

# Optimal Guard Synthesis for Memory Safety

Thomas Dillig<sup>1</sup>, Isil Dillig<sup>1</sup>, and Swarat Chaudhuri<sup>2</sup>

<sup>1</sup> UT Austin

<sup>2</sup> Rice University

**Abstract.** This paper presents a new synthesis-based approach for writing low-level memory-safe code. Given a partial program with missing guards, our algorithm synthesizes concrete predicates to plug in for the missing guards such that all buffer accesses in the program are memory safe. Furthermore, guards synthesized by our technique are the simplest and weakest among guards that guarantee memory safety, relative to the inferred loop invariants. Our approach is fully automatic and does not require any hints from the user. We have implemented our algorithm in a prototype synthesis tool for C programs, and we show that the proposed approach is able to successfully synthesize guards that closely match hand-written programmer code in a set of real-world C programs.

## 1 Introduction

Memory safety errors are a perennial source of crashes and vulnerabilities in programs written in unsafe languages, and even expert programmers often write erroneous code that accesses out-of-bounds buffers or invalid memory. Over the past few decades, there has been much research on helping programmers write memory safe code. Broadly speaking, existing approaches fall into two categories:

**Dynamic instrumentation.** Many approaches, such as those employed in memory managed languages like Java and C#, add run-time checks to guarantee the safety of each memory access. While such approaches prevent memory corruption and associated security vulnerabilities, they do not prevent run-time failures and often add significant performance overhead.

**Static verification.** Much recent research has focused on statically guaranteeing memory safety of programs written in unsafe languages [1–5]. While these techniques can uncover all potential memory safety errors, the errors identified by the verifier may be hard to understand, debug, and fix.

In this paper, we propose a new approach based on *program synthesis* to the design of memory-safe low-level code. Concretely, suppose that a programmer wishes to write a region of code  $R$  implementing a given functionality, but  $R$  can access out-of-bounds memory under certain assumptions about program inputs or previously taken branches. In our approach, the programmer embeds  $R$  within the scope of an unknown *guard* predicate whose sole purpose is to ensure the memory safety of  $R$ . This is done using a syntax of the form:

```
if(??) {R} else { /* handle error */ }
```

where the unknown guard is indicated by `??`. Our approach uses a new *guard synthesis* algorithm to compute a predicate  $P$  over program variables such that, when `??` is replaced by  $P$ , all memory accesses within  $R$  are provably memory-safe.

Unlike dynamic approaches, our method does not require run-time instrumentation to track allocation sizes or pointer offsets, thereby avoiding the associated performance overhead. Instead, we statically infer a single guard that guarantees the safety of *all* memory accesses within a code block. Furthermore, our approach goes beyond static verification: It not only guarantees memory safety, but also helps the programmer write safe-by-construction code. The programmer is only asked to tell us *which* code snippets must be protected by a guard, rather than the tedious, low-level details of *how* to protect them.

Our synthesis algorithm is based on the principle of *logical abduction*. Abduction is the problem of finding missing hypotheses in a logical inference task. In more detail, suppose we have a premise  $P$  and a desired conclusion  $C$  for an inference ( $P$  and  $C$  will be typically generated as constraints from a program) such that  $P \not\models C$ . Given  $P$  and  $C$ , abduction infers a simplest and most general explanation  $E$  such that  $P \wedge E \models C$  and  $P \wedge E \not\models \text{false}$ .

Previous work has shown how to use abduction for program verification, by framing unknown invariants as missing hypotheses in a logical inference problem [5–7]. While adapting abduction to synthesis is a nontrivial technical challenge, the end result is an algorithm with several appealing properties:

**Optimality of Synthesis.** Our algorithm gives a guarantee of *optimal synthesis* — i.e., the synthesized guards are optimal according to a quantitative criterion among all guards that guarantee memory safety. Optimality has been argued to be an important criterion in program synthesis. For instance, Alur et al. [8] argue that “Ideally, [in synthesis] we would like to associate a cost with each [synthesized] expression, and consider the problem of optimal synthesis which requires the synthesis tool to return the expression with the least cost among the correct ones. A natural cost metric is the size of the expression.” However, few existing approaches to software synthesis take on such an optimality goal.

The notion of costs used in this paper is two-dimensional: one dimension quantifies expression complexity (we use the number of variables as a proxy for complexity), and the other quantifies generality (weaker guards have lower costs). The guards we synthesize are *Pareto-optimal* with respect to this notion of costs — i.e., there is no solution that is weaker as well as less complex.

**Automation.** Unlike most recent approaches to program synthesis [9–11], our algorithm can synthesize expressions without the aid of user-specified structural hints. In particular, the programmer does not need to provide expression templates with unknown coefficients to be inferred.

**Practicality.** Our algorithm incorporates precise reasoning about array bounds and low-level pointer arithmetic, which are necessary ingredients for synthesizing guards to guarantee memory safety. Furthermore, as shown in our experimental evaluation, the proposed synthesis algorithm can successfully synthesize guards required for memory safety in real C applications and produces guards that closely match hand-written code.

```

1. int main(int argc, char** argv) {
2.     char *command = NULL;
3.     if (argc <= 1) {
4.         error (0, 0, _("too few arguments"));
5.         usage (EXIT_FAIL);
6.     }
7.     argv++; argc--;
8.     while ((optc = getopt(argc, argv, ...)) != -1) {
9.         switch(optc) {
10.            case 'c':
11.                command = optarg; break;
12.            ...
13.        }
14.    }
15.    if (??) usage (EXIT_CANCELED);
16.    timeout = parse (argv[optind++]);
17.    files = argv + optind;
18.    if (!target_dir) {
19.        if (! (mkdir_and_install ? install_in_parents(files[0], files[1])
20.              : install_in_file(files[0], files[1])))
21.            ...
22.    }
23. }

```

Fig. 1. Motivating example

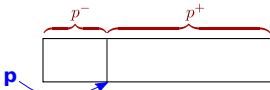
## 2 Motivating Example and Overview

We now present an overview of our approach using a motivating example. Consider the code snippet shown in Figure 1, which is based on the Unix `coreutils`. This program parses command line arguments with the help of a `clib` function called `getopt`. Specifically, lines 8-14 process the optional command line arguments while the code after line 16 performs the program's required functionality. Here, variable `optind` used at lines 16-17 is initialized by `getopt` to be the index of the next element to be processed in `argv`. Looking at lines 16-23, the programmer expects the user to pass some required arguments and accesses them at lines 16, 19, and 20. However, since the user may have forgotten to pass the required arguments, the programmer must explicitly check whether the memory accesses at lines 16, 19, 20 are safe in order to prevent potentially disastrous buffer overflow or underflow errors. If her assumptions are not met, the programmer wishes to terminate the program by calling the exit function called `usage`. However, coming up with the correct condition under which to terminate the program is tricky even on the small code snippet shown here: The programmer has performed pointer arithmetic on `argv` at line 7, and the variable `files` is an alias of `argv` at offset `optind` which has previously been modified at line 16.

Using our technique, the programmer can use the `??` predicate at line 15 to indicate the unknown check required for ensuring memory safety of the remainder of the program. Our technique then automatically synthesizes the guard (`argc`

- `optind`) > 2 as a sufficient condition for the safety of all buffer accesses in lines 16-23. Since the check inferred by our technique is correct-by-construction, the remainder of the program is guaranteed to be memory safe.

**Algorithm Overview.** Our algorithm proceeds in two phases, consisting of constraint generation and solving. During constraint generation, we represent the unknown guards using placeholder formulas  $\chi$  and then generate verification conditions over these unknown  $\chi$ 's. The constraint solving phase, which employs an iterative abduction-based algorithm, infers a concrete predicate for each  $\chi$  that makes all generated VCs valid. In addition to guaranteeing Pareto-optimality, this approach does not require the user to specify templates describing the shape of the unknown guards. Furthermore, since the abduced solutions imply the validity of the VCs, we do not need to externally validate the correctness of the synthesized program using a separate verifier or model checker.



**Fig. 2.** Auxiliary variables

The constraint generation phase consists of two key ingredients: First, to reason about out-of-bounds memory accesses, we introduce ghost variables that track allocation sizes and pointer offsets. Specifically, for each pointer  $p$ , a variable  $p^-$  indicates the offset of  $p$  in the block of memory it points to, and  $p^+$  tracks the size of  $p$  relative to  $p^-$ . This is shown in Figure 2. These ghost variables enable reasoning about pointer arithmetic in a precise way and allow us to generate symbolic verification conditions for memory safety.

The second key ingredient of constraint generation is a dual forwards and backwards static analysis that simultaneously computes strongest postconditions and weakest preconditions. For each unknown guard to be synthesized, the forwards analysis computes a formula  $\phi$  representing facts that are known at this program point, while the backwards analysis provides a weakest precondition  $\psi$  for the safety of the code protected by this unknown guard. Now, given a statement  $S$  involving an unknown guard and the two formulas  $\phi$  and  $\psi$ , our technique generates the VC  $(\phi \wedge \chi(\mathbf{v})) \rightarrow \psi$  where  $\chi$  is a predicate representing the unknown guard and  $\mathbf{v}$  represents all program variables in scope at this program point. Here, formulas  $\phi$  and  $\psi$  may also contain other unknowns.

In the constraint solving phase, we use an iterative, worklist-based algorithm that employs abduction to solve for the unknown  $\chi$  predicates. Given a set of constraints  $\mathcal{C}$  of the form  $(F_1(\chi_1, \dots, \chi_{i-1}) \wedge \chi_i) \rightarrow F_2(\chi_{i+1}, \dots, \chi_n)$  where  $F(\chi)$  denotes a formula over unknowns  $\chi$ , we show how to infer a solution for each  $\chi_i$  such that all constraints in  $\mathcal{C}$  become valid. Our algorithm guarantees the Pareto-optimality of the solution relative to the inferred loop invariants. That is, assuming a fixed set of loop invariants, if we pick any unknown guard and try to improve it according to our cost metric, then the resulting set of guards is no longer a solution to our synthesis problem.

**Example Redux.** We now go back to the code example from Figure 1 to illustrate our approach. Initially, we assume that `argv` points to the beginning of an allocated block of size `argc`; hence, our analysis starts with the fact:

$$\text{argv}^+ = \text{argc} \wedge \text{argv}^- = 0 \quad (1)$$

Next, we perform forward reasoning to compute the strongest postcondition of (1) right before line 20. Here, the forward analysis yields the condition:

$$\phi : \text{argc} > 0 \wedge \text{argv}^+ = \text{argc} \wedge \text{argv}^- = 1 \wedge \text{optind} \geq 0 \quad (2)$$

The first part of the conjunct ( $\text{argc} > 0$ ) comes from the condition at line 3: Since `usage` is an exit function, we know  $\text{argc} > 1$  at line 6, which implies  $\text{argc} > 0$  after line 7. The second part ( $\text{argv}^+ = \text{argc}$ ) states that the size of `argv` is still `argc`; this is because `argc` is decremented while `argv` is incremented at line 7. According to the third conjunct ( $\text{argv}^- = 1$ ), `argv` points to the second element in the original argument array due to the pointer increment at line 7. Finally, the last conjunct ( $\text{optind} \geq 0$ ) is a postcondition established by the call to `getopt`.

Next, we focus on the backwards analysis. To guarantee the safety of the buffer access `files[1]` at line 19, we need  $1 < \text{files}^+$  and  $1 \geq -\text{files}^-$  to ensure there are no buffer overflows and underflows respectively. Using similar reasoning for the accesses `files[0]` and `argv[optind++]`, our analysis generates the following necessary condition for the safety of the code after line 15:

$$\begin{aligned} & \text{optind} < \text{argv}^+ \wedge \text{optind} \geq -\text{argv}^- \wedge \\ \psi : & \text{target\_dir} = 0 \rightarrow (1 < \text{argv}^+ - \text{optind} - 1) \wedge \\ & \text{target\_dir} = 0 \rightarrow 0 \geq -\text{argv}^- - \text{optind} - 1 \end{aligned} \quad (3)$$

Observe that  $\text{files}^-$  and  $\text{files}^+$  do not appear in this formula because the backwards analysis relates the size and offset of `files` to those of `argv` when computing the weakest precondition of `files = argv + optind` at line 17. Now, to synthesize the unknown guard at line 15, we generate the following constraint:

$$(\phi \wedge \chi(\mathbf{v})) \rightarrow \psi \quad (4)$$

where  $\phi$  and  $\psi$  come from Equations 2 and 3,  $\chi$  is the unknown guard, and  $\mathbf{v}$  represents program variables in scope at line 15. Note that, since  $\text{argv}^-$ ,  $\text{argv}^+$  etc. are ghost variables, they are not allowed to appear in our solution for  $\chi$ .

Now, inferring a formula to plug in for  $\chi$  that makes Equation 4 valid is a logical abduction problem. By using abduction to solve for  $\chi$ , we obtain the solution  $\text{argc} - \text{optind} > 2$ . Observe that there are other guards that also guarantee memory safety in this example, such as:

- (S1)  $\text{argc} > \text{optind} \wedge (\text{target\_dir} = 0 \rightarrow \text{argc} - \text{optind} > 2)$ , or
- (S2)  $\text{argc} = 4 \wedge \text{optind} = 1$

However, both of these solutions are undesirable because (S1) is overly complicated, while (S2) is not sufficiently general.

### 3 Language and Preliminaries

We describe our techniques using the small imperative language given in Figure 3. Here, a program takes inputs  $\mathbf{v}$  and consists of one or more statements. We

Program  $P$      $:= \lambda v. S$   
 Guard  $G$         $:= ??_i \mid C$   
 Statement  $S$     $:= \text{skip} \mid v := E \mid S_1; S_2 \mid [p] := \text{alloc}(E) \mid [p_1] := [p_2] \oplus E$   
                    $\mid \text{access}([p], E) \mid \text{if}(G) \text{ then } S_1 \text{ else } S_2;$   
                    $\mid \text{while}(C) \text{ do } S \mid \text{while}(C \wedge ??_i) \text{ do } S$   
 Conditional  $C$   $:= E_1 \text{ comp } E_2$  (comp  $\in \{<, >, =\}$ )  $\mid C_1 \wedge C_2 \mid C_1 \vee C_2 \mid \neg C$   
 Expression  $E$   $:= \text{int} \mid v \mid E_1 + E_2 \mid E_1 - E_2 \mid E_1 \cdot E_2$

**Fig. 3.** Language used for the formal development

syntactically differentiate between scalar variables  $v$  and pointers  $[p]$ , which are always written inside brackets. Statements include skip, scalar assignments ( $v := E$ ), sequencing, memory allocations ( $[p] = \text{alloc}(E)$ ), and pointer arithmetic ( $[p_1] = [p_2] \oplus E$ ) which makes  $p_1$  point to offset  $E$  in the memory block pointed to by  $[p_2]$ . The statement  $\text{access}([p], E)$  accesses the  $E$ 'th offset of  $[p]$ . Since our main concern is to guarantee the safety of memory accesses, we use the access statement to model both array reads and writes. In particular,  $\text{access}([p], E)$  fails if  $E$  is not a valid offset in the memory region pointed to by  $E$ . We say that an access is *safe* if it can never fail in any execution; otherwise, it is *unsafe*. A program  $P$  is *memory-safe* if all accesses in  $P$  are safe.

In this language, unknown predicates  $??_i$  occur either as tests in if statements or as continuation conditions of while loops. We say a guard  $G_1$  is an *ancestor* of guard  $G_2$  if  $G_2$  is nested inside  $G_1$ ; conversely, we say  $G_2$  is a *descendant* of  $G_1$ . We call a program *complete* if it does not contain any unknown guards. Given a program  $P$  and a mapping  $\sigma$  from unknown guards to concrete predicates, we write  $P[\sigma]$  to denote the program obtained by substituting each  $??_i$  with  $\sigma(??_i)$ .

**Definition 1.** *Mapping  $\sigma$  is a solution to the guard synthesis problem defined by program  $P$  iff (i)  $P[\sigma]$  is a complete and memory-safe program, and (ii)  $\forall v \in \text{dom}(\sigma). \sigma(v) \not\equiv \text{false}$ .*

According to the second condition, a valid solution cannot instantiate any unknown predicate with false. Hence, the synthesis problem is unsolvable if we cannot guarantee memory safety without creating dead code.

**Definition 2.** *Given solutions  $\sigma$  and  $\sigma'$  to the synthesis problem, we say that  $\sigma$  refines  $\sigma'$ , written  $\sigma' \preceq \sigma$ , if for some unknown  $\chi \in \text{dom}(\sigma)$ , we have either (i)  $\sigma'(\chi) \Rightarrow \sigma(\chi)$  and  $\sigma(\chi) \not\equiv \sigma'(\chi)$ , or (ii)  $|\text{vars}(\sigma(\chi))| < |\text{vars}(\sigma'(\chi))|$ .*

In other words, solution  $\sigma$  refines  $\sigma'$  if it improves some guard either in terms of generality or simplicity.

**Definition 3.** *Solution  $\sigma$  is a Pareto-optimal solution to the synthesis problem if for all other solutions  $\sigma'$ , we have  $\sigma' \preceq \sigma$ .*

Intuitively, this means that, if we take solution  $\sigma$  and try to improve any guard in  $\sigma$  according to our cost metric, then the resulting mapping is no longer a solution. In the rest of the paper, we use the word “optimality” to mean Pareto-optimality in the above sense.

$$\begin{array}{l}
 (1) \quad \frac{}{\phi, \psi \vdash \text{skip} : \phi, \psi, \emptyset} \\
 (2) \quad \frac{\phi' = \exists v'. (v = E[v'/v] \wedge \phi[v'/v])}{\phi, \psi \vdash v := E : \phi', \psi[E/v], \emptyset} \\
 (3) \quad \frac{\phi, \psi_1 \vdash S_1 : \phi_1, \psi_2, \mathcal{C}_1 \quad \phi_1, \psi \vdash S_2 : \phi_2, \psi_1, \mathcal{C}_2}{\phi, \psi \vdash S_1; S_2 : \phi_2, \psi_2, \mathcal{C}_1 \cup \mathcal{C}_2} \\
 (4) \quad \frac{\phi, \psi \vdash p^- := 0; p^+ := E : \phi', \psi', \emptyset}{\phi, \psi \vdash [p] := \text{alloc}(E) : \phi', \psi', \emptyset} \\
 (5) \quad \frac{\phi, \psi \vdash (p_1^- := p_2^- + E) : \phi_1, \psi_1, \emptyset}{\phi_1, \psi_1 \vdash (p_1^+ := p_2^+ - E) : \phi_2, \psi_2, \emptyset} \\
 \frac{\phi, \psi \vdash [p_1] := [p_2] \oplus E : \phi_2, \psi_2, \emptyset}{} \\
 (6) \quad \frac{\varphi_{\text{safe}} = (E \geq -p^- \wedge E < p^+)}{\phi, \psi \vdash \text{access}([p], E) : \phi \wedge \varphi_{\text{safe}}, \psi \wedge \varphi_{\text{safe}}, \emptyset} \\
 (7a) \quad \frac{\phi \wedge C, \psi \vdash S_1 : \phi_1, \psi_1, \mathcal{C}_1 \quad \phi \wedge \neg C, \psi \vdash S_2 : \phi_2, \psi_2, \mathcal{C}_2}{\phi, \psi \vdash \text{if}(C) \text{ then } S_1 \text{ else } S_2 : \phi_1 \vee \phi_2, \psi', \mathcal{C}_1 \cup \mathcal{C}_2} \\
 \frac{\phi \wedge \chi_i(\mathbf{v}), \text{true} \vdash S_1 : \neg, \varphi, \mathcal{C}_1}{\text{VC} = (\phi \wedge \chi_i(\mathbf{v}) \rightarrow \varphi)} \\
 \frac{\phi \wedge \chi_i(\mathbf{v}), \psi \vdash \widetilde{S}_1 : \phi_1, \psi_1, \neg}{\phi \wedge \neg \chi_i(\mathbf{v}), \psi \vdash S_2 : \phi_2, \psi_2, \mathcal{C}_2} \\
 \frac{\psi' = (\chi_i(\mathbf{v}) \rightarrow \psi_1) \wedge (\neg \chi_i(\mathbf{v}) \rightarrow \psi_2)}{\phi, \psi \vdash \text{if}(\text{??}_i) \text{ then } S_1 \text{ else } S_2 : \phi_1 \vee \phi_2, \psi', \text{VC} \cup \mathcal{C}_1 \cup \mathcal{C}_2} \\
 (7b) \\
 (8a) \quad \frac{I \wedge C, I \vdash S : \neg, I', \mathcal{C} \quad I \wedge C \Rightarrow I'}{\phi, \psi \vdash \text{while}(C) \text{ do } S : I \wedge \neg C, I, \mathcal{C}} \\
 (8b) \quad \frac{\neg, I \vdash \widetilde{S} : \neg, I', \neg \quad I \wedge C \Rightarrow I'}{I \wedge C \wedge \chi_i(\mathbf{v}), \text{true} \vdash S : \neg, \psi, \mathcal{C}} \\
 \frac{\text{VC} = (I \wedge C \wedge \chi_i(\mathbf{v}) \rightarrow \psi)}{\phi, \psi \vdash \text{while}(C \wedge \text{??}_i) \text{ do } S : I \wedge \neg(C \wedge \chi_i(\mathbf{v})), I, \mathcal{C} \cup \text{VC}} \\
 (9) \quad \frac{\text{true}, \text{true} \vdash S : \phi, \psi, \mathcal{C}}{\vdash \lambda \mathbf{v}. S : \phi, \psi, \psi \cup \mathcal{C}}
 \end{array}$$

Fig. 4. Inference Rules for Constraint Generation

## 4 Constraint Generation

The constraint generation phase is shown in Figure 4 as inference rules of the form  $\phi, \psi \vdash S : \phi', \psi', \mathcal{C}$  where  $S$  is a statement,  $\phi, \psi, \phi', \psi'$  are formulas, and  $\mathcal{C}$  is a set of constraints. The meaning of this judgment is that, if all constraints in  $\mathcal{C}$  are valid, then  $\{\phi\}S\{\phi'\}$  and  $\{\psi'\}S\{\psi\}$  are valid Hoare triples. We call the

computation of postcondition  $\phi'$  from  $\phi$  the *forward analysis* and the computation of precondition  $\psi'$  from  $\psi$  the *backward analysis*. The constraints  $\mathcal{C}$  track assumptions about unknown predicates that must hold to ensure memory safety.

Since some of the rules in Figure 4 describe standard pre- and post-condition computation, we only explain some of these rules. Rule (4) for memory allocation  $[p] = \text{alloc}(E)$  uses ghost variables  $p^-$  and  $p^+$ . Since  $[p]$  points to the beginning of a memory block of size  $E$ , the allocation has the effect of assigning  $p^-$  to 0 and  $p^+$  to  $E$ . Hence,  $\phi'$  and  $\psi'$  are obtained by computing the strongest postcondition of  $\phi$  and weakest precondition of  $\psi$  with respect to the statement  $p^- := 0; p^+ := E$ .

Rule (5) for pointer arithmetic computes the effect of this statement on  $p_1^-$  and  $p_1^+$ . Since  $[p_2]$  points to offset  $p_2^-$  in memory block  $M$ ,  $[p_1]$  points to offset  $p_2^- + E$  within  $M$ . Hence, we obtain  $\phi_1$  as  $\text{sp}(p_1^- := p_2^- + E, \phi)$  and  $\psi_1$  as  $\text{wp}(p_1^- := p_2^- + E, \psi)$ . Similarly,  $\phi_2 = \text{sp}(p_1^+ := p_2^+ - E, \phi_1)$  and  $\psi_2 = \text{wp}(p_1^+ := p_2^+ - E, \psi_1)$ .

Rule (6) describes memory accesses. To guarantee that  $\psi$  holds after the memory access, expression  $E$  must evaluate to a valid offset in the memory block pointed to by  $[p]$ . Using ghost variables  $p^-$  and  $p^+$ , we can express this as  $\varphi_{\text{safe}} \equiv E < p^+ \wedge E \geq -p^-$ . Hence, the weakest precondition of  $\psi$  with respect to the access statement is  $\psi \wedge \varphi_{\text{safe}}$ .

Constraint generation for conditionals with unknown guards is given in Rule (7b). The first line of this rule computes a weakest sufficient condition for ensuring memory safety of statement  $S_1$ . Here, we compute the precondition of  $S_1$  with respect to true rather than  $\psi$  because the unknown guard is only required to guarantee the safety of  $S_1$  rather than the remainder of the entire program. Thus, the formula  $\varphi$  computed here represents the weakest sufficient condition for ensuring memory safety of  $S_1$ . When we analyze  $S_1$ , observe that the forward analysis propagates  $\phi \wedge \chi_i(\mathbf{v})$  as the precondition of  $S_1$ ; hence, statement preconditions computed by the forward analysis may refer to unknown predicates  $\chi_i$ . The constraints  $\mathcal{C}_1$  obtained in this rule describe the restrictions that must be satisfied by the unknown guards nested inside  $S_1$ .

The second line in rule (7b) generates a constraint (VC) on the unknown predicate  $\chi_i$ . Specifically, VC stipulates that the conjunction of the unknown guard  $\chi_i$  and the precondition  $\phi$  should be strong enough to imply the safety of memory accesses within  $S_1$ . Note that the generated VC may contain multiple unknown predicates since both  $\phi$  and  $\varphi$  may refer to other  $\chi_j$ 's.

The third line in rule (7b) uses the notation  $\tilde{S}$ , which denotes statement  $S$  with each  $\text{access}([p], E)$  statement within  $S$  replaced by  $\text{skip}$ . Here, we analyze statement  $S_1$  a second time but with two important differences from our previous analysis. First, since we consider  $\tilde{S}_1$ , we ignore any memory accesses within  $S_1$ . Second, we compute the weakest precondition of  $\tilde{S}_1$  with respect to  $\psi$  rather than true because we need to ensure the safety of memory accesses that come after  $S_1$ . However, we ignore all memory accesses within  $S_1$  because the synthesized guard already ensures the safety of these accesses. Also, observe that this rule discards constraints generated when analyzing  $\tilde{S}_1$ , which is sound because any constraints generated while analyzing  $\tilde{S}_1$  are trivially valid.



Another important point to note about rule (7b) is that the backwards analysis propagates the constraint  $(\chi_i(\mathbf{v}) \rightarrow \psi_1) \wedge (\neg\chi_i(\mathbf{v}) \rightarrow \psi_2)$  as the weakest precondition for the if statement. Hence, statement postconditions computed by the backwards analysis may also refer to unknown predicates.

*Example 1.* Consider the following code example:

1.  $[p] := \text{alloc}(n); [q] := \text{alloc}(n); [p] := [p] \oplus 1;$
2.  $\text{if}(\text{??}_1)$  then
3.      $n := n + 1; \text{access}([p], 3);$
4.      $\text{if}(\text{??}_2)$  then  $\text{access}([q], n - 2)$  else skip
5. else skip

In the forward direction, our analysis computes the following precondition  $\phi$  for the if statement at line 2:  $\phi : p^- = 1 \wedge p^+ = n - 1 \wedge q^- = 0 \wedge q^+ = n$ . Before the assignment at line 3, the forwards analysis computes the precondition  $\chi_1(n) \wedge p^- = 1 \wedge p^+ = n - 1 \wedge q^- = 0 \wedge q^+ = n$  where  $\chi_1$  denotes unknown guard  $\text{??}_1$  and  $n$  is the only scalar variable in scope at this point. For the if statement at line 4, we have the following precondition:

$$\begin{aligned} \exists n'. (\chi_1(n') \wedge n = n' + 1 \wedge p^- = 1 \wedge p^+ = n' - 1 \\ \wedge q^- = 0 \wedge q^+ = n' \wedge p^+ > 3 \wedge p^- \geq -3) \end{aligned}$$

Note that, since there is an assignment to  $n$  at line 3, the variable  $n$  inside the unknown predicate  $\chi_1(n)$  gets substituted by  $n'$ .

Now, in the backwards direction, the precondition for the then branch of the second if statement is  $(n - 2) < q^+ \wedge (n - 2) \geq -q^-$ . Hence, when analyzing the if statement at line 4, we generate the following VC:

$$\text{VC}_2 : (\chi_2(n) \wedge \exists n'. (\chi_1(n') \wedge n = n' + 1 \wedge p^- = 1 \wedge p^+ = n' - 1 \wedge q^- = 0 \\ \wedge q^+ = n' \wedge p^+ > 3 \wedge p^- \geq -3)) \rightarrow ((n - 2) < q^+ \wedge (n - 2) \geq -q^-)$$

where  $\chi_2$  represents  $\text{??}_2$  and the right-hand-side of the implication is the safety precondition for the then branch. For the if statement at line 2, the backwards analysis computes the precondition for the then branch as  $\varphi : 3 < p^+ \wedge 3 \geq -p^-$ . Using  $\varphi$  and formula  $\phi$  obtained through the forward analysis, we generate the following VC for the if statement at line 2:

$$\text{VC}_1 : (\chi_1(n) \wedge p^- = 1 \wedge p^+ = n - 1 \wedge q^- = 0 \wedge q^+ = n) \rightarrow (3 < p^+ \wedge 3 \geq -p^-)$$

Hence, our algorithm generates the constraints  $\text{VC}_1 \cup \text{VC}_2$ .

Continuing with the inference rules in Figure 4, rules (8b) and (8a) describe the analysis of while loops with and without unknown safety guards respectively. Rule (8a) gives the standard Hoare rule for while loops, asserting that  $I$  is an inductive loop invariant. Since the automatic inference of loop invariants is an orthogonal problem, this paper assumes that loop invariants are provided by an oracle, and our implementation uses standard abstract interpretation based techniques for loop invariant generation.

In rule (8b), our goal is to infer an additional guard as part of the loop continuation condition such that all memory accesses within the loop body are

safe. As in rule (8a), the first line of rule (8b) asserts that  $I$  is inductive. However, an important difference is that we check the inductiveness of  $I$  with respect to  $\tilde{S}$  rather than  $S$  because  $I$  is not required to be strong enough to prove the safety of memory accesses inside the loop body. In fact, if  $I$  was strong enough to prove memory safety, this would mean the additional unknown guard is unnecessary.

The last two lines of rule (8b) compute the safety precondition  $\psi$  for the loop body (i.e.,  $\psi = \text{wp}(\text{true}, S)$ ) and generate the constraint  $\text{VC} : I \wedge C \wedge \chi_i(\mathbf{v}) \rightarrow \psi$ . In other words, together with continuation condition  $C$  and known loop invariant  $I$ , unknown predicate  $\chi_i$  should imply the memory safety of the loop body.

*Example 2.* Consider the following code snippet:

1.  $[p] = \text{alloc}(n); i = 0;$
2.  $\text{while}(\text{true} \wedge ??_1)$  do
3.      $\text{access}([p], 1); [p] = [p] \oplus 1; i = i + 1;$

Assume we have the loop invariant  $I : p^- + p^+ = n \wedge i \geq 0 \wedge i = p^-$ . The safety precondition  $\psi$  for the loop body is  $1 < p^+ \wedge 1 \geq -p^-$ . Hence, rule (8b) from Figure 4 generates the following verification condition:

$$(\chi_1(i, n) \wedge p^- + p^+ = n \wedge i \geq 0 \wedge i = p^-) \rightarrow (1 < p^+ \wedge 1 \geq -p^-)$$

The last rule in Figure 4 generates constraints for the entire program. Since we add the program's weakest precondition to  $\mathcal{C}$ , *the synthesis problem has a solution only if all accesses that are not protected by unknown guards are safe.*

## 5 Constraint Solving

The rules described in Section 4 generate constraints of the form:

$$\mathcal{C}_i : (F_1(\chi_1, \dots, \chi_{i-1}) \wedge \chi_i(\mathbf{v})) \rightarrow F_2(\chi_{i+1}, \dots, \chi_n) \quad (5)$$

where  $F_1$  and  $F_2$  are arbitrary formulas containing program variables and unknowns. In each constraint, there is exactly one key unknown  $\chi_i$  that does not appear inside boolean connectives or quantifiers. Hence, we refer to  $\mathcal{C}_i$  as the constraint associated with  $\chi_i$  (or the  $\chi_i$ -constraint). Also, the only unknowns appearing on the right hand side of an implication (i.e., inside  $F_2$ ) in a  $\chi_i$ -constraint represent unknown guards that are syntactically nested inside  $\chi_i$ . Hence, we refer to  $\chi_{i+1}, \dots, \chi_n$  as the *descendants* of  $\mathcal{C}_i$ , denoted  $\text{DESCENDS}(\mathcal{C}_i)$ . In contrast, the unknowns that appear inside  $F_1$  are either ancestors of  $\chi_i$  or appear in the code before  $\chi_i$ . We say that  $\mathcal{C}_i$  *sequentially depends* on  $\chi_j$  if  $\chi_j$  appears inside  $F_1$  and is not an ancestor of  $\chi_i$ . We write  $\text{SEQDEP}(\mathcal{C}_i)$  to denote the set of  $\chi_j$ -constraints such that  $\mathcal{C}_i$  is sequentially dependent on  $\chi_j$ .

*Example 3.* Consider the following code snippet:

1.  $[a] := \text{alloc}(x);$
2.  $\text{if}(??_1)$  then  $\text{access}([a], 1)$  else skip
3.  $\text{if}(??_2)$  then
4.      $\text{if}(??_3)$  then  $\text{access}([a], x - 3); [b] := \text{alloc}(4);$  else  $[b] := \text{alloc}(2);$
5.      $\text{access}([b], 3);$
6. else skip

Let  $\chi_1, \chi_2, \chi_3$  denote the unknowns  $??_1, ??_2$ , and  $??_3$ , and let  $\mathcal{C}_i$  denote each  $\chi_i$  constraint. Here, we have:

$$\begin{aligned} \mathcal{C}_1 &: (a^- = 0 \wedge a^+ = x \wedge \chi_1(x)) \rightarrow (1 < a^+ \wedge 1 \geq -a^-) \\ \mathcal{C}_2 &: (a^- = 0 \wedge a^+ = x \wedge \chi_1(x) \wedge \chi_2(x)) \rightarrow ((\chi_3(x) \rightarrow 3 < 4) \wedge (\neg\chi_3(x) \rightarrow 3 < 2)) \\ \mathcal{C}_3 &: (a^- = 0 \wedge a^+ = x \wedge \chi_1(x) \wedge \chi_2(x) \wedge \chi_3(x)) \rightarrow (x - 3 < a^+ \wedge x - 3 \geq -a^-) \end{aligned}$$

Therefore,  $\text{DESCENDS}(\mathcal{C}_1) = \emptyset$ ,  $\text{DESCENDS}(\mathcal{C}_2) = \{\mathcal{C}_3\}$ , and  $\text{DESCENDS}(\mathcal{C}_3) = \emptyset$ . Also,  $\text{SEQDEPS}(\mathcal{C}_1) = \emptyset$ ,  $\text{SEQDEPS}(\mathcal{C}_2) = \{\mathcal{C}_1\}$ ,  $\text{SEQDEPS}(\mathcal{C}_3) = \{\mathcal{C}_1\}$ .

Our constraint solving algorithm is given in Figure 5. A key underlying insight is that we only solve a constraint  $\mathcal{C}_i$  when all sequential dependencies of  $\mathcal{C}_i$  are resolved. The intuition is that if  $\chi_i$  sequentially depends on  $\chi_j$ ,  $\chi_j$  will appear in a  $\chi_i$ -constraint, but not the other way around. Hence, by fixing the solution for  $\chi_j$  before processing  $\chi_i$ , we cannot break the optimality of the solution for  $\chi_j$ .

The SOLVE algorithm shown in Figure 5 takes as input constraints  $\mathcal{C}$  and returns a mapping  $S$  from each unknown  $\chi$  to a concrete predicate or  $\emptyset$  if no solution exists. Initially, we add all constraints  $\mathcal{C}$  to worklist  $W$  and initialize RESOLVED to  $\emptyset$ . In each iteration of the SOLVE loop, we dequeue constraints  $\Delta$  that have their sequential dependencies resolved (line (3)) and substitute any resolved unknowns in  $\Delta$  with the solution given by  $S$ , yielding a new set  $\Delta'$  (line 4). Hence, a  $\chi_i$ -constraint in  $\Delta'$  does not contain any unknowns that  $\chi_i$  sequentially depends on. Now, to solve  $\Delta'$ , we first obtain a sound, but not necessarily optimal, solution using the function SOLVEINIT. In particular, although the solutions returned by SOLVEINIT may be stronger than necessary, we iteratively weaken this initial solution using WEAKEN until we obtain an optimal solution.

The procedure SOLVEINIT processes constraints  $\mathcal{C}$  top-down, starting with the outermost guard  $\chi_i$ . In each iteration, we pick an unsolved constraint  $\mathcal{C}_i$  that has only one unknown on the left-hand side of the implication (line 13). However, since we don't yet have a solution for the unknowns  $\mathcal{V}$  on the right-hand side, we strengthen  $\mathcal{C}_i$  to  $\Phi$  by universally quantifying  $\mathcal{V}$  (line 16).<sup>1</sup> Observe that the universal quantification of  $\mathcal{V}$  has the same effect as treating any unknown guard inside  $\chi_i$  as a non-deterministic choice. The resulting constraint  $\Phi$  is of the form  $(\chi_i(\mathbf{v}) \wedge \phi_1) \rightarrow \phi_2$  where  $\phi_1$  and  $\phi_2$  do not contain any unknowns; hence, we can solve for unknown  $\chi_i$  using standard logical abduction. In the algorithm of Figure 5, this is done by calling an abduction procedure called ABDUCE (line 17). We do not describe the ABDUCE procedure in this paper and refer the interested reader to [12] for a description of an abduction algorithm which computes a logically weakest solution containing a fewest number of variables. Now, given solution  $\gamma$  for  $\chi_i$ , we add it to our solution set  $S$  and eliminate unknown  $\chi_i$  from other constraints in  $\mathcal{C}$  using the SUBSTITUTE procedure.

Because of the universal quantification of the unknowns on the right-hand side, the solution  $S_0$  returned by SOLVEINIT may be stronger than necessary. Hence, procedure WEAKEN iteratively weakens the initial solution until we obtain an optimal solution. In contrast to SOLVEINIT, WEAKEN processes the constraints

<sup>1</sup> Recall that  $\forall \chi_j. \Phi \equiv \Phi[\text{true}/\chi_j] \wedge \Phi[\text{false}/\chi_j]$ .

```

procedure SOLVE( $\mathcal{C}$ ):
  input: set of constraints  $\mathcal{C}$ 
  output: mapping  $S$  from each  $\chi_i$  to  $\gamma$  or  $\emptyset$  if no solution exists

(1)  RESOLVED :=  $\emptyset$ ;  $S := \emptyset$ ;  $W := \mathcal{C}$ 
(2)  while( $W \neq \emptyset$ )
(3)     $\Delta := \{\mathcal{C}_i \mid \mathcal{C}_i \in W \wedge \text{SEQDEP}(\mathcal{C}_i) \subseteq \text{RESOLVED}\}$ 
(4)     $W := W - G$ 
(5)     $\Delta' := \text{SUBSTITUTE}(\Delta, S)$ 
(6)     $S_0 := \text{SOLVEINIT}(\Delta')$ 
(7)     $S := \text{WEAKEN}(\Delta', S_0, S, \text{RESOLVED})$ 
(8)    if( $S = \emptyset$ ) return  $\emptyset$ 
(9)    RESOLVED := RESOLVED  $\uplus$   $\Delta$ 
(10) return  $S$ ;

procedure SOLVEINIT( $\mathcal{C}$ ):
(11)  $S := \emptyset$ 
(12) while( $\mathcal{C} \neq \emptyset$ )
(13)   let  $\mathcal{C}_i \in \mathcal{C}$  with one unknown  $\chi$  on LHS
(14)    $\mathcal{C} := \mathcal{C} - \mathcal{C}_i$ 
(15)    $\mathcal{V} := \text{unknowns of } \mathcal{C}_i \text{ on RHS}$ 
(16)    $\Phi := \forall \mathcal{V}. \mathcal{C}_i$ 
(17)    $\gamma := \text{ABDUCE}(\Phi)$ 
(18)    $S := S \uplus [\chi \mapsto \gamma]$ 
(19)    $\mathcal{C} := \text{SUBSTITUTE}(\mathcal{C}, S)$ 
(20) return  $S$ 

procedure WEAKEN( $\Delta, S, S_0, \text{RESOLVED}$ )
(21) DONE :=  $\{\Delta_i \mid \Delta_i \in \Delta \wedge \text{DESCENDS}(\Delta_i) = \emptyset\}$ 
(22) RESOLVED := RESOLVED  $\uplus$  DONE
(23)  $S := S \uplus \{[\chi_i \mapsto \gamma_i] \mid \mathcal{C}_i \in \text{DONE} \wedge S_0(\chi_i) = \gamma_i\}$ 
(24)  $\Delta := \Delta - \text{DONE}$ ;  $S_0 := S_0 - \text{DONE}$ 
(25) while( $\Delta \neq \emptyset$ )
(26)   let CUR  $\in \Delta$  s/t DESCENDS(CUR)  $\subseteq$  RESOLVED
(27)   let  $\chi = \text{UNKNOWN}(\text{CUR})$ 
(28)    $\Delta := \Delta - \text{CUR}$ ;  $S_0 := S_0 - \chi$ 
(29)    $\theta := \text{CUR} \uplus \{\mathcal{C}_i \mid \mathcal{C}_i \in \text{RESOLVED} \wedge \chi \in \text{UNKNOWN}(\mathcal{C}_i)\}$ 
(30)    $\theta' := \text{SUBSTITUTE}(S \uplus S_0)$ 
(31)    $\gamma := \text{ABDUCE}(\theta')$ 
(32)   if UNSAT( $\gamma$ ) return  $\emptyset$ 
(33)    $S := S \uplus [\chi \mapsto \gamma]$ 
(34)   RESOLVED := RESOLVED  $\uplus$  CUR
(35) return  $S$ 

```

**Fig. 5.** The constraint solving algorithm

bottom-up, starting with the innermost guard first. Specifically, the solution computed by SOLVEINIT for the innermost guard  $\chi_i$  cannot be further weakened, as it is not possible to obtain a weaker solution for  $\chi_i$  by plugging in weaker solutions for the unknowns appearing on the left-hand side of a  $\chi_i$ -constraint.

Hence, we add all constraints with no descendants in  $\Delta$  to RESOLVED and update  $S$  with the corresponding solutions given by  $S_0$  (lines 21-23).

The while loop in lines 25-34 iterates over the constraints bottom-up and, in each iteration, picks a  $\chi$ -constraint CUR all of whose descendants have been resolved (line 26). Now, the solution given by  $S_0$  for  $\chi$  could be stronger than necessary; thus, we would like to weaken it using the new optimal solutions for  $\chi$ 's descendants. However, since  $\chi$  will appear on the left-hand side of a constraint  $\mathcal{C}_i$  associated with a descendant  $\chi_i$  of  $\chi$ , we need to take care not to invalidate the solution for  $\chi_i$  as we weaken  $\chi$ . Hence, at line 29, we collect in set  $\theta$  all resolved constraints in which  $\chi$  appears. The idea here is that, when we abduce a new solution for  $\chi$ , we will simultaneously solve all constraints in  $\theta$  so that we preserve existing solutions for  $\chi$ 's descendants. Now, to solve for  $\chi$  using abduction, we first eliminate all unknowns in  $\theta$  except for  $\chi$ . For this purpose, we substitute all resolved unknowns in  $\theta$  with their solution given by  $S$  (line 30). However, observe that there may also be unknowns in  $\theta$  that have not yet been resolved; these unknowns correspond to ancestors of  $\chi$ , which only appear (unnegated) on the outerlevel conjunction of the left-hand side of the constraints in  $\theta$ . Hence, to eliminate the unresolved unknowns, we will use the initial solution given by  $S_0$  (also line 30). Observe that we can do this without undermining optimality because we will later only further weaken the solutions given by  $S_0$ , which cannot cause us to further weaken the solution for  $\chi$ .

After performing the substitution at line 30, the set of constraints  $\theta'$  only contains one unknown  $\chi$ , which we can now solve using standard abduction (line 31).<sup>2</sup> If the resulting solution  $\gamma$  is false, this means our synthesis problem does not have a solution. Otherwise, we add the mapping  $\chi \mapsto \gamma$  to our solution set and continue until all constraints in  $\theta$  have been processed.

Let us call a solution  $S$  to the synthesis problem *optimal relative to a set of loop invariants* if, among the set of solutions that any algorithm can generate using these loop invariants,  $S$  is optimal. We have:

**Theorem 1.** *Consider program  $P$  such that  $\vdash P : \phi, \psi, \mathcal{C}$  according to Figure 4, and let  $S$  be the result of  $\text{SOLVE}(\mathcal{C})$ . If  $S \neq \emptyset$ , then  $P[S]$  is memory safe. Furthermore,  $S$  is an optimal solution to the synthesis problem defined by  $P$  relative to the loop invariants used during constraint generation.*

*Example 4.* Consider the constraints VC<sub>1</sub> and VC<sub>2</sub> from Example 1. Here, neither VC<sub>1</sub> nor VC<sub>2</sub> have sequential dependencies, and since VC<sub>1</sub> contains only one unknown, we first solve for  $\chi_1$  in SOLVEINIT, for which abduction yields the solution  $n > 4$ . Next, we plug this solution into VC<sub>2</sub> (renaming  $n$  to  $n'$ ), which yields the following constraint in Prenex Normal Form:

$$\forall n'. (n' > 4 \wedge \chi_2(n) \wedge n = n' + 1 \wedge p^- = 1 \wedge p^+ = n' - 1 \wedge q^- = 0 \wedge q^+ = n' \wedge p^+ > 3 \wedge p^- \geq -3) \rightarrow (n - 2) < q^+ \wedge (n - 2) \geq -q^-)$$

<sup>2</sup> Observe that we can simultaneously solve all constraints in  $\theta$  using abduction because a pair of constraints of the form  $\chi \wedge \phi_1 \Rightarrow \phi_2$  and  $\chi \wedge \psi_1 \Rightarrow \psi_2$  (where the  $\phi$ 's and  $\psi$ 's are unknown free) can be rewritten as  $(\chi \wedge \phi_1 \wedge \psi_1) \Rightarrow (\phi_2 \wedge \psi_2)$ , which corresponds to a standard abduction problem.

Program	Lines	# holes	Time (s)	Memory	Synthesis successful?	Bug?
Coreutils hostname	160	1	0.15	10 MB	Yes	No
Coreutils tee	223	1	0.84	10 MB	Yes	Yes
Coreutils runcon	265	2	0.81	12 MB	Yes	No
Coreutils chroot	279	2	0.53	23 MB	Yes	No
Coreutils remove	710	2	1.38	66MB	Yes	No
Coreutils nl	758	3	2.07	80 MB	Yes	No
SSH - sshconnect	810	3	1.43	81 MB	Yes	No
Coreutils mv	929	4	2.03	42 MB	Yes	No
SSH - do_authentication	1,904	4	3.92	86 MB	Yes	Yes
SSH - ssh_session	2,260	5	4.35	81 MB	Yes	No

**Fig. 6.** Experimental benchmarks and results

Since  $VC_2$  has only one unknown left, we solve it using ABDUCE and obtain  $\chi_2(n) = \text{true}$ . Next, we attempt to weaken the solution for  $\chi_1$  in WEAKEN, but since  $\chi_2$  does not appear on the right hand side of  $VC_1$ , we cannot further weaken the solution for  $\chi_1$ . Hence, we obtain the solution  $[\chi_1 \mapsto n > 4, \chi_2 \mapsto \text{true}]$ .

*Example 5.* Consider constraints  $C_1, C_2$ , and  $C_3$  from Example 3. Since  $C_1$  does not have sequential dependencies, we first solve  $C_1$  and obtain the solution  $\chi_1 = (x > 1)$ . In the next iteration, both  $C_2$  and  $C_3$  have their sequential dependencies resolved; hence we plug in  $x > 1$  for  $\chi_1$  in  $C_2$  and  $C_3$ . In SOLVEINIT, we first solve  $C_2$  since it now contains only one unknown ( $\chi_2$ ) on the left hand side. When we universally quantify  $\chi_3$  on the right hand side, ABDUCE yields the solution  $\chi_2 = \text{false}$ . In the next iteration of SOLVEINIT, we obtain the solution *true* for  $\chi_3$ . Observe that our initial solution for  $\chi_2$  is stronger than necessary; hence we will weaken it. In the procedure WEAKEN, we simultaneously solve constraints  $C_2$  and  $C_3$ , using existing solutions for  $\chi_1$  and  $\chi_3$ . ABDUCE now yields  $\chi_2 = x > 2$ . Hence, the final solution is  $[\chi_1 = x > 1, \chi_2 = x > 2, \chi_3 = \text{true}]$ .

*Example 6.* For the constraint generated in Example 2, the SOLVE procedure computes the solution  $\chi_1(i, n) = i < n - 1$ .

## 6 Implementation and Evaluation

We implemented a prototype tool for synthesizing safety guards for C programs. Our tool is based on the SAIL infrastructure [13] and uses the Mistral SMT solver [12] for solving abduction problems in linear integer arithmetic.

We evaluated our tool on ten benchmark programs written in C. As shown in Figure 6, all of our benchmarks are taken either from the Unix coreutils, which implements basic command line utilities for Unix [14], or OpenSSH, which provides encrypted communication sessions based on the SSH protocol [15]. For each benchmark program, we manually removed 1-5 safety guards from the source code and then used our algorithm to infer these missing guards.<sup>3</sup> In total, we used our tool to synthesize 27 different safety guards.

<sup>3</sup> The URL <http://www.cs.utexas.edu/~tdillig/cav14-benchmarks.tar.gz> contains all benchmarks, where each missing guard is indicated with `_SYN`.

The results of our experimental evaluation are shown in Figure 6. Our tool was able to successfully synthesize all of the missing guards present in these benchmarks. For 23 of these 27 missing guards, our tool inferred the *exact same predicate* that the programmer had originally written, and for 4 out of the 27 missing guards, it inferred a syntactically different but semantically equivalent condition (e.g., our tool synthesized the guard  $x \neq 0$  when the programmer had originally written  $x > 0$  but  $x$  is already known to be non-negative). In two applications (Coreutils tee and SSH do\_authentication), the guards synthesized by our tool did not match the guards in the original program. However, upon further inspection, we found that both of these programs were in fact buggy. For example, in Coreutils tee, the program could indeed access the `argv` array out-of-bounds. We believe that the existence of such bugs in even extremely well-tested applications is evidence that writing memory safe code is hard and that many programmers can benefit from our guard synthesis technique.

As shown in Figure 6, the running time of our algorithm ranges from 0.15 seconds to 4.35 seconds with an average memory consumption of 49 MB. We believe these results suggest that our approach can be integrated into the development process, helping programmers write safe-by-construction low level code.

## 7 Related Work

**Program Synthesis.** The last few years have seen a flurry of activity in constraint-based software synthesis [9, 10, 16, 17]. As the first abduction-based approach to synthesis, our work is algorithmically very different from prior methods in this area. A concrete benefit is that, unlike prior constraint-based approaches to synthesis [10, 11, 18, 19], our method does not require a template for the expressions being synthesized. A second benefit is that we can show the synthesized expressions to be optimal relative to loop invariants.

There are a few approaches to synthesis that consider optimality as an objective [20–23]. However, in these papers, optimality is defined with respect to an explicit quantitative aspect of program executions, for example execution time. In contrast, in the current work, the cost metric is on the guards that we synthesize; we want to infer guards that are as simple and as general as possible.

**Program Analysis for Memory Safety.** Memory safety is a core concern in low-level programming, and there is a huge literature on program analysis techniques to guarantee memory safety [3, 24–33]. While many of these techniques can statically detect memory safety errors, they do not help the programmer write safe-by-construction code. Furthermore, unlike dynamic approaches to memory safety [27, 28, 32, 33], our technique guarantees the absence of runtime failures and does not require additional runtime book-keeping.

**Abduction-based Verification.** Many memory safety verifiers based on separation logic use abductive (or bi-abductive) reasoning for performing modular heap reasoning [5, 34, 35]. In these approaches, abductive reasoning is used to infer missing preconditions of procedures. A more algorithmic form of abduction for first-order theories is considered in [12]. The abduction algorithm described

in [12] computes a maximally simple and general solution and is used as a key building block in the constraint solving phase of our synthesis algorithm. This form of SMT-based abduction has also been used for loop invariant generation [6, 7] and for error explanation and diagnosis [36]. The contribution of the present paper is to show how abduction can be used in program synthesis.

## References

1. Berdine, J., Cook, B., Ishtiaq, S.: sLAYER: Memory safety for systems-level code. In: Gopalakrishnan, G., Qadeer, S. (eds.) CAV 2011. LNCS, vol. 6806, pp. 178–183. Springer, Heidelberg (2011)
2. Beyer, D., Henzinger, T.A., Jhala, R., Majumdar, R.: Checking memory safety with blast. In: Cerioli, M. (ed.) FASE 2005. LNCS, vol. 3442, pp. 2–18. Springer, Heidelberg (2005)
3. Chatterjee, S., Lahiri, S.K., Qadeer, S., Rakamarić, Z.: A reachability predicate for analyzing low-level software. In: Grumberg, O., Huth, M. (eds.) TACAS 2007. LNCS, vol. 4424, pp. 19–33. Springer, Heidelberg (2007)
4. Dillig, I., Dillig, T., Aiken, A.: Fluid updates: Beyond strong vs. Weak updates. In: Gordon, A.D. (ed.) ESOP 2010. LNCS, vol. 6012, pp. 246–266. Springer, Heidelberg (2010)
5. Calcagno, C., Distefano, D., O’Hearn, P., Yang, H.: Compositional shape analysis by means of bi-abduction. In: POPL, pp. 289–300 (2009)
6. Dillig, I., Dillig, T., Li, B., McMillan, K.: Inductive invariant generation via abductive inference. In: OOPSLA, pp. 443–456. ACM (2013)
7. Li, B., Dillig, I., Dillig, T., McMillan, K., Sagiv, M.: Synthesis of circular compositional program proofs via abduction. In: Piterman, N., Smolka, S.A. (eds.) TACAS 2013. LNCS, vol. 7795, pp. 370–384. Springer, Heidelberg (2013)
8. Alur, R., Bodik, R., Juniwal, G., Martin, M.M., Raghothaman, M., Seshia, S., Singh, R., Solar-Lezama, A., Torlak, E., Udupa, A.: Syntax-guided synthesis. In: FMCAD (2013)
9. Solar-Lezama, A.: The sketching approach to program synthesis. In: Hu, Z. (ed.) APLAS 2009. LNCS, vol. 5904, pp. 4–13. Springer, Heidelberg (2009)
10. Srivastava, S., Gulwani, S., Foster, J.S.: Template-based program verification and program synthesis. STTT 15(5-6), 497–518 (2013)
11. Srivastava, S., Gulwani, S., Chaudhuri, S., Foster, J.: Path-based inductive synthesis for program inversion. In: PLDI, pp. 492–503 (2011)
12. Dillig, I., Dillig, T.: EXPLAIN: A tool for performing abductive inference. In: Sharygina, N., Veith, H. (eds.) CAV 2013. LNCS, vol. 8044, pp. 684–689. Springer, Heidelberg (2013)
13. Dillig, I., Dillig, T., Aiken, A.: SAIL: Static Analysis Intermediate Language. Stanford University Technical Report
14. Unix coreutils, <http://www.gnu.org/software/coreutils/>:
15. Openssh 5.3p1, <http://www.openssh.com/>:
16. Vechev, M., Yahav, E., Yorsh, G.: Abstraction-guided synthesis of synchronization. In: POPL, pp. 327–338 (2010)
17. Seshia, S.: Sciduction: combining induction, deduction, and structure for verification and synthesis. In: DAC, pp. 356–365 (2012)
18. Srivastava, S., Gulwani, S., Foster, J.: From program verification to program synthesis. In: POPL, pp. 313–326 (2010)



19. Beyene, T., Chaudhuri, S., Popeea, C., Rybalchenko, A.: A constraint-based approach to solving games on infinite graphs. In: POPL (2014)
20. Bloem, R., Chatterjee, K., Henzinger, T.A., Jobstmann, B.: Better quality in synthesis through quantitative objectives. In: Bouajjani, A., Maler, O. (eds.) CAV 2009. LNCS, vol. 5643, pp. 140–156. Springer, Heidelberg (2009)
21. Jha, S., Seshia, S., Tiwari, A.: Synthesis of optimal switching logic for hybrid systems. In: 2011 Proceedings of the International Conference on Embedded Software (EMSOFT), pp. 107–116 (2011)
22. Chaudhuri, S., Solar-Lezama, A.: Smooth interpretation. In: PLDI (2010)
23. Chaudhuri, S., Clochard, M., Solar-Lezama, A.: Bridging boolean and quantitative synthesis using smoothed proof search. In: POPL (2014)
24. Younan, Y., Joosen, W., Piessens, F.: Runtime countermeasures for code injection attacks against c and c++ programs. *ACM Comput. Surv.* 44(3), 17 (2012)
25. Dhurjati, D., Kowshik, S., Adve, V., Lattner, C.: Memory safety without runtime checks or garbage collection. In: LCTES 2003, pp. 69–80 (2003)
26. Balakrishnan, G., Reps, T.: Analyzing memory accesses in x86 executables. In: Duesterwald, E. (ed.) CC 2004. LNCS, vol. 2985, pp. 5–23. Springer, Heidelberg (2004)
27. Condit, J., Harren, M., Anderson, Z., Gay, D.M., Necula, G.C.: Dependent types for low-level programming. In: De Nicola, R. (ed.) ESOP 2007. LNCS, vol. 4421, pp. 520–535. Springer, Heidelberg (2007)
28. Beyer, D., Henzinger, T.A., Jhala, R., Majumdar, R.: Checking memory safety with blast. In: Cerioli, M. (ed.) FASE 2005. LNCS, vol. 3442, pp. 2–18. Springer, Heidelberg (2005)
29. Yang, H., Lee, O., Berdine, J., Calcagno, C., Cook, B., Distefano, D., O’Hearn, P.W.: Scalable shape analysis for systems code. In: Gupta, A., Malik, S. (eds.) CAV 2008. LNCS, vol. 5123, pp. 385–398. Springer, Heidelberg (2008)
30. Berdine, J., Cook, B., Ishtiaq, S.: sLAYER: Memory safety for systems-level code. In: Gopalakrishnan, G., Qadeer, S. (eds.) CAV 2011. LNCS, vol. 6806, pp. 178–183. Springer, Heidelberg (2011)
31. Wagner, D., Foster, J.S., Brewer, E.A., Aiken, A.: A first step towards automated detection of buffer overrun vulnerabilities. In: Network and Distributed System Security Symposium, pp. 3–17 (2000)
32. Necula, G.C., McPeak, S., Weimer, W.: CCured: type-safe retrofitting of legacy code. In: POPL, pp. 128–139 (2002)
33. Jim, T., Morrisett, J.G., Grossman, D., Hicks, M.W., Cheney, J., Wang, Y.: Cyclone: A safe dialect of c. In: USENIX Annual Technical Conference, General Track, pp. 275–288 (2002)
34. Distefano, D., Filipović, I.: Memory leaks detection in java by bi-abductive inference. In: Rosenblum, D.S., Taentzer, G. (eds.) FASE 2010. LNCS, vol. 6013, pp. 278–292. Springer, Heidelberg (2010)
35. Botinčan, M., Distefano, D., Dodds, M., Grigore, R., Parkinson, M.J.: corestar: The core of jstar. In: Boogie. Citeseer (2011)
36. Dillig, I., Dillig, T., Aiken, A.: Automated error diagnosis using abductive inference. In: PLDI, pp. 181–192. ACM