Parameterized Construction of Program Representations for Sparse Dataflow Analyses

André Tavares, Benoit Boissinot, Fernando Pereira, Fabrice Rastello

Parameterized Construction of Program Representations for Sparse Dataflow Analyses

André Tavares, Benoit Boissinot, Fernando Pereira, Fabrice Rastello

RESEARCH REPORT

N° 8491

March 2014

Project-Teams GCG, Compsys

ISSN 0249-6399 ISRN INRIA/RR--8491--FR+ENG

Parameterized Construction of Program Representations for Sparse Dataflow Analyses

André Tavares*, Benoit Boissinot†, Fernando Pereira‡, Fabrice Rastello§

Project-Teams GCG, Compsys

Research Report n° 8491 — March 2014 — 27 pages

Abstract: Data-flow analyses usually associate information with control flow regions. Informally, if these regions are too small, like a point between two consecutive statements, we call the analysis dense. On the other hand, if these regions include many such points, then we call it sparse. This paper presents a systematic method to build program representations that support sparse analyses. To pave the way to this framework we clarify the bibliography about well-known intermediate program representations. We show that our approach, up to parameter choice, subsumes many of these representations, such as the SSA, SSI and e-SSA forms. In particular, our algorithms are faster, simpler and more frugal than the previous techniques used to construct SSI - Static Single Information - form programs. We produce intermediate representations isomorphic to Choi *et al.*'s Sparse Evaluation Graphs (SEG) for the family of data-flow problems that can be partitioned per variables. However, contrary to SEGs, we can handle - sparsely - problems that are not in this family.

Key-words: Sparse Data-Flow Analysis, Compiler, Static Single Assignment, Static Single Information, SSA, SSI, Static Single Use, SSU, Iterated Dominance Frontier, Control-Flow Graph

* UFMG
† ENS Lyon
‡ UFMG
§ Inria

RESEARCH CENTRE GRENOBLE – RHÔNE-ALPES

Inovallée
655 avenue de l'Europe Montbonnot
38334 Saint Ismier Cedex

Représentation de programmes pour l'analyse creuse de flots de données: construction paramétrée

Résumé : L'analyse de flot de données, associe en général l'information calculée, aux régions de flot de contrôle. Informellement cette analyse est dite dense, si ces régions sont trop petites, i.e. par exemple restreintes aux points de programme situés entre deux instructions. A l'opposé, cette analyse est dite creuse, si ces régions comprennent de nombreux points consécutifs. Cet article présente une méthode de construction systématique d'une représentation de programme qui permet de manière naturelle l'implémentation d'analyses creuses. Cette forme englobe plusieurs forme existante comme la forme SSA, la forme SSI, ou la forme e-SSA. En particulier, l'algorithme présenté est plus rapide, plus simple et moins gourmand que les méthodes existantes de construction de SSI –Static Single Information. Aussi, la représentation ainsi construite se trouve être isomorphe au graphe d'évaluation creux (Sparse Evaluation Graph — SEG in English) de Choi et al. dans le cas particulier ou le problème d'analyse de flot de données peut être partitionné par variable. Cela dit, contrairement aux SEG, l'approche ici décrite n'est pas restreinte à cette famille de problèmes.

Mots-clés : Analysis de flot de données, compilateur, forme à assignation unique, SSA, SSI, SSU, frontière de dominance itérée, graphe de flot de contrôle

1 Introduction

Many data-flow analyses bind information to pairs formed by a variable and a program point [1, 6, 10, 17, 25, 28, 30, 34, 36, 39, 40, 43, 44, 45, 46]. As an example, for each program point p, and each integer variable v live at p, Stephenson *et al.*'s [43] bit-width analysis finds the size, in bits, of v at p. Although well studied in the literature, this approach might produce redundant information. For instance, a given variable v may be mapped to the same bit-width along many consecutive program points. Therefore, a natural way to reduce redundancies is to make these analyses *sparser*, increasing the granularity of the program regions that they manipulate.

There exists different attempts to implement data-flow analyses sparsely. The Static Single Assignment (SSA) form [16], for instance, allows us to implement several analyses and optimizations, such as reaching definitions and constant propagation, sparsely. Since its conception, the SSA format has been generalized into many different program representations, such as the *Extended-SSA* form [6], the *Static Single Information* (SSI) form [2], and the *Static Single Use* (SSU) form [22, 27, 34]. Each of these representations extends the reach of the SSA form to sparser data-flow analyses; however, there is not a format that subsumes all the others. In other words, each of these three program representations fit specific types of data-flow problems. Another attempt to model data-flow analyses sparsely is due to Choi *et al.*'s *Sparse Evaluation Graph* (SEG) [12]. This data-structure supports several different analyses sparsely, as long as the abstract state of a variable does not interfere with the abstract state of other variables in the same program. This family of analyses is known as *Partitioned Variable Problems* in the literature [48].

In this paper, we propose a framework that includes all these previous approaches. Given a data-flow problem defined by (i) a set of control flow nodes, that produce information, and (ii) a direction in which information flows: forward, backward or both ways, we build a program representation that allows to solve the problem sparsely using def-use chains. The program representations that we generate ensure a key *single information property*: the data-flow facts associated with a variable are invariant along the entire live range of this variable.

2 Static Single Information

Our objective is to generate program representations that bestow the *Static Single Information property* (Definition 6) onto a given data-flow problem. In order to introduce this notion, we will need a number of concepts, which we define in this chapter. We start with the concept of a *Data-Flow System*, which Definition 1 recalls from the literature. We consider a *program point* a point between two consecutive instructions. If p is a program point, then $preds(p)$ (resp. $succs(p)$) is the set of all the program points that are predecessors (resp. successors) of p. A *transfer function* determines how information flows among these program points. Information are elements of a *lattice*. We find a solution to a data-flow problem by continuously solving the set of transfer functions associated with each program region until a fix point is reached. Some program points are *meet nodes*, because they combine information coming from two or more regions. The result of combining different elements of a lattice is given by a *meet* operator, which we denote by $\wedge$.

Definition 1 (Data-Flow System)**.** *A data-flow system E_{dense} is an equation system that associates, with each program point p, an element of a lattice $\mathcal{L}$, given by the equation $x^p = \bigwedge_{s\in preds(p)} F^{s,p}(x^s)$, where: x^p denotes the abstract state associated with program point p; $preds(p)$ is the set of control flow predecessors of p; $F^{s,p}$ is the transfer function from program point s to program point p. The analysis can alternatively be written as a constraint system that binds to* *each program point p and each $s \in preds(p)$ the equation $x^p = x^p \wedge F^{s,p}(x^s)$ or, equivalently, the inequation $x^p \sqsubseteq F^{s,p}(x^s)$.*

The program representations that we generate lets us solve a class of data-flow problems that we call *Partitioned Lattice per Variable* (PLV), and that we introduce in Definition 2. Constant propagation is an example of a PLV problem. If we denote by $\mathcal{C}$ the lattice of constants, the overall lattice can be written as $\mathcal{L} = \mathcal{C}^n$, where n is the number of variables. In other words, this data-flow problem ranges on a product lattice that contains a term for each variable in the target program.

Definition 2 (Partitioned Lattice per Variable Problem (PLV)). *Let $\mathcal{V} = \{v_1, \ldots, v_n\}$ be the set of program variables. The Maximum Fixed Point problem on a data-flow system is a* Partitioned Lattice per Variable Problem *if, and only if, $\mathcal{L}$ can be decomposed into the product of $\mathcal{L}_{v_1} \times \cdots \times \mathcal{L}_{v_n}$ where each $\mathcal{L}_{v_i}$ is the lattice associated with program variable v_i. In other words x^s can be writen as $([v_1]^s, \ldots, [v_n]^s)$ where $[v]^s$ denotes the abstract state associated with variable v and program point s. $F^{s,p}$ can thus be decomposed into the product of $F^{s,p}_{v_1} \times \cdots \times F^{s,p}_{v_n}$ and the constraint system decomposed into the inequalities $[v_i]^p \sqsubseteq F^{s,p}_{v_i}([v_1]^s, \ldots, [v_n]^s)$.*

The transfer functions that we describe in Definition 3 have no influence on the solution of a data-flow system. The goal of a sparse data-flow analysis is to shortcut these functions. We accomplish this task by grouping contiguous program points bound to these functions into larger regions.

Definition 3 (Trivial/Constant/Undefined Transfer functions). *Let $\mathcal{L}_{v_1} \times \mathcal{L}_{v_2} \times \cdots \times \mathcal{L}_{v_n}$ be the decomposition per variable of lattice $\mathcal{L}$, where $\mathcal{L}_{v_i}$ is the lattice associated with variable v_i. Let F_{v_i} be a transfer function from $\mathcal{L}$ to $\mathcal{L}_{v_i}$.*

- F_{v_i} *is* trivial *if* $\forall x = ([v_1], \ldots, [v_n]) \in \mathcal{L}$, $F_{v_i}(x) = [v_i]$
- F_{v_i} *is* constant with value $C \in \mathcal{L}_{v_i}$ *if* $\forall x \in \mathcal{L}$, $F_{v_i}(x) = C$
- F_{v_i} *is* undefined *if F_{v_i} is constant with value $\top$, e.g., $F_{v_i}(x) = \top$, where $\top \wedge y = y \wedge \top = y$.*

A sparse data-flow analysis propagates information from the control flow node where this information is created directly to the control flow node where this information is needed. Therefore, the notion of *dependence*, which we state in Definition 4, plays a fundamental role in our framework. Intuitively, we say that a variable v depends on a variable v_j if the information associated with v might change in case the information associated with v_j does.

Definition 4 (Dependence). *We say that F_v* depends on variable v_j *if:*

$$\begin{array}{l} \exists x = ([v_1], \ldots, [v_n]) \neq ([v_1]', \ldots, [v_n]') = x' \text{ in } \mathcal{L} \\ \text{such that } \left[F_v(x) \neq F_v(x') \text{ and } \forall k \neq j,\ [v_k] = [v_k]'\right] \end{array}$$

In a *backward* data-flow analysis, the information that comes from the predecessors of a node n is combined to produce the information that reaches the successors of n. A *forward* analysis propagates information in the opposite direction. We call meet nodes those places where information coming from multiple sources are combined. Definition 5 states this concept more formally.

Definition 5 (Meet Nodes). *Consider a forward (resp. backward) monotone PLV problem, where (Y^p_v) is the maximum fixed point solution of variable v at program point p. We say that a program point p is a meet node for variable v if, and only if, p has $n \geq 2$ predecessors (resp. successors), $s_1, \ldots, s_n$, and there exists $s_i \neq s_j$, such that $Y^{s_i}_v \neq Y^{s_j}_v$.*

Our goal is to build program representations in which the information associated with a variable is invariant along the entire live range of this variable. A variable v is *alive* at a program point p if there is a path from p to an instruction that uses v, and v is not re-defined along the way. The live range of v, which we denote by *live(v)*, is the collection of program points where v is alive.

Definition 6 (Static Single Information property)**.** *Consider a forward (resp. backward) monotone PLV problem* E_{dense} *stated as in Definition 1. A program representation fulfills the Static Single Information property if, and only if, it meets the following properties for each variable* v*:*

[SPLIT-DEF]: *for each two consecutive program points* s *and* p *(resp.* p *and* s*) such that* $p \in live(v)$*, and* $F_v^{s,p}$ *is non-trivial nor undefined, there should be an instruction between* s *and* p *that contains a definition (resp. last use) of* v*;*

[SPLIT-MEET]: *each meet node* p *with* n *predecessors* $\{s_1, \ldots, s_n\}$ *(resp. successors) should have a definition (resp. use) of* v *at* p*, and* n *uses (resp. definitions) of* v*, one at each* s_i*. We shall implement these defs/uses with* ϕ/σ*-functions, as we explain in Section 2.1.*

[INFO]: *each program point* $p \notin live(v)$ *should be bound to undefined transfer functions, e.g.,* $F_v^{s,p} = \lambda x.\top$ *for each* $s \in preds(p)$ *(resp.* $s \in succs(p)$*).*

[LINK]: *for each two consecutive program points* s *and* p *(resp.* p *and* s*) for which* $F_v^{s,p}$ *depends on some* $[u]^s$*, there should be an instruction between* s *and* p *that contains a (potentially pseudo) use (resp. def) of* u*.*

[VERSION]: *for each variable* v*,* $live(v)$ *is a connected component of the CFG.*

2.1 Special instructions used to split live ranges

We group control flow nodes in three kinds: interior nodes, forks and joins. At each place we use a different notation to denote live range splitting.

Interior nodes are control flow nodes that have a unique predecessor and a unique successor. At these control flow nodes we perform live range splitting via copies. If the control flow node already contains another instruction, then this copy *must* be done *in parallel* with the existing instruction. The notation,

$$inst \;\parallel\; v_1 = v_1' \;\parallel\; \ldots \;\parallel\; v_m = v_m'$$

denotes m copies $v_i = v_i'$ performed in parallel with instruction *inst*. This means that all the uses of *inst* plus all v_i' are read simultaneously, then *inst* is computed, then all definitions of *inst* plus all v_i are written simultaneously.

In forward analyses, the information produced at different definitions of a variable may reach the same meet node. To avoid that these definitions reach the same use of v, we merge them at the earliest control flow node where they meet; hence, ensuring [SPLIT-MEET]. We do this merging via special instructions called ϕ-functions, which were introduced by Cytron *et al.* to build SSA-form programs [16]. The assignment

$$v_1 = \phi(l^1 : v_1^1, \ldots, l^q : v_1^q) \;\parallel\; \ldots \;\parallel\; v_m = \phi(l^1 : v_m^1, \ldots, l^q : v_m^q)$$

contains m ϕ-functions to be performed in parallel. The ϕ symbol works as a multiplexer. It will assign to each v_i the value in v_i^j, where j is determined by l^j, the basic block last visited before reaching the ϕ-function. The above statement encapsulates m parallel copies: all the variables

$v_1^j, \ldots, v_m^j$ are simultaneously copied into the variables $v_1, \ldots, v_m$. Note that our notion of control flow nodes differs from the usual notion of nodes of the CFG. A join node actually corresponds to the entry point of a CFG node: to this end we denote as $\mathrm{In}(l)$ the point right before l. As an example in Figure 1(d), l_7 is considered to be an interior node, and the ϕ-function defining v_6 has been inserted at the join node $\mathrm{In}(l_7)$.

In backward analyses the information that emerges from different uses of a variable may reach the same meet node. To ensure Property [SPLIT-MEET], the use that reaches the definition of a variable must be unique, in the same way that in a SSA-form program the definition that reaches a use is unique. We ensure this property via special instructions that Ananian has called σ-functions [2]. The σ-functions are the symmetric of ϕ-functions, performing a parallel assignment depending on the execution path taken. The assignment

$$(l^1 : v_1^1, \ldots, l^q : v_1^q) = \sigma(v_1) \;\; \| \;\; \ldots \;\; \| \;\; (l^1 : v_m^1, \ldots, l^q : v_m^q) = \sigma(v_m)$$

represents m σ-functions that assign to each variable v_i^j the value in v_i if control flows into block l^j. These assignments happen in parallel, i.e., the m σ-functions encapsulate m parallel copies. Also, notice that variables live in different branch targets are given different names by the σ-function that ends that basic block. Similarly to join nodes, a fork node is the exit point of a CFG node: $\mathrm{Out}(l)$ denotes the point right after CFG node l. As an example in Figure 1(d), l_2 is considered to be an interior node, and the σ-function using v_1 has been inserted at the fork node $\mathrm{Out}(l_2)$.

2.2 Examples of PLV Problems

Many data-flow analyses can be classified as PLV problems. In this section we present some meaningful examples. Along each example we show the program representation that lets us solve it sparsely.

Class Inference: Some dynamically typed languages, such as Python, JavaScrip, Ruby or Lua, represent objects as hash tables containing methods and fields. In this world, it is possible to speedup execution by replacing these hash tables with actual object oriented virtual tables. A class inference engine tries to assign a virtual table to a variable v based on the ways that v is used. The Python program in Figure 1(a) illustrates this optimization. Our objective is to infer the correct suite of methods for each object bound to variable v. Figure 1(b) shows the control flow graph of the program, and Figure 1(c) shows the results of a dense implementation of this analysis. In a dense analysis, each program instruction is associated with a transfer function; however, some of these functions, such as that in label l_3, are trivial. We produce, for this example, the representation given in Figure 1(d). Because type inference is a backward analysis that extracts information from use sites, we split live ranges at these control flow nodes, and rely on σ-functions to merge them back. The use-def chains that we derive from the program representation, seen in Figure 1(e), lead naturally to a constraint system, which we show in Figure 1(f). A solution to this constraint system gives us a solution to our data-flow problem.

Constant Propagation: Figure 2 illustrates constant propagation, e.g., which variables in the program of Figure 2(a) can be replaced by constants? The CFG of this program is given in Figure 2(b). Constant propagation has a very simple lattice $\mathcal{L}$, which we show in Figure 2(c). In constant propagation, information is produced at the program points where variables are defined. Thus, in order to meet Definition 6, we must guarantee that each program point is reachable by a single definition of a variable. Figure 2(d) shows the intermediate representation that we create for the program in Figure 2(b). In this case, our intermediate representation is

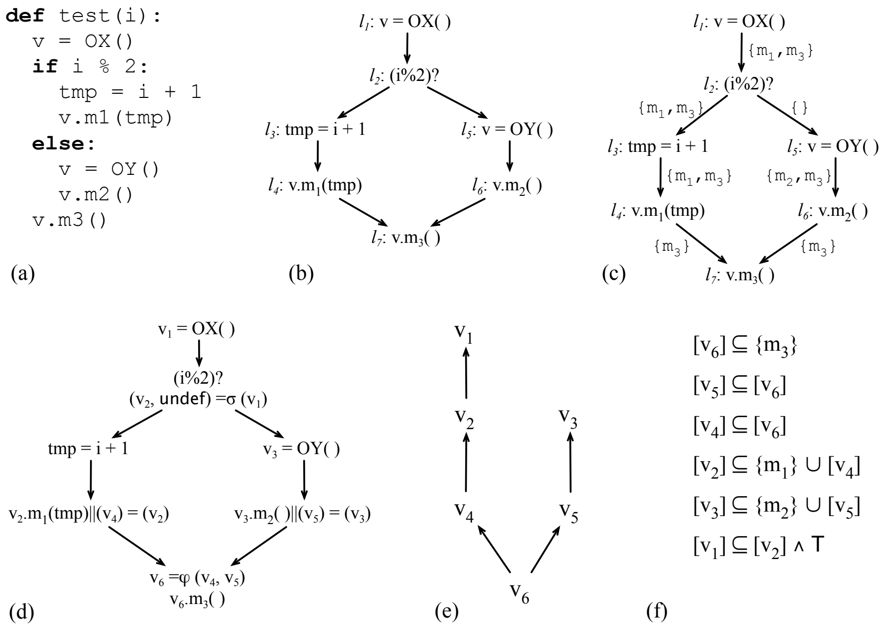

Figure 1: Class inference as an example of backward data-flow analysis that takes information from the uses of variables.

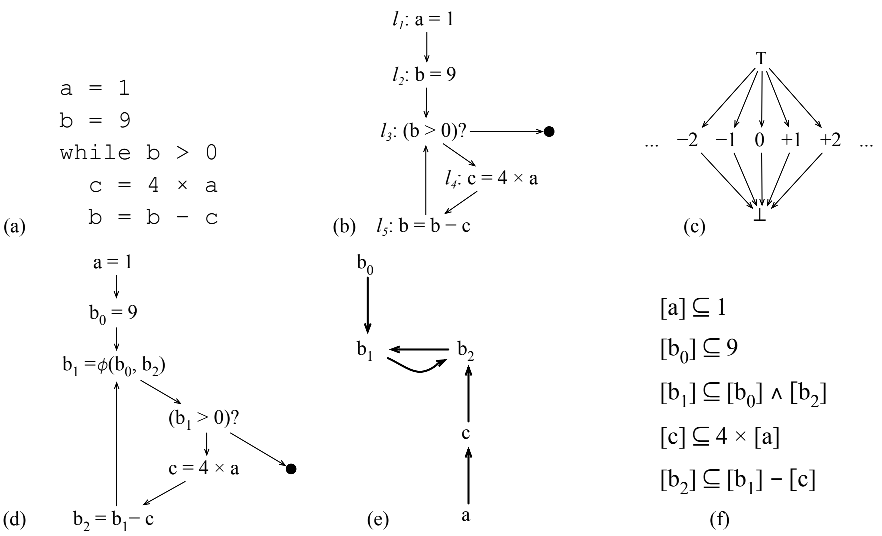

Figure 2: Constant propagation as an example of forward data-flow analysis that takes information from the definitions of variables.

equivalent to the SSA form. The def-use chains implicit in our program representation lead to the constraint system shown in Figure 2(f). We can use the def-use chains seen in Figure 2(e) to guide a worklist-based constraint solver, as Nielson *et al.* [31, Ch.6] describe.

Taint analysis: The objective of taint analysis [36, 37] is to find program vulnerabilities. In

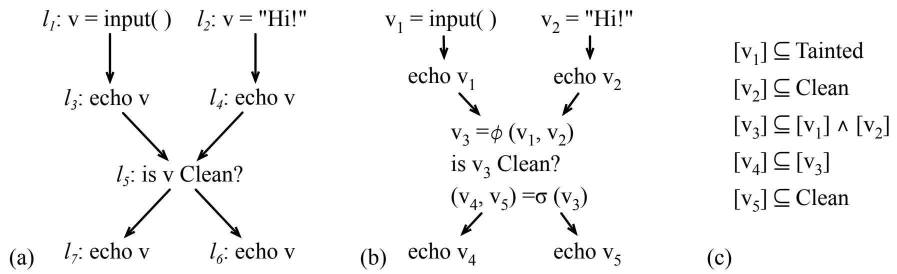

Figure 3: Taint analysis is a forward data-flow analysis that takes information from the definitions of variables and conditional tests on these variables.

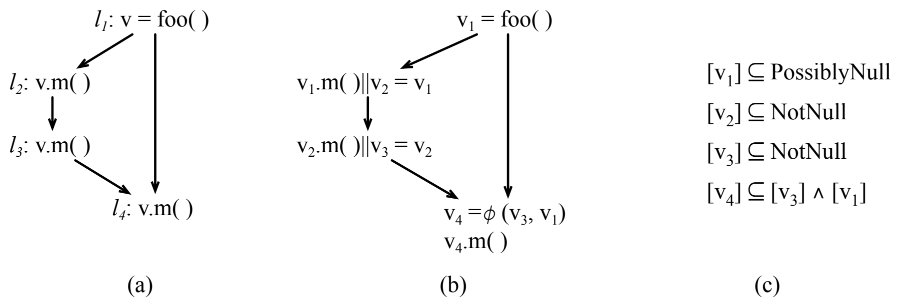

Figure 4: Null pointer analysis as an example of forward data-flow analysis that takes information from the definitions and uses of variables.

this case, a harmful attack is possible when input data reaches sensitive program sites without going through special functions called sanitizers. Figure 3 illustrates this type of analysis. We have used ϕ and σ-functions to split the live ranges of the variables in Figure 3(a) producing the program in Figure 3(b). Let us assume that *echo* is a sensitive function, because it is used to generate web pages. For instance, if the data passed to *echo* is a JavaScript program, then we could have an instance of cross-site scripting attack. Thus, the statement *echo* v_1 may be a source of vulnerabilities, as it outputs data that comes directly from the program input. On the other hand, we know that *echo* v_2 is always safe, for variable v_2 is initialized with a constant value. The call *echo* v_5 is always safe, because variable v_5 has been sanitized; however, the call *echo* v_4 might be tainted, as variable v_4 results from a failed attempt to sanitize v. The def-use chains that we derive from the program representation lead naturally to a constraint system, which we show in Figure 3(c). The intermediate representation that we create in this case is equivalent to the *Extended Single Static Assignment* (e-SSA) form [6]. It also suits the ABCD algorithm for array bounds-checking elimination [6], Su and Wagner's range analysis [44] and Gawlitza *et al.*'s range analysis [21].

Null pointer analysis: The objective of null pointer analysis is to determine which references may hold null values. Nanda and Sinha have used a variant of this analysis to find which method dereferences may throw exceptions, and which may not [30]. This analysis allows compilers to remove redundant null-exception tests and helps developers to find null pointer dereferences. Figure 4 illustrates this analysis. Because information is produced at use sites, we split live ranges after each variable is used, as we show in Figure 4(b). For instance, we know that the call $v_2.m()$ cannot result in a null pointer dereference exception, otherwise an exception would have been thrown during the invocation $v_1.m()$. On the other hand, in Figure 4(c) we notice that the state of v_4 is the meet of the state of v_3, definitely not-null, and the state of v_1, possibly null, and we must conservatively assume that v_4 may be null.

3 Building the Intermediate Program Representation

A *live range splitting strategy* $\mathcal{P}_v = I_\uparrow \cup I_\downarrow$ over a variable v consists of two sets of control flow nodes (see Section 2.1 for a definition of control flow nodes). We let $I_\downarrow$ denote a set of control flow nodes that produce information for a forward analysis. Similarly, we let $I_\uparrow$ denote a set of control flow nodes that are interesting for a backward analysis. The live-range of v must be split at least at every control flow node in $\mathcal{P}_v$. Going back to the examples from Section 2.2, we have the live range splitting strategies enumerated below. Further examples are given in Figure 5.

- **Class inference** is a backward analysis that takes information from the uses of variables. Thus, for each variable, the live-range splitting strategy contains the set of control flow nodes where that variable is used. For instance, in Figure 1(b), we have that $\mathcal{P}_v = \{l_4, l_6, l_7\}_\uparrow$.
- **Constant propagation** is a forward analysis that takes information from definition sites. Thus, for each variable v, the live-range splitting strategy is characterized by the set of points where v is defined. For instance, in Figure 2(b), we have that $\mathcal{P}_b = \{l_2, l_5\}_\downarrow$.
- **Taint analysis** is a forward analysis that takes information from control flow nodes where variables are defined, and conditional tests that use these variables. For instance, in Figure 3(a), we have that $\mathcal{P}_v = \{l_1, l_2, \mathrm{Out}(l_5)\}_\downarrow$.
- Nanda *et al.*'s **null pointer analysis** [30] is a forward flow problem that takes information from definitions and uses. For instance, in Figure 4(a), we have that $\mathcal{P}_v = \{l_1, l_2, l_3, l_4\}_\downarrow$.

The algorithm SSIfy in Figure 6 implements a live range splitting strategy in three steps: split, rename and clean, which we describe in the rest of this section.

Splitting live ranges through the creation of new definitions of variables: To implement $\mathcal{P}_v$, we must split the live ranges of v at each control flow node listed by $\mathcal{P}_v$. However, these control flow nodes are not the only ones where splitting might be necessary. As we have pointed out in Section 2.1, we might have, for the same original variable, many different sources of information reaching a common meet point. For instance, in Figure 3(b), there exist two definitions of variable v: v_1 and v_2, that reach the use of v at l_5. Information that flows forward from l_3 and l_4 collide at l_5, the meet point of the if-then-else. Hence the live-range of v has to be split at the entry of l_5, e.g., at $\mathrm{In}(l_5)$, leading to a new definition v_3. In general, the set of control flow nodes where information collide can be easily characterized by join sets [16]. The join set of a group of nodes P contains the CFG nodes that can be reached by two or more nodes of P through disjoint paths. Join sets can be over-approximated by the notion of iterated dominance frontier [47], a core concept in SSA construction algorithms, which, for the sake of completeness, we recall below:

- **Dominance**: a CFG node n dominates a node n' if every program path from the entry node of the CFG to n' goes across n. If $n \neq n'$, then we say that n *strictly* dominates n'.
- **Dominance frontier** (DF): a node n' is in the dominance frontier of a node n if n dominates a predecessor of n', but does not strictly dominate n'.

Client	**Splitting strategy $\mathcal{P}$**
Alias analysis, reaching definitions cond. constant propagation [46]	$\mathit{Defs}_{\downarrow}$
Partial Redundancy Elimination [2, 41]	$\mathit{Defs}_{\downarrow} \bigcup \mathit{LastUses}_{\uparrow}$
ABCD [6], taint analysis [36], range analysis [44, 21]	$\mathit{Defs}_{\downarrow} \bigcup \mathrm{Out}(\mathit{Conds})_{\downarrow}$
Stephenson's bitwidth analysis [43]	$\mathit{Defs}_{\downarrow} \bigcup \mathrm{Out}(\mathit{Conds})_{\downarrow} \bigcup \mathit{Uses}_{\uparrow}$
Mahlke's bitwidth analysis [28]	$\mathit{Defs}_{\downarrow} \bigcup \mathit{Uses}_{\uparrow}$
An's type inference [23], class inference [11]	$\mathit{Uses}_{\uparrow}$
Hochstadt's type inference [45]	$\mathit{Uses}_{\uparrow} \bigcup \mathrm{Out}(\mathit{Conds})_{\uparrow}$
Null-pointer analysis [30]	$\mathit{Defs}_{\downarrow} \bigcup \mathit{Uses}_{\downarrow}$

Figure 5: Live range splitting strategies for different data-flow analyses. We use *Defs* (resp. *Uses*) to denote the set of instructions that define (resp. use) the variable; *Conds* to denote the set of instructions that apply a conditional test on a variable; Out(*Conds*) the exits of the corresponding basic blocks; *LastUses* to denote the set of instructions where a variable is used, and after which it is no longer live.

```
function SSIfy(var v, Splitting_Strategy P_v)
    split(v, P_v)
    rename(v)
    clean(v)
```

Figure 6: Split the live ranges of v to convert it to SSI form

- **Iterated dominance frontier** (DF^{+}): the iterated dominance frontier of a node n is the limit of the sequence:

$$\begin{array}{rcl} DF_1 & = & DF(n) \\ DF_{i+1} & = & DF_i \cup \{DF(z) \mid z \in DF_i\} \end{array}$$

Similarly, split sets created by the backward propagation of information can be over-approximated by the notion of *iterated post-dominance frontier* (pDF^{+}), which is the DF^{+} [3] of the CFG where orientation of edges have been reverted. If $e = (u, v)$ is an edge in the control flow graph, then we define the dominance frontier of e, i.e., $DF(e)$, as the dominance frontier of a fictitious node n placed at the middle of e. In other words, $DF(e)$ is $DF(n)$, assuming that (u, n) and (n, v) would exist. Given this notion, we also define $DF^{+}(e)$, $pDF(e)$ and $pDF^{+}(e)$.

Figure 7 shows the algorithm that creates new definitions of variables. This algorithm has three phases. First, in lines 3-9 we create new definitions to split the live ranges of variables due to *backward* collisions of information. These new definitions are created at the iterated post-dominance frontier of control flow nodes that originate information. Notice that if the control flow node is a join (entry of a CFG node), information actually originate from each incoming edges (line 6). In lines 10-16 we perform the inverse operation: we create new definitions of variables due to the forward collision of information. Finally, in lines 17-23 we actually insert the new definitions of v. These new definitions might be created by σ functions (due exclusively to the splitting in lines 3-9); by ϕ-functions (due exclusively to the splitting in lines 10-16);

```
1  function split(var v, Splitting_Strategy P_v = I_↓ ∪ I_↑)
2      "compute the set of split points"
3      S_↑ = ∅
4      foreach i ∈ I_↑:
5          if i.is_join:
6              foreach e ∈ incoming_edges(i):
7                  S_↑ = S_↑ ⋃ Out(pDF⁺(e))
8          else:
9              S_↑ = S_↑ ⋃ Out(pDF⁺(i))
10     S_↓ = ∅
11     foreach i ∈ S_↑ ⋃ Defs(v) ⋃ I_↓:
12         if i.is_fork:
13             foreach e ∈ outgoing_edges(i)
14                 S_↓ = S_↓ ⋃ In(DF⁺(e))
15         else:
16             S_↓ = S_↓ ⋃ In(DF⁺(i))
17     S = P_v ⋃ S_↑ ⋃ S_↓
18     "Split live range of v by inserting φ, σ, and copies"
19     foreach i ∈ S:
20         if i does not already contain any definition of v:
21             if i.is_join: insert "v = φ(v, ..., v)" at i
22             elseif i.is_fork: insert "(v, ..., v) = σ(v)" at i
23             else: insert a copy "v = v" at i
```

Figure 7: Live range splitting. We use $\text{In}(l)$ to denote a control flow node at the entry of l, and $\text{Out}(l)$ to denote a control flow node at the exit of l. We let $\text{In}(S) = \{\text{In}(l) \mid l \in S\}$. $\text{Out}(S)$ is defined in a similar way.

or by parallel copies. Contrary to Singer's algorithm, originally designed to produce SSI form programs, we do not iterate between the insertion of ϕ and σ functions.

The Algorithm split preserves the SSA property, even for data-flow analyses that do not require it. As we see in line 11, the loop that splits meet nodes forwardly include, by default, all the definition sites of a variable. We chose to implement it in this way for practical reasons: the SSA property gives us access to a fast liveness check [7], which is useful in actual compiler implementations. This algorithm inserts ϕ and σ functions conservatively. Consequently, we may have these special instructions at control flow nodes that are not true meet nodes. In other words, we may have a ϕ-function $v = \phi(v_1, v_2)$, in which the abstract states of v_1 and v_2 are the same in a final solution of the data-flow problem.

Variable Renaming: The algorithm in Figure 8 builds def-use and use-def chains for a program after live range splitting. This algorithm is similar to the standard algorithm used to rename variables during the SSA construction [3, Algorithm 19.7]. To rename a variable v we traverse the program's dominance tree, from top to bottom, stacking each new definition of v that we find. The definition currently on the top of the stack is used to replace all the uses of v that we find during the traversal. If the stack is empty, this means that the variable is not defined at that point. The renaming process replaces the uses of undefined variables by undef (line 3). We have two methods, *stack*.set_use and *stack*.set_def to build the chain relations between the variables. Notice that sometimes we must rename a single use inside a ϕ-function, as in lines 10-11 of the algorithm. For simplicity we consider this single use as a simple assignment when calling *stack*.set_use, as one can see in line 11. Similarly, if we must rename a single definition inside a σ-function, then we treat it as a simple assignment, like we do in lines 8-9 of the algorithm.

```
function rename(var v)
      "Compute use-def & def-use chains"
      "We consider here that stack.peek() = undef if stack.isempty(),
        and that Def(undef) = entry"
      stack = ∅
      foreach CFG node n in dominance order:
            foreach m that is a predecessor of n:
                  if exists d_m of the form "l^m : v = ..." in a σ-function in Out(m):
                        stack.set_def(d_m)
                  if exits u_m of the form "··· = l^m : v" in a ϕ-function in In(n):
                        stack.set_use(u_m)
            if exists a ϕ-function d in In(n) that defines v:
                  stack.set_def(d)
            foreach instruction u in n that uses v:
                  stack.set_use(u)
            if exists an instruction d in n that defines v:
                  stack.set_def(d)
            foreach σ-function u in Out(n) that uses v:
                  stack.set_use(u)

function stack.set_use(instruction inst):
      while Def(stack.peek()) does not dominate inst: stack.pop()
      v_i = stack.peek()
      replace the uses of v by v_i in inst
      if v_i ≠ undef: set Uses(v_i) = Uses(v_i) ⋃ inst

function stack.set_def(instruction inst):
      let v_i be a fresh version of v
      replace the defs of v by v_i in inst
      set Def(v_i) = inst
      stack.push(v_i)
```

Figure 8: Versioning

Dead and Undefined Code Elimination: The algorithm in Figure 9 eliminates ϕ-functions that define variables not actually used in the code, σ-functions that use variables not actually defined in the code, and parallel copies that either define or use variables that do not reach any actual instruction. "Actual" instructions are those instructions that already existed in the program before we transformed it with split. In line 3 we let "*web*" be the set of versions of v, so as to restrict the cleaning process to variable v, as we see in lines 4-6 and lines 10-12. The set "*active*" is initialized to actual instructions in line 4. Then, during the loop in lines 5-8 we add to active ϕ-functions, σ-functions, and copies that can reach actual definitions through use-def chains. The corresponding version of v is then marked as *defined* (line 8). The next loop, in lines 11-14 performs a similar process to add to the active set the instructions that can reach actual uses through def-use chains. The corresponding version of v is then marked as *used* (line 14). Each non live variable (see line 15), i.e. either undefined or dead (non used) is replaced by undef in all ϕ, σ, or copy functions where it appears. This is done in lines 15-18. Finally useless ϕ, σ, or copy functions are removed in lines 19-20. As a historical curiosity, Cytron *et al.*'s procedure to build SSA form produced what is called *the minimal representation* [16]. Some of the ϕ-functions in the minimal representation define variables that are never used. Briggs *et al.* [8] remove these

```
function clean(var v)
    let web = {v_i | v_i is a version of v}
    let defined = ∅
    let active = { inst | inst is actual instruction and web ∩ inst.defs ≠ ∅}
    while exists inst in active s.t. web ∩ inst.defs \ defined ≠ ∅:
        foreach v_i ∈ web ∩ inst.defs\defined:
            active = active ∪ Uses(v_i)
            defined = defined ∪ {v_i}
    let used = ∅
    let active = {inst |inst is actual instruction and web ∩ inst.uses ≠ ∅}
    while exists inst ∈ active s.t. inst.uses\used ≠ ∅:
        foreach v_i ∈ web ∩ inst.uses\used:
            active = active ∪ Def(v_i)
            used = used ∪ {v_i}
    let live = defined ∩ used
    foreach non actual inst ∈ Def(web):
        foreach v_i operand of inst s.t. v_i ∉ live:
                replace v_i by undef
        if inst.defs = {undef} or inst.uses = {undef}
            eliminate inst from the program
```

Figure 9: Dead and undefined code elimination. Original instructions not inserted by split are called *actual* instruction. We let *inst*.defs denote the set of variables defined by *inst*, and *inst*.uses denote the set of variables used by *inst*.

variables; hence, producing what compiler writers normally call *pruned SSA-form*. We close this section stating that the SSIfy algorithm preserves the semantics of the modified program [1]:

Theorem 1 (Semantics)**.** SSIfy *maintains the following property: if a value n written into variable v at control flow node i' is read at a control flow node i in the original program, then the same value assigned to a version of variable v at control flow node i' is read at a control flow node i after transformation.*

The Propagation Engine: Def-use chains can be used to solve, sparsely, a PLV problem about any program that fulfills the SSI property. However, in order to be able to rely on these def-use chains, we need to derive a sparse constraint system from the original - dense - system. This sparse system is constructed according to Definition 7. Theorem 2 states that such a system exists for any program, and can be obtained directly from the Algorithm SSIfy. The algorithm in Figure 10 provides worklist based solvers for backward and forward sparse data-flow systems built as in Definition 7.

Definition 7 (SSI constrained system)**.** *Let E^{ssi}_{dense} be a forward (resp. backward) constraint system extracted from a program that meets the SSI properties. Hence, for each pair (variable v, program point p) we have equations $[v]^p = [v]^p \wedge F_v^{s,p}([v_1]^s, \ldots, [v_n]^s)$. We define a system of sparse equations E^{ssi}_{sparse} as follows:*

- *Let $\{a, \ldots, b\}$ be the variables used (resp. defined) at control flow node i, where variable v is defined (resp. used). Let s and p be the program points around i. The LINK property ensures that $F_v^{s,p}$ depends only on some $[a]^s \ldots [b]^s$. Thus, there exists a function G_v^i defined as the projection of $F_v^{s,p}$ on $\mathcal{L}_a \times \cdots \times \mathcal{L}_b$, such that $G_v^i([a]^s, \ldots, [b]^s) = F_v^{s,p}([v_1]^s, \ldots, [v_n]^s)$.*

[1] The theorems in the main part of this paper are proved in the appendix

```
function forward_propagate(transfer_functions 𝒢)
    worklist = ∅
    foreach variable v: [v] = ⊤
    foreach instruction i: worklist += i
    while worklist ≠ ∅:
        let i ∈ worklist
        worklist −= i
        foreach v ∈ i.defs:
            [v]_new = [v] ∧ G_v^i([i.uses])
            if [v] ≠ [v]_new:
                worklist += Uses(v)
                [v] = [v]_new
```

Figure 10: Forward propagation engine under SSI. For backward propagation, we replace i.defs by i.uses, i.uses by i.defs, and Uses(v) by Def(v)

- *The sparse constrained system associates with each variable v, and each definition (resp. use) point i of v, the corresponding constraint $[v] \sqsubseteq G_v^i([a], \ldots, [b])$ where $a, \ldots, b$ are used (resp. defined) at i.*

Theorem 2 (Correctness of SSIfy)**.** *The execution of SSIfy(v, $\mathcal{P}_v$), for every variable v in the target program, creates a new program representation such that:*

1. *there exists a system of equations E_{dense}^{ssi}, isomorphic to E_{dense} for which the new program representation fulfills the SSI property.*
2. *if E_{dense} is monotone then E_{dense}^{ssi} is also monotone.*

4 Our Approach vs Other Sparse Evaluation Frameworks

There have been previous efforts to provide theoretical and practical frameworks in which data-flow analyses could be performed sparsely. In order to clarify some details of our contribution, this section compares it with three previous approaches: Choi's Sparse Evaluation Graphs, Ananian's Static Single Information form and Oh's Sparse Abstract Interpretation Framework.

Sparse Evaluation Graphs: Choi's *Sparse Evaluation Graphs* [12] are one of the earliest data-structures designed to support sparse analyses. The nodes of this graph represent program regions where information produced by the data-flow analysis might change. Choi *et al.*'s ideas have been further expanded, for example, by Johnson *et al.*'s *Quick Propagation Graphs* [25], or Ramalingan's *Compact Evaluation Graphs* [35]. Nowadays we have efficient algorithms that build such data-structures [24, 33]. These graphs improve many data-flow analyses in terms of runtime and memory consumption. However, they are more limited than our approach, because they can only handle sparsely problems that Zadeck has classified as *Partitioned Variable* (PVP). In these problems, a program variable can be analyzed independently from the others. Reaching definitions and liveness analysis are examples of PVPs, as this kind of information can be computed for one program variable independently from the others. For these problems we can build intermediate program representations isomorphic to SEGs, as we state in Theorem 3. However, many data-flow problems, in particular the PLV analyses that we mentioned in Section 2.2, do not fit into this category. Nevertheless, we can handle them sparsely. The SEGs can still support

PLV problems, but, in this case, a new SEG vertex would be created for every control flow node where new information is produced, and we would have a dense analysis.

Theorem 3 (Equivalence SSI/SEG)**.** *Given a forward Sparse Evaluation Graph (SEG) that represents a variable v in a program representation Prog with CFG G, there exists a live range splitting strategy that once applied on v builds a program representation that is isomorphic to SEG.*

Static Single Information Form and Similar Program Representations: Scott Ananian has introduced in the late nineties the *Static Single Information* (SSI) form, a program representation that supports both forward and backward analyses [2]. This representation was later revisited by Jeremy Singer [41]. The σ-functions that we use in this paper is a notation borrowed from Ananian's work, and the algorithms that we discuss in Section 3 improve on Singer's ideas. Contrary to Singer's algorithm we do not iterate between the insertion of phi and sigma functions. Consequently, as we will show in Section 5, we insert less phi and sigma functions. Nonetheless, as we show in Theorem 2, our method is enough to ensure the SSI properties for any combination of unidirectional problems. In addition to the SSI form, we can emulate several other different representations, by changing our parameterizations. Notice that for SSI we have $\{Defs_\downarrow \cup LastUses_\uparrow\}$. For Bodik's e-SSA [6] we have $Defs_\downarrow \bigcup \mathrm{Out}(Conds)_\downarrow$. Finally, for SSU [22, 27, 34] we have $Uses_\uparrow$.

The SSI constrained system might have several inequations for the same left-hand-side, due to the way we insert phi and sigma functions. Definition 6, as opposed to the original SSI definition [2, 41], does not ensure the SSA or the SSU properties. These guarantees are not necessary to every sparse analysis. It is a common assumption in the compiler's literature that "data-flow analysis (...) can be made simpler when each variable has only one definition", as stated in Chapter 19 of Appel's textbook [3]. A naive interpretation of the above statement could lead one to conclude that data-flow analyses become simpler as soon as the program representation enforces a single source of information per live-range: SSA for forward propagation, SSU for backward, and the *original* SSI for bi-directional analyses. This premature conclusion is contradicted by the example of dead-code elimination, a backward data-flow analysis that the SSA form simplifies. Indeed, the SSA form fulfills our definition of the SSI property for dead-code elimination. Nevertheless, the corresponding constraint system may have several inequations, with the same left-hand-side, i.e., one for each use of a given variable v. Even though we may have several sources of information, we can still solve this backward analysis using the algorithm in Figure 10. To see this fact, we can replace G_v^i in Figure 10 by "*i is a useful instruction or one of its definitions is marked as useful*" and one obtains the classical algorithm for dead-code elimination.

Sparse Abstract Interpretation Framework: Recently, Oh *et al.* [32] have designed and tested a framework that sparsifies flow analyses modelled via abstract interpretation. They have used this framework to implement standard analyses on the interval [14] and on the octogon lattices [29], and have processed large code bodies. We believe that our approach leads to a sparser implementation. We base this assumption on the fact that Oh *et al.*'s approach relies on standard def-use chains to propagate information, whereas in our case, the merging nodes combine information before passing it ahead. As an example, lets consider the code `if () then a=•; else a=•; endif if () then •=a; else •=a; endif` under a forward analysis that generates information at definitions and requires it at uses. We let the symbol • denote unimportant values. In this scenario, Oh et al.'s framework creates four dependence links between the two control flow nodes where information is produced and the two control flow nodes where it is consumed.

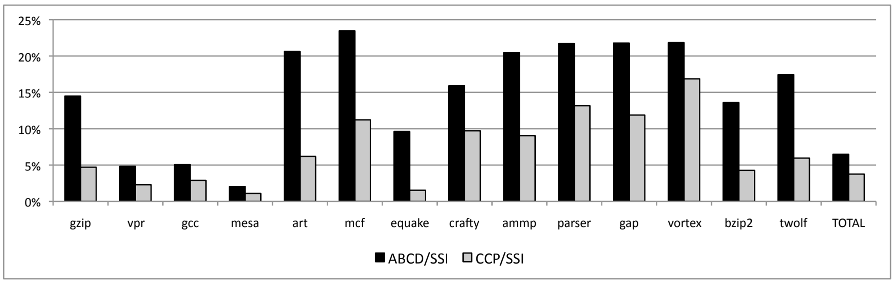

Figure 11: Comparison of the time taken to produce the different representations. 100% is the time to use the SSI live range splitting strategy. The shorter the bar, the faster the live range splitting strategy. The SSI conversion took 1315.2s in total, the ABCD conversion took 85.2s, and the CCP conversion took 49.4s.

Our method, on the other hand, converts the program to SSA form; hence, creating two names for variable a. We avoid the extra links because a ϕ-function merges the data that comes from these names before propagating it to the use sites.

5 Experimental Results

This section describes an empirical evaluation of the size and runtime efficiency of our algorithms. Our experiments were conducted on a dual core `Intel Pentium D` of 2.80GHz of clock, 1GB of memory, running `Linux Gentoo`, version 2.6.27. Our framework runs in LLVM 2.5 [26], and it passes all the tests that LLVM does. The LLVM test suite consists of over 1.3 million lines of C code. In this paper we show results for SPEC CPU 2000. To compare different live range splitting strategies we generate the program representations below. Figure 5 explains the sets *Defs, Uses* and *Conds*.

1. *SSI*: Ananian's Static Single Information form [2] is our baseline. We build the SSI program representation via Singer's iterative algorithm.

2. *ABCD*: $(\{Defs, Conds\}_{\downarrow})$. This live range splitting strategy generalizes the ABCD algorithm for array bounds checking elimination [6]. An example of this live range splitting strategy is given in Figure 3.

3. *CCP*: $(\{Defs, Conds_{eq}\}_{\downarrow})$. This splitting strategy, which supports Wegman *et al.*'s [46] conditional constant propagation, is a subset of the previous strategy. Differently of the ABCD client, this client requires that only variables used in equality tests, e.g., `==`, undergo live range splitting. That is, $Conds_{eq}(v)$ denotes the conditional tests that check if v equals a given value.

Runtime: The chart in Figure 11 compares the execution time of the three live range splitting strategies. We show only the time to perform live range splitting. The time to execute the optimization itself, removing array bound checks or performing constant propagation, is not shown. The bars are normalized to the running time of the SSI live range splitting strategy. On the average, the ABCD client runs in 6.8% and the CCP client runs in 4.1% of the time of SSI. These two forward analyses tend to run faster in benchmarks with sparse control flow graphs,

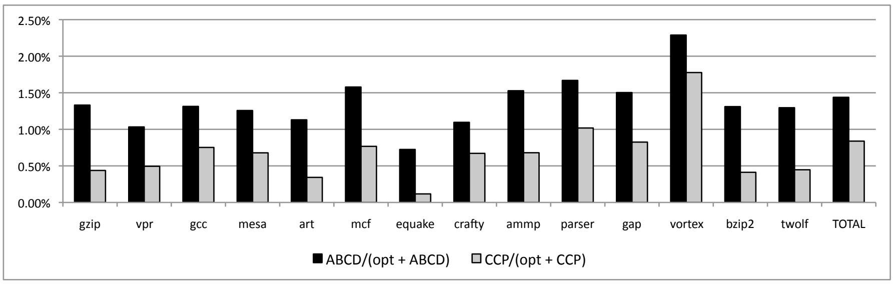

Figure 12: Execution time of two different live range splitting strategies compared to the total time taken by machine independent LLVM optimizations (`opt` -O1). 100% is the time taken by `opt`. The shorter the bar, the faster the conversion.

which present fewer conditional branches, and therefore fewer opportunities to restrict the ranges of variables.

In order to put the time reported in Figure 11 in perspective, Figure 12 compares the running time of our live range splitting algorithms with the time to run the other standard optimizations in our baseline compiler[2]. In our setting, LLVM -O1 runs 67 passes, among analysis and optimizations, which include partial redundancy elimination, constant propagation, dead code elimination, global value numbering and invariant code motion. We believe that this list of passes is a meaningful representative of the optimizations that are likely to be found in an industrial strength compiler. The bars are normalized to the optimizer's time, which consists of the time taken by machine independent optimizations plus the time taken by one of the live range splitting clients, e.g, ABCD or CCP. The ABCD client takes 1.48% of the optimizer's time, and the CCP client takes 0.9%. To emphasize the speed of these passes, we notice that the bars do not include the time to do machine dependent optimizations such as register allocation.

Space: Figure 13 outlines how much each live range splitting strategy increases program size. We show results only to the ABCD and CCP clients, to keep the chart easy to read. The SSI conversion increases program size in 17.6% on average. This is an absolute value, i.e., we sum up every ϕ and σ function inserted, and divide it by the number of bytecode instructions in the original program. This compiler already uses the SSA-form by default, and we do not count as new instructions the ϕ-functions originally used in the program. The ABCD client increases program size by 2.75%, and the CCP client increases program size by 1.84%.

An interesting question that deserves attention is "What is the benefit of using a sparse dataflow analysis in practice?" We have not implemented dense versions of the ABCD or the CCP clients. However, previous works have shown that sparse analyses tend to outperform equivalent dense versions in terms of time and space efficiency [12, 35]. In particular, the e-SSA format used by the ABCD and the CCP optimizations is the same program representation adopted by the tainted flow framework of Rimsa *et al.* [36, 37], which has been shown to be faster than a dense implementation of the analysis, even taking the time to perform live range splitting into consideration.

[2]To check the list of LLVM's target independent optimizations try `llvm-as < /dev/null | opt -std-compile-opts -disable-output -debug-pass=Arguments`

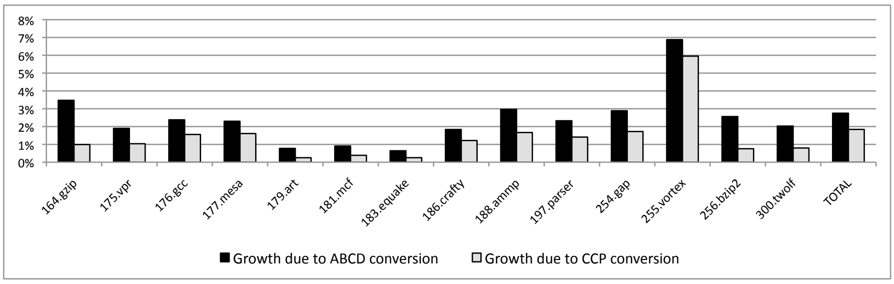

Figure 13: Growth in program size due to the insertion of new ϕ and σ functions to perform live range splitting.

6 Conclusion

This paper has presented a systematic way to build program representations that suit sparse data-flow analyses. We build different program representations by splitting the live ranges of variables. The way in which we split live ranges depends on two factors: (i) which control flow nodes produce new information, e.g., uses, definitions, tests, etc; and (ii), how this information propagates along the variable live range: forwardly or backwardly. We have used an implementation of our framework in LLVM to convert programs to the Static Single Information form [2], and to provide intermediate representations to the ABCD array bounds-check elimination algorithm [6] and to Wegman *et al.*'s Conditional Constant Propagation algorithm [46]. Our framework has been used by Couto *et al.* [19] and by Rodrigues *et al.* [38] in different implementations of range analyses. We have also used our live range splitting algorithm, implemented in the `phc` PHP compiler [4, 5], to provide the Extended Static Single Assignment form necessary to solve the tainted flow problem [36, 37].

Extending our Approach. For the sake of simplicity, in this paper we have restricted our discussion to: non relational analysis (PLV), intermediate-representation based appoach, and scalar variables without aliasing.

(1) *non relation analysis.* In this paper we have focused on PLV problems, i.e. solved by analyses that associate some information with each variable individually. For instance, we bind i to a range $0 \leq i < \texttt{MAX\_N}$, but we do not relate i and j, as in $0 \leq i < j$. A relational analysis that provides a all-to-all relation between all variables of the program is dense by nature, as any control flow node both produces and consumes information for the analysis. Nevertheless, our framework is compatible with the notion of *packing*. Each pack is a set of variable groups selected to be related together. This approach is usually adopted in practical relational analyses, such as those used in Astrée [15, 29].

(2) *IR based approach.* Our framework constructs an intermediate representation (IR) that preserves the semantic of the program. Like the SSA form, this IR has to be updated, and prior to final code generation, destructed. Our own experience as compiler developers let us believe that manipulating an IR such as SSA has many engineering advantages over building, and afterward dropping, a separate sparse evaluation graph (SEG) for each analysis. Testimony of this observation is the fact that the SSA form is used in virtually every modern compiler. Although this opinion is admittedly arguable, we would like to point out that updating and destructing our SSI form is equivalent to the update and destruction of SSA form. More importantly, there is no fundamental limitation in using our technique to build a separate SEG without modifying

the IR. This SEG will inherit the sparse properties as his corresponding SSI flavor, with the benefit of avoiding the quadratic complexity of direct def-use chains ($|\text{Defs}(v)| \times |\text{Uses}(v)|$ for a variable v) thanks to the use of ϕ and σ nodes. Note that this quadratic complexity becomes critical when dealing with code with aliasing or predication [32, pp.234].

(3) *analysis of scalar variables without aliasing or predication.* The most successful flavor of SSA form is the minimal and pruned representation restricted to scalar variables. The SSI form that we describe in this paper is akin to this flavor. Nevertheless, there exists several extensions to deal with code with predication (e.g. ψ-SSA form [18]) and aliasing (e.g. Hashed SSA [13] or Array SSA [20]). Such extensions can be applied without limitations to our SSI form allowing a wider range of analyses involving object aliasing and predication.

References

[1] W. B. Ackerman. *Efficient Implementation of Applicative Languages.* PhD thesis, MIT, 1984.

[2] Scott Ananian. The static single information form. Master's thesis, MIT, September 1999.

[3] Andrew W. Appel and Jens Palsberg. *Modern Compiler Implementation in Java.* Cambridge University Press, 2nd edition, 2002.

[4] Paul Biggar. *Design and Implementation of an Ahead-of-Time Compiler for PHP.* PhD thesis, Trinity College Dublin, 2009.

[5] Paul Biggar, Edsko de Vries, and David Gregg. A practical solution for scripting language compilers. In *SAC*, pages 1916–1923. ACM, 2009.

[6] Rastislav Bodik, Rajiv Gupta, and Vivek Sarkar. ABCD: eliminating array bounds checks on demand. In *PLDI*, pages 321–333. ACM, 2000.

[7] Benoit Boissinot, Sebastian Hack, Daniel Grund, Benoit Dupont de Dinechin, and Fabrice Rastello. Fast liveness checking for SSA-form programs. In *CGO*, pages 35–44. IEEE, 2008.

[8] Preston Briggs, Keith D. Cooper, and Linda Torczon. Improvements to graph coloring register allocation. *TOPLAS*, 16(3):428–455, 1994.

[9] Zoran Budimlic, Keith D. Cooper, Timothy J. Harvey, Ken Kennedy, Timothy S. Oberg, and Steven W. Reeves. Fast copy coalescing and live-range identification. In *PLDI*, pages 25–32. ACM, 2002.

[10] Robert Cartwright and Mattias Felleisen. The semantics of program dependence. *SIGPLAN Not.*, 24(7):13–27, 1989.

[11] Craig Chambers and David Ungar. Customization: optimizing compiler technology for self, a dynamically-typed object-oriented programming language. *SIGPLAN Not.*, 24(7):146–160, 1989.

[12] Jong-Deok Choi, Ron Cytron, and Jeanne Ferrante. Automatic construction of sparse data flow evaluation graphs. In *POPL*, pages 55–66. ACM, 1991.

[13] Fred Chow, Sun Chan, Shin-Ming Liu, Raymond Lo, and Mark Streich. Effective representation of aliases and indirect memory operations in SSA form. In *CC*, pages 253–267. Springer, 1996.

[14] P. Cousot and R. Cousot. Abstract interpretation: a unified lattice model for static analysis of programs by construction or approximation of fixpoints. In *POPL*, pages 238–252. ACM, 1977.

[15] Patrick Cousot, Radhia Cousot, Jérôme Feret, Laurent Mauborgne, Antoine Miné, and Xavier Rival. Why does astrée scale up? *Form. Methods Syst. Des.*, 35(3):229–264, 2009.

[16] Ron Cytron, Jeanne Ferrante, Barry K. Rosen, Mark N. Wegman, and F. Kenneth Zadeck. Efficiently computing static single assignment form and the control dependence graph. *TOPLAS*, 13(4):451–490, 1991.

[17] Luis Damas and Robin Milner. Principal type-schemes for functional programs. In *POPL*, pages 207–212, New York, NY, USA, 1982. ACM.

[18] François de Ferrière. Improvements to the ψ-SSA representation. In *SCOPES*, pages 111–121. ACM, 2007.

[19] Douglas do Couto Teixeira and Fernando Magno Quintao Pereira. The design and implementation of a non-iterative range analysis algorithm on a production compiler. In *SBLP*, pages 45–59. SBC, 2011.

[20] Stephen J. Fink, Kathleen Knobe, and Vivek Sarkar. Unified analysis of array and object references in strongly typed languages. In *SAS*, pages 155–174. Springer, 2000.

[21] Thomas Gawlitza, Jerome Leroux, Jan Reineke, Helmut Seidl, Gregoire Sutre, and Reinhard Wilhelm. Polynomial precise interval analysis revisited. *Efficient Algorithms*, 1:422 – 437, 2009.

[22] Lal George and Blu Matthias. Taming the IXP network processor. In *PLDI*, pages 26–37. ACM, 2003.

[23] Jong hoon An, Avik Chaudhuri, Jeffrey S. Foster, and Michael Hicks. Dynamic inference of static types for ruby. In *POPL*, pages 459–472. ACM, 2011.

[24] R. Johnson, D. Pearson, and K. Pingali. The program tree structure. In *PLDI*, pages 171–185. ACM, 1994.

[25] Richard Johnson and Keshav Pingali. Dependence-based program analysis. In *PLDI*, pages 78–89. ACM, 1993.

[26] Chris Lattner and Vikram S. Adve. LLVM: A compilation framework for lifelong program analysis & transformation. In *CGO*, pages 75–88. IEEE, 2004.

[27] Raymond Lo, Fred Chow, Robert Kennedy, Shin-Ming Liu, and Peng Tu. Register promotion by sparse partial redundancy elimination of loads and stores. In *PLDI*, pages 26–37. ACM, 1998.

[28] S. Mahlke, R. Ravindran, M. Schlansker, R. Schreiber, and T. Sherwood. Bitwidth cognizant architecture synthesis of custom hardware accelerators. *TCAD*, 20(11):1355–1371, 2001.

[29] Antoine Miné. The octagon abstract domain. *Higher Order Symbol. Comput.*, 19:31–100, 2006.

[30] Mangala Gowri Nanda and Saurabh Sinha. Accurate interprocedural null-dereference analysis for java. In *ICSE*, pages 133–143, 2009.

[31] Flemming Nielson, Hanne Riis Nielson, and Chris Hankin. *Principles of program analysis*. Springer, 2005.

[32] Hakjoo Oh, Kihong Heo, Wonchan Lee, Woosuk Lee, and Kwangkeun Yi. Design and implementation of sparse global analyses for c-like languages. In *PLDI*, pages 229–238. ACM, 2012.

[33] Keshav Pingali and Gianfranco Bilardi. Optimal control dependence computation and the roman chariots problem. In *TOPLAS*, pages 462–491. ACM, 1997.

[34] John Bradley Plevyak. *Optimization of Object-Oriented and Concurrent Programs*. PhD thesis, University of Illinois at Urbana-Champaign, 1996.

[35] G. Ramalingam. On sparse evaluation representations. *Theoretical Computer Science*, 277(1-2):119–147, 2002.

[36] Andrei Alves Rimsa, Marcelo D'Amorim, and Fernando M. Q. Pereira. Tainted flow analysis on e-SSA-form programs. In *CC*, pages 124–143. Springer, 2011.

[37] Andrei Alves Rimsa, Marcelo D'Amorim, Fernando M. Q. Pereira, and Roberto Bigonha. Efficient static checker for tainted variable attacks. *Science of Computer Programming*, 80:91–105, 2014.

[38] Raphael Ernani Rodrigues, Victor Hugo Sperle Campos, and Fernando Magno Quintao Pereira. A fast and low overhead technique to secure programs against integer overflows. In *CGO*, pages 1–11. ACM, 2013.

[39] Subhajit Roy and Y. N. Srikant. The hot path ssa form: Extending the static single assignment form for speculative optimizations. In *CC*, pages 304–323, 2010.

[40] Bernhard Scholz, Chenyi Zhang, and Cristina Cifuentes. User-input dependence analysis via graph reachability. Technical report, Sun, Inc., 2008.

[41] Jeremy Singer. *Static Program Analysis Based on Virtual Register Renaming*. PhD thesis, University of Cambridge, 2006.

[42] Vugranam C. Sreedhar, Roy Dz ching Ju, David M. Gillies, and Vatsa Santhanam. Translating out of static single assignment form. In *SAS*, pages 194–210. Springer-Verlag, 1999.

[43] Mark Stephenson, Jonathan Babb, and Saman Amarasinghe. Bitwidth analysis with application to silicon compilation. In *PLDI*, pages 108–120. ACM, 2000.

[44] Zhendong Su and David Wagner. A class of polynomially solvable range constraints for interval analysis without widenings. *Theoretical Computeter Science*, 345(1):122–138, 2005.

[45] Sam Tobin-Hochstadt and Matthias Felleisen. The design and implementation of typed scheme. *POPL*, pages 395–406, 2008.

[46] Mark N. Wegman and F. Kenneth Zadeck. Constant propagation with conditional branches. *TOPLAS*, 13(2), 1991.

[47] Michael Weiss. The transitive closure of control dependence: the iterated join. *TOPLAS*, 1(2):178–190, 1992.

[48] Frank Kenneth Zadeck. *Incremental Data Flow Analysis in a Structured Program Editor*. PhD thesis, Rice University, 1984.

A Isomorphism to Sparse Evaluation Graphs

Given a control flow graph G, Choi *et al.* define a sparse evaluation graph as a tuple $\langle N_{SG}, E_{SG}, M \rangle$, such that:

- N_{SG} is a set of nodes defined as follows:
  1. N_{SG} contains a node n_s representing the entry control flow node $s \in G$;
  2. N_{SG} contains a node n_p for each control flow node $p \in G$ that is associated with a non-identity transfer function.
  3. N_{SG} contains a node n_m for each point m in the iterated dominance frontier of the control flow nodes of G used to build the nodes in step (1) and (2). These are called *meet* nodes.
- We let P denote the set of control flow nodes $p \in G$ used in step 2 above, plus the control flow node $s \in G$ used in step 1 above; we let M denote the set of control flow nodes $m \in G$ used in step 3 above; if we let $S = P \bigcup M$ then we define E_{SG} as follows:
  1. there is an edge $(n_q, n_m) \in N_{SG}^2$ whenever $m \in M$ and q is, among all the nodes in S, the immediate dominator of one of the CFG predecessors of m. See `search(3b)` and `link(2b)` in Choi *et al* [12];
  2. there is an edge $(n_q, n_p) \in N_{SG}^2$ whenever $p \in P$, and q is, among all the nodes in S, the immediate dominator of p. See `search(1)` and `link(2b)` [12];
- The mapping function $M : E_G \mapsto N_{SG}$ associates to each edge (u, v) of the CFG the node $n_q \in N_{SG}$, whenever $q \in S$ is the immediate dominator of $u \in G$. See `search(3a)` [12]. This is done through the recursive function `search` that performs a topological traversal of the CFG (DFS of the dominance tree; See `search(4)` [12]).

Theorem 3 states that, for forward partitioned variable data-flow problems (PVP), the algorithm in Figure 6 can build program representations isomorphic to Sparse Evaluation Graphs. The proof that this result holds for backward data-flow problems, is analogous, and we omit it.

Lemma 1 (CFG cover)**.** *Let Prog be a program with its corresponding CFG G with start node s, and exit node x. Let $Prog'$ be the program that we obtain from Prog by:*

1. *adding a pseudo-definition of each variable to s;*
2. *adding a pseudo-use of each variable to x;*
3. *placing a pseudo-use of a variable v at each control flow node where v is defined;*
4. *converting the resulting program into SSA form.*

If v is a variable in Prog, then the live ranges of the different names of v in $Prog'$ completely partition the program points of G. In other words, each program point of G belongs to exactly one live range of v in $Prog'$.

Proof. First, v is alive at every program point of G, due to transformations (1), (2) and (3). Therefore, if V is the set of the different names of v after the conversion to SSA form in step (4), then any program point of G belongs to the live range of at least one $v' \in V$. The result follows from a well-know property of Cytron's SSA-form conversion algorithm [16], which, as observed by Sreedhar *et al.* [42], creates variables with non-intersecting live ranges. In other words, after the SSA renaming, two different names of v cannot be simultaneously alive at a program point p. □

[Equivalence SSI/SEG - See Theorem 3] Given a forward Sparse Evaluation Graph (SEG) that represents a variable v in a program representation $Prog$ with CFG G, there exits a live range splitting strategy that once applied on v builds a program representation that is isomorphic to SEG.

Proof. We argue that the SEG of v is isomorphic to the representation of v in $Prog'$, the program representation that we derive from $Prog$ by applying the transformations 1-3 listed in Lemma 1 in addition to a pass of SSIfy. If we let P, as before, be defined as the set of CFG nodes associated with non-identity transfer functions, plus the start node s of the CFG, then after we apply the splitting strategy $P_\downarrow$, we have that:

1. there will be exactly one definition per node of P and one definition per node of $DF^+(P)$. So there is an one-to-one correspondence between SSA definitions and SG nodes.

2. From Lemma 1 the live-ranges of the different names of v provides a partitioning of the program points of G. If v' is a new name of v, then each program point where v' is alive is dominated by v''s definition[3]. Each program point belongs to the live-range of the name of v whose definition immediately dominates it (among all definitions). Thus, live ranges give origin to a function that maps SSA definitions to program points. Consequently, there is an isomorphism between the live-ranges and the mapping function M.

3. def-use chains on $Prog'$ are isomorphic to the edges in E_{SG}: indeed a SEG node n_p is linked to n_q whenever (i) n_p immediately dominates n_q if $q \in P$; or (ii) n_q is in the dominance frontier of n_p if $q \in M$. In the former case the definition of v at p reaches the (pseudo-)use of v at q. In the latter this definition reaches the use of v at the ϕ-function placed at q by SSIfy$(v, P_\downarrow)$.

□

In the proof of Theorem 3 we had to augment the program with a pseudo-definition of v at the CFG's entry node and a pseudo-use at every actual definition of v and at the CFG's exit node. The difference between a code with or without pseudo uses/defs is related to the necessity to compute data-flow information beyond the live-ranges of variables or not. This necessity exists for optimizations such as partial redundancy elimination, which may move, create or delete code.

Figure 14 compares SEG and the forward live range splitting strategy in the example taken from Figure 11 of Choi *et al.* [12], which shows the reaching uses analysis. In the left we see the original program, and in the middle the SEG built for a forward flow analysis that extracts information from uses of variables. We have augmented the edges in the left CFG with the mapping M of SEG nodes to CFG edges. In the right we see the same CFG, augmented with pseudo defs and uses, after been transformed by SSIfy applied on the control flow nodes $\{S, 4, 5, 7, 11, 12\}_\downarrow$. The edges of this CFG are labeled with the definitions of v live there.

B Correctness of our SSIfication

In this section we consider a unidirectional forward (resp. backward) PLV problem stated as a set of equations $[v]^p = [v]^p \wedge F_v^{s,p}(\dots)$ for every variable v, each program point p, and each $s \in preds(p)$ (resp. $s \in succs(p)$). We rely on the nomenclature introduced by Definition 3 in order to prove Theorem 2.

[3]This is a classical result of SSA-form. See Budimlic *et al.* [9] for a proof

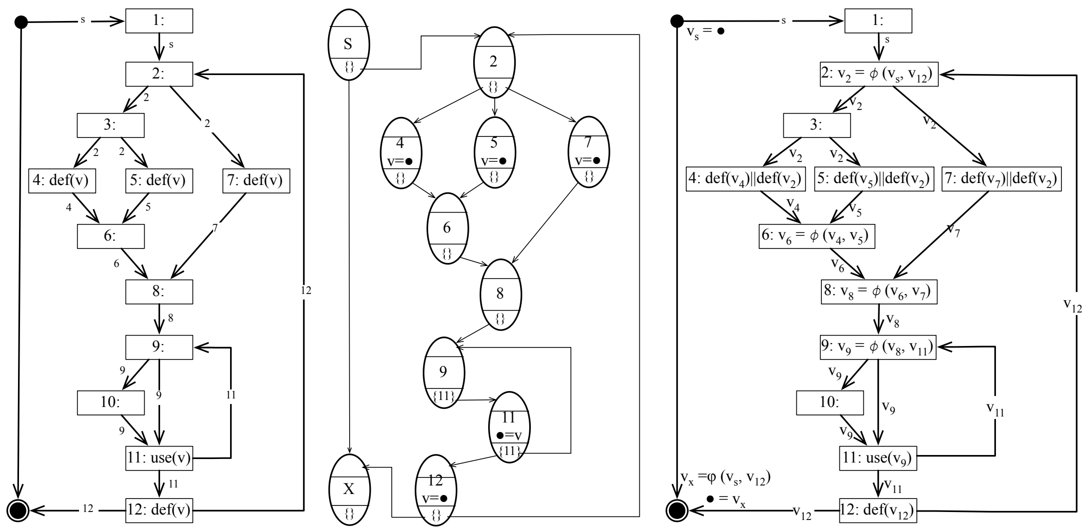

Figure 14: Example of equivalence between SEGs and our live range splitting strategy for reaching uses.

Lemma 2 (Live range preservation)**.** *If variable v is live at a program point p, then there is a version of v live at p after we run* SSIfy*.*

Proof. Split cannot remove any live range of v, as it only inserts "copies" from v to v, e.g., each copy has the same source and destination. Rename removes live ranges of v, but it replaces them with the live ranges of new versions of this variable whenever a use of v is renamed. Clean only removes "copies"; hence, all the original instructions remain in the code. □

Lemma 3 (Non-Overlapping)**.** *Two different versions of v, e.g., v_k and v_j cannot both be live at a program point p transformed by* SSIfy*.*

Proof. The only algorithm that creates new versions of v is rename. Each new version of v is unique, as we ensure in lines 28-30 of the algorithm. If rename changes the use of v to v_k at a control flow node i, then there exists a definition of v_k at some control flow node i' that dominates i, as we ensure in line 22 of the algorithm. Let us assume that we have two versions of v, e.g., v_k and v_j, live at a program point p, in order to derive a contradiction. In this case, there exist control flow nodes i_k where v_k is used, and i_j where v_j is used, reachable from p. Also there exists a control flow node i'_k where v_k is defined, and a control flow node i'_j where v_j is defined. i'_k dominates p, and i'_j dominates p. Thus, either i'_k dominates i'_j or vice-versa. Without loss of generality, let us assume that i'_k dominates i'_j. In this case, rename visits i'_k first, and upon visiting i'_j, places the definition of v_j on top of the definition of v_k in the stack in line 31. Thus, i'_k cannot dominate i'_j, or we would have, at i_k, a use of v_j, instead of v_k. □

[Semantics - Theorem 1] SSIfy maintains the following property: if a value n written to variable v at control flow node i' is read at a control flow node i in the original program, then the same value assigned to a version of variable v at control flow node i' is read at a control flow node i after transformation.

Proof. For simplicity, we will extend the meaning of "copy" to include not only the parallel copies placed at interior nodes, but also ϕ and σ-functions. Split cannot create new values, as it only inserts "copies". Clean cannot remove values, as it only removes "copies". From the hypothesis we know that the definition of v that reaches i is live at i. From Lemma 2 we know that there is a version of v live at i. From Lemma 3 we know that only one version of v can be live at i, and so rename cannot send new values to i. □

Now suppose that the program, not necessarily under SSI form, fulfills INFO and LINK from Definition 6 for a system of monotone equations E_{dense}, given as a set of constraints $[v]^p \sqsubseteq F_v^{s,p}([v_1]^s, \ldots, [v_n]^s)$. Consider a live range splitting strategy $\mathcal{P}_v$ that *includes* for each variable v the set of control flow nodes $I_\downarrow$ (resp. $I_\uparrow$) where $F_v^{s,p}$ is non-trivial. The following theorem states that Algorithm SSIfy creates a program form that fulfills the Static Single Information property.

[Correctness of SSIfy - Theorem 2] Given the conditions stated above, Algorithm SSIfy(v, $\mathcal{P}_v$) creates a new program representation such that:

1. there exists a system of equations E_{dense}^{ssi}, isomorphic to E_{dense} for which the new program representation fulfills the SSI property.
2. if E_{dense} is monotone then E_{dense}^{ssi} is also monotone.

Proof. We derive from this new program representation a system of equations isomorphic to the initial one by associating trivial transfer functions with the newly created "copies". The INFO and LINK properties are trivially maintained. As only trivial and constant functions have been added, monotonicity is maintained.

To show that we provide SPLIT-DEF, we must first show that each $i \in \text{live}(v)$ where F_v^s is non-trivial contains a definition (resp. last use) of v. The function split separates these program points in lines 9 and 16, and later, in line 23, inserts definitions in those control flow nodes. To show that we provide SPLIT-MEET, we must prove that each join (resp. split) node for which E_{dense} has possibly different values on its incoming edges should have a ϕ-function (resp. σ-function) for v. These program points are separated in lines 7 and 14 of split. To see why this is the case, notice that line 7 separates the program points in the iterated dominance frontier of program points that originate information that flows forward. These are, as a direct consequence of the definition of iterated dominance frontier, the control flow nodes where information collide. Similarly, line 14 separates the program points in the post-dominance frontier of regions which originate information that flows backwardly.

We ensure VERSION as a consequence of the SSA conversion. All our program representations preserve the SSA representation, as we include the definition sites of v in line 11 of split. Function rename ensures the existence of only one definition of each variable in the program code (line 27), and that each definition dominates all its uses (consequence of the traversal order). Therefore, the newly created live ranges are connected on the dominance tree of the source program. Function rename also creates a new program representation for which it is straightforward to build a system of equations E_{dense}^{ssi} isomorphic to E_{dense}: Firstly, the constraint variables are renamed in the same way that program variables are. Secondly, for each program variable, new system variables bound to $\perp$ are created for each program point outside of its live-range.

□

C Equivalence between sparse and dense analyses.

We have shown that SSIfy transforms a program P into another program P^{ssi} with the same semantics. Furthermore, this representation provides the SSI property for a system of equations E^{ssi}_{dense} that we extract from P^{ssi}. This system is isomorphic to the system of equations E_{dense} that we extract from P. From the so obtained program under SSI for the constrained system E^{ssi}_{dense}, Definition 7 shows how to construct a sparse constrained system E^{ssi}_{sparse}. When transfer functions are monotone and the lattice has finite height, Theorem 4 states the equivalence between the sparse and the dense systems. The purpose of this section is to prove this theorem. We start by introducing the notion of *coalescing*. Let E be a constraint system that associates with each $1 \leq i \leq n$ the constraint $a_i \sqsubseteq H_i(a_1, \dots, a_n)$, where each a_i is an element of a lattice $\mathcal{L}$ of finite height, and H_i is a monotone function from $\mathcal{L}^n$ to $\mathcal{L}$. Let $(A_1, \dots, A_n)$ be the maximum solution to this system, and let $1 \leq m \leq n$ such that $\forall i$, $1 \leq i \leq m$, $A_i = A_m$. We define a "coalesced" constraint system E_{coal} in the following way: for each $1 \leq i \leq m$ we create the constraint $b_m \sqsubseteq H_i(b_m, \dots, b_m, b_{m+1}, \dots, b_n)$; for each $m < i \leq n$ we create the constraint $b_i \sqsubseteq H_i(b_m, \dots, b_m, b_{m+1}, \dots, b_n)$. Lemma 4 shows that coalescing preserves the maximum solution of the original system.

Lemma 4 (Equivalence with coalescing)**.** *If E is a constraint system with maximum solution $(A_1, \dots, A_m, \dots, A_n)$, for any $i, j, 1 \leq i, j \leq m$ we have that $A_i = A_j$, and E_{coal} is the "coalesced" system that we derive from E, then the maximum solution of E_{coal} is $(A_m, \dots, A_n)$.*

Proof. Both system have a (unique) maximum solution (see e.g. [31]), although the solution of the "coalesced" system has smaller cardinality, e.g., n-m+1. Now, as $(A_m, \dots, A_m, A_{m+1}, \dots, A_n)$ is a solution to E, by definition of E_{coal}, $(A_m, \dots, A_n)$ is a solution to E_{coal}. Let us prove that this solution is maximum, i.e. for any solution $(B_m, \dots, B_n)$ of E_{coal}, we have $(B_m, \dots, B_n) \sqsubseteq (A_m, \dots, A_n)$. By definition of E_{coal}, we have that $(B_m, \dots, B_m, B_{m+1}, \dots, B_n)$ is a solution to E. As $(A_1, \dots, A_n)$ is maximum, we have $(B_m, \dots, B_m, B_{m+1}, \dots, B_n) \sqsubseteq (A_1, \dots, A_n)$. So $(B_m, \dots, B_n) \sqsubseteq (A_m, \dots, A_n)$. □

We now prove Theorem 4, which states that there exists a direct mapping between the maximum solution of a dense constraint system associated with a SSI-form program, and the sparse system that we can derive from it, according to Definition 7.

Theorem 4 (sparse ≡ dense)**.** *Consider a program in SSI-form that gives origin to a constraint system $E^{\mathrm{ssi}}_{\mathrm{dense}}$ associating with each variable v the constraints $[v]^p = [v]^p \wedge F^{s,p}_v([v_1]^s, \dots, [v_n]^s)$. Suppose that each $F^{s,p}_v$ is a monotone function from $\mathcal{L}^n$ to $\mathcal{L}$ where $\mathcal{L}$ is of finite height. Let $(Y_v)_{v \in variables}$ be the maximum solution of the corresponding sparse constraint system.*

Then, $(X^p_v)_{(v,i) \in variables \times prog\_points}$ with $\begin{cases} X^p_v = Y_v & \text{for } p \in \mathrm{live}(v) \\ X^p_v = \bot & \text{otherwise} \end{cases}$ is the maximum solution to $E^{\mathrm{ssi}}_{\mathrm{dense}}$.

Proof. The constraint systems E^{ssi}_{dense} and E^{ssi}_{sparse} have a maximum unique solution, because the transfer functions are monotone and $\mathcal{L}$ has finite height

The idea of the proof is to modify the constraint system E^{ssi}_{dense} into a system equivalent to E^{ssi}_{sparse}. To accomplish this transformation, we (i) replace each $F^{s,p}_v$ by G^i_v, where G^i_v is constructed as in Definition 7; (ii) for each v, coalesce $\{[v]^p\}_{p \in live(v)}$ into $[v]$; (iii) coalesce all other constraint variables into [undef].

The LINK property allows us to replace $F^{s,p}_v$ by G^i_v. Due to SPLIT-DEF, a new variable is defined at each control flow node where information is generated, and due to VERSION there is only one live range associated with each variable. Hence, $\{[v]^p\}_{p \in live(v)}$ is invariant. Due to INFO, we have that $\{[v]^p\}_{p \notin live(v)}$ is bound to $\bot$. Due to Lemma 4, we know that this new constraint system has a maximum solution $(Y_v)_{v \in variables \cup \mathsf{undef}}$: X_v^p equals Y_v for all $p \in live(v)$, and Y_{undef} otherwise.

We translate each constraint $[v]^p \sqsubseteq F_v^{s,p}([v_1]^s, \ldots, [v_n]^s)$ (with i the control flow node between p and s), in the original system, to a constraint in the "coalesced" one in the following way:

$$\begin{cases} \text{if } p \in live(v): & \text{if } i \in defs(v) & : [v] \sqsubseteq G_v^i([a], \ldots, [b]) & (1) \\ & \text{else} & : [v] \sqsubseteq [v] & (2) \\ \text{otherwise} & & : [\mathsf{undef}] \sqsubseteq \bot & (3) \end{cases}$$

Case (1) follows from LINK, case (2) follows from SPLIT-DEF, and case (3) follows from INFO. By ignoring undef that appears only in (3), and by removing the constraints produced by (2), which are useless, we obtain E_{sparse}^{ssi}. □

RESEARCH CENTRE
GRENOBLE – RHÔNE-ALPES

Inovallée
655 avenue de l'Europe Montbonnot
38334 Saint Ismier Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399