug optimized programs, debuggers may thus give misleading information about the value of variables at breakpoints. Hence, the programmer has the difficult task of attempting to unravel the optimized code and determine what values the variables should have [20]. When global optimizations can cause the run-time value of a variable to be inconsistent with the source-level value expected at the breakpoint, the variable is called endangered [2].

A symbolic debugger has two ways of presenting meaningful information for an optimized program [47]: a) it can provide expected behavior by hiding the effects of optimizations, presenting the program state consistently with what the user expects from the source; or b) it can provide truthful behavior by making them aware of the effects of the optimizations, and warning them of possibly surprising outcomes.

Constraining optimizations or adding extra code during compilation to aid debugging does not solve the problem of debugging the optimized translation of a program, as the user debugs suboptimal code [1]. Source-level debuggers should thus explore techniques to recover expected behavior without relying on intrusive compiler extensions.

7.2 Using reconstruct for State Recovery

Dynamic deoptimization was pioneered in the SELF VM to provide expected behavior with globally optimized code [23].
Debugging information was supplied by the compiler at discrete interrupt points that acted as a barrier for optimizations, letting the compiler run unhindered between them. Motivated by the observation that our algorithms do not limit applications with different sizes, we analyze each method of SPEC CPU2006 and may need recovery.

As in most recent works on the topic, we focus on identifying and recovering scalar source variables in the presence of global optimizations. LLVM front-ends encode debugging information as metadata attached to IR objects (e.g., instructions, global variables, functions). Metadata are transparent to passes (i.e., they do not affect optimization) and agnostic to both the source language and the target debugging data format. Two intrinsics (llvm.dbg.value and llvm.dbg.declare) are used to associate source variables with a virtual register or the address of an alloca buffer.

We extend TinyVM to rebuild this mapping and identify which locations in the unoptimized IR f_base correspond to source-level locations (i.e., possible breakpoints) for a function. An OSR mapping is constructed as well when we apply LVE transformations to generate f_opt. For each location in f_opt that might correspond to (i.e., have as OSR landing pad) a source-level location in f_base, we determine which live variables at the destination are live also at the source (and thus yield the same value), and which ones are instead endangered and may need recovery.

### 7.3 The SPEC CPU2006 Benchmarks

To capture a variety of programming patterns and styles from applications with different sizes, we analyze each method of each C benchmark from the SPEC CPU2006 suite, applying the OSR-aware optimization passes from Section 5.4 to the baseline IR version f_base obtained with clang −O0 followed by mem2reg.

Table 4 reports for each benchmark the code size (LOC), the total number of functions in it (|F|_{tot}), the number of functions modified by the applied optimizations (|f|_{opt}), and, in turn, how many optimized functions are endangered (|f|_{end}), i.e., contain endangered user variables.

We observe that 11% (libquantum) to 54% (gobmk) of the optimized functions are endangered, while for 10% to 33% of the functions in each benchmark, the applied passes do not kick in. For endangered functions, on average at more than 25% of program points there is at least a user variable whose source value might not be reported correctly. For most functions in the benchmarks, the average number of affected user variables at such points ranges between 1 and 2, although we sometimes observe higher peaks at specific points (e.g., as high as 9 for gobmk and 14 for gcc and h264ref).

To investigate possible correlations between the size of a function and the number of user variables affected by source-level debugging issues, we analyze the corpus of functions for the three largest benchmarks in our suite, i.e., gcc, gobmk, and perlbench. Our findings suggest that, although larger functions might be more prone to have a large number of affected variables, such issues frequently arise for smaller functions as well.

### 7.4 Experimental Results

We evaluate the ability of reconstruct to recover the source-level expected value for endangered user variables in the SPEC CPU2006 experiments. For each function, we measure the average recoverability ratio, defined as the average across all IR points corresponding to source-level locations of the ratio between recoverable and endangered user variables at the point. Two versions of reconstruct can be used here.

Implementing l l i v e should not be difficult for debuggers such as gdb and LLDB that can evaluate expressions over the current program state, as it only needs to access the live state of the optimized program at the breakpoint. a v a i l could use invisible breakpoints to spill a number of available values before they are overwritten. Such breakpoints are largely employed in source-level debuggers [25, 47, 50]. Using spilled values and the current live state, expected values for endangered user variables can be reconstructed as for live. In a VM with debugging features, one can recompile a function when the user inserts a breakpoint in it, extending the liveness range for available values possibly needed by reconstruct.

Figure 9 shows for each benchmark the global average recoverability ratio achieved by live and avail on the set of affected functions |f|_{end}. We observe that avail performs particularly well on all benchmarks, with a global ratio higher than 95% for half of the benchmarks, and higher than 90% for 10 out of 12 benchmarks. In the worst case (gobmk), we observe a global ratio slightly higher than 83%. Results thus
suggest that reconstruct can recover expected values for the vast majority of source-level endangered variables.

To estimate how many values should be preserved to integrate available in a debugger, we collect for each function the “keep” set of non-live available values to be saved to support weighted average of each function’s average recoverability ratio. As in Table 4, we use \(|F_{\text{final}}|\) as weight for the average.

Observe that values in the keep set do not necessarily need to be preserved all simultaneously or at all points, as the minimal set to be maintained may change across function regions. Typically when debugging, values are saved using an invisible breakpoint before they are overwritten, and deleted as soon as they are no longer needed [25].

8 Related Work

On-Stack Replacement. The SELF VM has pioneered dynamic deoptimization for source-level debugging [23] and OSR [24] for efficient dynamic recompilation. The rise of the Java language has then brought OSR technology to the mass market, employing it in the most sophisticated runtimes.

HotSpot Server [37] instruments function entry points and backward branches to optimize performance–critical methods, and transfers execution to the interpreter when an instruction triggers deoptimization, e.g., after class loading.

Jikes RVM places instrumentation as in HotSpot to enable a profile-driven deferred compilation mechanism [15]. OSR enables recovery from speculative inlining too, using a stub to divert execution to a newly generated function. An OSRBarrier can be injected when lowering bytecode for interruptible methods to capture the JVM-level program state before the bytecode is executed.

Graal [49] aggressively optimizes code via partial evaluation, tracking the interpreter (i.e., unoptimized) state throughout the compilation: FrameState metadata are thus used to reconstruct interpreter frames during deoptimization. [45] proposes an alternate scheme by letting the optimizing compiler generate also the deoptimization target code. This simplifies the implementation of a deoptimization handler, as frames are now described in a uniform format. The Truffle language implementation framework [46] relies on Graal and exposes deoptimization to the programmer, allowing them to transition back to the interpreter and throw away the machine code for a specialization state of a method [48].

We have discussed OSR in LLVM in Section 5.4. Also, an earlier work [29] provides support in the legacy JIT for transitions at loop headers when no state adjustment is required. V8 uses multiple compilers with the recent addition of an interpreter. The IR graph is processed in an abstract fashion, incrementally tracking changes to program state performed by single instructions. In the lowering phase this information is materialized as deoptimization data where needed. Its highly optimizing TurboFan compiler supports OSR at loop headers, generating a continuation function specialized for the current variable values (as in Jikes) at the loop entry.

Most aforementioned works have an interpreter as deoptimization target and focus on efficiency aspects related to where to support bidirectional OSR transitions between interpreted and compiled code; our goal has been to investigate relations between program points and variables across code versions. We share similarities with [48] in allowing the same compiler for both code versions, and we hope that our algorithms could help [48] support optimizations that make new values alive across deoptimization entry points. We share the goal of crystallizing OSR abstractions and explore reusable solutions with [43], which prototypes in the Mu micro VM [44] a frame replacement API with an abstract view of stacks that can easily be used in any runtime, and has been implemented even on concrete hardware.

Correctness of Compiler Optimizations. Translation validation [35, 39] aims at verifying semantic equivalence for an optimized translation of a specific program. [27, 28] use rewrite rules with CTL-based side conditions to express transformations and prove them correct, while [30, 31] investigate automated soundness proofs for transformations expressed in a similar manner. Alive [32] has emerged as a domain-specific language for writing sound LLVM peephole optimizations. While these works focus on proving optimizations sound, in this paper we aim at proving OSR correct in the presence of optimizations that we know to be sound.
[26] aims at proving the equivalence of parameterized programs, which yields correctness of transformation rules once for all. This approach deserves further investigation in our setting, as it could provide a principled approach to computing mappings between equivalent points in different versions in the face of complex optimizations. [16] studies the correctness of speculative JIT optimizations, showing how incorporating assumptions in the IR makes it easier to reason about equivalence. We look forward to investigating how our techniques and theirs might benefit from each other.

**Optimized-Code Debugging.** We only know of one work that supports full symbolic debugging with expected behavior. TARDIS [5] is a time-traveling debugger for VMs that saves program state on a regular basis, and runs the unoptimized code in a restored snapshot to answer queries. Our solution is different, as it lies at the performance-preserving end of the spectrum [1], and in some ways more general, as it can be applied to statically compiled languages such as C.

Wu et al. [47] propose a framework to selectively take control of the execution by inserting four kinds of breakpoints, and perform a forward recovery process in an emulator that executes the optimized instructions mimicking their ordering at the source level. The emulation scheme however cannot report values whose reportability is path-sensitive. FULLDOC [25] makes a step further, as it can provide truthful behavior for deleted values, and expected behavior for the other values. The authors remark that FULLDOC can be integrated with techniques for reconstructing deleted values: our reconstruct might be an ideal candidate.

Hennessy [20] presents value recovery algorithms for local (with weaker extensions to global) optimizations, but advances in modern compilers make a revision of the assumptions behind them necessary [11]. Adl-Tabatabai [1] presents algorithms that identify aliases for source variables in compiler temporaries introduced by global optimizations. This idea is captured by our technique, which can also use facts recorded during IR manipulation when recursively recon-structing portions of the original program’s state.

**Other Related Work.** Product programs [6] used in verifi-cation of relational and k-safety properties are orthogonal to multi-version programs, which embody the notion of OSR and rely on CTL and model checking. [8] discusses loop tiling in the face of throw statements that thwart optimizations: to roll back out-of-order updates during deoptimization, an algorithm identifies a minimal number of elements to back up. Our work does not log incremental state changes, and may resort to liveliness extension for available variables only.

We share similarities with the formalism from [19] for verifiable dynamic software updating (DSU): patches come with heap transformer expressions, while a language construct indicates where updates can take place; symbolic execution [4] and Thor [34] are used for verification. DSU [22] faces more complex requirements than OSR, for instance involving different object versions [33], but can take place once execution reaches a safe point. [41, 42] discuss how to integrate DSU in a VM and devise extensions to the OSR support of Jikes.

Of a different flavor, but in a similar spirit as ours, [18] uses bisimulation to study what trace optimizations are sound. A dynamic compiler uses OSR to replace whole methods; a tracing JIT aggressively optimizes linear recorded instruction sequences, which control flow can leave through guarded side exits only: for instance, RPython [40] uses trampolines to analyze resume information for a guard and runs a compensation code to leave the trace, while SPUR [7] uses a transfer-tail JIT to bridge the execution to the baseline JIT.

## 9 Conclusions

This paper aims at making a first step towards a more general applicability of OSR. We hope that our ideas can provide the foundation for exploring novel applications of OSR and shed light on its flexibility. Optimizations can significantly affect the live state of a program across its locations: we show how to generate compensation code that runs in O(1) time by recursively reassembling portions of the state for the target function. Preliminary evidence suggests that our techniques can provide useful building blocks for symbolic debuggers. Also, we enable bidirectional transitions at most program locations in a dynamic compilation setting. While we have prototyped them in LLVM, our ideas are general and do not depend on a specific platform or IR representation.

A number of directions for future work are possible. Currently, we only study static transformations that are often enabled by dynamic optimizations: we would like to explore handling purely dynamic ones as well, for instance to materialize objects that have been optimized by scalar replacement. We also plan to model frame replacement and identify other classes of transformations for which a special-purpose replacement can be automatically generated. Heavy-duty transformations may likely require state logging as in the tiling case [8], leaving interesting performance/flexibility trade-offs for exploration. Compensation code with control flow (e.g., using gating functions) might enable state reconstruction at points that our live reconstruction algorithm currently does not support. More flexibility might come from observational equivalence as well, by allowing for execution transfers to the last state-changing instruction in lieu of the equivalent program point, letting the target code re-execute side-effect free instructions as in Graal. reconstruct may also be used to rematerialize dead values in a VM, instead of keeping them alive in optimized code.

**Acknowledgements.** We are grateful to Gustavo Petri for suggesting the use of bisimulation in our theoretical framework, to Jan Vitek for some enlightening discussions, and to the anonymous PLDI reviewers and our shepherd Iulian Neamtiu for their many useful comments.
References


On-Stack Replacement, Distilled

PLDI ’18, June 18–22, 2018, Philadelphia, PA, USA

https://doi.org/10.1023/B:LISP.0000029444.99264.c0


