Partial Evaluation of Machine Code *

Venkatesh Srinivasan
University of Wisconsin–Madison, USA
venk@cs.wisc.edu

Thomas Reps
University of Wisconsin–Madison
and GrammaTech, Inc., USA
reps@cs.wisc.edu

Abstract
This paper presents an algorithm for off-line partial evaluation of machine code. The algorithm follows the classical two-phase approach of binding-time analysis (BTA) followed by specialization. However, machine-code partial evaluation presents a number of new challenges, and it was necessary to devise new techniques for use in each phase.

• Our BTA algorithm makes use of an instruction-rewriting method that “decouples” multiple updates performed by a single instruction. This method counters the cascading imprecision that would otherwise occur with a more naïve approach to BTA.

• Our specializer specializes an explicit representation of the semantics of an instruction, and emits residual code via machine-code synthesis. Moreover, to create code that allows the stack and heap to be at different positions at run-time than at specialization-time, the specializer represents specialization-time addresses using symbolic constants, and uses a symbolic state for specialization.

Our experiments show that our algorithm can be used to specialize binaries with respect to commonly used inputs to produce faster binaries, as well as to extract an executable component from a bloated binary.

Categories and Subject Descriptors F.3.2 [Semantics of Programming Languages]: Partial evaluation

Keywords Partial evaluation, machine code, BTA, specialization, machine-code synthesis, IA-32 instruction set

1. Introduction
The analysis and rewriting of binaries has gotten an increasing amount of attention from the academic community in the last decade (e.g., see references in [44, §7], [9, §1], [13, §1], [18, §7]). One of the potential applications of machine-code\(^1\) analysis and rewriting is to facilitate reuse of software binaries in the absence of source code. In particular, it would be desirable to have a binary-rewriting tool that can (i) optimize a binary for the common case (e.g., a file-write routine optimized for a certain file descriptor), or (ii) extract an executable component from a bloated binary (e.g., a small text encoder embedded in a web server).

Partial evaluation [25] is a program-specialization framework that can perform the aforementioned tasks. However, there are no tools that partially evaluate machine code. Existing binary-rewriting tools either de-obfuscate [16, 42, 48], superoptimize [11, 39], or harden [4, 20, 43] binaries, but do not perform partial evaluation. Moreover, machine-code partial evaluation presents a number of new challenges, and source-code partial-evaluation algorithms [6, 15, 26] would not produce satisfactory results if they were applied to machine code in a straightforward manner (see §3).

This paper presents an algorithm for off-line partial evaluation of machine code. The inputs to the algorithm are a binary B, and a division of B’s inputs as S (static inputs) and D (dynamic inputs). Values for static inputs are known at specialization time; values for dynamic inputs are unknown at specialization time. Our partial-evaluation algorithm produces a binary B\(_S\) that satisfies the equation

\[
[B](S, D) = [B_S](D),
\]

where \([ \cdot ]\) denotes the meaning function for the instruction set in which the binary is written. Executing B\(_S\) with input D produces the same results as executing B with inputs S and D. However, B\(_S\) is specialized with respect to S, and can be significantly faster than B.

Partially evaluating a binary with respect to commonly used inputs produces specialized versions of the binary that are optimized for the common inputs. (See §6.1.)

\(^1\) We use the term “machine code” to refer generically to low-level code, and do not distinguish between actual machine-code bits/bytes and assembly code to which it is disassembled.
evaluating with respect to a fixed value of an input flag can extract a smaller executable component from a bloated binary. (See §6.2.) The extracted component can be used in embedded systems where executable size matters, or be linked against other programs that require the component. (See the experiment with Iezfx in §6.2.)

We have implemented our algorithm in a tool, called WiPER (an acronym for the “Wisconsin Partial Evaluator”), that partially evaluates Intel IA-32 binaries. WiPER works on the Intermediate Representations (IRs) recovered from machine code by an existing tool, CodeSurfer/x86 [10]. WiPER follows the classical two-phase approach of binding-time analysis (BTA) followed by specialization. WiPER’s BTA uses forward slicing to determine a priori which instructions can be specialized away (static instructions), and which cannot (dynamic instructions). Given the values for static inputs in the form of a partial state, WiPER’s specializer evaluates static instructions, residuates code for static values that might be used by dynamic instructions, and emits unmodified dynamic instructions.

As mentioned earlier, there are several new challenges that arise in the context of machine-code partial evaluation. These issues, along with the strategies that WiPER uses to overcome them, distinguish WiPER from other partial evaluators that have been described in the literature.

- **Machine-code instructions are usually multi-assignments:** they have several inputs, and several outputs (e.g., registers, flags, and memory locations). This aspect of the language introduces a granularity issue during slicing: in some cases, although we would like the slice to follow only a subset of an instruction’s semantics, the slicing algorithm is forced to include the entire instruction. This effect can cascade, and cause BTA to be very imprecise. The BTA algorithm in WiPER makes use of an instruction-rewriting method that decouples multiple updates performed by a single instruction by splitting the assignments across multiple instructions. The decoupling transformation increases the precision of WiPER’s BTA significantly.

- **An instruction’s abstract-syntax tree (AST) is often not parameterized by all of the operands of the instruction:** some operands are “baked into” the semantics. In contrast with source-code partial evaluators that specialize the AST of a dynamic statement, WiPER specializes an explicit representation of the semantics of a dynamic instruction, and uses a machine-code synthesizer [46] to produce residual code.

- **In machine code, values in memory are accessed by explicit address computations, followed by loading.** The residual code produced by a naïve machine-code partial evaluator will contain specialization-time addresses, and consequently would not allow the stack and heap to be at different positions at run-time than at specialization-time. WiPER represents specialization-time addresses using symbolic constants, and uses a symbolic state for specialization. The resulting code produced by WiPER is independent of the layout of the specialization-time address space, and can be run with, e.g., the stack base at a different address.²

The contributions of our work include the following:

- **We present an algorithm for machine-code partial evaluation.** To the best of our knowledge, WiPER is the first partial evaluator that works on machine code.³

- **We have developed an instruction-decoupling transformation to counter cascading imprecision caused by the granularity issue in machine-code slicing.**

- **We have introduced an instruction-specialization technique that specializes an explicit representation of the semantics of an instruction with respect to a partial state, and residuates code via machine-code synthesis.**

- **We have developed a specialization algorithm that represents specialization-time addresses using symbolic constants, and uses a symbolic state for specialization.** The residual code produced by our specializer is independent of the specialization-time memory layout, and allows the stack and heap to be at different positions at run-time than at specialization-time.

Our methods have been implemented in WiPER, a partial evaluator for Intel IA-32. We partially evaluated seven test applications using WiPER, and found that, on average (computed via geometric mean), the specialized binaries have a speedup of 1.3. We also used WiPER to extract executable components from bloated binaries such as bzip2.

## 2. Background

In this section, we briefly describe source-code partial evaluation, slicing, syntax and semantics of the IA-32 instruction-set architecture (ISA), and a logic to express the semantics of IA-32 instructions.

### 2.1 Partial Evaluation

Partial evaluation is a framework for specializing and optimizing programs [25]. Given a program that takes some inputs, and given values for some of the inputs, the goal of partial evaluation is to produce a program that is specialized (optimized) with respect to the input values provided. Fig. 1 shows a program power that computes \((a + b)^n\), where \(a\), \(b\), and \(n\) are the inputs to the program. Suppose that power is frequently called with the values \(b = 1\), and \(n = 4\). Given power, and the input values \(b = 1\) and \(n = 4\), a partial evaluator produces the residual program \(\text{power}_{R}\) shown in Fig. 2. As we can see in \(\text{power}_{R}\), partial evaluation has compiled data \((n = 4)\) into control. (See Eqn. (1).) In effect, partial evaluation performs optimizations such as loop unrolling.

² Note that we are referring here to the ability to reposition the stack at the beginning of run-time, not the relocatability of the residual code per se. The residual code produced WiPER is also relocatable.

³ Confirmed by personal communication with Neil Jones, Robert Glück, and Saumya Debray.
int power(int a, int b, int n) {
    int prod = 1;
    while (n--) {
        prod *= (a + 1);
    }
    return prod;
}

int power_r(int a) {
    int prod = 1;
    while (n--) {
        prod *= (a + 1);
    }
    return prod;
}

Figure 1: Input power program. Computes \((a + b)^n\).

Figure 2: Residual power program. Computes \((a + 1)^n\).

can be used in a dynamic context. For example, consider the program \texttt{foo} given in Fig. 3. Because \texttt{c}’s value is established dynamically in line 3, both of the formal parameters of \texttt{add} are classified dynamic by BTA. Although both actual parameters of the call to \texttt{add} in line 4 of Fig. 3 are static, they appear in a dynamic context—the static actuals will be bound to dynamic formals in \texttt{add}. To produce a residual program that compiles and/or does not access uninitialized locations, the partial evaluator must residuate code for the static actuals. Thus the partial evaluator “lifts” static expressions that appear in dynamic contexts, and the specialister residuates code for lifted expressions (e.g., line 3 in Fig. 4). The term “lifting” refers to the process of changing the binding time of an expression from static to dynamic.

Suppose that each statement in the original program is uniquely identified by a label \(l\), and that each concrete state \(\sigma\) that arises during an execution is partitioned into \(\sigma_S\) and \(\sigma_D\) according to the division established by BTA. Then an execution trace of the original program is a sequence of elements of the form \(l : (\sigma_S, \sigma_D)\). (Each element of the trace records the state \((\sigma_S, \sigma_D)\) observed at an execution of a statement \(l\).) The effect of partial evaluation on an execution trace of the specialized program can be expressed by the following reparenthesization:

\[
l : (\sigma_S, \sigma_D) \rightarrow (l, \sigma_S) : \sigma_D
\]

Operationally, the partial evaluator creates a new residual statement for every unique \((l, \sigma_S)\) pair it observes at specialization time. Eqn. (1) shows how each element \((l, \sigma_S) : \sigma_D\) in an execution trace of the specialized program can be mapped back to the corresponding element \(l : (\sigma_S, \sigma_D)\) in the trace of the original program.

### 2.2 Slicing

Slicing [24, 47] computes the set of program points that affect (or are affected by) a given program point called the slicing criterion. (Backward slicing computes the set of program points that might affect the slicing criterion; forward slicing computes the set of program points that might be affected by the slicing criterion.) Slicing is typically performed using an IR called a system dependence graph (SDG) [21, 24]. An SDG consists of a set of program dependence graphs (PDGs), one for each procedure in the program. A node in a PDG corresponds to a construct in the program, such as a statement, a condition, a call to a procedure, a procedure entry/exit, an actual parameter of a call, or a formal parameter of a procedure. The edges correspond to data and control dependences between the nodes [21]. The PDGs are connected together with interprocedural control-dependence edges between call-site nodes and procedure-entry nodes, and interprocedural data-dependence edges between actual parameters and formal parameters/return values. Given an SDG representation of the program and a slicing criterion, an interprocedural-slicing algorithm includes in the slice the
INT ∈ Integer, Reg ∈ Register, I ∈ Instruction, O ∈ Operand,
R ∈ RegisterOperand, C ∈ ImmediateOperand,
M ∈ IndirectOperand, RC ∈ RegisterImmediateOperand
INT ::= {..., -1, 0, 1, ...} → Reg ::= eax | esp | esp ... 
R ::= Direct(Reg) → C ::= Imm(INT) → RC ::= R | C
M ::= Indirect(R, R, INT) → O ::= RC | M
I ::= push(O) | mov(R, O) | mov(M, RC) | lea(R, M) |
add(R, O) | add(M, RC) | sub(R, O) | sub(M, RC) |
cmp(M, C) | jz(C) | jmp(C)

Figure 5: Abstract syntax for a subset of IA-32.

SDG nodes that reach (or can be reached from) the slicing criterion by following data- and control-dependence edges.

A context-sensitive interprocedural-slicing algorithm uses context-free language (CFL) reachability [35] to reduce the number of nodes in the slice [24, 36].

2.3 Syntax and Semantics of IA-32

Syntax. The abstract syntax for a subset of IA-32 instructions that will be used in this paper is given in Fig. 5. Indirect operands are of the form ⟨base, index, scale, offset⟩, where base and index are registers, and scale and offset are integers. The effective address is computed as base + index * scale + offset. In instructions with two operands, the operand on the left is the destination (except the cmp instruction, in which both operands are source operands).

Semantics. The primitive domains of the IA-32 semantics include 32-bit integers, Booleans, registers, and flags. The primitive domains and their operators are given below.

\[\begin{align*}
i & \in \text{INT} = \mathbb{Z}_{32} \\
b & \in \text{BOOL} = \{\text{True}, \text{False}\} \\
r & \in \text{Register} = \{\text{EAX, ESP, ...}\} \\
f & \in \text{Flag} = \{\text{CF, SF, ...}\} \\
op & \in \text{ArithOp} = \{+, -, ..., \} \\
bop & \in \text{BoolOp} = \{\land, \lor, ...\} \\
rop & \in \text{RelOp} = \{=, \neq, <, ...\} \\
\end{align*}\]

\[\begin{align*}
\text{Store} & \text{ is a compound domain denoting an IA-32 store.}
\text{Store} & \text{ is a triple consisting of three maps: a register map, a flag map, and a memory map. \text{Store} has operators to access and update the maps. \text{Store} domain is defined below.}
\end{align*}\]

\[\begin{align*}
\rho & \in \text{Store} = \text{RegMap} \times \text{FlagMap} \times \text{MemMap} \\
\text{RegMap} : \text{Flag} & \rightarrow \text{BOOL} \\
\text{MemMap} : \text{INT} & \rightarrow \text{INT}
\end{align*}\]

The valuation function \(I\) has the type \(I : \text{Instruction} \rightarrow \text{Store} \rightarrow \text{Store}\). \(I\) takes an instruction \(I\) and a pre-store, and returns a post-store that reflects the updates made by the execution of \(I\). For example, the valuation function for the \(\text{mov eax, [ebp]}\) instruction is given below. The instruction copies a 32-bit value from the memory location pointed to by the frame-pointer register \(EBP\) to the \(EAX\) register. The overloaded function \([\cdot]\) returns primitive semantic objects for their syntactic counterparts. The instruction also increments the program counter \(EIP\) by the length of the instruction. For brevity, we do not show this increment explicitly.

\[\begin{align*}
T & \in \text{Term}, \varphi \in \text{Formula}, \text{FE} \in \text{FuncExpr} \\
c & \in \text{Int32} = \{\ldots, -1, 0, 1, \ldots\} \\
b & \in \text{BOOL} = \{\text{True, False}\} \\
I_{\text{Inst32}} & \in \text{Int32Id} = \{\text{EAX, ESP, EBP, ...}\} \\
I_{\text{Bool}} & \in \text{BoolId} = \{\text{CF, SF, ...}\} \\
F & \in \text{FuncId} = \{\text{Mem}\} \\
op & \in \text{ArithOp} = \{+, -, ..., \} \\
bop & \in \text{BoolOp} = \{\land, \lor, ...\} \\
rop & \in \text{RelOp} = \{=, \neq, <, ...\} \\
T & ::= c | I_{\text{Inst32}} | T_1 \op T_2 | \text{ite} (\varphi, T_1, T_2) | F (T_1) \\
\varphi & ::= b | I_{\text{Bool}} | T_1 \rop T_2 | \neg \varphi_1 | \varphi_1 \bop \varphi_2 | F = \text{FE} \\
\text{FE} & ::= F | \text{FE}_1 [T_1 \mapsto T_2]
\end{align*}\]

Figure 6: Syntax of L[IA-32].

\[\begin{align*}
\text{I[mov eax, [ebp]]} \rho = \text{update}_{\text{reg}}(\rho, [\text{eax}], \\
\text{access}_{\text{mem}}(\rho, \text{access}_{\text{reg}}(\rho, [\text{ebp}])))
\end{align*}\]

2.4 QFBV Formulas for IA-32

The semantics of IA-32 instructions can also be formally expressed by formulas in a logic. Consider a quantifier-free bit-vector logic \(L\) over finite vocabularies of constant symbols and function symbols. We will be dealing with a specific instantiation of \(L\), denoted by \(L[\text{IA-32}]\). (L can also be instantiated for other ISAs.) In \(L[\text{IA-32}]\), some constants represent IA-32’s registers (\(\text{EAX, ESP, EBP, ...}\)), and some represent flags (\(\text{CF, SF, ...}\)). \(L[\text{IA-32}]\) has only one function symbol “\(\text{Mem}\)”, which denotes memory. Note that syntactic constructs of \(L[\text{IA-32}]\) are boldfaced to distinguish them from their counterparts in IA-32 semantics. The syntax of \(L[\text{IA-32}]\) is defined in Fig. 6.

A term of the form \(\text{ite}(\varphi, T_1, T_2)\) represents an if-then-else expression. A \(\text{FuncExpr}\) of the form \(\text{FE}[T_1 \mapsto T_2]\) denotes a function-update expression.

The function \(\langle \cdot \rangle\) converts an IA-32 instruction sequence into a QFBV formula. While others have created such encodings by hand (e.g., [38]), we use a method that takes a specification of the concrete operational semantics of IA-32 instructions and creates a QFBV encoder automatically. The method reinterprets each semantic operator as a QFBV formula-creator or term-creator. (See [30].) To write formulas that express store transitions, all \(\text{Int32Ids}, \text{BoolIds},\) and \(\text{FuncIds}\) can be qualified by primes (e.g., \(\text{Mem}'\)). The QFBV formula for an instruction sequence is a restricted 2-valuation formula that specifies a store transformation (also known as a “transition formula”). It has the form

\[\begin{align*}
\bigwedge_n (t'_n = t_n) \land \bigwedge_n (j'_n = \varphi_n) \land \text{Mem}' = \text{FE}
\end{align*}\]

where \(t'_n\) and \(j'_n\) range over the constant symbols for registers and flags, respectively. The primed vocabulary is the post-store vocabulary, and the unprimed vocabulary is the pre-store vocabulary. We also refer to them as “vocabulary 1” and “vocabulary 0,” respectively. The QFBV formulas for a few IA-32 instructions are given in Fig. 7. The second instruction pushes the 32-bit constant value 0 on the stack. If the zero flag \(ZF\) is set, the third instruction updates \(EIP\)
\[\langle\langle \text{mov eax, [ebp]} \rangle\rangle \equiv EAX' = \text{Mem}(EBP)\]
\[\langle\langle \text{push 0} \rangle\rangle \equiv ESP' = ESP - 4 \land \text{Mem}' = \text{Mem}[ESP - 4 \rightarrow 0]\]
\[\langle\langle \text{jz 1000} \rangle\rangle \equiv \text{ite}(ZF = 0, \text{EIP}' = 1000, \text{EIP}' = \text{EIP} + 4)\]
\[\langle\langle \text{lea eax, [ebp+4]} \rangle\rangle \equiv EAX' = EBP + 4\]

Figure 7: QFBV formulas for example IA-32 instructions.

to the value 1000; otherwise, it increments \text{EIP} by four. The fourth instruction loads \text{EAX} with the value \text{EBP}+4 (without modifying the value of any flag).

In this section, and in the rest of the paper, we show only the portions of QFBV formulas that express how the store is modified. QFBV formulas actually contain identity conjuncts of the form \(I_i = I_i' = I_j\) and \(\text{Mem}' = \text{Mem}\) for constants and functions that are unmodified.

3. Overview

At a very high level, WiPER is similar to the off-line source-code partial evaluator described in §2.1; WiPER’s algorithm works in two phases, BTA and specialization. However, because WiPER is dealing with machine code, WiPER’s algorithm differs significantly from that of a source-code partial evaluator. In this section, we present an example to illustrate WiPER’s algorithm, while highlighting the following: (i) the issues that arise in machine code, (ii) why techniques used in a source-code partial evaluator are unsatisfactory to resolve the issues, and (iii) how WiPER resolves the issues.

Fig. 8 shows an IA-32 program \text{sum} that takes, \(a\), \(v\), and \(n\) as inputs, and computes \(a + b[n]\). (Array \(b\) is statically allocated and has 5 elements. Lines 6–12 in Fig. 8 initialize the elements of \(b\) to \(v\).) \(n\) is assumed to be between 0 and 4. The inputs \(a\), \(v\), and \(n\) are available in the eax, ebx, and ecx registers, respectively, at the beginning of the program. The output is available in the eax register at the end of the program. Let us use WiPER to partially evaluate \text{sum} with respect to the input value \(v = 1\).

No source-code variables. Recall from §2.1 that the goal of BTA is to compute the division of all program variables as either static or dynamic. The abstraction of a source-code "variable" is absent at the machine-code level. (We assume that the input binary lacks symbol-table and debugging information.) Consequently, a division of source-code variables cannot be obtained at the machine-code level.

However, there are tools [10] that recover "variable-like" abstractions [9] from machine code, and use them to construct IRs such as an SDG, a control flow graph (CFG), etc. WiPER performs BTA via forward slicing over the SDG recovered from the binary [17].

The inputs to WiPER’s BTA phase are: (i) the SDG of the binary, and (ii) the instructions that initialize the dynamic inputs (the slicing criterion). The output is an annotated program, whose instructions are annotated with binding times: static (S), or dynamic (D). In our example, the dynamic inputs \(a\) and \(n\) are available in the eax and ecx registers at instructions 1 and 3, respectively. Consequently, instructions 1 and 3 constitute the slicing criterion. All instructions that are in the forward slice depend on \(a\) and \(n\), and hence are classified dynamic; the remaining instructions are classified static.

Another feature of ISAs that makes them different from typical high-level languages is that most instructions are multi-assignments: they have several inputs, and several outputs. The next two issues discussed below arise in machine-code partial evaluation because instructions have multiple outputs, and multiple inputs, respectively.

The granularity issue in slicing. Consider instruction 1 in the \text{sum} program. 1 modifies the stack-pointer register ESP in addition to copying the value in the EAX register to a memory location. Because all of the remaining instructions in \text{sum} either directly or transitively use ESP, the slice with respect to instruction 1 consists of all the instructions in \text{sum}, and thus every instruction in \text{sum} would be classified dynamic. This imprecision in slicing is caused by a granularity problem: the \text{push} instruction independently updates ESP and the memory location. Because the dynamic input is in EAX, the slice requires only the memory update that \text{push} performs and not the update to the value of ESP. However, a slice cannot include only a part of an instruction; it has to include the entire instruction. Consequently, the entire \text{push} instruction, and the remaining instructions in \text{sum} are added to the slice. This issue is illustrated by the SDG snippet on the left in Fig. 9.

To resolve this issue, WiPER rewrites \text{sum} by decoupling each instruction that independently updates the stack pointer along with another register or memory location \(l\) (e.g., \text{push}, \text{pop}, \text{leave}, etc.). For such instructions, WiPER uses a machine-code synthesizer [46] (a tool that synthesizes an instruction sequence from a semantic specification) to synthe-
size an equivalent instruction sequence that updates ESP in one instruction and l in a different instruction. The effect of decoupling on BTA is illustrated by the SDG snippet on the right in Fig. 9.

The rewritten binary \(\sum'\) is shown in Fig. 10. In \(\sum'\), the new slicing criterion includes instructions \(2'\) and \(5'\). The new forward slice includes only the instructions highlighted in Fig. 10 in light gray, which are classified dynamic. The remaining instructions are classified static.\(^4\)

The inputs to the specialization phase of WiPER are: (i) the interprocedural CFG of the binary, (ii) the binding-time annotations, and (iii) an input partial static-store. The output is the residual program. Recall from §2.3 that an IA-32 store is a triple that consists of a register map, a flag map, and a memory map. The input partial static-store \(\rho_0\) is

\[
\rho_0 \equiv \langle \text{ESP} \mapsto 1000 \rangle \langle \text{EBX} \mapsto 1 \rangle \langle v, [\text{ESP}] \rangle.
\]

In the rest of the paper, when we use the term “store,” we are referring to a partial static-store. The input store has the value \(l\) for \(v\), and 1000 for the stack pointer. Let us specialize \(\sum'\) with respect to \(\rho_0\).

Non-parameterized operands in Instruction ASTs. Recall from §2.1 that \(\text{PE}_S\) substitutes values for operands while specializing ASTs of dynamic expressions and statements. However, WiPER cannot use this simple approach for specializing dynamic instructions. For an instruction that has multiple input operands, the instruction AST may not always be parameterized by all operands. For example, the instruction \(\text{adc eax, ebx}\) adds the values in registers EAX and EBX, and the value of the carry flag CF. The flag operand is “baked into” the semantics, and is not an explicit syntactic operand. Another example is the instruction \(\text{push eax}\), which has the stack-pointer operand (ESP) baked into the semantics.

To handle this issue, WiPER specializes a representation of the semantics of an instruction—instead of its syntax—with respect to a store. WiPER uses a QFBV formula to represent an instruction’s semantics.

Now let us use the following approach to specialization:

\(^4\)The division produced by WiPER’s BTA is pointwise, and monovariant for procedures (i.e., an instruction in a procedure is classified as static or dynamic for all calling contexts).

1. Static instructions: Evaluate.
2. Dynamic instructions: Specialize the instruction’s QFBV formula with respect to the pre-store, and residue code. WiPER uses a machine-code synthesizer to synthesize an instruction-sequence that is equivalent to the specialized QFBV formula. The synthesized instruction-sequence constitutes the residual code.

Let us specialize \(\sum'\) with respect to \(\rho_0\) using this simple approach. The specialization approach did not residuate code because it is static. The post-store is \(\rho_1\).

\[
\rho_1 \equiv \langle \text{ESP} \mapsto 996 \rangle \langle \text{EBX} \mapsto 1 \rangle \langle v, [\text{ESP}] \rangle.
\]

The QFBV formula for instruction \(2'\) is

\[
\langle z' \rangle \equiv \langle v \rangle \langle \text{Mem} \rangle \langle \text{ESP} \rangle \langle \text{EAX} \rangle.
\]

Because \(z'\) is dynamic, the specializer specializes \(\langle z' \rangle\) with respect to \(\rho_1\) to obtain the specialized formula

\[
\langle \text{Mem}' \rangle \equiv \langle v \rangle \langle \text{Mem} \rangle \langle 996 \rangle \langle \text{EAX} \rangle,
\]

and uses the synthesizer to synthesize the instruction “mov [996], eax.” After evaluating static instructions \(3'\) and \(4'\), the resulting store is

\[
\rho_4 \equiv \langle [\text{ESP} \mapsto 992] \langle \text{EBX} \mapsto 1 \rangle \langle v, [\text{ESP}] \rangle \langle 992 \rangle \rangle.
\]

The residual program produced by this simple approach is given below.

\[
\begin{align*}
\text{mov} \ [996], \text{eax} & \quad \text{mov} \ [998], \text{ecx} \\
\text{mov} \ [988], \text{ecx} & \quad \text{mov} \ [964+\text{ecx}], \text{ebx} \\
\text{mov} \ [996], \text{eax} & \quad \text{add} \ \text{eax}, \text{ebx}
\end{align*}
\]

Lifting. The approach described above produces code that can access uninitialized locations. For example, one can see that the instruction “mov [964+ecx], eax” in the residual program dereferences a dynamic pointer to access an element in the static array \(b\). However, the static array is not initialized by the residual code. This issue arises because our simple specialization approach did not residuate code for static instructions that produce values that might be consumed by a downstream dynamic instruction (a.k.a., a dynamic context).

To resolve this issue, WiPER conservatively lifts static predecessors of a dynamic instruction. In Fig. 10, the boxed instructions are the lifted instructions. After lifted instructions are identified, one can use the following approach to specialization:

Approach \(\text{PE}_1\)

- Static instructions: Evaluate.
- Lifting instructions: Specialize the instruction’s QFBV formula with respect to the pre-store, and residue code. Then update the store by evaluating the instruction.
- Dynamic instructions: Emit. Because all static locations that might be used by a dynamic instruction \(d\) have been initialized by upstream residual lifted instructions, there is no static data that needs to be infused into \(d\) via specialization, and the unmodified \(d\) can be emitted.
Let us specialize $\sum'$ with respect to $\rho_0$ using PE$_1$. Because $1'$ is lifted, the specializer residuates code for it. The QFBV formula for $1'$ is $\langle \langle 1' \rangle \rangle \equiv ESP' = ESP - 4$. The specializer specializes $\langle \langle 1' \rangle \rangle$ with respect to $\rho_0$ to obtain the specialized formula $ESP' = 996$, and uses the synthesizer to synthesize instruction $1_1$ in $\sum_1$ (shown in Fig. 11). The specializer also evaluates $1'$ to obtain $\rho_1$ (Eqn. (2)). Because $2'$ is dynamic, the specializer emits it as $2_1$. Because $3'$ is static, the specializer evaluates the instruction without residuating code. The new residual program $\sum_1$ is shown in Fig. 11. In $\sum_1$, one can see that all the static values required by emitted dynamic instructions ($2_1$, $6_1$, etc.) have been set up by residual lifted instructions ($1_1$, $5_1$, etc.).

**Residual specialization-time addresses.** In $\sum_1$, one can see that the specialization-time values of the stack pointer have been residuated as constants. Although $\sum_1$ is a correct residual program for $\sum'$ partially evaluated with respect to $\rho_0$, the specialization-time stack layout is hardwired into the residual code. (The stack begins at address 1000, and grows toward lower addresses.) As a result, the residual code might crash if a different address is used as the base of the runtime stack.

One way to prevent the specializer from emitting stack-pointer values is to make the stack pointer dynamic by including instruction $1'$ in the slicing criterion during BTA. However, that would make all instructions in $\sum'$ dynamic.

To resolve this issue, WiPER uses a *symbolic* store during specialization, instead of a *concrete* store. In the initial symbolic store, specialization-time addresses (e.g., the initial value of ESP) are represented using symbolic constants. Static non-address values are represented using integer and Boolean constants. WiPER uses PE$_2$ for specialization.

**Approach PE$_2$**

- Static instructions: Symbolically evaluate. Although specialization-time addresses are symbolic, WiPER can still access and update static values in memory.
- Lifted instructions: Specialize the instruction’s QFBV formula with respect to the symbolic pre-store, and residuate code. Then update the symbolic store by symbolically executing the instruction.
- Dynamic instructions: Emit.

Let us illustrate PE$_2$ on $\sum'$. The new initial store is $\rho_0^{sym} \equiv \langle \langle ESP \rightarrow c [EBX \rightarrow I] \rangle \rangle$. Note that the values in the store are boldface to indicate that they are logical terms, and not semantic values. The value of the stack pointer ESP is the symbolic constant $c$ to indicate that we are not restricting ESP to a concrete value known at specialization-time.

Before starting specialization, WiPER residuates instruction $0_2$ in $\sum_2$ (shown in Fig. 12) to save the initial value of the stack pointer ESP in a dedicated location. (In this example, WiPER uses the ESI register because ESI is not used in $\sum'$.) This location will be used whenever downstream residual instructions need the initial value of ESP. To residuate code for the lifted instruction $1'$, WiPER specializes $\langle \langle 1' \rangle \rangle$ with respect to $\rho_1^{sym}$. The specialized formula is $ESP' = c - 4$. Because at run-time the value of $c$ will be available in ESI, WiPER replaces $c$ with ESI to produce $ESP' = ESI - 4$. WiPER uses the synthesizer to synthesize instruction $1_2$ for the specialized formula. WiPER also symbolically evaluates $1'$ to produce $\rho_1^{sym}$.

$\rho_1^{sym} \equiv \langle \langle ESP \rightarrow c - 4 [EBX \rightarrow I] \rangle \rangle$. $2'$ is dynamic, and WiPER emits it as $2_2$. After symbolically evaluating static instructions $3'$ and $4'$, we obtain the store $\rho_2^{sym} \equiv \langle \langle ESP \rightarrow c - 8 [EBX \rightarrow I] \rangle \rangle$.

The final residual program $\sum_2$ produced by WiPER is shown in Fig. 12. One can see that WiPER arranges for the initial stack-pointer value to be retrieved from ESI ($1_2$, $5_2$, $9_2$, $13_2$, etc.) instead of residuating the stack-pointer value at residual lifted instructions (cf. $1_1$, $5_1$, $9_1$, $13_1$, etc. in $\sum_1$); consequently, $\sum_2$ will not crash when the stack base is initialized to different addresses. In contrast, static non-address quantities are residuated at lifted instructions (e.g., the value $1$ in instructions $13_2$ through $13_2$).

**Why do PE$_1$ and PE$_2$ work?**

To formalize the effect of PE$_1$ and PE$_2$, we introduce the notion of an *environment*, which makes explicit the starting address of the stack and heap blocks. An environment maps a symbolic constant denoting the starting address of the stack or a heap block to a concrete address (or “location”). A machine-code state $\sigma$ consists of an environment $\eta$, a store $\rho$, which maps locations to values. For example, $\rho_4$ in Eqn. (3) is actually part of a state $\sigma_4$

$$
\sigma_4 \equiv \langle \eta_4, \rho_4 \rangle \equiv \langle \text{Stack} \rightarrow 1000,
\langle \langle ESP \rightarrow 992 \rangle [EBX \rightarrow I] \rangle, [\langle 992 \rightarrow I \rangle] \rangle.
$$

$\eta_4$ maps the start of the stack, denoted by the symbolic constant Stack, to the address 1000.

Suppose that $\eta$ and $\rho$ can be partitioned into $\eta_S$ and $\eta_D$, and $\rho_S$ and $\rho_D$, respectively. The effect of PE$_1$ on
4.1 Pre-processing: Decoupling Instructions

Prior to performing BTA, WiPER rewrites the input binary by decoupling each instruction $i$ that updates the stack-pointer register $ESP$ along with another location $l$. WiPER uses the machine-code synthesizer MC SYNTH [46] for decoupling. MC SYNTH uses a divide-and-conquer strategy that splits $i$’s QFBV formula into two independent subformulas $\varphi_1$ and $\varphi_2$, such that the updates to $ESP$ and $l$ are split between $\varphi_1$ and $\varphi_2$. Then, MC SYNTH synthesizes code for $\varphi_1$ and $\varphi_2$, concatenates the results, and returns the resulting instruction sequence $I$. $ESP$ and $l$ are updated in different instructions in $I$. The rewriting of the input binary via instruction decoupling is shown as Alg. 1. In the algorithm, UpdatesSpanAndLoc returns true if an instruction updates the stack pointer along with another register or memory location; InstrToQFBV converts an instruction into a QFBV formula; Synth invokes MC SYNTH to synthesize an instruction sequence; Replace replaces an instruction in a binary with an instruction sequence.

### Algorithm 1 Instruction Decoupling (Pre-processing step)

**Input:** Binary B

**Output:** Rewritten binary B'

1. for each $i \in B$ do
2. if UpdatesSpanAndLoc($i$) then
3. $I \leftarrow$ Synth(InstrToQFBV($i$))
4. $B' \leftarrow$ Replace($B, i, I$)
5. end if
6. end for
7. return $B'$

Algorithm 1 shows how each execution trace of the binary can be expressed by the following reparenthesized:

$$l : (\eta_s, \rho_s, \eta_d, \rho_d) \rightarrow (l, \eta_s, \rho_s) : (\eta_d, \rho_d) \quad (4)$$

Operationally, the partial evaluator creates a new residual instruction for every unique $(l, \eta_s, \rho_s)$ triple it observes at specialization time. Eqn. (4) shows how each element $(l, \eta_s, \rho_s) : (\eta_d, \rho_d)$ in an execution trace of a binary specialized using $PE_1$ can be mapped back to the corresponding element $l : (\eta_s, \rho_s, \eta_d, \rho_d)$ in the trace of the original binary. Consequently, the specialization-time memory layout $(\eta_s)$ is hard-wired into the residual code. If the run-time memory layout is different, the residual code produced by $PE_1$ might crash.

To describe the symbolic techniques used in $PE_2$, we introduce a static symbolic store $\rho^sym_S$, which maps symbolic expressions to symbolic expressions, where each such symbolic expression is parameterized on the symbolic constants found in $\eta_s$. (Note that the domain of $\rho^sym_S$ also contains registers and flags.) $\rho^sym_S$ mirrors the static concrete store $\rho_S$, i.e., $[\rho^sym_S](\eta_s) = \rho_S$, where the left-hand side denotes the simplification of each symbolic expression in $\rho^sym_S$ with respect to the values obtained from $\eta_s$. The principal effect of $PE_2$ on each execution trace of the binary can be expressed by the following reparenthesized:

$$l : (\eta_s, \rho_s, \eta_d, \rho_d) \rightarrow (l, \rho^sym_S) : (\eta_s, \rho_s, \eta_d, \rho_d) \quad (5)$$

Eqn. (5) shows how each element $(l, \rho^sym_S) : (\eta_s, \rho_s, \eta_d, \rho_d)$ in an execution trace of a binary specialized using $PE_2$ can be mapped back to the corresponding element $l : (\eta_s, \rho_s, \eta_d, \rho_d)$ in the trace of the original binary. In effect, the residual code is independent of $\eta_s$. (See also §4.4.)

4. Algorithm

In this section, we describe the algorithms used by WiPER. First, we present the algorithm for a pre-processing step used prior to partial evaluation. Second, we present WiPER’s intraprocedural partial-evaluation algorithm. Third, we present extensions to handle multiple procedures. Finally, we present the threats to the validity of our algorithms.

4.1 Pre-processing: Decoupling Instructions

Prior to performing BTA, WiPER rewrites the input binary by decoupling each instruction $i$ that updates the stack-pointer register $ESP$ along with another location $l$. WiPER

4.2 Intraprocedural Partial Evaluation

In this section, we present WiPER’s intraprocedural partial-evaluation algorithm. First, we present the BTA algorithm; then, we present the specialization algorithm.

4.2.1 BTA

The slicing-based BTA algorithm used in WiPER is shown as Alg. 2. In the algorithm, $I_D$ is the set of instructions that initialize dynamic inputs; ForwardSlice computes a forward slice with respect to a slicing criterion; BTAMap is a data structure that maps each instruction to its binding time. If a static instruction is control dependent on a dynamic branch, the specialization algorithm might not terminate [25, Chap. 14]. To avoid the possibility of non-termination, ForwardSlice follows both data and control dependences during slicing. Lines 4–6 in Alg. 2 show the computation of lifted instructions during BTA. WiPER lifts all static calls to malloc (lines 9–11). This step is performed so that heap blocks that are allocated at specialization-time also get allocated at run-time. (See §4.2.2.3.)

### Algorithm 2 BTA algorithm in WiPER

**Input:** SDG, $I_D$

**Output:** BTAMap

1. slice ← ForwardSlice(SDG, $I_D$)
2. for each $I \in$ slice do
3. BTAMap ← BTAMap[$I \mapsto$ Dynamic]
4. for each static predecessor $S$ of $I$ do
5. BTAMap ← BTAMap[$S \mapsto$ Lifted]
6. end for
7. end for
8. for each $I \notin$ Dom(BTAMap) do
9. if $I$ is a call to malloc or an actual argument of a call to malloc then
10. BTAMap ← BTAMap[$I \mapsto$ Lifted]
11. else
12. BTAMap ← BTAMap[$I \mapsto$ Static]
13. end if
14. end for
15. return BTAMap
4.2.2 Specialization

This section presents the specialization algorithm used in WiPER. We first present the skeleton of WiPER’s specialization loop. We then present two versions of the core specialization function specialize, which gets called in the specialization loop. While the first version produces correct residual code, the residual code cannot be executed with the stack and heap blocks at different positions at run-time than at specialization-time. The second version does not have this limitation, and is used in WiPER.

4.2.2.1 Specialization loop. The skeleton of WiPER’s specialization loop is similar to that of a standard partial evaluator [25, p. 87], and is given as Alg. 3. specialize is the core specialization function (line 13 of Alg. 3), which specializes an instruction with respect to a pre-store to obtain the residual instruction and a post-store. In the algorithm, ε denotes the empty sequence of instructions; NewCFG creates a new CFG; FirstBB returns the first basic-block in a CFG; FinalBB returns true if the argument is the final basic-block in its parent CFG; RemoveItem removes an item from the worklist; Emit appends an instruction to a CFG. (We assume that Emit creates and connects nodes in the CFG.)

Recall that the register EIP is the IA-32 program counter. Alg. 3 follows CFG edges, rather than the value of EIP, because we have no way to give consistent EIP values to new instructions created via decoupling. Alg. 3 does not even track EIP in stores, and consequently, it can propagate the same store to the two successors of a branch instruction (lines 21–23).

4.2.2.2 PE1 Specializer. specialize1 is a specialization function that implements PE1. specialize1 has two constituent functions, eval1 and reduce1. eval1 evaluates instructions, and reduce1 residuates code. The algorithm for specialize1 is given as Alg. 4.

eval1. eval1 can be implemented by writing an interpreter for IA-32 instructions using the concrete operational-semantics of IA-32. eval1 evaluates an instruction with respect to a pre-store, and returns the post-store.

reduce1. reduce1 can be defined as follows:

reduce1 : Instruction → Store → Instruction-sequence

reduce1(I, ρ) = Synth(Subst0(InstrToQFBV(I, ρ))
reduce1 converts an instruction into a QFBV formula, specializes the formula with respect to the pre-store (using Subst0), and uses a machine-code synthesizer to synthesize an instruction sequence that is equivalent to the specialized formula. The synthesized instruction-sequence constitutes the residual code. Recall from §2.4 that vocabulary-0 terms and formulas in a QFBV formula represent pre-store locations and predicates, respectively. Subst0 replaces vocabulary-0 terms and formulas in I’s QFBV formula with static values from ρ. Because lifted instructions are originally static, they access only static locations. Consequently, Subst0 replaces all vocabulary-0 terms and formulas in I’s QFBV formula with static values from the store.

For example, suppose that I is the instruction “add [esp], eax.” I’s QFBV formula ϕ is given below. (For brevity, we only show one of the flags updated by add.)

ϕ ≡ Mem′ = Mem[ESP → Mem(ESP) + EAX] ∧
SF′ = (Mem(ESP) + EAX < 0)

Suppose that ρ is ρ ≡ ⟨[ESP → 1000][EAX → 2], [], [1000 → I]⟩. Subst0 produces the following formula:

Mem′ = Mem[1000 → 3] ∧ SF′ = False,

and reduce1(I, ρ) produces the residual instruction-sequence “mov [1000], 3; add [1000], 0.” (The second instruction in the sequence is necessary to set the SF flag.)
4.2.2.3 **PE₂ specializer.** As discussed in §3, Alg. 4 specializes the binary with respect to static addresses as well as static values (e.g., sum₁ in Fig. 11). In this section, we present the specialization algorithm that WiPER uses to resynchronize code that is independent of the specialization-time memory layout. In particular, we illustrate how the algorithm resynchronizes code that is independent of the specialization-time annotations for the instructions are shown as comments in L[IA-32] by using symbolic constants (m, n, etc.). A different constant is used to represent the starting address of the stack and each heap block allocated at specialization-time.

\[
\text{Int32Id} = \{\text{EAX, ESP, EBP, ..., m, n, ...}\}
\]

**specialize₂** uses symbolic addresses and a symbolic store to access and update memory at specialization time. To represent symbolic addresses, we enhance Int32Id in L[IA-32] by using symbolic constants (m, n, etc.). A different constant is used to represent the starting address of the stack and each heap block allocated at specialization-time.

\[
\text{specialize₂} uses a symbolic store Store^{sym} instead of Store.
\]

\[
\rho^{sym} \in \text{Store}^{sym} = \text{RegMap}^{sym} \times \text{FlagMap}^{sym} \times \text{MemMap}^{sym}
\]

\[
\text{RegMap}^{sym} : \text{Int32Id} \rightarrow \text{Term}, \text{MemMap}^{sym} : \text{Term} \rightarrow \text{Term}, \text{FlagMap}^{sym} : \text{BoolId} \rightarrow \text{Formula}
\]

The residual code generated by **specialize₂** will use specific run-time addresses to access and update memory using an extra level of indirection. To achieve this effect, **specialize₂** uses a memory-layout map \( \mu \) in MemLayout. A memory-layout map \( \mu \) maps the symbolic starting-address \( m \) of the stack or a heap block to a location \( l \) (register or global address) that holds the run-time counterpart of \( m \).

\[
\mu \in \text{MemLayout} : \text{Int32Id} \rightarrow (\text{Register} \cup \text{INT})
\]

We will illustrate **specialize₂**, and its constituent functions eval₂ and reduce₂, using the code snippet shown in Fig. 14, which allocates and uses a heap block. (We describe how **specialize₂** handles the stack and stack locations at the end of this subsection.) The snippet has a static call to malloc at instruction 2, and static values are written to the allocated heap-block at instructions 3 and 4. The binding-time annotations for the instructions are shown as comments on the left-hand column of Fig. 14. Because WiPER lifts all static calls to malloc, instructions 1 and 2 are lifted.

The algorithms for **specialize₂**, eval₂, and reduce₂ are given as Algs. 5, 6, and 7, respectively. In Alg. 5, we overload \( \epsilon \) to denote “no symbolic constant” (lines 4 and 13) and “no location” (line 5), respectively. In our example, because instruction 2 allocates memory (checked by Alloc), **specialize₂** adds \( [m \rightarrow EDI] \) to \( \mu \), where \( m \) is a fresh symbolic constant (obtained using NewConst) that is used to represent the specialization-time starting address of the heap block, and EDI is a location that is not used in the residual binary (obtained using NewLoc). This location can also be a global memory location that is otherwise unused in the residual binary. **specialize₂** passes \( m \) to eval₂, which stores \( m \) in \( \rho \) under the key NextBlock. eval₂ symbolically executes the instruction using \( T^{sym} \), which reinterprets the semantics of an instruction \( I \) to symbolically execute \( I \). If I allocates memory, \( T^{sym} \) uses the symbolic constant bound to NextBlock in \( \rho \) as
Every static instruction that accesses (updates) memory
reduce the next address in the free list. Consequently, in \( \rho' \) returned by \( \text{eval}_2 \) for the call to \text{malloc}, \text{EAX} \) holds the symbolic constant \( m \)—i.e., \( m \) is the starting address of the heap block allocated at instruction 2.

\text{specialize}_2 \) then replaces I, \( \rho, \mu, \) and \( \text{EDI} \) to \( \text{reduce}_2 \). Because instruction 2 allocates memory, \( \text{reduce}_2 \) resuates two instructions. The first instruction (\( z_R^L \)) is the call to \text{malloc}; the second instruction (\( z_R^H \)) saves the \text{EAX} register in \text{EDI} (emitted using \text{EmrMove}). After the residual code executes instruction \( z_R^H \), the run-time starting address of the heap block will be available in \text{EDI}.

\text{specialize}_2 \) symbolically executes the static instruction 3, producing the post-store \( \langle \text{EAX} \mapsto [m], [], [m \mapsto 1] \rangle \). While residuating code for lifted instruction 4, \( \text{reduce}_2 \) uses \text{Subst}_R \) to replace all vocabulary-0 terms and formulas in 4’s QFBV formula with terms and formulas, respectively, from the symbolic store. This step produces the formula
\[
\text{Mem}' = \text{Mem} m + 4 \mapsto 2,
\]
reduce_2 then replaces each symbolic constant \( m \) in the resulting formula with \( [\mu(m)]_L \) using \text{Subst}. For our example, for Eqn. (6), \text{Subst} produces \( \text{Mem}' = \text{Mem} \langle \text{EDI} \mapsto [m + 4, m \mapsto 2] \rangle \). Finally, \( \text{reduce}_2 \) uses the synthesizer to resuate instruction 4_R for the specialized formula. One can see that when the residual code executes, instructions 4_R and 5_R uses the correct starting address of the heap block.

Handling the stack. Suppose that \( n \) is the symbolic constant used to denote the specialization-time base of the stack, and \( l \) is a dedicated location that will hold the address of the base of the stack at run-time. Because the stack is typically allocated at the beginning of the program, \text{WiPER} \) adds \( n \mapsto l \) to \( \mu \) before specialization begins, and emits “move \( l, \text{esp} \)” as the first instruction in the residual program. Additionally, in the initial symbolic store, \text{ESP} \) is bound to \( n \). In our running example from §3, \text{WiPER} \) adds \( n \mapsto \text{ESI} \) to \( \mu \). Consequently, \( \text{reduce}_2 \) uses \text{ESI} \) instead of \( n \) in the QFBV formulas of all downstream lifted instructions.

In summary, to residuate code that is independent of the specialization-time memory layout, \text{specialize}_2 \) uses symbolic addresses, and residuates code that obtains the corresponding run-time addresses from designated locations, using one level of indirection. In effect, the binary is partially evaluated with respect to static values, but not specialization-time addresses.

\text{PE}_2\text{-Safety.} \) For \text{PE}_2 \) to produce correct residual code for a binary \( B, \) \( B \) should be \text{PE}_2\text{-safe}. A binary \( B \), along with a specific division of instructions in \( B \) as static and dynamic, are \text{PE}_2\text{-safe} if they satisfy the following properties:

- **P1** \( B \) does not contain a static branch condition that depends on the starting address of the stack or heap blocks (e.g., \text{cmp esp,1000} \).

- **P2** Every static instruction that accesses (updates) memory in block \( M \) only accesses (updates) memory at a static offset from the base of \( M \), and within the bounds of \( M \).

In Figure 15, we present the extensions to the algorithm to handle binaries with multiple procedures.

**4.3 Interprocedural Partial Evaluation**

In this section, we present the extensions to the algorithm to handle binaries with multiple procedures.

**4.3.1 BTA**

For interprocedural partial-evaluation, \text{WiPER} \) uses a monovariant division, i.e., the classification of each instruction as static/dynamic holds for all calling contexts. This approach to BTA introduces the two issues discussed below, which will be illustrated using the code snippet shown in Fig. 15. Procedure \text{main} \) contains two calls to \text{foo}. The actual parameter of the first call is static, and that of the second call is dynamic. \text{foo} \) merely returns the formal parameter. (For possible future work on polyvariant BTA, see §8.)
Algorithm 8 specializeFn

Input: CFG, ρ, BTAMap, specializedCFGs
Output: ρ′
1: if (CFG, ρ) ∈ specializedCFGs then
2: return ⟨specializedCFGs[(CFG, ρ)], specializedCFGs⟩
3: end if
4: if Recursive(CFG) then
5: specializedCFGs[(CFG, ρ)] = ρ // See explanation in the text
6: end if
7: ⟨ρ′, specializedCFGs⟩ ← IntraPE(CFG, ρ, BTAMap, specializedCFGs)
8: specializedCFGs[(CFG, ρ)] = ρ′
9: return (ρ′, specializedCFGs)

Parameter mismatches. A forward slice can include multiple calls to the same procedure, with different subsets of actual parameters at different call-sites. However, the slice contains the union of the corresponding formal-parameter sets, which causes a mismatch between the actual parameters at a call-site and the procedure’s formal parameters [12, 24].

For the moment, assume that the forward slice used by BTA is context-sensitive. In the code snippet shown in Fig. 15, suppose that the slice includes instructions A2, C2, R2, and the entire body of foo. Note that there is a mismatch between the actual parameter A1, and foo’s formal parameter. (The latter is in the slice, but the latter is not.) WiPER lifts instructions A1 and C1 because they are interprocedural predecessors of foo’s instructions; this step ensures that there will be no parameter mismatches in the residual code.

Return mismatches. If the forward slice used during BTA is context-sensitive, there is also a return mismatch between the return value of foo, and the use of the return value at instruction R1. The former is in the slice, but the latter is not, which violates congruence. (At instruction R1, the specialization expects to find a return value for foo in the static store, but there is no value.) To prevent return mismatches, WiPER uses context-insensitive slicing, which causes instruction R1 to be included in the slice.

In summary, for interprocedural BTA, WiPER (i) lifts interprocedural static-predecessors of dynamic instructions, and (ii) uses context-insensitive forward slicing. In addition, to ensure that partial evaluation always terminates, WiPER conservatively classifies all instructions in a recursive procedure as dynamic. (This point is discussed further in §4.3.2.)

4.3.2 Specialization

To perform specialization for interprocedural partial-evaluation, the following lines are added after line [16] of Alg. 3. IsCall returns true if the argument is a call instruction; Callee returns the callee CFG for a call instruction.

In addition, Alg. 3 takes an additional argument, specializedCFGs, which is a map whose entries are of the form (CFG, ρ) → ρ′, where ρ is a partial static pre-store and ρ′ is a partial static post-store. Alg. 3 also returns specializedCFGs as an additional return value. Procedure SpecializeFn is given as Alg. 8. Alg. 8 implements function caching—if CFG has already been specialized with respect to ρ, Alg. 8 returns ρ′ (lines 1–3).

To ensure that partial evaluation always terminates, WiPER does not attempt to specialize recursive functions. Recall from §4.3.1 that all instructions in a recursive procedure are classified as dynamic (and the binding-time classification computed by BTA is congruent with respect to that classification). Thus, for a recursive procedure P, the correct values for all variables that are classified static at the return from P are found in the input partial static pre-store ρ. Consequently, Alg. 8 can use ρ as the partial static post-store in the entry added to specializedCFGs before IntraPE is called (lines 4–6). This trick ensures that Alg. 8 cannot get into an infinite loop while specializing a recursive function.

If a function has not already been specialized, Alg. 8 calls Alg. 3 to specialize the CFG with respect to ρ, and adds the partial static post-store to the map (lines 7–8).

4.4 Correctness

In this section, we present two theorems concerning: (i) termination of WiPER, and (ii) correctness of the residual code produced by WiPER. The first theorem applies to binaries with instructions from the entire IA-32 instruction set. Because it would be difficult to prove the correctness theorem for the entire IA-32 ISA, we prove correctness only for binaries containing instructions from a small instruction subset. (See App. A.1.)

Theorem 1. Suppose that B is a PE2-safe binary, and ρ′sym is a partial static symbolic-store. Then WiPER(B, ρ′sym) terminates.

Proof. WiPER treats recursive functions conservatively, so WiPER cannot enter into an infinite loop while attempting to specialize a recursive function. WiPER unrolls static loops finitely many times depending upon the static loop bound. Moreover, the forward slice performed during BTA follows both data and control dependences, which rules out the possibility of there being a static instruction that is control dependent on a dynamic branch. Consequently, WiPER cannot create infinitely many specialized versions of the same basic block. Thus WiPER(B, ρ′sym) is guaranteed to terminate.

In a standard semantics that formalizes the behavior of an IA-32 processor, an evaluation function would take an IA-32 Store as an input, and produce an IA-32 Store as an output. In §3, we introduced the notion of a state σ consisting of an environment τ and an IA-32 store ρ. In this section, we use a non-standard semantics that uses an environment along with an IA-32 store. The environment is really a ghost variable.
that makes explicit the starting address of the stack and heap blocks, while the concrete part of the state is the store.

Based on the congruent division established by BTA, a state can be partitioned into \( \eta_S, \rho_S, \eta_D, \) and \( \rho_D \). Thus the type for the evaluation function \([\cdot]\) (the function that evaluates an instruction or a sequence of instructions with respect to a state, and returns the post-state) can be written as

\[
[\cdot] : \text{Instruction} \times \text{Env}_S \times \text{Store}_S \times \text{Env}_D \times \text{Store}_D \rightarrow \text{Env}_S \times \text{Store}_S \times \text{Env}_D \times \text{Store}_D.
\]

If we execute an instruction in a partially evaluated binary, \( \rho_S \) is never accessed or updated. Thus we will overload \([\cdot]\) to also denote the meaning function for instructions in a partially evaluated binary:

\[
[\cdot] : \text{Instruction} \times \text{Env}_S \times \text{Env}_D \times \text{Store}_D \rightarrow \text{Env}_S \times \text{Env}_D \times \text{Store}_D
\]

It will always be clear from context which meaning of \([\cdot]\) is intended.

We assume that the memory map for a program that has been partially evaluated is equipped with a special area in which the starting addresses of the stack and heap blocks allocated by lifted instructions are stored. Thus, strictly speaking, the memory maps for a binary \( B \) and some residual binary \( B' \) of \( B \) will be different. However, the only differences are in the special area, and we denote equality of two stores \( \rho \) and \( \rho' \) modulo the special area by \( \rho \approx \rho' \).

For an input binary \( B \) that is \( \text{PE}_2 \)-safe, and an IA-32 state \( \sigma \) that can be partitioned into \( \eta_S, \rho_S, \eta_D, \) and \( \rho_D \) based on a division of inputs to \( B \), \( \text{WiPER} \) produces a specialized binary that, when executed, produces an answer that matches the answer that would be produced by \( B \) (modulo the special area). The precise statement of the property is given in the following theorem:

**Theorem 2.** Suppose that \( B \) is \( \text{PE}_2 \)-safe. Also suppose that \( \eta_S \) is an environment such that there are no overlaps among (a) any of the chunks in \( \text{Dom}(\eta_S) \), (b) the chunk in which the residual code produced by \( \text{WiPER} \) is loaded, and (c) any of the chunks allocated during the execution of the residual code. For all \( \sigma = (\eta_S, \rho_S, \eta_D, \rho_D) \) such that \( [B](\eta_S, \rho_S, \eta_D, \rho_D) \) terminates, suppose that \( [B](\eta_S, \rho_S, \eta_D, \rho_D) = (\eta_S', \rho_S', \eta_D', \rho_D') \).

If \( \rho_S^{\text{sym}} \) is a symbolic store such that \( \|\rho_S^{\text{sym}}\|_{\eta_S} = \rho_S \), then \( \|\text{WiPER}(B, \rho_S^{\text{sym}})\|_{(\eta_S', \eta_D', \rho_D')} = (\eta_S', \rho_S', \eta_D', \rho_D') \), where \( \rho_D' \approx \rho_D' \).

**Proof.** The proof for Thm. 2 is given in App. A.1. \( \square \)

Thm. 2 does not include \( \rho_S \) while comparing states because the residual code produced by \( \text{WiPER}(B, \rho_S^{\text{sym}}) \) does not contain static instructions, and consequently never accesses or updates the static store \( \rho_S \).

### 4.5 Threats to Validity

There are two threats to the validity of our algorithms.

1. If the input binary is not \( \text{PE}_2 \)-safe, the residual code might violate Thm. 2.

2. The accuracy of BTA depends on the accuracy of the SDG for the binary. If the methods used to construct the SDG are overly conservative, BTA can classify instructions as “dynamic” that do not depend on dynamic inputs. If the methods used to construct the SDG are underapproximative, BTA can classify instructions as “static” that actually do depend on dynamic inputs. In our experiments, we observed the former behavior. The SDG that \( \text{WiPER} \) uses is built using “variable-like abstractions” recovered by machine-code variable-recovery analyses [9]. Each recovered variable-like abstraction could be imprecise—it is often an agglomeration of some source-code variables. Consequently, the SDG recovered from the program binary could be more imprecise than an SDG that is obtained from the program’s source code. As a result, \( \text{WiPER} \)’s BTA classifies more instructions dynamic than an analogous BTA that starts from source code.

### 5. Implementation

\( \text{WiPER} \) uses CodeSurfer/x86 [10] to obtain the SDG, CFG, and call graph for the binary. \( \text{WiPER} \)’s BTA uses CodeSurfer/x86’s forward-slicing operation. \( \text{WiPER} \)’s specialization uses CodeSurfer/x86’s rewriting API to create the sections, CFGs, and instructions in the residual binary. \( \text{WiPER} \) uses Transformer Specification Language (TSL) [29] to build its concrete and symbolic interpreters. The concrete operational semantics of the integer subset of IA-32 is written in TSL, and the semantics is interpreted for concrete evaluation, and reinterpreted for symbolic evaluation. \( \text{WiPER} \) also uses TSL [30] to convert an instruction into a QFBV formula. \( \text{WiPER} \) uses \( \text{McSYNTH} \) [46] to synthesize an IA-32 instruction sequence from a QFBV formula. \( \text{McSYNTH} \) sometimes requires a set of scratch registers to hold results of intermediate calculations in the residual code. \( \text{WiPER} \) supplies dead registers to \( \text{McSYNTH} \) to be used as scratch registers. If the number of dead registers at a point is not sufficient, \( \text{WiPER} \) supplies global addresses that are unused in the residual program as scratch locations to \( \text{McSYNTH} \).

In the examples presented in this paper, we have treated memory as if each memory location holds a 32-bit integer. However, our implementation supports the actual IA-32 memory model, in which each memory location holds an 8-bit integer.

The compiler might add instructions at the beginning of each procedure so that stack variables are aligned on 4-byte, 8-byte, or 16-byte boundaries. For example, consider the code snippet shown in Fig. 16. In procedures \( \text{foo} \) and \( \text{main} \), the compiler aligns the stack variables on a 16-byte boundary. To accommodate such alignment, we relax P2 of...
Table 1: Characteristics of applications for optimization via specialization.

<table>
<thead>
<tr>
<th>Application</th>
<th>Domain</th>
<th>LOC</th>
<th>Description</th>
<th>Static Input</th>
</tr>
</thead>
<tbody>
<tr>
<td>Static Input</td>
<td>Mathematical library functions</td>
<td>19</td>
<td>Computes $x^n$</td>
<td>$n = 100$</td>
</tr>
<tr>
<td>Dotproducl</td>
<td>Linear algebra operations</td>
<td>29</td>
<td>Computes the dot product of two vectors with $n$ dimensions</td>
<td>$n = 3$, $n = 3$, and elements of the filter</td>
</tr>
<tr>
<td>Interpreter</td>
<td>Language interpreters</td>
<td>71</td>
<td>Interpreter for the minimalist language “Brail*ck”</td>
<td>Input program</td>
</tr>
<tr>
<td>Filter</td>
<td>Image processing</td>
<td>107</td>
<td>Applies a convolution filter of size $m \times m$ on an image of size $n \times n$</td>
<td>$n = 256$, and the string</td>
</tr>
<tr>
<td>uuencode</td>
<td>Text utilities</td>
<td>129</td>
<td>Base-64 encodes a string of size $n$</td>
<td>$n = 1024$, and contents of the first 512 bits</td>
</tr>
<tr>
<td>SHA1</td>
<td>Cryptographic operations</td>
<td>140</td>
<td>Computes the sha1 digest of a message of size $n$ bits</td>
<td>$n = 256$, and the encoded string</td>
</tr>
<tr>
<td>uuencode</td>
<td>Text utilities</td>
<td>167</td>
<td>Decodes a base-64-encoded string of size $n$</td>
<td>$n = 256$, and the encoded string</td>
</tr>
</tbody>
</table>

```c
foo:
  main:
    push ebp
    mov ebp,esp
    and esp,0xFFFFFFFF00
    sub esp,170
    sub esp,150

Figure 16: Code snippet to illustrate variable alignment.
```

PE2-safety properties to allow terms of the form $m + p_0 + p_1 + \ldots + c$, where $m$ is a symbolic constant representing the specialization-time starting address of the stack or a heap block $M$, $c$ is an integer constant, and $p_0, p_1, \ldots$ etc. are symbolic constants that represent the alignment adjustments performed by the procedures Proc0, Proc1, \ldots on the current call stack.

If the binary is statically linked, WiPER does not need any additional input from the user for partial evaluation. If the binary is not statically linked, the user needs to provide a model for each library function as additional inputs to WiPER. A model consists of the following information: (i) an input-output dependence relation, and (ii) a procedure to compute return values of library-function calls that are classified as static. WiPER specializes library-function calls in a bimodal manner: if any argument to a library-function call is dynamic, all of the static arguments are lifted along with the call; if all arguments are static, WiPER uses the user-supplied model to compute a return value for the call (i.e., a fully static call is specialized away.)

We are generally unable to use WiPER’s fully automated end-to-end partial-evaluation algorithm to extract components from a large binary. Binaries of large applications do not have clearly demarcated inputs. They often have only one input—a buffer that stores the entire command line. The binary interprets the contents of the buffer to assign values to variables. If the buffer is classified dynamic, Alg. 2 classifies all instructions in the binary dynamic.

For component extraction, we run WiPER in an alternative experimental mode, called Component Extraction (CE) mode. In CE-mode, WiPER performs a restricted BTA. (The specialization algorithm is the same in CE and non-CE modes.) In CE-mode, the user specifies a subset $I_5$ of instructions in the original binary to be treated as static instructions. The remaining instructions in the binary are treated as dynamic instructions. WiPER identifies lifted instructions, and then specializes the binary using WiPER’s original specialization algorithm.

CE-mode is related to generalized partial computation [22], in which a program is specialized with respect to con-

Figure 17: Comparison of the execution times of the original and specialized binaries.

straints. As will become clear below, in CE-mode we are specializing a program $P$ with respect to a condition, such as $\varphi \equiv (EAX = 1)$, which must hold at a given point $p$ in the middle of the program. The goal is to produce a residual program that is (i) a specialization of $P$, in which (ii) $\varphi$ holds each time the points residuated from $p$ are reached.

For simple examples of the kind discussed below, CE-mode successfully extracts components, even though congruence is violated. It remains for future work to devise a theory of program specialization that covers CE-mode.

6. Experiments

The goal of our experiments was to test and evaluate the two primary use cases of WiPER: (i) optimization via specialization, and (ii) component extraction. Our experiments were run on a system with a 64-core, 2.3 GHz AMD Opteron 6376 processor; however, WiPER’s algorithm is single-threaded. The system has 264 GB of memory and runs RHEL 6.6.

6.1 Optimization via Specialization

For this use case, we wanted to answer the following question:

- What is the speedup produced by partial evaluation for different applications?

For the first set of experiments, we used a suite of applications from application domains that were suitable for partial evaluation. Table 1 presents the characteristics of the applications. The first two applications listed in the table have been used in prior work [34, 41]. There are four versions of the interpreter application in our test suite: (i) a standard version, (ii) a version that uses indirect jumps (via switch statements), (ii) a version that uses indirect calls (via function pointers), and (iv) a version that uses heap-allocated storage. uuencode and uudecode are FreeBSD utilities [3]. We used CE-mode for partially evaluating the FreeBSD utilities. The
results are shown in Fig. 17. For these applications, the average speedup produced by partial evaluation, computed via the geometric mean, is 1.3.

For Power, partial evaluation results in a slight slowdown. Power has a tight loop, and we believe that the aggressive loop unrolling caused by partial evaluation disrupts instruction caching.

At the other end of the spectrum, we measured a speedup in Filter of 1.9. Filter has four nested for-loops. The inner two loops iterate over the filter elements. Partial evaluation unrolls all four loops, and—because the filter elements are static—replaces almost all instructions in the bodies of the inner two loops with just a few residual instructions.

### 6.2 Component Extraction

For this use case, we wanted to answer the following questions:

- How can we extract an executable component from a binary via partial evaluation?
- How does the size of the extracted component compare to the size of the original binary?

For this set of experiments, we used three applications that bundle several components into a bloated executable. Table 2 presents the characteristics of the applications. The first and second applications were obtained from Google Code [1] and SourceForge [2], respectively.

CE-mode of WiPER was used for this set of experiments. The sizes of the original binary and the extracted components are shown in Table 3.

### 6.2.1 Component Extraction

<table>
<thead>
<tr>
<th>Application</th>
<th>Size of binary (KB)</th>
<th>Size of component (KB)</th>
<th>No. of procedures in binary</th>
<th>No. of procedures in component</th>
</tr>
</thead>
<tbody>
<tr>
<td>b64</td>
<td>3.4</td>
<td>5.9</td>
<td>16</td>
<td>2</td>
</tr>
<tr>
<td>lzfx</td>
<td>7.6</td>
<td>6</td>
<td>24</td>
<td>8</td>
</tr>
<tr>
<td>bzip2</td>
<td>90.2</td>
<td>64.6</td>
<td>117</td>
<td>78</td>
</tr>
</tbody>
</table>

**Table 2: Characteristics of applications for component extraction.**

**Table 3: Comparison of sizes and number of procedures in original binary and extracted component.**

For this use case, we wanted to answer the following questions:

- How can we extract an executable component from a binary via partial evaluation?
- How does the size of the extracted component compare to the size of the original binary?

For this set of experiments, we used three applications that bundle several components into a bloated executable. Table 2 presents the characteristics of the applications. The first and second applications were obtained from Google Code [1] and SourceForge [2], respectively.

CE-mode of WiPER was used for this set of experiments. The sizes of the original binary and the extracted components are shown in Table 3.

#### b64

The decoder extracted from b64 is bigger than the b64 binary because our current implementation does not optimize the layout of basic blocks in the residual code to reduce the number of jump instructions. For example, b64 has 30 jmp instructions, and the extracted decoder has 164 jmp instructions. The issues in partially evaluating b64 are similar to those of bzip2, which is discussed below.

#### lzfx

lzfx is a text-compression utility. In this case study, we present how WiPER extracts the compression component of lzfx. The decision about whether to call `fx_create` (which compresses) or `fx_read` (which decompresses) is made in the following code snippet from the binary:

```asm
_text:08049A3A mov eax,dword [esp+44] ;L
_text:08049A3E mov dword [esp+12],eax ;D
_text:08049A5E call fx_create ;D
loc_80499E2:
_text:08049A84 call fx_read; D
```

If the value in the eax register is 0 at instruction 80499D4, lzfx calls `fx_create`. The user specifies that instructions 80499D4 through 8049A3A should be treated as static instructions. WiPER lifts instruction 8049A3A because it supplies a static value to the dynamic instruction 8049A3E. WiPER performs specialization and evaluates the static and lifted instructions with respect to the partial store

\[ \rho \equiv \langle [EAX \mapsto 0][ESP \mapsto n], [[, [, ]] \rangle, \]

and residues the instruction `mov eax,0` for the lifted instruction 8049A3A, and emits the remaining dynamic instructions. Because `eax is 0` at instruction 80499D4, the fallback branch is taken at 80499D6, which forces WiPER to enter `fx_create` and to bypass `fx_read`. Consequently, WiPER residuates an executable that only performs compression.

With lzfx, we also went further by embedding the extracted compression component in a different application. We created a small header file with the signature of `fx_create`, and linked the new application with the compression component extracted by WiPER.

#### bzip2

bzip2 is a large application that takes several auxiliary command-line inputs in addition to one that specifies whether to perform compression or decompression. Ultimately, the decision about whether to call compress or uncompress is made in the following code snippet of the binary:

```asm
_text:0805B539 mov eax,[esp+40] ;S
_text:0805B53E cmp eax,1 ;S
_text:0805B541 jnz loc_805B55D ;S
_text:0805B558 call compress ;D
loc_805B55B:
_text:0805B55E4 call uncompress ;D
```

The memory location whose address is `ESP + 40` holds a value computed from the compression/decompression flag on the command line. If the value is 1, bzip2 calls compress; otherwise, it calls uncompress.

The user specifies that instructions 805B539 through 805B541 should be treated as static instructions. WiPER performs specialization and evaluates the aforementioned instructions with respect to the partial store

\[ \rho \equiv \langle [ESP \mapsto n], [[, [n + 40 \mapsto 1]] \rangle. \]

This partial store forces WiPER to enter `compress` and bypass `uncompress`, and consequently WiPER residuates an executable that only performs compression.
7. Related Work

**Partial evaluation.** Partial evaluation has been used as a general framework for the specialization of programs belonging to different languages, and different domains. In addition to several functional languages [25, 32], partial evaluation has been used to specialize a flow-chart language [25], C [6, 15], FORTRAN [26], and Java [40, 41]. Partial evaluators can either be on-line (performed in a single phase) [23], off-line (performed in two phases), or hybrid [41]. WiPER’s high-level architecture resembles that of an off-line partial evaluator for an imperative language. However, the need to handle several machine-code-specific issues makes WiPER’s algorithm significantly different from those of source-code partial evaluators.

**Run-time specialization techniques.** Run-time specialization (or dynamic compilation) generates optimized code during program execution by partially evaluating user-annotated regions of the program with respect to invariant data computed at run time [8, 14, 19, 28].

Like WiPER, these tools perform specialization on machine code. However, the partial-evaluator-like component in such tools is part of the compiler. BTA is performed with respect to source-code variables, and carried out on an IR that is constructed from source code. Run-time-specialization tools have to address an issue that does not arise in WiPER, namely, the need to create code templates that can be reused at run-time. Some tools accomplish this task by jury-rigging an existing compiler’s code generator to inhibit optimizations, such as code motion, by using the volatile type qualifier of C.

In contrast, WiPER faced other issues, such as the need to decouple multiple updates performed by a single instruction, and to create code that allows the stack and heap to be at different positions at run-time than at specialization-time.

**Specialization of Java bytecode.** WiPER is not the first partial evaluator that works on low-level code. Lancet [37] performs partial evaluation of Java bytecode snippets that need to be compiled by a just-in-time (JIT) compiler. JScp [27] performs supercompilation on Java bytecode. Like WiPER, the specializers in Lancet and JScp make use of a static partial store. For Lancet and JScp, the static partial store is a Java virtual-machine (JVM) store, while WiPER’s is an IA-32 store. However, while those specializers work on low-level code, WiPER works at an “even lower level.” The following issues that arise in IA-32 and not in bytecode justify the novel design choices made in WiPER:

1. Because the JVM is a stack-based machine, specialization of bytecode instructions often involves loading constant values instead of locals/fields. In contrast, the operands of IA-32 instructions vary widely, and the specialization of an instruction with respect to a partial store need not be a variant of the instruction. Moreover, IA-32 has around 43,000 unique opcodes, and WiPER’s specializer must be able to specialize any instruction with respect to any partial store. WiPER addressed this issue by using machine-code synthesis.

2. Address computation is not explicit in bytecode (variables are referenced using offsets). Consequently, bytecode specialization does not face the problem of residuating specialization-time addresses. However in machine code, instructions compute addresses, and a naïve specializer, such as PE, is unsatisfactory. For this reason, WiPER uses symbolic techniques for specialization.

**Superoptimization and link-time optimization.** Superoptimization aims at finding an optimal instruction-sequence for a target instruction-sequence [11, 31, 39]. Peephole superoptimization [11] uses “peepholes” to harvest target instruction-sequences, and replace them with equivalent instruction-sequences that have a lower cost. Link-time optimizers carry out whole-program optimizations at link time [33]. While these optimization techniques optimize a binary for all possible inputs, WiPER optimizes a binary for a specific input.

8. Conclusion and Future Work

In this paper, we presented an algorithm for machine-code partial evaluation. We presented WiPER, a partial evaluator for IA-32. We used WiPER to partially evaluate the binaries of seven test applications with respect to static inputs, and obtained specialized binaries that had an average speedup of 1.3 (computed via the geometric mean). We also presented two case studies that describe how WiPER can be used to extract an executable component from a bloated binary.

One possible direction for future work is to investigate polyvariant BTA via specialization slicing [7]. For every call to a procedure P with a different combination of static and dynamic actual parameters, specialization slicing can be used to produce a different binding-time annotation for P’s instructions. This technique could increase the number of static instructions during specialization, leading to residual code that is more specialized than the code currently residuated by WiPER. A second direction is to use liveness information to reduce the number of specialized copies of a basic block produced by WiPER.

**Acknowledgments**

We thank Suan Yong, Junghee Lim, Tushar Sharma, and Brian Alliet for answering several questions during the implementation of WiPER. We thank Suan Yong for implementing a skeleton binary rewriter using CodeSurfer/x86. We thank the anonymous reviewers for their valuable feedback.

**References**

Partial Evaluation and Automatic Program Generation

C. Consel and F. Noël. A general approach for run-time specialization and its application to C. In POPL, 1996.


A. Appendix

Theorem 2. Suppose that $B$ is PE$_2$-safe. Also suppose that $\eta_S$ is an environment such that there are no overlaps among (a) any of the chunks in Dom($\eta_S$), (b) the chunk in which the residual code produced by $\text{WIPER}$ is loaded, and (c) any of the chunks allocated during the execution of the residual code. For all $\sigma = (\eta_S, \rho_S, \eta_D, \rho_D)$ such that $[B](\eta_S, \rho_S, \eta_D, \rho_D)$ terminates, suppose that $[B](\eta_S, \rho_S, \eta_D, \rho_D) = (\eta_S', \rho_S', \eta_D', \rho_D')$. If $\rho_S' = \text{sym}$ is a symbolic store such that $[\rho_S' \eta_S](\eta_S) = \rho_S$, then $[\text{WIPER}(B, \rho_S')](\eta_S, \rho_D) = (\eta_S', \rho_D')$, where $\rho_D' \approx \rho_D'$.

A.1 Proof of Thm. 2

We prove Thm. 2 for binaries whose instructions come from a small instruction set, TinyIA32. TinyIA32 consists of a small subset of the IA-32 instruction set, augmented with (i) an extra register, and (ii) two new instructions. We decided to exclude instructions related to branches and function calls from TinyIA32 because if we were to prove Thm. 2 for the entire IA-32 ISA, the part of the proof concerning branches and function calls would be similar to that of an analogous proof for a source-code partial evaluator. We prove Thm. 2 using induction on the length of the trace of a run of the binary, and it is straightforward to extend our trace-based proof to an ISA that includes branches and function calls.

TinyIA32 has three registers, no flags, and seven instructions. The abstract syntax of TinyIA32 is defined in Fig. 18. In an indirect operand, the register holds the address of the memory location. For instructions with two operands, the operand on the left is the destination operand. For the
We assume the following: 

\[ \text{we assume the following:} \]

block to a concrete address.) As the induction hypothesis, 

\[ \text{as the induction hypothesis,} \]

that SubjThm. 2 holds on a partial trace 

\[ \text{that Thm. 2 holds on a partial trace} \]

\[ T \]

\[ A.1.2 \text{ Induction Hypothesis} \]

\[ \text{Let us assume that Thm. 2 holds on a partial trace} \]

\[ T \]

\[ A.1.3 \text{ Induction Step} \]

\[ \text{We now prove Thm. 2 on a partial trace} \]

\[ I_1, I_2, \ldots, I_{n+1} \]

\[ \text{with} \]

\[ n+1 \text{ instructions. That is, the goal is to show} \]

\[ \text{We prove Thm. 2 by induction on the length} \]

\[ n \]

\[ \text{of} \]

\[ \text{A.1.1 Base Case} \]

\[ n = 0 \]

\[ \text{Thm. 2 trivially holds on a partial trace with no instructions.} \]

\[ \text{A.1.2 Induction Hypothesis} \]

\[ \text{Let us assume that Thm. 2 holds on a partial trace} \]

\[ T \]

\[ I_1, I_2, \ldots, I_n \]

\[ \text{with} \]

\[ n \text{ instructions. Suppose that the corresponding residual trace produced by} \]

\[ \text{WiPER for} \]

\[ T' \]

\[ I_1, I_2, \ldots, I_n \]

\[ \text{where} \]

\[ I_i \]

\[ \text{is the residual instruction sequence produced by the} \]

\[ \text{Specialize phase of} \]

\[ \text{WiPER for} \]

\[ I_i \]

\[ \text{(Note that} \]

\[ J_i \]

\[ \text{can be an empty sequence of instructions or can consist of several instructions.) Suppose that the initial state is} \]

\[ (\eta_S, \rho_S, \eta_D, \rho_D) \]

\[ \text{(Recall from §3 that} \]

\[ \eta \]

\[ \text{maps a symbolic constant denoting the starting address of the stack or a heap block to a concrete address.) As the induction hypothesis, we assume the following:} \]

\[ \begin{align*}
[I_1, I_2, \ldots, I_n](\eta_S, \rho_S, \eta_D, \rho_D) & = (\eta_S^{n+1}, \rho_S^{n+1}, \eta_D^{n+1}, \rho_D^{n+1}) \\
\wedge [J_1, J_2, \ldots, J_{n+1}](\eta_S, \rho_S, \eta_D, \rho_D) & = (\eta_S^{n+1}, \rho_S^{n+1}, \eta_D^{n+1}, \rho_D^{n+1}) \quad (\text{Induction Hypothesis}) \\
\wedge \rho_D^{n+1} & \approx \rho_D^{n+1}
\end{align*} \]

\[ \text{To account for the actions of} \]

\[ \text{WiPER on static and lifted instructions, we need to introduce yet one more ghost variable,} \]

\[ \text{ghostStaticStore} (gSS) \]

\[ \text{At each point} \]

\[ k \]

\[ \text{in the trace} \]

\[ T' \]

\[ \text{this variable holds the static symbolic store that was used by} \]

\[ \text{WiPER to create the next residual instruction-sequence in} \]

\[ T' \]

\[ \text{i.e.,} \]

\[ J_{k+1} = \text{Specialize}(J_{k+1}, gSS^k) \]

\[ \text{We strengthen the induction hypothesis with the following conjunct:} \]

\[ \begin{align*}
gSS^n(\eta_S) & = \rho_S^n, \\
\end{align*} \]

\[ \text{where the left-hand side denotes the simplification of each symbol} \]

\[ \text{ic expression in} \]

\[ gSS \]

\[ \text{with respect to the values obtained from} \]

\[ \eta_S^n. \]
A.1.3.2 Lifted and dynamic instructions. We now make an observation that is relied on in many of the cases—namely, because static-allocation sites are lifted, if instruction \( I_i \) in \( T \) is an allocation instruction, then \( J_i \) in \( T' \) also contains an allocation instruction, i.e., the allocations in the two traces are in lock-step. For purposes of the proof, we can assume that \( B \) and \( \text{WiPER}(B, \rho_S) \) are run in environments that will perform storage allocation in the same way. Consequently, if \( I_i \) allocates the stack or a heap block \( M \) at address \( m \), an instruction in \( J_i \) also allocates \( M \) at \( m \). In the respective semantics, such allocations will be recorded by updating \( \eta \) to \( \eta(m \mapsto m) \), where \( m \) is a symbolic constant that denotes the starting address of \( M \). (Lifted allocation instructions update \( \eta_S \) and \( \eta_D \); dynamic allocation instructions update \( \eta_D \).)

**Dynamic instructions.** The argument for each of the dynamic-instruction cases for \( I_{n+1} \), except `stackalloc` and `heapalloc`, has the following form: Suppose that instruction \( I_{n+1} \) is dynamic. Then

\[
[I_{n+1}](\eta^n_S, \rho^n_S, \eta^n_D, \rho^n_D) = (\eta^n_S, \rho^n_S, \eta^n_D, \rho^n_D+1),
\]

where some update is made to \( \rho^n_D \) to create \( \rho^n_D+1 \).

Specialize \( I_{n+1} \), \( gSS^n \) \( \equiv J_{n+1} = I_{n+1} \)

Because \( J_{n+1} \) is the same as \( I_{n+1} \), we have

\[
[I_{n+1}](\eta^n_S, \rho^n_D, \eta^n_D, \rho^n_D+1) = (\eta^n_S, \rho^n_D, \eta^n_D, \rho^n_D+1),
\]

where \( \rho^n_D+1 \approx \rho^n_D+1 \). Eqsns. (13) and (14), together with Eqn. (7), show that Eqn. (9) holds.

The static components of the state, \( \eta^n_S \), \( gSS^n \), and \( \rho^n_S \), are left unchanged, so Eqn. (10) holds.

The cases for `stackalloc` and `heapalloc` are similar, except that those instructions update \( \rho^n_D \) in addition to \( \rho^n_D \). Moreover, because of our assumption about the storage allocator, the updates to \( \eta^n_D \) by \( [I_{n+1}] \) and \( [J_{n+1}] \) are identical, and the updates to \( \rho^n_D \) by \( [I_{n+1}] \) and \( [J_{n+1}] \) are identical. Consequently, the induction hypothesis holds on a trace with \( n+1 \) instructions, where the \( (n+1)^{\text{st}} \) instruction is dynamic.

**Lifted instructions.** The semantics we use for evaluating static and dynamic instructions is fairly standard—static instructions access and update \( \rho_S \), and dynamic instructions access and update \( \rho_D \) (in addition to updating \( \eta_D \)). However, the evaluation of lifted instructions is non-standard: lifted instructions were originally classified static, so they must be evaluated just like static instructions; however, they also take “snapshots” of the current static state and dump them into \( \eta_D \) and \( \rho_D \) for use by downstream dynamic instructions.

Lifted instructions access \( \rho_S \), but produce values that might be consumed by dynamic successors, so they must update \( \rho_D \). In addition, lifted instructions can also have static successors, so they must also update \( \rho_S \). Thus, in the semantics used for the original program, a lifted instruction performs the same update on both \( \rho_S \) and \( \rho_D \). In the semantics used for the residual program, it updates only \( \rho_D \) because \( \rho_S \) is not present in states. However, for the purposes of this proof, an appropriate update is performed on \( gSS \).

Similarly, if a lifted instruction allocates memory, both static and dynamic instructions downstream might access or update locations in the allocated memory. Thus in both semantics, if a lifted instruction allocates memory, it performs identical updates on \( \eta_S \) and \( \eta_D \).

The non-overlap hypothesis of Thm. 2, together with property P2 of \( \text{PE}_2 \)-safety, imply that each specialization-time address can be represented canonically by a term of the form \( m + c \), where \( m \) is the symbolic constant that denotes the starting address of the allocated stack or heap block \( M \), and \( c \) is a constant offset.

The actual symbolic constants used are essentially arbitrary (as long as each new symbolic constant is fresh). However, for the purposes of the proof, we need to relate the symbolic constants used by the original program to those in the residual program. Thus, for the purposes of the proof, we assume that when \( \text{WiPER} \) partially evaluates a lifted instruction \( I_i \) in \( T \) that allocates memory, the symbolic constant \( m \) that should be used in \( gSS \) to denote the starting address of the allocated stack or heap block is supplied by an oracle. This oracle ensures that the same symbolic constant \( m \) gets added to the ghost variables \( \eta_S \) and \( \eta_D \) when \( I_i \) and the allocation instruction in \( J_i \) are evaluated. (If the oracle did not ensure this concordance, we cannot establish the equality in Eqn. (8).) \( \text{WiPER} \) binds \( m \) to a dedicated global address \( M-\text{addr} \), and uses \( M-\text{addr} \) in the residual code to save the concrete starting address \( m \) of the allocated memory. \( M-\text{addr} \) is located in the special area in the memory map of the residual program discussed in §4.4. Recall from §4.2.2.3 that \( \text{WiPER} \) uses a memory-layout map \( \mu \) to bind a symbolic constant to a dedicated location.)

We handle the lifted-instruction cases according to the form of the instruction. In the cases below, we abbreviate an update to the register map as \( \rho[R1 \mapsto v] \) rather than let \( (rm, mm) = \rho \text{ in } (rm[R1 \mapsto v], mm) \), and similarly use \( \rho[a \mapsto v] \), to denote an update to the memory map at address \( a \). Which kind of map-update operation is intended should be clear from context. We present the most important cases below. The remaining cases can be found in an extended version of the paper [45].

1. \( I_{n+1} = \text{mov} \ (\text{Direct}(R1), \text{Imm}(v)) \):

\[
[I_{n+1}](\eta^n_S, \rho^n_S, \eta^n_D, \rho^n_D) = (\eta^n_S, \rho^n_S, \eta^n_D, \rho^n_D) \]

\[
[I_{n+1}](\eta^n_S, \rho^n_D, \eta^n_D, \rho^n_D+1) = (\eta^n_S, \rho^n_D, \eta^n_D, \rho^n_D+1)
\]

Specialize \( I_{n+1} \), \( gSS^n \) \( \equiv J_{n+1} = \text{mov} \ r1, v \)

\[
[I_{n+1}](\eta^n_S, \rho^n_D, \eta^n_D, \rho^n_D+1) = (\eta^n_S, \eta^n_D, \rho^n_D \ R1 \mapsto v])
\]

Eqns. (15), (16), and (7) show that Eqn. (9) holds.

\( gSS^n+1 = gSS^n \ R1 \mapsto v \)

\( \eta^n_S+1 = \eta^n_S \) and \( \rho^n_D+1 = \rho^n_D \) (From (15))
Eqns. (17), (18), and (8) show that Eqn. (10) holds.

2. $I_{n+1} = \text{mov} (\text{Direct}(R1), \text{Direct}(R2))$: There are two cases that can arise.

(a) Suppose that $\rho_2^S(R2) = u$ and $gSS^n(R2) = v$.

\[
[I_{n+1}][\eta_S^a, \rho_2^S, \eta_D^a, \rho_D^a] = \\
(\eta_S^u, \rho_2^S[R1 \mapsto u], \eta_D^u, \rho_D^u[R1 \mapsto u])
\]

specialize($I_{n+1}, gSS^n) = \{J_{n+1} = \text{mov} \ r1, v \}$

\[
[J_{n+1}][\eta_S^u, \eta_D^u, \rho_D^u] = (\eta_S^n, \eta_D^n, \rho_D^n[R1 \mapsto v])
\]

\[
u = \rho_2^S(R2) = ([gSS^n]^2)(R2) \quad \text{(By Eqn. (8))}
\]

\[= ([gSS^n(R2)][\eta_S^n] = [v] \eta_S^n = v \tag{21}
\]

Eqns. (19)–(21), and (7) show that Eqn. (9) holds.

\[
gSS^{n+1} = gSS^n[R1 \mapsto v] \tag{22}
\]

\[
\eta_S^{n+1} = \eta_S^n \text{ and } \rho_S^{n+1} = \rho_S^n[R1 \mapsto u] \quad \text{(From (19))}
\]

Eqns. (21)–(23), and (8) show that Eqn. (10) holds.

(b) Suppose that $\rho_2^S(R2) = u$ and $gSS^n(R2) = m + c$.

specialize($I_{n+1}, gSS^n) = \{J_{n+1} = \text{mov} \ r1, \text{M-addr}\}$

\[
\text{mov} \ r1, [\text{rwl}], \text{mov} \ rwl, c; \text{add} \ r1, \text{rwl}; \text{mov} \ rwl, 0
\]

\[
\{J_{n+1}][\eta_S^n, \eta_D^n, \rho_D^n] = (\eta_S^n, \eta_D^n, \rho_D^n[R1 \mapsto m + c])
\]

\[
u = \rho_2^S(R2) = ([gSS^n]^2)(R2) \quad \text{(By Eqn. (8))}
\]

\[= ([gSS^n(R2)][\eta_S^n] = [m + c] \eta_S^n = m + c \tag{25}
\]

Eqns. (19), (24), (25), and (7) show that Eqn. (9) holds.

\[
gSS^{n+1} = gSS^n[R1 \mapsto m + c] \tag{26}
\]

\[
\eta_S^{n+1} = \eta_S^n \text{ and } \rho_S^{n+1} = \rho_S^n[R1 \mapsto u] \quad \text{(From (19))}
\]

Eqns. (25)–(27), and (8) show that Eqn. (10) holds. Note that $J_{n+1}$ has the mov rwl,0 instruction at the end of the residual code to restore the default value 0 in work register rwl.

3. $I_{n+1} = \text{mov} (\text{Indirect}(R1), \text{Direct}(R2))$: Because $R1$ must hold an address, there are two cases that can arise.

(a) Suppose that $\rho_2^S(R1) = a$ and $\rho_2^S(R2) = u$, and $gSS^n(R1) = m + c$ and $gSS^n(R2) = v$.

\[
[I_{n+1}][\eta_S^a, \rho_2^S, \eta_D^a, \rho_D^a] = \\
(\eta_S^u, \rho_2^S[a \mapsto u], \eta_D^u, \rho_D^u[a \mapsto u])
\]

specialize($I_{n+1}, gSS^n) = \{J_{n+1} = \text{mov} \ rwl, \text{M-addr}\}$

\[
\text{mov} \ r1, [\text{rwl}], \text{mov} \ rwl, c; \text{add} \ r1, \text{rwl}; \text{mov} \ rwl, 0
\]

\[
\{J_{n+1}][\eta_S^n, \eta_D^n, \rho_D^n] = (\eta_S^n, \eta_D^n, \rho_D^n[R1 \mapsto m + c])
\]

Similar to the proof of Eqn. (21), we can prove that $u = v$. Similar to the proof of Eqn. (25), we can prove that $a = m + c$. These equalities, together with Eqns. (28), (29), and (7), show that Eqn. (9) holds.

\[
gSS^{n+1} = gSS^n[m + c \mapsto v] \tag{30}
\]

\[
\eta_S^{n+1} = \eta_S^n \text{ and } \rho_S^{n+1} = \rho_S^n[a \mapsto u] \quad \text{(From (28))}
\]

The equalities $u = v$ and $a = m + c$, along with Eqns. (30), (31), and Eqn. (8), show that Eqn. (10) holds.

(b) Suppose that $\rho_2^S(R1) = a$ and $\rho_2^S(R2) = u$, and $gSS^n(R1) = m_1 + c_1$ and $gSS^n(R2) = m_2 + c_2$.

specialize($I_{n+1}, gSS^n) = \{J_{n+1} = \text{mov} \ r1, \text{M-addr}\}$

\[
\text{mov} \ r1, [\text{rwl}], \text{mov} \ rwl, c_1; \text{add} \ r1, \text{rwl}; \text{mov} \ rwl, 0
\]

\[
\{J_{n+1}][\eta_S^n, \eta_D^n, \rho_D^n] = (\eta_S^n, \eta_D^n, \rho_D^n[m_1 + c_1 \mapsto m_2 + c_2])
\]

Similar to the proof of Eqn. (25), we can prove that $a = m_1 + c_1$ and $u = m_2 + c_2$. These equalities, together with Eqns. (28), (32), and (7), show that Eqn. (9) holds.

\[
gSS^{n+1} = gSS^n[m_1 + c_1 \mapsto m_2 + c_2] \tag{33}
\]

\[
\eta_S^{n+1} = \eta_S^n \text{ and } \rho_S^{n+1} = \rho_S^n[a \mapsto u] \quad \text{(By Eqn. (28))}
\]

The equalities $a = m_1 + c_1$ and $u = m_2 + c_2$, along with Eqns. (33), (34), and Eqn. (8), show that Eqn. (10) holds.

4. $I_{n+1} = \text{stackalloc}(\text{Direct}(R1))$

\[
[I_{n+1}][\eta_S^n, \rho_2^S, \eta_D^n, \rho_D^n] = (\eta_S^n[m \mapsto m], \rho_2^S[R1 \mapsto m], \eta_D^n[m \mapsto m], \rho_D^n[R1 \mapsto m])
\]

specialize($I_{n+1}, gSS^n) = \{J_{n+1} = \text{stackalloc} \ r1; \text{mov} \ r1, \text{M-addr}; \text{mov} \ [\text{rwl}], \text{r1}; \text{mov} \ rwl, 0 \}$

\[
\{J_{n+1}][\eta_S^n, \eta_D^n, \rho_D^n] = (\eta_S^n[m \mapsto m], \eta_D^n[m \mapsto m], \rho_D^n[R1 \mapsto m][\text{M-addr} \mapsto m])
\]

$\rho_D^{n+1} \approx \rho_D^n$ because $\text{M-addr}$ is in the special area of the memory map of the residual binary. This observation, along with Eqns. (35), (36), and (7), show that Eqn. (9) holds.

\[
gSS^{n+1} = gSS^n[R1 \mapsto m] \tag{37}
\]

\[
\eta_S^{n+1} = \eta_S^n[m \mapsto m] \text{ and } \rho_S^{n+1} = \rho_S^n[R1 \mapsto m] \quad \text{(By Eqn. (35))}
\]

Note that the oracle supplies the correct $m$ while updating $gSS$, and WIPER uses $m$ to obtain the correct $\text{M-addr}$ to use in the residual code. Eqns. (37), (38), and (8) show that Eqn. (10) holds.