Scalability-First Pointer Analysis with Self-Tuning Context-Sensitivity

Yue Li, Tian Tan, Anders Møller
Aarhus University, Denmark
{yueli,tiantan,amoeller}@cs.au.dk

Yannis Smaragdakis
University of Athens, Greece
smaragd@di.uoa.gr

ABSTRACT
Context-sensitivity is important in pointer analysis to ensure high precision, but existing techniques suffer from unpredictable scalability. Many variants of context-sensitivity exist, and it is difficult to choose one that leads to reasonable analysis time and obtains high precision, without running the analysis multiple times.

We present the SCALER framework that addresses this problem. SCALER efficiently estimates the amount of points-to information that would be needed to analyze each method with different variants of context-sensitivity. It then selects an appropriate variant for each method so that the total amount of points-to information is bounded, while utilizing the available space to maximize precision.

Our experimental results demonstrate that SCALER achieves predictable scalability for all the evaluated programs (e.g., speedups can reach 10x for 2-object-sensitivity), while providing a precision that matches or even exceeds that of the best alternative techniques.

CCS CONCEPTS
• Theory of computation → Program analysis;

KEYWORDS
static analysis, points-to analysis, Java

ACM Reference Format:

1 INTRODUCTION
Pointer analysis is a family of static analysis techniques that provide a foundation for many other analyses and software engineering tasks, such as program slicing [36, 39], reflection analysis [19, 31], bug detection [13, 26], security analysis [1, 23], program verification [8, 27], and program debugging and comprehension [5, 21].

The goal of pointer analysis is to statically compute a set of objects (abstracted as their allocation sites) that a program variable may point to during run time. Although stating this goal is simple, it is challenging to produce precise analysis results without sacrificing scalability [12, 30, 35]. Thus, for decades, researchers have continued to develop sophisticated pointer analysis techniques [2, 14–16, 18, 22, 24, 25, 32, 33, 37, 38].

One of the key mechanisms for achieving high analysis precision is context sensitivity, which allows each program method to be analyzed differently according to the context it is used in [17]. Context sensitivity has different variants, depending on the kind of context information used. For Java programs, object-sensitivity [25] and type-sensitivity [32] have proven to be quite effective. The former is strictly more precise but less efficient than the latter [15, 37]. However, with any available variant, although the analysis can gain in precision, scalability is known to be unpredictable [33, 38], in the sense that programs very similar in size and other complexity metrics may have completely different scalability profiles.

Figure 1 shows time spent analyzing 10 real-world Java programs\(^1\) under 2-object-sensitivity (2obj) [25], which is among the most precise variants of context sensitivity, 2-type-sensitivity (2type) [32], and context-insensitive (CI). We observe that

- 2obj is not scalable for 6 out of 10 programs within 3 hours, while it can finish running for 3 programs within 5 minutes;
- program size is far from a reliable predictor—for example, jython (12 718 methods) is smaller than briss (26 582 methods), however, 2type is not scalable for the former but scalable for the latter;

\(^1\)These are all popular open-source applications, including the heaviest (jython and eclipse) of the DaCapo benchmarks [3].
context insensitivity exhibits very good and stable scalability for all programs (but it is much less precise).

Scalability of context-sensitivity is not only unpredictable, but also tends to be bimodal \cite{32}: when analyzing a given program with existing context-sensitivity techniques, usually the analysis either terminates relatively quickly (we say that the analysis is scalable in this case), or it blows up and does not terminate even with very high time limits.

Consider a scenario where the task is to analyze a set of programs within a given time budget, for example as part of a large-scale security analysis. Should one pick a precise context-sensitive pointer analysis and take the risk that it may not scale for several programs, or pick a scalable, context-insensitive, analysis that sacrifices precision for all programs? One answer is introspective analysis \cite{33}, which tunes context sensitivity according to the results of a first-pass, context-insensitive analysis. However, that approach is, as the authors put it, an “if all else fails” analysis that should only be used if traditional context-sensitive algorithms fail. Thus, computing resources are wasted, because one has to wait until the context-sensitive analysis reaches a timeout, before the introspective analysis is run as a fallback.

In this paper, we present a pointer analysis framework named \textsc{Scaler} that has the following desirable properties.

(i) Users only have to apply \textsc{Scaler} once for each program, without a need to experiment with different variants of context sensitivity.

(ii) \textsc{Scaler} prioritizes scalability. Given a reasonable time budget, \textsc{Scaler} can be expected to finish analyzing any given program \( P \) within the budget. More specifically, if a context-insensitive pointer analysis is scalable for \( P \), users can confidently expect that \textsc{Scaler} will also scale for \( P \).

(iii) \textsc{Scaler} is able to achieve precision comparable to, or better than that of the most precise context-sensitivity variant that is scalable for \( P \). That is, the user does not need to manually pick a context-sensitivity variant a priori, but can expect to effectively match the best option that one could have picked, on a single analysis run. Experimentally, this precision is much greater than prior introspective analyses \cite{33}.

The key insight of \textsc{Scaler} is that the size of context-sensitive points-to information (or, equivalently, the total memory consumed) is the critical factor that determines whether analysis of a given program using a particular context-sensitivity variant is scalable or not, and that it is possible to estimate the size of context-sensitive points-to information produced by different kinds of context sensitivity without conducting the context-sensitive analysis. This estimate can be computed using a cheap, context-insensitive pre-analysis, by leveraging the notion of object allocation graphs \cite{37}.

\textsc{Scaler} is parameterized by a number called the \textit{total scalability threshold} (TST), which can be selected based on the available memory. For a given TST, \textsc{Scaler} automatically adapts different kinds of context sensitivity (e.g., object-sensitivity, type-sensitivity) and context insensitivity, at the level of individual methods in the program, to stay within the desired bound and thereby retain scalability, while utilizing the available space to maximize precision.

In summary, this paper makes the following contributions:

- We propose the \textsc{Scaler} pointer analysis framework that, given a total scalability threshold, automatically selects an appropriate variant of context sensitivity for each method in the program being analyzed (Section 3). The approach relies on the concept of \textit{scalability-critical methods} that helps explain why context-sensitive pointer analysis is sometimes unscalable.

- We present a novel technique to efficiently estimate the amount of points-to information that would be needed to analyze each method with different variants of context sensitivity, using object allocation graphs (Section 4).

- We describe our open-source implementation (Section 5).

- We conduct extensive experiments by comparing \textsc{Scaler} with state-of-the-art pointer analyses (Section 6). The evaluation demonstrates that \textsc{Scaler} achieves predictable scalability for all the evaluated programs in one shot, while providing a precision that matches or even exceeds that of the best alternatives. As an example, the \texttt{jython} benchmark from DaCapo is known to cause problems for context-sensitive pointer analysis \cite{15, 38}: 1\texttt{type} is the most precise conventional pointer analysis that is scalable for \texttt{jython}, taking around 33 minutes on an ordinary PC, while 2\texttt{obj} and 2\texttt{type} do not complete within 3 hours. In comparison, \textsc{Scaler} achieves significantly better precision than both 1\texttt{type} and the state-of-the-art introspective analysis \cite{33} and takes under 8 minutes.

2 BACKGROUND

Pointer analysis typically consists of computing the points-to sets of variables in the program text.\(^2\) Points-to sets form a relation between variables and abstract objects, i.e., a subset of

\[ \text{Var} \times \text{Obj} \]

with \text{Var} being the set of program variables and \text{Obj} the set of abstract objects. Abstract objects are typically represented as \textit{allocation sites}, i.e., instructions that allocate objects (e.g., \texttt{new} in Java) \cite{6}. An allocation site stands for all the run-time objects it can possibly allocate. This representation by nature loses significant precision. A program variable corresponds to many run-time incarnations during program execution—not just for executions under different inputs but also for different instances of local variables during distinct activations of the same method.

To combat the loss of precision, context-sensitive pointer analysis enhances the computed relations to maintain a more precise abstraction of variables and objects \cite{30, 35}. Variables get qualified with contexts, to distinguish their different incarnations. The analysis effectively computes a subset of

\[ \text{Ctx} \times \text{Var} \times \text{Obj} \]

where \text{Ctx} is a set of contexts for variables.\(^3\) This precision is valuable for intermediate analysis computations, even though the final analysis results get collapsed in the easily-understandable \text{Var} \times \text{Obj} relation: distinguishing the behavior of a much-used variable or

\(^2\)A second formulation is that of \textit{alias analysis}, which computes the pairs of variables or expressions that can be aliased. For most published algorithms, computing points-to sets and computing alias sets are equivalent problems: one can be mapped to the other without affecting the algorithm’s fundamental precision or scalability.

\(^3\)For simplicity of exposition, we ignore context sensitivity for abstract objects (a.k.a., \textit{heap cloning}) which also qualifies \text{Obj} with a set of heap contexts, \text{HCtx}; in much the same way as qualifying variables.
object (e.g., in a library method) according to the context in which it is used helps the precision at all use sites of the common code.

Two observations on context sensitivity will be important in our subsequent discussion. First, contexts qualify variables but are generally chosen per method. The main use of context sensitivity is to distinguish different activations of the same method (e.g., “stack frames” for the same procedure, in traditional stack-based languages), which create fresh instances of all local variables at run time. Therefore, all local variables of the same method have the same set of contexts.

Second, the worst-case complexity of a context-sensitive pointer analysis is much higher than that of a context-insensitive one: the number of computed points-to facts, and hence the complexity of the analysis, increases in the worst case multiplicatively by the size of the set $Ctx$. However, in the common case, the precision of context sensitivity often compensates, reducing the analysis complexity: different contexts divide the points-to sets of variables into non-overlapping subsets. In the ideal case, if a context-insensitive analysis computes a points-to relation $pt \subseteq Var \times Obj$, a context-sensitive analysis would compute a relation $pt' \subseteq Ctx \times Var \times Obj$ that is not greater in cardinality than $Var \times Obj$. The extra precision of the $Ctx$ information would be enough to split the original points-to sets into disjoint subsets. This observation also hints at why context sensitivity has unpredictable scalability. When it works well to maintain precision, its cost is modest to none. When it fails to do so, however, the cost is orders-of-magnitude higher. The actual definitions of the set $Ctx$ can vary widely. Three main categories are call-site sensitivity, object sensitivity, and type sensitivity:

- **Call-site sensitivity** has $Ctx$ be a sequence of call sites, i.e., instructions that call the method in which the qualified variable has been declared.
- **Object sensitivity** has $Ctx$ be a sequence of abstract objects: they are the receiver object of the method containing the qualified variable, the receiver object of the method that allocated the previous receiver object, etc.
- **Type sensitivity** keeps the same information as object sensitivity, but objects used as contexts are collapsed not per allocation instruction but per class that contains it.

We refer the reader to surveys that discuss the options in full detail [50, 35].

### 3 THE SCALER FRAMEWORK

In this section, we present an overview of the Scaler approach. We first describe the idea of scalability thresholds, which is the key to predictable scalability (Section 3.1). Next, we explain how Scaler automatically chooses a suitable scalability threshold for each program, based on a single parameter called the total scalability threshold that depends only on the available space for storing points-to information (Section 3.2). These ideas are then collected in the overall Scaler framework (Section 3.3).

#### 3.1 Scalability Thresholds

We begin by introducing the concept of **scalability-critical methods** (Section 3.1.1), and we then leverage this concept to address unscalability (Section 3.1.2).
where will be analyzed with different context sensitivity variants. As discussed earlier, the overall cost of a context-sensitive pointer analysis is closely linked to the size of the points-to relation being computed. An analysis that fails to terminate within realistic time limits often does so because the space needed to represent the points-to relation exhausts the available RAM. The scalability of pointer analysis is therefore related to the features of the analysis environment, most importantly the memory size.

We now show how SCALER determines a scalability threshold for a given program, based on the concept of a total scalability threshold that represents the analysis capacity and can be selected based on, for example, the available memory, independently of the program being analyzed.

**Definition 3.2 (Total Scalability Threshold).** A total scalability threshold (TST) is a number that satisfies the inequality $E(ST_P) \leq TST$ where $E(ST_P)$ is the estimated cost for the given value of $ST_P$, computed as the sum of the sizes of the points-to relations (cf. Definition 3.1) for all the methods in the program:

$$E(ST_P) = \sum_m \#ctx_m \cdot \#pts_m$$

Figure 3 illustrates the concept. The complexity of a pointer analysis (running on a given program) is closely related to the size of the points-to information, which is upper-bounded by $E(ST_P)$. The TST of the environment can be seen as analogous to the volume of a container. The volume of the points-to information computed by the analysis has to fit in the TST volume for the analysis to scale, i.e., $E(ST_P) \leq TST$. This effectively normalizes the analysis capacity for all programs, regardless of the number of methods they contain, or the variants of context sensitivity employed. We discuss in Section 6 the actual TST values used in our experiments.

Importantly, in the TST inequality, $\#ctx_m$, $\#pts_m$, $E(ST_P)$ and $\sum_m \#ctx_m \cdot \#pts_m$ depend on the choice of context sensitivity, which can vary based on the approach described in Section 3.1. The self tuning approach of SCALER consists of computing, given the TST, an appropriate value of $ST_P$ for a program $p$ that will be used to assign appropriate context-sensitivity variants to $p$’s methods.

To maximize precision while ensuring scalability, SCALER needs to pick the largest $ST_P$ that satisfies the TST inequality. In this way, as many methods as possible are analyzed with the most precise context sensitivity that the total analysis capacity can afford. That is, SCALER needs to pick $\max(ST_P \mid E(ST_P) \leq TST)$.

Figure 4 illustrates the mechanism using the distribution of Figure 2. This time, $ST_P$ is not given by the user in advance. Instead, its value is computed to satisfy the inequality shown in Figure 4. For a candidate $ST_P$ value, each method is classified as detailed in Section 3.1.2: it is assigned the most precise context sensitivity that keeps it from being scalability-critical.

In the case of Figure 4, the value of $E(ST_P)$ for this program corresponds to the sum of the areas of $A_1$, $A_2$, and $A_3$. A $1$ $A_2$ $A_3$ under the candidate $ST_P$ is below TST, then $ST_P$ is viable, but
Scalability-First Pointer Analysis with Self-Tuning Context-Sensitivity

Figure 4: Selecting the scalability threshold (ST_p) based on the total scalability threshold (TST).

higher ST_p values may also be viable. The maximum such ST_p is computed using binary search in a range between 0 and the maximum value of \( \#ctx_m \cdot \#pts_m \) for all \( m \).

3.3 Overall Approach

Figure 5 shows the overall structure of the Scaler framework. Scaler is a stand-alone system that automatically selects a per-method context-sensitivity variant that can subsequently be used to configure an existing pointer analysis tool like Doop, WalA, or Chord.

Given a program \( p \), a fast (but imprecise) context-insensitive pointer analysis is run first to produce some basic points-to information that is used for estimating \( \#pts_m \) (Section 3.1.1) and \( \#ctx_m \) (Section 4). The core of Scaler consists of one component that decides the scalability threshold based on a given total scalability threshold (Section 3.2) and another component that chooses context sensitivity strategies for all methods in the program (Section 3.1.2).

The framework relies on the collection \( C \) of context-sensitivity variants. All that is needed for each variant is a mechanism for estimating the number of contexts, as presented in Section 4 for object-sensitivity and type-sensitivity.

4 ESTIMATING THE NUMBER OF CONTEXTS

In Section 3.1 we postponed the discussion of how to compute an upper bound on the number of contexts (\( \#ctx_m \)), for a given variant of context sensitivity \( c \), using only context-insensitive analysis results. This computation is one of the key elements of Scaler and distinguishes it from prior approaches that have also tried to adapt context sensitivity on a per-method basis [14, 33, 41].

The computation of the possible contexts for the context-sensitive analysis of a method, from only context-insensitive analysis results, is relatively easy for simple variants of context sensitivity, such as call-site sensitivity [28]: the possible contexts are call sites, readily identifiable in the program code. This computation is nontrivial for object- and type-sensitivity [24, 32], however: the contexts of a method are abstract objects, determined by the analysis mechanism. We are not aware of any prior work that performs a similar computation of possible contexts for object-sensitive or type-sensitive analyses, without running the analyses themselves.

Scaler performs this computation by leveraging the object allocation graph (OAG) structure proposed by Tan et al. [37]. With the OAG, the context computation problem can be formulated as a graph traversal problem. For any program, based on an OAG derived from pre-analysis (context-insensitive pointer analysis), Scaler computes the number of contexts (\( \#ctx_m \)) of every method for each kind of context sensitivity \( c \) used (2obj, 2type and 1 type in our setup) by enumerating all contexts of the method.

The OAG of a program is a directed graph. A node of the OAG represents an abstract object, which is identified by its allocation site in the program. An edge of the OAG, say \( o_1 \rightarrow o_2 \), represents an object-allocation relation between \( o_1 \) and \( o_2 \), i.e., \( o_1 \) is a receiver object of the method that contains the allocation site of \( o_2 \). Scaler leverages the pre-analysis to build the OAG for the given program. The OAG provides a graphical perspective of object- and type-sensitivity, i.e., a \( k \)-depth context in object-sensitivity corresponds to a \( k \)-node path in the OAG [37]. Thus, to compute \( \#obj \) and \( \#type \) contexts of a method \( m \), Scaler simply enumerates \( k \)-node paths in the OAG, leading to the receiver objects of \( m \).

Figure 6 illustrates the mechanism with a simple example. The allocation sites are labeled 1, 2, and 3, respectively. Suppose we compute 2obj contexts for method \( m(\) (line 12) and its receiver object is \( C \) (allocated at line 16). Further, 1 and 2 are two allocator objects of \( C \) according to object-sensitivity [25, 32]. The possible (2obj) contexts of \( m(\) are \( [B_1, C_1] \) and \( [B_2, C_1] \). The corresponding OAG is given in Figure 6, which shows two 2-node paths that are exactly the 2obj contexts for \( m(\). Since a context-insensitive analysis over-approximates the fully precise OAG, some edges may be spurious. However, this is fine, since we only need an upper bound of the number of possible contexts, to establish scalability.

Type sensitivity is an isomorphic approximation of object sensitivity [30], thus we can easily derive the contexts for type sensitivity with the OAG. Following the definition [32], Scaler computes the contexts for type sensitivity by merely replacing objects in contexts (as computed from the OAG) by the types that contain the allocation sites of the objects. For example, to compute the contexts for method \( m(\) (Figure 6) under 2type, Scaler first obtains the 2obj contexts, i.e., \( [B_1, C_1] \) and \( [B_2, C_1] \), then replaces \( B_1 \) and \( B_2 \) by type \( A \) and \( C_1 \) by type \( B \). As a result, there is only one context of \( m(\) under 2type, i.e., \( [A, B] \).

5 IMPLEMENTATION

We have implemented Scaler as a stand-alone open-source tool in Java, available at [Link]. Scaler is designed to work with various pointer analysis frameworks such as Doop [4], WalA [7], Chord [26], and Soot [40]. To demonstrate its effectiveness, we have integrated Scaler with Doop [4], a state-of-the-art pointer analysis framework for Java.

In practice, we found that the \( \#ctx_m \cdot \#pts_m \) values of some Java collection methods under package java.util.* are very large, because (1) the collection methods are frequently called, thus their \( \#ctx_m \) values are large, and (2) many objects are passed to the

We do not consider call-site sensitive analyses (1call, 2call) or 1-object sensitivity (1obj) in our evaluation, as these analysis variants are typically both less precise and less scalable than at least one of the analyses in our setup. The Scaler approach can be adapted to any collection of context sensitivity variants, as long as a relative ordering in both increasing precision and cost is possible.
6 EVALUATION

In the evaluation, we investigate the following research questions:

RQ1. Does SCALER consistently achieve scalability, while matching or exceeding the precision of the most precise conventional pointer analyses [25, 32] that use the same variant of context-sensitivity for the entire program?

RQ2. How does SCALER fare against introspective analysis [33] that also applies context-sensitivity selectively for the different methods of the program?

RQ3. What is the overhead of running SCALER? What are the computed values of \( \text{TST}_p \) and what are the resulting distributions of SelectCtx in practice?

RQ4. How does SCALER perform with different TST values and different memory sizes?

Experimental Setup. All pointer analyses were performed using Doop [29] (with the version published in the artifact of [33], which contains the exact setup for different analyses). The time budget for all analyses is 3 hours (10,800 seconds). To demonstrate the generality of SCALER’s scalability and precision, we consider 10 real-world Java programs that cover different levels of program complexity. Five of these programs come from the DaCapo 2006 benchmark suite and include the toughest-to-analyze programs (jython, eclipse). The rest are well-known open-source applications. Concretely, as shown in Table 1, 6 programs (jython, soot, pmd, briss, jedit and eclipse) represent complex applications for which 2obj is not scalable within 3 hours; 2 programs (findbugs and chart) represent medium-complexity programs for which 2obj costs thousands or hundreds of seconds; 2 programs (luindex and lusearch) represent simple programs for which 2obj costs only dozens of seconds. These programs are all analyzed with a large Java library: Open JDK 1.6.0_24.

SCALER uses a default TST value of 30M (million) which exhibits uniform scalability on all machines in our experimental setup. To answer RQ1–RQ3 (Section 6.1–6.3), we run the experiments on our default machine with a Xeon E5-2697A 2.6GHz CPU and 48GB of RAM, with the default TST (30M). To answer RQ4, different TST values are considered, and the experiments are also run on machines with smaller and larger memory sizes (Section 6.4).

6.1 Scalability and Precision of SCALER-Guided Pointer Analysis

In this section, we examine the scalability and precision of SCALER-guided pointer analysis (SCALER for short) by comparing it with conventional context-sensitive pointer analyses for Java: object-sensitivity [25] and type-sensitivity [32], which are the mainstream variants adopted in recent analysis clients. For example, object-sensitivity is widely adopted in recent Android static analysis frameworks [1, 9] and type-sensitivity is used in recent reflection analysis [20] and security analysis [10] tools.

Table 1 shows the results for all analyses. Each program has five rows of data, respectively representing context-insensitive (CI), conventional context-sensitive, SCALER, and two introspective (IntroA and IntroB) pointer analyses. The last two analyses will be discussed in Section 6.2.) For conventional context-sensitive pointer analyses, we consider 2obj, 2type, and 1type. Call-site-sensitivity and 1obj are not considered as the former is not effective for Java programs [15, 17, 37] and the latter is usually both less efficient and less precise than 2type [15, 32]. As it is virtually impossible to predict the scalability of a precise context-sensitive pointer analysis in advance, to get the most precise conventional analysis results we run 2obj, 2type, and 1type in the given order (from the most to the least precise one) until one of them terminates within 3 hours. So in Table 1, the second row for each program shows the results of the most precise conventional pointer analysis that is scalable.

6.1.1 Scalability of SCALER-Guided Pointer Analysis. In Table 1, the third column for each program demonstrates SCALER’s extremely good scalability. 2obj is not scalable for the first six programs, and even the fast 2type also fails to scale for jython and soot. However, SCALER scales for all of them, with its one-shot principle. In addition, for the first seven complex and large programs, SCALER runs (sometimes significantly) faster than the most precise conventional pointer analysis that is scalable, even if we add SCALER’s pre-analysis time (CI).

For example, according to past literature and extensive experience [15, 32, 37, 38], jython is considered the most troublesome program in terms of scalability, among the DaCapo benchmarks [3].
To get the most precise results for a conventional pointer analysis, 23,597 seconds (3h + 3h + 1 997s) are spent before one discovers that 1type is scalable for jython on our machine; however, SCALER only costs 452 seconds for jython, and with better precision (in terms of all precision metrics) as described in the next section.

### 6.1.2 Precision of SCALER-Guided Pointer Analysis

To measure precision, we use four independently useful client analyses that are often (although rarely all together) used as precision metrics in existing literature [14, 15, 32, 33, 38]. Together they paint a fairly accurate picture of precision analysis, as it impacts clients. The clients are: a cast-resolution analysis (metric: the number of cast operations that may fail—#may-fail casts), a devirtualization analysis (metric: the number of virtual call sites that cannot be disambiguated into monomorphic calls—#poly calls), a method reachability analysis (metric: the number of reachable methods—#reach methods), and a call-graph construction analysis (metric: the number of call graph edges—#call edges). In Table 1, columns (4–7) for each program list precision results. In all cases, lower is better. We find that, overall, SCALER achieves comparable, and typically better, precision than the most precise conventional pointer analysis overwhelmingly, with the exception of eclipse. In the case of eclipse, the precision of SCALER is marginally lower than that of IntroA and IntroB, which is comparably scalable (e.g., 1type, 2type), for IntroA and IntroB, the running time can be considered similar. However, the running time without also using heuristics (such as “total points-to information”) that do not, however, have the upper-bound guarantee, scalability emphasis, or context-number-estimation ability of SCALER.

There are two published heuristics leading to different variants of introspective analyses, IntroA and IntroB. Generally, IntroA is faster but less precise than IntroB. Like SCALER, introspective analysis also relies on a context-insensitive analysis (CI) as its pre-analysis. Unlike SCALER, which decides on each method what context-sensitivity it needs (e.g., 2obj, 2type, 1type or CI), introspective analysis also decides on each method whether it needs contexts. Despite the difference, the computation time of producing the context selection information is very similar (a few seconds for each program on average). As a result, the overhead of both analyses’ decision making can be considered similar.

In Table 1, the last three rows for each program show the comparison results. In most programs (except soot and eclipse), IntroA

### 6.2 Comparison with Introspective Analyses

The closest relative of our work in past literature is introspective analysis [33]: a technique that also attempts to tune context sensitivity per-method based on a pre-analysis. Introspective analysis uses heuristics (such as “total points-to information”) that do not, however, have the upper-bound guarantee, scalability emphasis, or context-number-estimation ability of SCALER.

Table 1: Efficiency and precision metrics for all programs and pointer analyses. In all cases, lower is better.

<table>
<thead>
<tr>
<th>Program</th>
<th>Analysis</th>
<th>Time (seconds)</th>
<th>Precision metrics</th>
<th>Program</th>
<th>Analysis</th>
<th>Time (seconds)</th>
<th>Precision metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>jython</td>
<td>CI</td>
<td>112</td>
<td>20 726</td>
<td>CI</td>
<td>135</td>
<td>135</td>
<td>3h=10 800s</td>
</tr>
<tr>
<td></td>
<td>2obj → 2type → 1type</td>
<td>&lt;3h + 3h + 1 997</td>
<td>2 234</td>
<td>2type</td>
<td>3 536</td>
<td>3 536</td>
<td>#may-fail casts</td>
</tr>
<tr>
<td></td>
<td>soot</td>
<td>452</td>
<td>20 726</td>
<td>SCALER</td>
<td>652</td>
<td>652</td>
<td>#poly calls</td>
</tr>
<tr>
<td></td>
<td>IntroA</td>
<td>381</td>
<td>20 726</td>
<td>IntroA</td>
<td>1 369</td>
<td>1 369</td>
<td>#reach methods</td>
</tr>
<tr>
<td></td>
<td>IntroB</td>
<td>&lt;3h</td>
<td>20 726</td>
<td>IntroB</td>
<td>528</td>
<td>528</td>
<td>#call edges</td>
</tr>
<tr>
<td>pmd</td>
<td>CI</td>
<td>984</td>
<td>20 726</td>
<td>CI</td>
<td>49</td>
<td>49</td>
<td>3h=10 800s</td>
</tr>
<tr>
<td></td>
<td>2obj → 2type → 1type</td>
<td>&lt;3h + 3h + 3 812</td>
<td>2 234</td>
<td>2type</td>
<td>2 458</td>
<td>2 458</td>
<td>#may-fail casts</td>
</tr>
<tr>
<td></td>
<td>SCALER</td>
<td>1 236</td>
<td>20 726</td>
<td>SCALER</td>
<td>272</td>
<td>272</td>
<td>#poly calls</td>
</tr>
<tr>
<td></td>
<td>IntroA</td>
<td>1 295</td>
<td>20 726</td>
<td>IntroA</td>
<td>188</td>
<td>188</td>
<td>#reach methods</td>
</tr>
<tr>
<td></td>
<td>IntroB</td>
<td>4 885</td>
<td>20 726</td>
<td>IntroB</td>
<td>397</td>
<td>397</td>
<td>#call edges</td>
</tr>
<tr>
<td>briss</td>
<td>CI</td>
<td>67</td>
<td>20 726</td>
<td>CI</td>
<td>48</td>
<td>48</td>
<td>3h=10 800s</td>
</tr>
<tr>
<td></td>
<td>2obj → 2type</td>
<td>&lt;3h + 1 203</td>
<td>2 234</td>
<td>2type</td>
<td>2 458</td>
<td>2 458</td>
<td>#may-fail casts</td>
</tr>
<tr>
<td></td>
<td>SCALER</td>
<td>705</td>
<td>20 726</td>
<td>SCALER</td>
<td>254</td>
<td>254</td>
<td>#poly calls</td>
</tr>
<tr>
<td></td>
<td>IntroA</td>
<td>356</td>
<td>20 726</td>
<td>IntroA</td>
<td>128</td>
<td>128</td>
<td>#reach methods</td>
</tr>
<tr>
<td></td>
<td>IntroB</td>
<td>2 986</td>
<td>20 726</td>
<td>IntroB</td>
<td>189</td>
<td>189</td>
<td>#call edges</td>
</tr>
<tr>
<td>jedit</td>
<td>CI</td>
<td>228</td>
<td>20 726</td>
<td>CI</td>
<td>22</td>
<td>22</td>
<td>3h=10 800s</td>
</tr>
<tr>
<td></td>
<td>2obj → 2type</td>
<td>&lt;3h + 5 374</td>
<td>2 234</td>
<td>2type</td>
<td>2 458</td>
<td>2 458</td>
<td>#may-fail casts</td>
</tr>
<tr>
<td></td>
<td>SCALER</td>
<td>1 194</td>
<td>20 726</td>
<td>SCALER</td>
<td>53</td>
<td>53</td>
<td>#poly calls</td>
</tr>
<tr>
<td></td>
<td>IntroA</td>
<td>497</td>
<td>20 726</td>
<td>IntroA</td>
<td>45</td>
<td>45</td>
<td>#reach methods</td>
</tr>
<tr>
<td></td>
<td>IntroB</td>
<td>&lt;3h</td>
<td>20 726</td>
<td>IntroB</td>
<td>48</td>
<td>48</td>
<td>#call edges</td>
</tr>
<tr>
<td>lusearch</td>
<td>CI</td>
<td>117</td>
<td>20 726</td>
<td>CI</td>
<td>22</td>
<td>22</td>
<td>3h=10 800s</td>
</tr>
<tr>
<td></td>
<td>2obj → 2type</td>
<td>&lt;3h + 2 950</td>
<td>2 234</td>
<td>2type</td>
<td>2 458</td>
<td>2 458</td>
<td>#may-fail casts</td>
</tr>
<tr>
<td></td>
<td>SCALER</td>
<td>1 769</td>
<td>20 726</td>
<td>SCALER</td>
<td>93</td>
<td>93</td>
<td>#poly calls</td>
</tr>
<tr>
<td></td>
<td>IntroA</td>
<td>300</td>
<td>20 726</td>
<td>IntroA</td>
<td>94</td>
<td>94</td>
<td>#reach methods</td>
</tr>
<tr>
<td></td>
<td>IntroB</td>
<td>6 942</td>
<td>20 726</td>
<td>IntroB</td>
<td>96</td>
<td>96</td>
<td>#call edges</td>
</tr>
</tbody>
</table>

To get the most precise results for a conventional pointer analysis, 23,597 seconds (3h + 3h + 1 997s) are spent before one discovers that 1type is scalable for jython on our machine; however, SCALER only costs 452 seconds for jython, and with better precision (in terms of all precision metrics) as described in the next section.

### 6.1.2 Precision of SCALER-Guided Pointer Analysis

To measure precision, we use four independently useful client analyses that are often (although rarely all together) used as precision metrics in existing literature [14, 15, 32, 33, 38]. Together they paint a fairly accurate picture of precision analysis, as it impacts clients. The clients are: a cast-resolution analysis (metric: the number of cast operations that may fail—#may-fail casts), a devirtualization analysis (metric: the number of virtual call sites that cannot be disambiguated into monomorphic calls—#poly calls), a method reachability analysis (metric: the number of reachable methods—#reach methods), and a call-graph construction analysis (metric: the number of call graph edges—#call edges). In Table 1, columns (4–7) for each program list precision results. In all cases, lower is better. We find that, overwhelmingly, SCALER achieves comparable, and typically better, precision than the most precise conventional pointer analysis that is comparably scalable (e.g., 1type for jython, 2type for pmd and 2obj for 1uindex). In the case of findbugs, the precision of SCALER is marginally lower than that of 2obj, but the running time is an order of magnitude lower. The chart benchmark is the only one for which SCALER is slightly less precise than 2obj without also being much faster, yet SCALER still attains most of the precision gains of 2obj relative to a context-insensitive analysis. Moreover, as shown in Section 6.4, the precision of SCALER can improve by simply increasing the TST value.

Answer to RQ1: SCALER-guided pointer analysis exhibits extremely good and uniform scalability while matching or even exceeding the precision of the most precise conventional pointer analysis that is scalable.
Table 2: Performance of Scaler.

<table>
<thead>
<tr>
<th>Program</th>
<th>jython</th>
<th>soot</th>
<th>pmd</th>
<th>briss</th>
<th>jedit</th>
<th>eclipse</th>
<th>findbugs</th>
<th>chart</th>
<th>luindex</th>
<th>lusearch</th>
<th>avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ScALER time (seconds)</td>
<td>35.0</td>
<td>1.8</td>
<td>0.8</td>
<td>2.6</td>
<td>1.4</td>
<td>1.0</td>
<td>0.5</td>
<td>0.4</td>
<td>0.2</td>
<td>0.3</td>
<td>4.4</td>
</tr>
</tbody>
</table>

Answer to RQ2: Comparing with state-of-the-art adaptive analyses, introspective analyses IntroA and IntroB, both ScALER and IntroA exhibit extremely good scalability while ScALER’s precision is significantly better than IntroA’s; and ScALER’s scalability is better than IntroB’s while being more precise in most cases.

6.3 ScALER’s Effectiveness as a Pre-Analysis

This section answers RQ3 by examining the overhead of ScALER’s adaptivity as well as the ST\textsubscript{P} values selected by ScALER for each program, and the corresponding distribution of different kinds of context sensitivity based on the selected ST\textsubscript{P}.

6.3.1 Overhead of ScALER. The overall overhead of the ScALER framework (Figure 5) consists of two components: (1) a context-insensitive pointer analysis CI, which provides points-to information to ScALER, and (2) ScALER logic itself, which performs TST-based ST\textsubscript{P} selection and ST\textsubscript{P}-based context-sensitivity. ScALER’s pre-analysis time (CI) is given in Table 1, and Table 2 shows the overhead of ScALER itself. The average overall overhead of ScALER for each program is 183.8 seconds, of which CI pre-analysis costs account for the vast majority (179.4 seconds), while ScALER logic costs only 4.4 seconds. Considering the significant scalability improvements achieved by ScALER, its overhead is negligible.

ScALER spends 35 seconds for jython, which is markedly longer than for other programs (Table 2). The reason is that jython is especially complicated in the sense that many of its methods have enormous numbers of contexts, thus ScALER spends much time on \#ctx\textsubscript{m} computation. For instance, ScALER’s \#ctx\textsubscript{m} computation shows that 130 methods of jython have more than 1 000 000 contexts under 2obj. (For comparison, the maximal \#ctx\textsubscript{m} value of a single method under 2obj in soot is only 88 473.) In this way, ScALER also reveals the reason why many context-sensitive pointer analyses fail to analyze jython scalably as reported in previous work [15, 32, 37, 38]. ScALER avoids the problem due to its ST\textsubscript{P}-based context-sensitivity design (Section 3.1.2).

6.3.2 ST Values and Context-Sensitivity Distributions. Figure 7 gives the ST\textsubscript{P} values selected by ScALER for each program according to the default TST (30M), and the distribution of different kinds of context sensitivity over the methods of the programs based on the selected ST\textsubscript{P}. Generally, given the same TST, ScALER automatically selects small ST\textsubscript{P} values for complex programs (e.g., soot and briss), medium ST\textsubscript{P} values for medium-complexity programs (findbugs and chart), and large ST\textsubscript{P} values for simple programs (luindex and lusearch).

ScALER automatically selects 2obj, the most precise context-sensitivity in our experiments, for most methods (80.6% per program on average). This is the reason why ScALER-guided pointer analysis achieves very good precision as shown in Table 1. This also demonstrates that our insight of scalability-critical methods (Section 3.1.1) holds in practice: in most cases, only a small set of scalability-critical methods make the whole analysis unscalable.

We next discuss two outlier cases, at both ends of the spectrum, since they are informative of ScALER’s limit behavior.

soot. ScALER selects 0 as ST\textsubscript{P} value for soot, so most of its methods are analyzed with context-insensitivity (CI). The reason is that soot is very large. The pre-analysis of ScALER reports that the total size of points-to information of soot (i.e., the sum of \#pts\textsubscript{m} of all methods) is 110 901 529, which already exceeds our default TST (30M). As a result, ScALER automatically selects 2obj for only 1 792 methods in package java.util.* (as explained in Section 5) and CI for all other methods according to the definition of SelectCtx in Section 3.1.2, to ensure the scalability of pointer analysis.

lusearch and lusearch. ScALER selects two very large ST\textsubscript{P} values for luindex and lusearch (the two simplest programs in our evaluation). Accordingly, it assigns 2obj for all methods in the two programs which are all classified as non-scalability-critical methods under 2obj. There is only one exception: the method \texttt{<java.lang.Object: void <init>()>}. In Java’s type hierarchy, Object is the subtype of all classes; thus its \texttt{<init>} method will be called whenever a constructor is called, yielding too many contexts (\#ctx\textsubscript{m}), and a this variable that abstractly points to all the objects created in the program, which makes the method’s \#pts\textsubscript{m} value very large.

As a result, the \#ctx\textsubscript{m}, \#pts\textsubscript{m} value for this \texttt{<init>} method exceeds the selected ST\textsubscript{P} value, so ScALER selects less precise context
sensitivity for it. Since java.lang.Object’s constructor <init> is an empty method, analyzing it with or without 2obj does not affect the precision of pointer analysis. Thus SCALER-guided pointer analysis achieves exactly the same precision as the conventional 2obj analysis for jython and lusearch (see Table 1). This again demonstrates the self-tuning ability of SCALER which can automatically obtain maximal precision for simple programs.

### 6.4 SCALER-Guided Pointer Analysis with Different TSTs and Memory Sizes

As real-world analysis settings may differ widely, we conduct experiments by running SCALER with different TST and memory sizes, to further evaluate the adaptiveness and usability of SCALER in practice. We run SCALER for the top six costliest-to-analyze programs in Table 1 (jython, soot, pmd, briss, jedit, and eclipse), with TST values (20M, 30M, 60M, 80M, and 150M) on three machines with different memory sizes: 12GB, 48GB, and 368GB, representing typical memory sizes of a personal laptop, a commodity server, and a large server, respectively.

Conventional pointer analyses that are not scalable for these six programs in Table 1 (with memory size 48GB) are also all unscalable on another machine with a much larger memory size, 368GB. However, under SCALER’s default TST of 30M (as used in previous experiments), SCALER-guided pointer analysis still scales for all six programs even on a machine with only 12GB memory. This result further demonstrates the effectiveness of SCALER in making pointer analysis scalable in practice.

Since the six programs under different TST and memory settings exhibit similar trends, for space reasons we only show two representative programs, eclipse and pmd, in Figure 8. We illustrate precision changes via the #may-fail casts metric, which is arguably the most common precision metric in Java pointer-analysis literature [14, 15, 17, 32–34, 37, 38].

**SCALER-Guided Pointer Analysis under Different TSTs.** Figure 8 encodes a lot of information: it shows how both precision and scalability vary as a function of TST values, for three different RAM configurations. As the figure demonstrates, there is nothing special about SCALER’s default 30M TST: the system adapts appropriately for both larger and smaller values. Users can simply increase or decrease the TST value to achieve better precision (up to the level the analysis setup can support) or better efficiency, respectively. This simple design with TST being the only tuning knob is able to drive complicated pointer analyses to adapt to different programs to achieve their preferred scalability and precision.

**Answer to RQ3:** SCALER automatically selects appropriate TST values and context-sensitivity to ensure the scalability of pointer analysis for programs of varying complexity. In addition, the cost of such good adaptivity is very low.

**Answer to RQ4:** Better precision or better efficiency of SCALER-guided pointer analysis can be achieved by simply increasing or decreasing the TST value. The scalability is related to memory sizes: with the same TST, SCALER-guided pointer analysis is more likely to scale under larger memory; thus a small (large) TST is recommended for a small (large) memory size for good scalability (precision).

### 7 RELATED WORK

Context sensitivity, despite bringing great precision benefits, introduces many uncertainties to scalability, which may render a pointer analysis useless in practice. We focus on related work that leverages pre-analysis to achieve good efficiency and precision trade-offs for context-sensitive pointer analysis.

Introspective analysis [33] attempts to achieve precision and scalability trade-offs by refining a context-sensitive analysis while avoiding its worst-case cost. Similar to SCALER, it first performs a pre-analysis (context-insensitive pointer analysis) to extract necessary information to guide later pointer analysis. Unlike SCALER, it relies on a set of six manually-selected metrics (e.g., the maximum
field points—to set over all fields) to define two heuristics, resulting in two introspective analyses, IntroA and IntroB, which are compared with SCALER in Section 6.2. Benefiting from the new insight of SCALER (TST-based self-tuning context sensitivity), SCALER outperforms IntroB on both precision and scalability and achieves the same level of scalability as IntroA, while being significantly more precise. Moreover, the six different metrics in introspective analysis need appropriate values to be set in advance to produce effective analysis results; SCALER’s insights enable its methodology to be quite simple: it only needs one value (TST) for users to achieve better precision or better efficiency as desired, resulting in better usability in practice. Finally, introspective analysis is not one-shot: using it will always incur a cost in precision, even if the program could be analyzed more precisely. Therefore, its user will deploy it only after first attempting a precise analysis and failing.

Hassanshahi et al. [11] leverage similar metrics as introspective analysis to guide selective object-sensitive pointer analysis for large codebases. However, their pre-analysis involves several phases (that need different metrics and heuristics): a context-insensitive analysis is first performed to extract the program kernel where a context-insensitive or fixed object-sensitive analysis is still not sufficiently precise; then a fixed (heavy) object-sensitive pointer analysis is applied to the extracted (smaller) kernel to determine appropriate context depth for each selected object. After these pre-analyses, the selected object-sensitive information is used to guide the main pointer analysis which is demonstrated to work well for the OpenJDK library. However, unlike introspective analysis [33] and SCALER, the overhead of the pre-analysis is uncertain, as it heavily relies on the complexity of the extracted kernel, which further depends on various metric values selected by users. Thus, it is unclear if the technique can exhibit general effectiveness for arbitrary Java programs in practice.

Both of the above approaches [11, 33] involve metrics and heuristics that are defined manually. An alternative is to use machine learning techniques, as in the two approaches we describe next.

Wei and Ryder [41] present an adaptive context-sensitive analysis for JavaScript. They first use a machine learning algorithm to obtain the relationship between some user-defined method characteristics (extracted from a pre-analysis) and context-sensitivity choice (1-call-site-, 1-object-, or 1-parameter-sensitivity), and express the results as a decision tree. Based on domain knowledge, the decision tree is further manually adjusted to produce heuristics that are easy to interpret while the classifications can still maintain good accuracy. Finally, based on the heuristics, methods are analyzed with different context sensitivities, resulting in better precision achieved than single context-sensitive analysis.

Jeong et al. [14] propose a data-driven approach to guiding context sensitivity for Java. Unlike SCALER, where various kinds of context sensitivity with different lengths are applied to different methods, only a single kind of context sensitivity is applied to the program, and each method is finally assigned an appropriate context length, including zero (i.e., context insensitivity). As deep contexts are finally properly applied to only a subset of the methods, more efficient context-sensitive analysis can be achieved with still good precision. To select a context length for each method, 25 metrics (atomic features) are selected and a machine learning approach is used to learn heuristics based on these metrics. However, unlike SCALER’s lightweight pre-analysis, the learning phase is heavy and costs 54 hours in the Jeong et al. experimental setting.

Generally, a machine learning approach is sensitive to the training on input programs, and its learned results are usually difficult to explain, for example to discern why the learning algorithm selects a given context for a method. Instead, SCALER is a principled, rigorously-modeled, approach derived from simple insights; thus its guided results are tractable and interpretable, leading to more stable and uniform effectiveness.

Unlike conventional context-sensitive pointer analysis, which uses consecutive context elements for each context, the Bean approach by Tan et al. [37] identifies and skips the redundant context elements that are useless for improving the precision of context-sensitive analysis. As a result, the saved space allows more precision-useful context elements to be involved to distinguish more contexts, making the analysis more precise with a small efficiency overhead. Precision is the focus of Bean while SCALER’s is scalability. In addition, as explained in Section 4, rather than identifying redundant precision-useless context elements, SCALER leverages the OAG from Tan et al. [37] to compute the context numbers in advance.

MAHjong [38], a recent heap abstraction for pointer analysis of Java, is also based on a cheap pre-analysis, like SCALER. It enables an allocation-site-based pointer analysis to run significantly faster while achieving nearly the same precision for type-dependent clients, such as call graph construction. Differently, SCALER works for general pointer analysis, including alias analysis (i.e., not just type-dependent clients) which cannot be handled by MAHjong effectively. In addition, SCALER is able to scale for trouble programs such as jython where even the very fast MAHjong analysis fails.

8 CONCLUSIONS

Good scalability is hard to achieve for precise context-sensitive pointer analysis. To tackle this problem, we have introduced the SCALER framework, which automatically chooses a suitable context-sensitivity variant for each method in the given program, based on a fast, context-insensitive pre-analysis. The key insight is that it is possible to efficiently identify scalability-critical methods and that scalability can be predicted using the ideas of scalability thresholds and total scalability thresholds. The focus of SCALER is scalability, but at the same time it aims to maximize precision, relative to a given total scalability threshold that can be selected based on the available memory.

The experimental evaluation of SCALER demonstrates that it is able to achieve extremely good scalability while producing highly precise points-to results, in one shot, regardless of the programs being analyzed. This may directly benefit many other program analyses and software engineering tools that require scalable and precise pointer analysis. Moreover, we expect the ideas behind SCALER may help other kinds of static analyses to become more scalable with good precision for real-world programs.

ACKNOWLEDGMENTS

This work was supported by the European Research Council (ERC) under the FP7 and Horizon 2020 research and innovation programs (grant agreements 307334 and 647544).


