# CIL to Java-bytecode Translation for Static Analysis Leveraging

Pietro Ferrara
Julia SRL, Italy
pietro.ferrara@juliasoft.com

Agostino Cortesi
Università Ca' Foscari Venezia, Italy
cortesi@unive.it

Fausto Spoto
Università di Verona, Italy
spoto@univr.it

## ABSTRACT

A formal translation of CIL (*i.e.*, .Net) bytecode into Java bytecode is introduced and proved sound with respect to the language semantics. The resulting code is then analyzed with Julia, an industrial static analyzer of Java bytecode. The overall process of translation and analysis is fast, scales up to industrial programs, and introduces a negligible number of false alarms. The main result of this work is to leverage existing, mature, and sound analyzers for Java bytecode by applying them to the (translated) CIL bytecode.

## KEYWORDS

Static Analysis, bytecode, JVM, .Net, CIL.

## 1 INTRODUCTION

Static analysis infers, at compile-time, properties about the runtime behavior of computer programs. It allows one to verify, for instance, the absence of runtime errors or security breaches. Static analysis applies also to compiled code in assembly or bytecode format. This is particularly interesting for applications distributed on the Internet, or downloaded from public (and possibly unsafe) application repositories (*e.g.*, the Google Play Store), when the source code is not available, but the user would like to statically check some safety or security properties.

The analysis of Java bytecode for the Java Virtual Machine (from now on, JB) has a long research tradition and many analyzers exist [24]. Some analyses build on formal mathematical roots, such as abstract interpretation [13, 20, 23]. Moreover, JB makes the design of static analysis easier by requiring bytecode to be type-checkable [19] and without unsafe operations such as free pointer operations. On the contrary, CIL bytecode, that is, the compiled bytecode used for the .Net platform (from now on, just CIL) , has not received much attention from the static analysis community yet. Moreover, CIL can be used in an *unsafe* way, that is, allowing free pointer operations, which makes its static analysis harder. However, these operations are very often used in very controlled contexts, hence, in most cases, a static analyzer could possibly capture their actual behavior anyway.

Despite clear differences, JB and CIL share strong similarities, being both low-level object-oriented languages where objects are stored and shared in the heap. Hence, it is tempting to leverage mature existing static analyses and tools for JB by translating CIL into *equivalent* JB and running the tools on the latter. Obviously, this introduces issues about the exact meaning of *equivalence* between CIL and its translation into JB. Moreover, the translation should not introduce code artifacts that confuse the analyzer and should work on industrial-size CIL applications, supporting as many unsafe pointer operations as possible.

The main contribution of this work is the introduction of a translation of CIL to JB that is (i) theoretically sound, and (ii) effective in practice, so that an industrial static analyzer for JB can be applied to .Net (and in particular C#) programs. More languages compile into JB (*e.g.*, Scala) and CIL (*e.g.*, VB.Net and F#), with distinct features and code structures. Here, we focus on Java and C#, that have similar structure and compile into comparable bytecode.

We start by formalizing the concrete semantics of a representative subset of CIL and JB, and the translation of CIL into JB. Then, we prove this translation sound, that is, the concrete semantics of the initial CIL program is equivalent to that of the translated JB program. This guarantees that, if we prove a property of the JB program, then such property holds also for the original CIL program. Then we present a deep experimental evaluation over industrial-size open source popular programs, by applying the Julia static analyzer [23] to the translated JB.

We focus on three main research issues about scalability, precision, and coverage of the overall approach:

Research Question 1 (Scalability). *Does the CIL to JB translation scale up, that is, (i) can it deal with libraries of industrial size (100KLOCs) in a few minutes, and (ii) is its computational time comparable to that required by the static analysis phase?*

Research Question 2 (Precision). *Does the CIL to JB translation introduce less than 10% of the false alarms produced by the static analyzer?*

Research Question 3 (Libraries). *Despite supporting only a subset of CIL, does the CIL to JB translation succeed on at least 95% of the system libraries?*

Notice that it is crucial that a static analyzer understands the behaviour of system libraries, as otherwise it could only rely on manual annotations or on (possibly unsound) assumptions on their execution. System libraries need to access memory through unsafe pointer. Java allows such behaviour through native methods (written in languages other than Java and bound through the Java Native Interface), while .Net allows unsafe pointers in its code. In these cases, our translation produces Java native methods. We ensure that the effort of manually annotating .Net libraries is comparable to that needed for Java.

### 1.1 Related Work

Few attempts have been made in the past to translate CIL to JB. Grasshopper is probably the most popular one. However, it is not available any more[1]. As far as we can see, it was abandoned about a decade ago, and we cannot make any comparison with our translation. A similar tool was CLR2JVM [2]: it translates CIL to an intermediate $XML_{CLR}$ representation, that can be then translated into $XML_{JVM}$, and finally to JB. As far as we can see[2],

[1]We were unable to access the website http://dev.mainsoft.com, that seems to be the website of the tool from past forum discussions (http://stackoverflow.com/questions/95163/differences-between-msil-and-java-bytecode)
[2]http://xmlvm.org/documentation/

the tool should read .NET executables, but it failed parsing all the executable files of our experiments (see Sec. 5 for the complete list). This probably happened because CLR2JVM is not maintained any more (the last commit to the repository https://sourceforge.net/p/xmlvm/code/HEAD/tree/trunk/xmlvm/src/clr2jvm/ was more than six years ago), and it does not support the last CIL versions. Neither Grasshopper nor CLR2JVM has any documentation or discussion about how the translation is performed (in particular, how they handle instructions that are different from CIL to JB such as direct references). Therefore, as far as we can see our translation from CIL to JB is the only one that (i) works on recent releases of CIL and JB, and (ii) is formalized and proved sound.

Other translations between low-level languages exist, justified by the need of applying verification tools that work on a specific language only. For instance, [9] defines a translation from Boogie into WhyML and proves it sound, as we have done from CIL to JB. Similar translations work also at runtime, in particular inside a just-in-time compiler, as in [12]. However, we did not find any literature on the translation of CIL into JB for industrial-size software.

Many other static analysis tools for the .Net platform exist, in particular for C# code. There are tools that verify compliance to some guideline, such as Fxcop [22] and Coverity Prevent [15]. Other tools, such as NDepend [5] and CodeMetrics [6], provide metrics about the code under analysis. ReSharper [18] applies syntactical code inspections, finds code smells and guarantees compliance to coding standards. As far as we can see, there exist only two main fundamental tools with scientific base: (i) Spec# [11], an extension of C# with static checking of various kinds of manual specifications, and (ii) CodeContracts [21], an abstract interpretation-based static analyzer for CIL. In the Java world, the number of static analyzers based on syntactic reasoning, such as Checkstyle [1], FindBugs [4] and PMD [7], is comparable to that for .Net. However, Java attracted much more attention from the scientific community, and more semantic analyzers have been introduced during the last decade, such as CodeSonar [3], ThreadSafe [10] and Julia [23]. Few semantic analyzers, such as WALA [8], have been applied to various languages (*e.g.*, Java and JavaScript), but with ad-hoc source translations.

Our approach lets us apply all the Java analyzers on .Net programs (almost) *for free*, that is, by translating CIL into JB and using the analyzers as they are (we expect that few manual annotations are needed to improve the precision of the analysis, in particular when dealing with library calls). We have also studied performance and results with Julia. As far as we know, our work is the first translation of CIL into JB for static analysis that is proven to be *sound* and comes with evidence that this translation applies to *industrial-size software* with results that are comparable in terms of precision and efficiency to those obtained on JB.

## 2 BACKGROUND

We provide here a basic introduction to CIL and JB, with a running example and discussion on the architecture of the Julia static analyzer. For an exhaustive definition of JB and CIL, see [19] and [16], respectively.

### 2.1 CIL and JB

Bytecode is a machine-independent low-level programming language, used as target of the compilation of high-level languages, that hence becomes machine-independent. Bytecode languages are interpreted by their corresponding virtual machine, specific to each execution architecture. Both .Net and Java compile into bytecode. However, they use distinct instructions and virtual machines. .Net compiles into CIL, while Java compiles into JB. These have strong similarities: both use an operand stack for temporary values and an array of local variables standing for source code variables; both are object-oriented, with instructions for object creation, field access and virtual method dispatch. Despite these undeniable similarities, CIL and JB differ for the way of performing parameter passing (CIL uses a specific array of variables for the formal parameters, while JB merges them into the array of local variables); they handle object creation differently (CIL creates and initializes the object at the same time, while these are distinct operations in JB); they allocate memory slots differently (in CIL each value uses a slot, while JB uses 1 or 2 slots per 32- or 64- bit values, respectively); finally, CIL uses pointers explicitly, also in type-unsafe ways, while JB has no notion of pointer.

| **JB** | **CIL** | |
|---|---|---|
| `iadd`<br>`ladd` | `add` | *(arith. op.)* |
| `iload i`<br>`lload i`<br>`aload i`<br>`istore i`<br>`lstore i`<br>`astore i` | `ldloc i`<br>`stloc i`<br>`ldarg i` | *(local vars)* |
| `invokevirtual`<br>`invokestatic` | `call` | *(meth. call)* |
| `new T`<br>`getfield f`<br>`putfield f` | `newobj T(···)`<br>`ldfld f`<br>`stfld f` | *(objects)* |
| `if_icmpgt` | `bgt` | *(cond. branch)* |
| `dup`<br>`dup2` | `dup` | *(stack)* |
| | `ldloca i`<br>`stind`<br>`ldind` | *(pointers)* |

**Figure 1: JB and CIL minimal bytecode languages.**

We focus our formalization on a minimal representative subset of JB and CIL, as defined in Fig. 1. That figure presents bytecode instructions for:

**arithmetic:** JB has type-specific operations, such as `iadd` and `ladd` to add two integer or long values, respectively. Instead, CIL has generic operations, such as `add` to add two numerical values of the same type;

**local variables access:** JB has a single array of variables for both local variables and method arguments, and reads and writes values from this array through `xload` and `xstore`, where `x` is `i` for integer values, `l` for long values and `a` for references, respectively. In this array, 64 bits values use two subsequent slots. CIL, instead, uses two arrays: one for method's arguments (`ldarg i` loads the value of the i-th argument) and one for local variables (`ldloc i` and `stloc i` read and write the i-th local variable, respectively). In addition, it uses one slot both for 32- and 64-bit variables;

**method call:** JB has several kinds of method call instructions, such as `invokevirtual` and `invokestatic`. Instead, CIL has a unique `call` instruction;

**object manipulation:** in JB, instructions `new`, `getfield`, and `putfield` allocate a new object, read, and write its fields, respectively. CIL has similar instructions `newobj`, that also calls the constructor, `ldfld` and `stfld`;
**conditional branch:** JB has type-specific conditional branch instructions such as `if_icmpgt` (to branch if the greater than operator returns true on the topmost two integer values of the operand stack); CIL has generic instructions such as `bgt`;
**stack:** CIL duplicates the top value of the stack through the `dup` instruction. JB does the same with `dup` for 32 bits values and `dup2` for 64 bits values;
**pointers:** CIL contains some instructions to load the address of a local variable (`ldloca i`), and to store and load a value into the memory cell pointed by a reference (`stind` and `ldind`, respectively). Instead, JB has no direct pointer manipulation.

In the rest of this article, $\mathrm{St}_{\mathrm{CIL}}$ and $\mathrm{St}_{\mathrm{JB}}$ denote CIL and JB instructions or statements, respectively. A method (both in CIL and JB) is represented by (i) a sequence of (possibly conditional and branching) statements, and (ii) the number and static types of arguments and local variables.

## 2.2 Running Example

Fig. 2 shows the running example that Sec. 3 and 4 use to clarify the formalization. The C# code in Fig. 2a defines a class `Wrapper` that wraps an integer value, and a static method `WrappersCollection` that, given an integer `n`, returns a collection of $n$ wrappers containing values from 0 to $n-1$. Fig. 2b presents the (simplified) CIL obtained from its compilation: as usual with CIL, code is unstructured (*e.g.*, there are branches at lines 7 and 20), and each source code statement could be translated into many bytecode statements (*e.g.*, line 7 in Fig. 2a compiles into lines 3 and 4 in Fig. 2b). Fig. 2c presents the results of our translation of CIL into JB. Next sections explain the steps of the translation. First of all, notice some of the differences highlighted in Sec. 2.1. Namely, the type-generic CIL statement `stloc.1` at line 6 of Fig. 2b is translated into the type-specific `istore_2` at line 7 of Fig. 2c. Similarly, `newobj` (line 11) is translated into multiple JB statements (line 12-16). The running example contains few instructions that are not part of the minimal language defined in Sec. 2.1, and in particular (i) `ldc` and `i_const` that load constant (integer) values, (ii) `blt` and `if_icmplt` for conditional branching when an (integer) value is strictly less than another, (iii) `invokespecial` to invoke a specific method in JB, and (iv) `ret` and `areturn` to return a (reference) value.

### 2.2.1 *Example with Direct References.*

Safe C# code adopts direct pointers only for `out` and `ref` method parameters. These parameters can be assigned (and read as well in case of `ref`) inside the method, and "any change to the parameter in the called method is reflected in the calling method"[3]. In our translation, we build up wrapper objects to soundly represent their semantics.

Consider for instance the C# code in Fig. 3a. Method `init` receives a `ref` parameter and it increments it by one. This is compiled (Fig. 3b) into a method reading the value pointed by the direct reference (line 5), and writing it (line 8). Our goal is to translate this code into the Java code in Fig. 3c: we simulate the direct reference by constructing a wrapper object (line 7), assigning the value of the local variable to field `value` of the wrapper (line 8), and then propagating back the results of the call to `init` (line 9) by assigning the value of the local variable with the one stored in the wrapper (line 10). In addition, the `ref` parameter is replaced by the type of the wrapper object.

[3]https://msdn.microsoft.com/en-us/library/14akc2c7.aspx

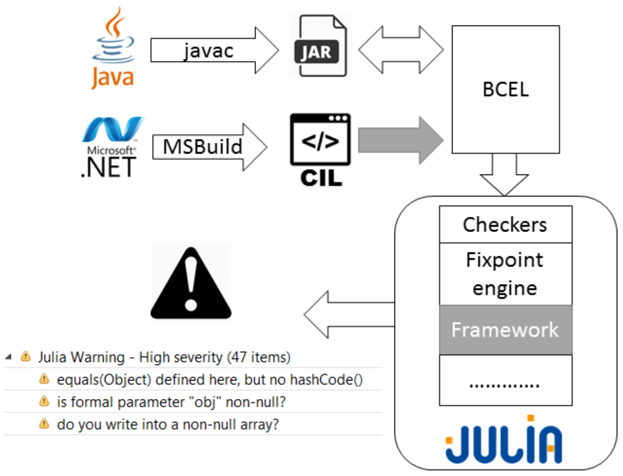
Java
javac
JAR
BCEL
Microsoft .NET
MSBuild
CIL
Checkers
Fixpoint engine
Framework
............
JULIA
Julia Warning - High severity (47 items)
equals(Object) defined here, but no hashCode()
is formal parameter "obj" non-null?
do you write into a non-null array?

**Figure 4: Julia's architecture.**

## 2.3 Julia

Julia [23] is a static analyzer for JB, based on abstract interpretation [14]. It transforms JB to basic blocks of code, that analyzes through a fixpoint algorithm. Analyses are constraint-based or denotational. Currently, Julia features around 70 checkers, including nullness, termination, synchronization and taint analysis. Since Julia works on JB, it may analyse any programming language that compiles into that bytecode. In particular, we apply Julia to the compilation of CIL into JB.

Fig. 4 is a high level view of Julia's architecture. Java code is compiled by `javac` into a `jar` file, then parsed through the BCEL library [17]. Julia receives this latter format, applies its analysis (by using many components such as the checkers, that define what properties to check, a fixpoint engine, and the framework specifying the semantics of some specific components of the programming language), and outputs a list of warnings. The added component in our approach is the translation of CIL into BCEL format (grey arrow in the upper part of Fig. 4). Since a program in BCEL format can be dumped into a jar file, we can dump a .dll file in this format. We added few annotations in a new framework of Julia to specify the main components of the .Net framework (*e.g.*, the signatures in the `Object` class).

## 3 CONCRETE SEMANTICS

We formalize here the concrete semantics of CIL and JB.

## 3.1 Notation

Let Ref and Num be the set of reference and numerical values, respectively, and let Val = Ref ∪ Num. A stack $s$ of elements in $T$ is a function in $\mathbb{N} \to T$ such that $\exists i \in \mathbb{N} : \forall i_1 \le i : i_1 \in dom(s) \wedge \forall i_2 > i : i_2 \notin dom(s)$. We will refer to $i$ as the height of $s$ ($height(s)$). Given a stack $s$ and an element $e$, $s :: e$ denotes a

```
public class Wrapper
{
  int f;
  Wrapper(int f) { this.f = f; }
  static ICollection<Wrapper> WrappersCollection(int n)
  {
    ICollection<Wrapper> result = new List<Wrapper>();
    for (int i = 0; i < n; i++)
      result.Add(new Wrapper(i));
    return result;
  }
}
```

**(a) The C# source code of the running example.**

```
1  static ICollection<Wrapper> WrappersCollection(int n)
2  {
3    newobj List<Wrapper>::.ctor()
4    stloc.0
5    ldc.0
6    stloc.1
7    br #18
8
9    ldloc.0
10   ldloc.1
11   newobj Wrapper::.ctor(int32)
12   call Generic.ICollection<Wrapper>::Add
13   ldloc.1
14   ldc.1
15   add
16   stloc.1
17
18   ldloc.1
19   ldarg.0
20   blt #9
21
22   ldloc.0
23   ret
24 }
```

**(b) The compilation into CIL of the code in (a).**

```
1  static WrappersCollection(I)LSystem/Collections/Generic/ICollection {
2    new <System/Collections/Generic/List>
3    dup
4    invokespecial <System/Collections/Generic/List.<init>>
5    astore_1
6    iconst_0
7    istore_2
8    goto 23
9
10   aload_1
11   iload_2
12   istore 3
13   new <Wrapper>
14   dup
15   iload 3
16   invokespecial <Wrapper.<init>>
17   invokevirtual <System/Collections/Generic/ICollection.Add>
18   iload_2
19   iconst_1
20   iadd
21   istore_2
22
23   iload_2
24   iload_0
25   if_icmplt 10
26
27   aload_1
28   areturn
29 }
```

**(c) The translation into JB of the CIL in (b).**

**Figure 2: The C# code, CIL, and JB of the running example.**

```
void init(ref int i)
{
  i++;
}

void run()
{
  int i=0;
  init(ref i);
}
```

**(a) C# code**

```
static void init(ref A& a)
{
  ldarg.0
  ldarg.0
  ldind.i4
  ldc.i4.1
  add
  stind.i4
}

public static int run()
{
  ldc.i4.0
  stloc.0
  ldloca.s 0
  call void Temporary.Foo::'init'(int32&)
}
```

**(b) CIL**

```
public static final void init (WrapRef i) {
  i.value = i.value + 1;
}

public static final int run() {
  int i = 0;
  WrapRef wrapper = new WrapRef();
  wrapper.value = i;
  init (wrapper);
  i = wrapper.value;
}
```

**(c) Java code**

**Figure 3: CIL code using `ref` parameters.**

stack whose top element (that is, the one with the highest index) is $e$ followed by the stack $s$.

As usual for object-oriented programming languages, an object is a map from field names to values, and a heap is a map from references to objects. Formally, Heap : Ref $\rightarrow$ Field $\rightarrow$ Val, where Field is the set containing all field names. $fresh(T, \mathrm{h}) = (r, \mathrm{h}')$ allocates an object of type $T$ in heap h and returns (i) the reference $r$ of the freshly allocated object, and (ii) the heap $\mathrm{h}'$ resulting from the allocation of memory on h.

For simplicity, we consider only integer (Int) and long (Long) numerical types (NumTypes = {Int, Long}), and references (Ref). Given a value $v$, *typeOf*($v$) returns its type (Int, Long, or Ref). Since JB instructions often append a prefix to distinguish instructions dealing with different types (*e.g.*, `iadd` and `ladd`), we define a support function *JVMprefix* that, given a type $t$, returns the prefix of the given type (*i.e.*, `i` if $t$ = Int, `l` if $t$ = Long, and `a` if $t$ = Ref). We define by WRef an object type with a unique field value.

Given a method signature `m` and a list of arguments $L$ (with the receiver in the first argument if `m` is not static), $body(\mathtt{m}, L) : \mathbb{N} \rightarrow \mathsf{St}$ returns the body of the method resolving the call, that is, a sequence of statements (represented by a function mapping indexes to statements). Similarly, each statement belongs to a method; hence $getBody(\mathtt{st}) = b$ is the body of the method where `st` occurs. Finally, *isStatic*(`m`) means that `m` is static.

## 3.2 CIL

First, we define the concrete semantics of our CIL fragment (Sec. 2.1).

*Concrete State* A local state in CIL is composed by a stack of values or reference to local variables $\mathsf{Stack} : \mathbb{N} \rightarrow \mathsf{Val} \cup \mathsf{Ref_{Loc}}$ (where $r_i \in \mathsf{Ref_{Loc}}$ represents the cell's reference of the $i$-th local variable), an array of local variables $\mathsf{Loc} : \mathbb{N} \rightarrow \mathsf{Val}$, and an array of method arguments $\mathsf{Arg} : \mathbb{N} \rightarrow \mathsf{Val}$. A concrete CIL state consists of a local state and a heap, that is, $\Sigma_{\mathsf{CIL}} = \mathsf{Stack} \times \mathsf{Loc} \times \mathsf{Arg} \times \mathsf{Heap}$.

*Concrete Semantics* Fig. 5 shows the concrete CIL semantics $\langle \mathtt{st}, \sigma \rangle \rightarrow_{\mathsf{CIL}} \sigma'$. For a statement `st` and an entry state $\sigma$, it yields the state $\sigma'$ resulting from the execution of `st` over $\sigma$; or a program label `l`, meaning that the next instruction to execute is that at `l`. Otherwise, the next instruction to execute is implicitly assumed to be the subsequent one, sequentially (if any).

For the most part, the concrete semantics just formalizes the runtime semantics of the CIL ECMA Standard [16]. For instance, rule `add` pops the two topmost values of the operand stack and replaces them with their addition. However, its semantics is defined iff the two values have the same type. Instead, `ldloc i` pushes to the operand stack the value of the $i$-th local variables, while `stloc i` stores the top of the operand stack into the $i$-th local variable. Statements working with objects, such as `ldfld` and `stfld`, read from and write into the heap, if their receiver is not `null`. `call` and `callstatic` create a frame (*i.e.*, an array of arguments, an empty array of local variables and an empty operand stack), execute the callee and leave its returned value on the stack, if any. For simplicity, the formalization assumes that there is no returned value. Finally, `ldloca i` loads the reference to the $i$-th local variable to the stack (represented by $r_i$), `stind` stores the given value to the given reference, and `ldind` loads the value pointed by the given reference.

**Running example:** Consider the running example in Fig. 2b and apply the concrete semantics when `n` is 1, that is, when the entry state consists of an empty operand stack and of an array of local variables, while the value of the arguments is $[0 \mapsto 1]$. Assume that `newobj` allocates the object at address #1. Then after the first block (lines 3-7) the address of the object is stored in local variable 0, local variable 1 (representing variable `i` of the source program) holds 0, and address #1 in the heap holds an object of type `List⟨Collection⟩`. Formally, the concrete state at line 7 is $(\emptyset, [0 \mapsto \#1, 1 \mapsto 0], [0 \mapsto 1], [\#1 \mapsto <>])$, where <> stands for the empty list. Then the body of the `for` loop (lines 9-16) is executed once and creates a new `Wrapper` object (assume at address #2) wrapping 0, adds it to the list at address #1 and increments counter `i` (*i.e.*, local variable 0) by 1. Hence the execution of the concrete semantics of the body of the loop leads to the concrete state $\sigma = (\emptyset, [0 \mapsto \#1, 1 \mapsto 1], [0 \mapsto 1], [\#1 \mapsto < \#2 >, \#2 \mapsto \{f \mapsto 0\}])$ at line 16. The condition at line 20 will then route the program to line 22 (since both argument 0 and local variable 1 hold 1). In conclusion, the concrete semantics will reach statement `ret` at line 23 with a state $\sigma_{\mathsf{CIL}}$ equal to $\sigma$, but where the operand stack contains reference #1.

## 3.3 JB

Let us turn to the concrete semantics of the JB fragment (Sec. 2.1).

*Concrete State* A local state in JB is composed by a stack of values $\mathsf{Stack} : \mathbb{N} \rightarrow \mathsf{Val}$ and an array of local variables $\mathsf{Loc} : \mathbb{N} \rightarrow \mathsf{Val}$. A concrete JB state consists of a local state and a heap, that is, $\Sigma_{\mathsf{JB}} = \mathsf{Stack} \times \mathsf{Loc} \times \mathsf{Heap}$.

*Concrete Semantics* Fig. 6 reports the concrete JB semantics $\rightarrow_{\mathsf{JB}}$. For a statement `st` and an entry state $\sigma$, it yields the state $\sigma'$ resulting from the execution of `st` over $\sigma$.

For the most part, the behavior of this semantics is identical to that for CIL. The main differences are that JB instructions work on specific types (*e.g.*, while CIL `add` statement adds two values of the same type, JB `iadd` and `ladd` statements add the values iff they are both `int` or `long`, respectively), and `new` only allocates a new object, while CIL `newobj` statements also calls a constructor.

**Running example:** The application of the JB concrete semantics is similar to that for CIL, but there are two minor differences: (i) there is only one array of local variables representing both CIL arguments and local variables (*e.g.*, CIL local variable 0 is represented by JB local variable 1, since the first local variable holds the argument of the method); and (ii) there is an *instrumentation* local variable at index 3. Therefore, after we apply the JB concrete semantics from the entry state that maps the argument to 1, we obtain the concrete state $([0 \mapsto \#1], [0 \mapsto 1, 1 \mapsto \#1, 2 \mapsto 1, 3 \mapsto 0], [\#1 \mapsto < \#2 >, \#2 \mapsto \{f \mapsto 0\}])$.

## 4 FROM CIL TO JB

This section formalizes and proves correct the translation of CIL statements into JB statements.

## 4.1 Concrete States

Function $\mathbb{T}_\sigma[\![\ ]\!] : \Sigma_{\mathsf{CIL}} \rightarrow \Sigma_{\mathsf{JB}}$ translates CIL concrete states into JB concrete states: $\mathbb{T}_\sigma[\![(\mathsf{s}, \mathsf{l}, \mathsf{a}, \mathsf{h})]\!] = (\mathsf{s}', cnvrtLoc(\mathsf{l}, \mathsf{a}), \mathsf{h}'_l)$

This function (i) replaces direct reference with wrapper objects, and (ii) merges the array of local variables and arguments, adjusting variable indexes for 64 bits values. Formally:

$$
\begin{array}{l}
i \in dom(\mathsf{s}), l = height(\mathsf{s}) \\
\mathsf{s}' = \left[ i \mapsto \begin{cases} \mathsf{s}(i) & \text{if } \mathsf{s}(i) \in \mathsf{Val} \\ r_i & \text{if } \mathsf{s}(i) \in \mathsf{Ref_{Loc}} \end{cases} \right] \\
\text{where } \mathsf{h}'_{-1} = \mathsf{h}, \text{ and } (\mathsf{h}'_i, \mathsf{r}_i) = allocWrp(\mathsf{h}'_{i-1}, i) \\
allocWrp(\mathsf{h}, j) = \begin{cases} (\mathsf{h}, \mathtt{null}) \text{ if } \mathsf{s}(j) \in \mathsf{Val} & \\ \quad (\mathsf{h}'[r \mapsto \mathsf{h}(r)[\mathtt{value} \mapsto \mathsf{l}(j)]]) & \text{if } \mathsf{s}(i) \in \mathsf{Ref_{Loc}} \\ \qquad \text{where } (r, \mathsf{h}') = fresh(\mathsf{WRef}, \mathsf{h}) & \end{cases}
\end{array}
$$

$$\frac{\mathit{typeOf}(v_1) = \mathit{typeOf}(v_2)}{\langle \mathsf{add}, (s :: v_1 :: v_2, \mathsf{l}, \mathsf{a}, \mathsf{h})\rangle \rightarrow_{\mathsf{CIL}} (s :: (v_1 + v_2), \mathsf{l}, \mathsf{a}, \mathsf{h})}\ (\mathsf{add}) \qquad \frac{}{\langle \texttt{ldloc i}, (s, \mathsf{l}, \mathsf{a}, \mathsf{h})\rangle \rightarrow_{\mathsf{CIL}} (s :: \mathsf{l}(\mathtt{i}), \mathsf{l}, \mathsf{a}, \mathsf{h})}\ (\texttt{ldloc})$$

$$\frac{}{\langle \texttt{stloc i}, (s :: \mathsf{v}, \mathsf{l}, \mathsf{a}, \mathsf{h})\rangle \rightarrow_{\mathsf{CIL}} (s :: \mathsf{l}[\mathtt{i} \mapsto \mathsf{v}], \mathsf{a}, \mathsf{h})}\ (\texttt{stloc}) \qquad \frac{}{\langle \texttt{ldarg i}, (s :: \mathsf{l}, \mathsf{a}, \mathsf{h})\rangle \rightarrow_{\mathsf{CIL}} (s :: \mathsf{a}(\mathtt{i}), \mathsf{l}, \mathsf{a}, \mathsf{h})}\ (\texttt{ldarg})$$

$$\frac{\begin{array}{c}\mathit{isStatic}(\mathsf{m}(\mathtt{arg_0}, \cdots, \mathtt{arg_i})) = \mathsf{false} \wedge t \neq \mathtt{null} \wedge \\ \langle \mathit{body}(\mathsf{m}(\mathtt{arg_0}, \cdots, \mathtt{arg_i}), (t, v_1, \cdots, v_i)), (\varepsilon, \emptyset, [0 \mapsto t, j \mapsto v_j : j \in [1..i]], \mathsf{h})\rangle \rightarrow_{\mathsf{CIL}} (\mathsf{s}', \mathsf{l}', \mathsf{a}', \mathsf{h}')\end{array}}{\langle \mathsf{call}\ \mathsf{m}(\mathtt{arg_1}, \cdots, \mathtt{arg_i}), (s :: t :: v_1 :: \cdots :: v_i, \mathsf{l}, \mathsf{a}, \mathsf{h})\rangle \rightarrow_{\mathsf{CIL}} (s, \mathsf{l}, \mathsf{a}, \mathsf{h}')}\ (\texttt{call})$$

$$\frac{\begin{array}{c}\mathit{isStatic}(\mathsf{m}(\mathtt{arg_0}, \cdots, \mathtt{arg_i})) = \mathsf{true} \wedge \\ \langle \mathit{body}(\mathsf{m}(\mathtt{arg_0}, \cdots, \mathtt{arg_i}), (v_1, \cdots, v_i)), (\varepsilon, \emptyset, [j-1 \mapsto v_j : j \in [1..i]], \mathsf{h})\rangle \rightarrow_{\mathsf{CIL}} (\mathsf{s}', \mathsf{l}', \mathsf{a}', \mathsf{h}')\end{array}}{\langle \mathsf{call}\ \mathsf{m}(\mathtt{arg_1}, \cdots, \mathtt{arg_i}), (s :: v_1 :: \cdots :: v_i, \mathsf{l}, \mathsf{a}, \mathsf{h})\rangle \rightarrow_{\mathsf{CIL}} (s, \mathsf{l}, \mathsf{a}, \mathsf{h}')}\ (\texttt{callstatic})$$

$$\frac{\mathit{fresh}(\mathsf{T}, \mathsf{h}) = (r, \mathsf{h}_1) \wedge \langle \mathit{body}(\mathtt{ctor}(\mathtt{arg_1}, \cdots, \mathtt{arg_i}), (v_1, \cdots, v_i)), (\varepsilon, \emptyset, [0 \mapsto r, j \mapsto v_j : j \in [1..i]], \mathsf{h}_1)\rangle \rightarrow_{\mathsf{CIL}} (\mathsf{s}', \mathsf{l}', \mathsf{a}', \mathsf{h}')}{\langle \mathsf{newobj}\ \mathsf{T}(\mathsf{a}_1, \cdots, \mathsf{a}_i), (s :: v_1 :: \cdots :: v_i, \mathsf{l}, \mathsf{a}, \mathsf{h})\rangle \rightarrow_{\mathsf{CIL}} (s :: r, \mathsf{l}, \mathsf{a}, \mathsf{h}')}\ (\mathsf{newobj})$$

$$\frac{o \neq \mathtt{null}}{\langle \texttt{ldfld}\ \mathsf{f}, (s :: o, \mathsf{l}, \mathsf{a}, \mathsf{h})\rangle \rightarrow_{\mathsf{CIL}} (s :: \mathsf{h}(o)(\mathsf{f}), \mathsf{l}, \mathsf{a}, \mathsf{h})}\ (\texttt{ldfld}) \qquad \frac{o \neq \mathtt{null} \quad s' = \mathsf{h}(o)[\mathsf{f} \mapsto v]}{\langle \texttt{stfld}\ \mathsf{f}, (s :: o :: v, \mathsf{l}, \mathsf{a}, \mathsf{h})\rangle \rightarrow_{\mathsf{CIL}} (s, \mathsf{l}, \mathsf{a}, \mathsf{h}[o \mapsto s'])}\ (\texttt{stfld})$$

$$\frac{\mathit{typeOf}(v_1) = \mathit{typeOf}(v_2) \wedge v_1 > v_2}{\langle \mathsf{bgt}\ \mathtt{l}, (s :: v_1 :: v_2, \mathsf{l}, \mathsf{a}, \mathsf{h})\rangle \rightarrow_{\mathsf{CIL}} \langle \mathtt{l}, (s, \mathsf{l}, \mathsf{a}, \mathsf{h})\rangle}\ (\mathsf{bgt}\ \texttt{true}) \qquad \frac{\mathit{typeOf}(v_1) = \mathit{typeOf}(v_2) \wedge v_1 \leq v_2}{\langle \mathsf{bgt}\ \mathtt{l}, (s :: v_1 :: v_2, \mathsf{l}, \mathsf{a}, \mathsf{h})\rangle \rightarrow_{\mathsf{CIL}} (s, \mathsf{l}, \mathsf{a}, \mathsf{h})}\ (\mathsf{bgt}\ \texttt{false})$$

$$\frac{}{\langle \texttt{ldloca i}, (s, \mathsf{l}, \mathsf{a}, \mathsf{h})\rangle \rightarrow_{\mathsf{CIL}} (s :: r_i, \mathsf{l}, \mathsf{a}, \mathsf{h})}\ (\texttt{ldloca}) \qquad \frac{}{\langle \texttt{stind}, (s :: r_i :: v, \mathsf{l}, \mathsf{a}, \mathsf{h})\rangle \rightarrow_{\mathsf{CIL}} (s, \mathsf{l}, \mathsf{a}, \mathsf{h}[r_i \mapsto v])}\ (\texttt{stind})$$

$$\frac{}{\langle \mathsf{dup}, (s :: v, \mathsf{l}, \mathsf{a}, \mathsf{h})\rangle \rightarrow_{\mathsf{CIL}} (s :: v :: v, \mathsf{l}, \mathsf{a}, \mathsf{h})}\ (\mathsf{dup}) \qquad \frac{}{\langle \texttt{ldind}, (s :: r_i, \mathsf{l}, \mathsf{a}, \mathsf{h})\rangle \rightarrow_{\mathsf{CIL}} (s :: \mathsf{h}(r_i), \mathsf{l}, \mathsf{a}, \mathsf{h})}\ (\texttt{ldind})$$

**Figure 5: Concrete CIL semantics.**

$$\frac{\mathit{typeOf}(v) \neq \mathsf{Long}}{\langle \mathsf{dup}, (s :: v, \mathsf{l}, \mathsf{h})\rangle \rightarrow_{\mathsf{JB}} (s :: v :: v, \mathsf{l}, \mathsf{h})}\ (\mathsf{dup})$$

$$\frac{\mathit{typeOf}(v_1) \neq \mathsf{Long}}{\langle \mathsf{dup2}, (s :: v_1 :: v_2, \mathsf{l}, \mathsf{h})\rangle \rightarrow_{\mathsf{JB}} (s :: v_1 :: v_2 :: v_1 :: v_2, \mathsf{l}, \mathsf{h})}\ (\mathsf{dup2\ 32}) \qquad \frac{\mathit{typeOf}(v) = \mathsf{Long}}{\langle \mathsf{dup2}, (s :: v, \mathsf{l}, \mathsf{h})\rangle \rightarrow_{\mathsf{JB}} (s :: v :: v, \mathsf{l}, \mathsf{h})}\ (\mathsf{dup2\ 64})$$

$$\frac{\mathit{typeOf}(v_1) = \mathsf{Int} \wedge \mathit{typeOf}(v_2) = \mathsf{Int}}{\langle \texttt{iadd}, (s :: v_1 :: v_2, \mathsf{l}, \mathsf{h})\rangle \rightarrow_{\mathsf{JB}} (s :: (v_1 + v_2), \mathsf{l}, \mathsf{h})}\ (\texttt{iadd}) \qquad \frac{\mathit{typeOf}(v_1) = \mathsf{Long} \wedge \mathit{typeOf}(v_2) = \mathsf{Long}}{\langle \texttt{ladd}, (s :: v_1 :: v_2, \mathsf{l}, \mathsf{h})\rangle \rightarrow_{\mathsf{JB}} (s :: (v_1 + v_2), \mathsf{l}, \mathsf{h})}\ (\texttt{ladd})$$

$$\frac{x = \mathit{JVMprefix}(\mathit{typeOf}(\mathsf{l}(\mathtt{i})))}{\langle x\texttt{load i}, (s, \mathsf{l}, \mathsf{h})\rangle \rightarrow_{\mathsf{JB}} (s :: \mathsf{l}(\mathtt{i}), \mathsf{l}, \mathsf{h})}\ (x\texttt{load}) \qquad \frac{x = \mathit{JVMprefix}(\mathit{typeOf}(\mathsf{v}))}{\langle x\texttt{store i}, (s :: \mathsf{v}, \mathsf{l}, \mathsf{h})\rangle \rightarrow_{\mathsf{JB}} (s, \mathsf{l}[\mathtt{i} \mapsto \mathsf{v}], \mathsf{h})}\ (x\texttt{store})$$

$$\frac{\begin{array}{c}\mathit{isStatic}(\mathsf{m}(\mathtt{arg_0}, \cdots, \mathtt{arg_i})) = \mathsf{false} \wedge t \neq \mathtt{null} \wedge \\ \langle \mathit{body}(\mathsf{m}(\mathtt{arg_0}, \cdots, \mathtt{arg_i}), (t, v_1, \cdots, v_i)), ([], [0 \mapsto t, j \mapsto v_j : j \in [1..i]], \mathsf{h})\rangle \rightarrow_{\mathsf{JB}} (\mathsf{s}', \mathsf{l}', \mathsf{h}')\end{array}}{\langle \texttt{invokevirtual}\ \mathsf{m}(\mathtt{arg_1}, \cdots, \mathtt{arg_i}), (s :: t :: v_1 :: \cdots :: v_i, \mathsf{l}, \mathsf{h})\rangle \rightarrow_{\mathsf{JB}} (s, \mathsf{l}, \mathsf{h}')}\ (\texttt{invokevirtual})$$

$$\frac{\begin{array}{c}\mathit{isStatic}(\mathsf{m}(\mathtt{arg_0}, \cdots, \mathtt{arg_i})) = \mathsf{true} \wedge \\ \langle \mathit{body}(\mathsf{m}(\mathtt{arg_0}, \cdots, \mathtt{arg_i}), (v_1, \cdots, v_i)), ([], [j-1 \mapsto v_j : j \in [1..i]], \mathsf{h})\rangle \rightarrow_{\mathsf{JB}} (\mathsf{s}', \mathsf{l}', \mathsf{h}')\end{array}}{\langle \texttt{invokestatic}\ \mathsf{m}(\mathtt{arg_1}, \cdots, \mathtt{arg_i}), (s :: v_1 :: \cdots :: v_i, \mathsf{l}, \mathsf{h})\rangle \rightarrow_{\mathsf{JB}} (s, \mathsf{l}, \mathsf{h}')}\ (\texttt{invokestatic})$$

$$\frac{\mathit{fresh}(\mathsf{T}, \mathsf{h}) = (r, \mathsf{h}')}{\langle \mathsf{new}\ \mathsf{T}, (s, \mathsf{l}, \mathsf{h})\rangle \rightarrow_{\mathsf{JB}} (s :: r, \mathsf{l}, \mathsf{h}')}\ (\mathsf{new}) \qquad \frac{o \neq \mathtt{null}}{\langle \texttt{getfield}\ \mathsf{f}, (s :: o, \mathsf{l}, \mathsf{h})\rangle \rightarrow_{\mathsf{JB}} (s :: \mathsf{h}(o)(\mathsf{f}), \mathsf{l}, \mathsf{h})}\ (\texttt{getfield})$$

$$\frac{o \neq \mathtt{null} \quad s' = \mathsf{h}(o)[\mathsf{f} \mapsto v]}{\langle \texttt{putfield}\ \mathsf{f}, (s :: o :: v, \mathsf{l}, \mathsf{h})\rangle \rightarrow_{\mathsf{JB}} (s, \mathsf{l}, \mathsf{h}[o \mapsto s'])}\ (\texttt{putfield}) \qquad \frac{\mathit{typeOf}(v_1) = \mathsf{Int} \wedge \mathit{typeOf}(v_2) = \mathsf{Int} \wedge v_1 > v_2}{\langle \texttt{if\_icmpgt l}, (s :: v_1 :: v_2, \mathsf{l}, \mathsf{h})\rangle \rightarrow_{\mathsf{JB}} \langle \mathtt{l}, (s, \mathsf{l}, \mathsf{h})\rangle}\ \begin{pmatrix}\texttt{if\_icmpgt} \\ \texttt{true}\end{pmatrix}$$

$$\frac{\mathit{typeOf}(v_1) = \mathsf{Int} \wedge \mathit{typeOf}(v_2) = \mathsf{Int} \wedge v_1 \leq v_2}{\langle \texttt{if\_icmpgt l}, (s :: v_1 :: v_2, \mathsf{l}, \mathsf{h})\rangle \rightarrow_{\mathsf{JB}} (s, \mathsf{l}, \mathsf{h})}\ \begin{pmatrix}\texttt{if\_icmpgt} \\ \texttt{false}\end{pmatrix}$$

**Figure 6: Concrete JB semantics.**

Intuitively, each direct reference in the operand stack (that is, $s(i) \in \mathsf{Ref}_{\mathsf{Loc}}$) is replaced by another reference pointing to a wrapper object freshly allocated and containing in its field the value pointed by the original direct reference.

Then, for an array of values $b$ and an index $i$, the following function counts the 64 bits types among the first $i$:

$$64_b^i = |\{j : 0 \leq j < i \text{ and } b[j] \text{ is a 64 bit value}\}|$$

Then the array *cnvrtLoc*(l, a) is defined as follows:

$$\forall 0 \le i < |\mathsf{a}| : \mathit{cnvrtLoc}(\mathsf{l},\mathsf{a})[i + 64^i_{\mathsf{a}}] = \mathsf{a}[i]$$

$$\forall 0 \le i < |\mathsf{l}| : \mathit{cnvrtLoc}(\mathsf{l},\mathsf{a})[|\mathsf{a}| + 64^{|\mathsf{a}|}_{\mathsf{a}} + i + 64^i_{\mathsf{l}}] = \mathsf{l}[i]$$

**Running example:** Consider the CIL exit state computed in Sec. 3.2, that is, $\sigma_{\mathsf{CIL}} = ([0 \mapsto \#1], [0 \mapsto \#1, 1 \mapsto 1], [0 \mapsto 1], [\#1 \mapsto < \#2 >, \#2 \mapsto \{f \mapsto 0\}])$. The CIL to JB translation computes $\mathbb{T}_\sigma[\![\sigma_{\mathsf{CIL}}]\!] = ([0 \mapsto \#1], [0 \mapsto 1, 1 \mapsto \#1, 2 \mapsto 1], [\#1 \mapsto < \#2 >, \#2 \mapsto \{f \mapsto 0\}])$ (it merges CIL arguments and local variables). This state is almost identical to the exit state of the JB concrete semantics applied to the running example (Sec. 3.3) but for the instrumentation local variable at index 3.

## 4.2 Statements

Fig. 7 formalizes the translation $\mathbb{T}[\![\mathtt{st}_{\mathsf{CIL}}, \mathsf{K}]\!] = \mathtt{st}_{\mathsf{JB}}$ of a single CIL statement into a sequence of JB statements. *K* is the static type information about locals, arguments and stack elements, computed at $\mathtt{st}_{\mathsf{CIL}}$ by a standard algorithm [16]. In particular, types and height of the stack are fixed and statistically known at each bytecode. In addition, the forth component $\overline{\mathsf{w}}$ is a stack of elements in $\bot \cup (\mathbb{N}\times\mathbb{N})$ that, for each element in the operand stack, tells (i) $\bot$ if it is not a direct reference, or (ii) $(i, j)$ where $i$ is the index of the local variables pointed by the direct reference[4], and $j$ the index of the local instrumentation variable containing a pointer to the wrapper simulating the direct reference.

Few CIL statements (namely, `ldfld` and `stfld`) have a one-to-one translation into a JB statement (`getfield` and `putfield`). The statements reading and writing local variables and arguments (`ldarg`, `ldloc`, and `stloc`) are translated into their JB counterpart (*x*`load`, *x*`store`, respectively) taking into account the type of the value at the top of the stack, and adjusting the index of the variable taking into account arguments and 64 bit variables. Some CIL statements (`dup`) get translated into different JB statements on the basis of contextual information such as the type of values in the operand stack (`dup` and `dup2`). Other CIL statements can be translated only if the type of the values in the operand stack is numeric: (i) `add` can be translated into `ladd` and `iadd`, and (ii) `bgt` to `is_icmpgt`; if they are applied to references (as in generic CIL code), then the code is unsafe and we do not support its translation.

`call` requires to (i) translate the method call to the corresponding static or dynamic invocation statement in JB, and (ii) to propagate the side effects on direct pointers passed to the method as `out`/`ref` parameters to the local variables of the callee.

The translation of `newobj` is tricky because of the different patterns used in CIL and JB for object creation[5]. While CIL creates and initializes the object (*i.e.*, calls its constructor) with a single instruction, JB splits these operations and requires the newly created object to occur below the arguments on the stack, before calling the constructor. Hence, the translation relies on a function *freshIdx* to store and load the values of the constructor arguments through instrumentation local variables. In particular, given a CIL method *m*, the number and types of arguments and local variables of the method are known (Sec. II.15.4 of [16]). Therefore, function *cnvrtLoc* tells which local variables the translated JB method already uses. Then, for each argument of each `newobj` statement in *m*, it is possible to allocate a fresh local variable to store and load its value. In this way, the translation allocates a new object and puts its address below the constructor arguments.

Instructions dealing with direct pointers (namely, `ldloca`, `stind`, and `ldind` in our minimal language) are translated through equivalent CIL instructions dealing with wrapper objects (and their field `value`). Therefore, `stind` and `ldind` are simply translated through equivalent write and read of field `value`, respectively. `ldloca` instead requires to allocate a wrapper object `newobj`, stores a reference to the wrapper (`stloc`) in an instrumentation variable obtained through *freshIdx*, stores the value pointed by the direct reference in the local variable to its field `value` (`ldloc` and `stfld`), and leaves a reference to the wrapper object in the operand stack (`dup`).

In general, each CIL statement is translated into one or more JB statements, hence offsets are not preserved. Thus, function *statementIdx* : St → ℕ yields the JB offset of the first statement in the translation of the given CIL statement. In addition, since direct references are replaced by wrapper objects, when a method parameter has a direct reference type `&T` (and this happens when it is a `ref` or `out` parameter in safe C#), this is replaced by a wrapper object WRef.

**Running example:** Consider the running example in Fig. 2. Most CIL statements are translated into a single JB statement (*e.g.*, lines 18-20 and 22-23 of Fig. 2b are translated into lines 23-25 and 27-28 of Fig. 2c), with the noticeable exception of the CIL `newobj` statements at line 3 and 11, translated into lines 2-4 and 12-16, respectively. The former passes no argument to the constructor; the latter (that instantiates a `Wrapper`) calls a constructor with an argument, hence requiring an instrumentation variable at index 3.
**Direct references** As sketched in Sec. 2.2.1, we model the semantics of pointers in safe C# code through wrapper objects. In particular, `ldind` (line 5 of Fig. 3b) is translated into the field access `i.value` (right side of the assignment at line 2 of Fig. 3c), while `stind` (line 8 of Fig. 3b) is translated into the assignment of `i.value` (left side of the assignment at line 2 of Fig. 3c). In addition, `ldloca` (line 15 of Figure 3b) leads to the construction and assignment of a wrapper object (lines 7 and 8 of Figure 3c), while after the method call the value contained in the wrapper object is written into the local variable (line 10 of Figure 3c).

## 4.3 Correctness

We prove the translation from CIL to JB being correct. Namely, given a concrete CIL state $\sigma_{\mathsf{CIL}}$, by applying the operational semantics for a statement `st`, we get a state that, when translated into JB, is exactly the one resulting from the translation of $\sigma_{\mathsf{CIL}}$ into JB and the application of the JB semantics to it:

$$\begin{gathered}\forall \mathtt{st} \in \mathsf{St}_{\mathsf{CIL}}, \sigma_{\mathsf{CIL}} \in \Sigma_{\mathsf{CIL}} : \\ \langle \mathtt{st}, \sigma_{\mathsf{CIL}} \rangle \rightarrow_{\mathsf{CIL}} \sigma'_{\mathsf{CIL}} \text{ and } \langle \mathbb{T}[\![\mathtt{st}, \mathsf{K}]\!], \mathbb{T}_\sigma[\![\sigma_{\mathsf{CIL}}]\!] \rangle \rightarrow_{\mathsf{JB}} \sigma'_{\mathsf{JB}} \\ \Downarrow \\ \mathbb{T}_\sigma[\![\sigma'_{\mathsf{CIL}}]\!] =_\bullet \sigma'_{\mathsf{JB}}\end{gathered}$$

where $\sigma_1 =_\bullet \sigma_2$ means that the two states are equal up to instrumentation variables introduced by the translation process. Formally, let

[4]Since the language we introduced in Fig. 1 supports only `ldloca` to get a direct pointer, we need to track only this information in the formalization.
[5]For sake of simplicity, we assume the constructor does not have `out`/`ref` parameters. In the implementation, they are treated as for `call` statements

$$
\begin{aligned}
\mathbb{T}[\![\mathsf{dup}, \bar{\mathsf{s}} :: \mathsf{t}, \bar{\mathsf{l}}, \bar{\mathsf{a}}, \overline{\mathsf{w}}]\!] &= \begin{cases} \mathsf{dup} & \text{if } t \neq \mathsf{Long} \\ \mathsf{dup2} & \text{if } t = \mathsf{Long} \end{cases} \\
\mathbb{T}[\![\mathsf{add}, \bar{\mathsf{s}} :: \mathsf{t}_1 :: \mathsf{t}_2, \bar{\mathsf{l}}, \bar{\mathsf{a}}, \overline{\mathsf{w}}]\!] &= \begin{cases} \mathtt{iadd} & \text{if } t_1 = t_2 = \mathsf{Int} \\ \mathtt{ladd} & \text{if } t_1 = t_2 = \mathsf{Long} \end{cases} \\
\mathbb{T}[\![\mathtt{ldloc\ i}, \bar{\mathsf{s}}, \bar{\mathsf{l}}, \bar{\mathsf{a}}, \overline{\mathsf{w}}]\!] &= x\mathtt{load}\ j \text{ where } j = |a| + 64_a^{|a|} + i + 64_l^i \land x = \mathit{JVMprefix}(\mathit{typeOf}(\bar{\mathsf{l}}(\mathtt{i}))) \\
\mathbb{T}[\![\mathtt{stloc\ i}, \bar{\mathsf{s}} :: \mathsf{t}, \bar{\mathsf{l}}, \bar{\mathsf{a}}, \overline{\mathsf{w}}]\!] &= x\mathtt{store}\ j \text{ where } j = |a| + 64_a^{|a|} + i + 64_l^i \land x = \mathit{JVMprefix}(\mathit{typeOf}(\bar{\mathsf{l}}(\mathtt{i}))) \\
\mathbb{T}[\![\mathtt{ldarg\ i}, \bar{\mathsf{s}}, \bar{\mathsf{l}}, \bar{\mathsf{a}}, \overline{\mathsf{w}}]\!] &= x\mathtt{load}\ j \text{ where } j = i + 64_a^i \land x = \mathit{JVMprefix}(\mathit{typeOf}(\bar{\mathsf{a}}(\mathtt{i}))) \\
\mathbb{T}[\![\mathtt{call}\ \mathsf{m}(\mathsf{arg}_1, \cdots, \mathsf{arg}_\mathsf{i}), &= \mathtt{invoke}\ ;\ \mathtt{aload}\ p^1_{idx_1}\ ;\ \mathtt{getfield\ value}\ ;\ x_{idx_1}\mathtt{store}\ p^2_{idx_1}\ ;\ \cdots \\
\bar{\mathsf{s}} :: t_1 :: \cdots :: t_i, \bar{\mathsf{l}}, \bar{\mathsf{a}}, \overline{\mathsf{w}} :: p_1 :: \cdots :: p_i]\!] &\qquad \cdots \mathtt{aload}\ p^1_{idx_j}\ ;\ \mathtt{getfield\ value}\ ;\ x_{idx_j}\mathtt{store}\ p^2_{idx_j}\ ; \\
&\quad \text{where } \mathtt{invoke} = \begin{cases} \mathtt{invokestatic}\ \mathsf{m}(\mathsf{arg}_1, \cdots, \mathsf{arg}_\mathsf{i}) & \text{if } \mathit{isStatic}(\mathsf{m}(\mathsf{arg}_1, \cdots, \mathsf{arg}_\mathsf{i})) \\ \mathtt{invokevirtual}\ \mathsf{m}(\mathsf{arg}_1, \cdots, \mathsf{arg}_\mathsf{i}) & \text{otherwise} \end{cases} \\
&\quad \{idx_1, \cdots, idx_j\} = \{k : \mathsf{arg}_\mathsf{k} \in \mathsf{Ref}_{\mathsf{Loc}}\} \\
&\quad \forall k \in [1..j] : x_{idx_k} = \mathit{JVMprefix}(\mathit{typeOf}(\bar{\mathsf{l}}(p^2_{idx_k}))) \land \forall r \in [1..i] : p_i = (p^1_i, p^2_j) \\
\mathbb{T}[\![\mathsf{newobj}\ \mathsf{T}(\mathsf{a}_1, \cdots, \mathsf{a}_\mathsf{i}), &= x_i\mathtt{store}\ idx_i\ ;\ \cdots\ ;\ x_1\mathtt{store}\ idx_1\ ;\ \mathsf{new\ T}\ ;\ \mathsf{dup}\ ; \\
\bar{\mathsf{s}} :: t_1 :: \cdots :: t_i, \bar{\mathsf{l}}, \bar{\mathsf{a}}, \overline{\mathsf{w}}]\!] &\qquad x_1\mathtt{load}\ idx_1\ ;\ \cdots\ ;\ x_i\mathtt{load}\ idx_i\ ;\ \mathtt{invokevirtual}\ <\mathtt{init}>(\mathsf{arg}_1, \cdots, \mathsf{arg}_\mathsf{i}) \\
&\quad \text{where } \forall j \in [1..i] : x_j = \mathit{JVMprefix}(a_j) \land idx_j = \mathit{freshIdx}(\mathsf{newobj\ T}(\mathsf{a}_1, \cdots, \mathsf{a}_\mathsf{i}), j) \\
\mathbb{T}[\![\mathtt{ldfld}\ \mathsf{f}, \bar{\mathsf{s}} :: \mathsf{t_o}, \bar{\mathsf{l}}, \bar{\mathsf{a}}, \overline{\mathsf{w}}]\!] &= \mathtt{getfield}\ \mathsf{f} \\
\mathbb{T}[\![\mathtt{stfld}\ \mathsf{f}, \bar{\mathsf{s}} :: \mathsf{t_o} :: \mathsf{t_v}, \bar{\mathsf{l}}, \bar{\mathsf{a}}, \overline{\mathsf{w}}]\!] &= \mathtt{putfield}\ \mathsf{f} \\
\mathbb{T}[\![\mathsf{bgt\ k}, \bar{\mathsf{s}} :: \mathsf{t}_1 :: \mathsf{t}_2, \bar{\mathsf{l}}, \bar{\mathsf{a}}, \overline{\mathsf{w}}]\!] &= \mathtt{if_icmpgt}\ \mathsf{k}' \text{ where } k' = \mathit{statementIdx}(\mathit{getBody}(\mathsf{bgt\ k})(k)) \text{ if } t_1 = t_2 = \mathsf{Int} \\
\mathbb{T}[\![\mathtt{ldloca\ i}, \bar{\mathsf{s}}, \bar{\mathsf{l}}, \bar{\mathsf{a}}, \overline{\mathsf{w}}]\!] &= \mathbb{T}[\![\mathtt{newobj\ WrapRef()}\ ; \mathtt{dup2}\ ; \mathtt{stloc\ j}; \mathtt{ldloc\ i}; \mathtt{stfld\ value}, \bar{\mathsf{s}}, \bar{\mathsf{l}}, \bar{\mathsf{a}}, \overline{\mathsf{w}}]\!] \\
&\quad \text{where } j = \mathit{freshIdx}(\mathtt{ldloca\ i}, 0) \\
\mathbb{T}[\![\mathtt{stind}, \bar{\mathsf{s}}, \bar{\mathsf{l}}, \bar{\mathsf{a}}, \overline{\mathsf{w}}]\!] &= \mathbb{T}[\![\mathtt{stfld\ value}, \bar{\mathsf{s}}, \bar{\mathsf{l}}, \bar{\mathsf{a}}, \overline{\mathsf{w}}]\!] \\
\mathbb{T}[\![\mathtt{ldind}, \bar{\mathsf{s}}, \bar{\mathsf{l}}, \bar{\mathsf{a}}, \overline{\mathsf{w}}]\!] &= \mathbb{T}[\![\mathtt{ldfld\ value}, \bar{\mathsf{s}}, \bar{\mathsf{l}}, \bar{\mathsf{a}}, \overline{\mathsf{w}}]\!]
\end{aligned}
$$

**Figure 7: Translation of CIL statements into JB.**

$\sigma_1 = (\mathsf{s}_1, \mathtt{l}_1^1 :: \ldots :: \mathtt{l}_1^\mathtt{n}, \mathsf{h}_1)$ and $\sigma_2 = (\mathsf{s}_2, \mathtt{l}_2^1 :: \ldots :: \mathtt{l}_2^\mathtt{n}, :: \mathtt{l}_2^{\mathtt{n}+1} :: \ldots :: \mathtt{l}_2^{\mathtt{n}+\mathtt{k}}, \mathsf{h}_2)$, then $\sigma_1 =_\bullet \sigma_2$ iff $\mathsf{s}_1 = \mathsf{s}_2$, and $\forall \mathtt{i} \leq \mathtt{n} : \mathtt{l}_1^\mathtt{i} = \mathtt{l}_2^\mathtt{i}$, and $\mathsf{h}_1 = \mathsf{h}_2$. Note that instrumentation variables are present only in the JB state, hence in the right hand-side of the equality. [6]

**Running example:** Sec. 3.2 showed that, starting from $\sigma_{\mathsf{CIL}} = (\emptyset, \emptyset, [0 \mapsto 1], \emptyset)$, the concrete semantics on the program in Fig. 2b ends up in $\sigma'_{\mathsf{CIL}} = ([0 \mapsto \#1], [0 \mapsto \#1, 1 \mapsto 1], [0 \mapsto 1], [\#1 \mapsto < \#2 >, \#2 \mapsto \{f \mapsto 0\}])$. Then Sec. 3.3 showed that, starting from the corresponding $\mathbb{T}_\sigma[\![\sigma_{\mathsf{CIL}}]\!]$ state, the JB concrete semantics leads to $\sigma'_{\mathsf{JB}} = ([0 \mapsto \#1], [0 \mapsto 1, 1 \mapsto \#1, 2 \mapsto 1, 3 \mapsto 0], [\#1 \mapsto < \#2 >, \#2 \mapsto \{f \mapsto 0\}]$. So, by definition of $\mathbb{T}_\sigma[\![\ ]\!]$, we get $\mathbb{T}_\sigma[\![\sigma'_{\mathsf{CIL}}]\!] = \bullet \sigma'_{\mathsf{JB}}$ since the two stacks and the two heaps are equal, the values of three local variables of the JB state correspond to the values of the argument and the two local variables of the CIL state, respectively, and $=_\bullet$ projects out the fourth variable of the JB state $\sigma'$.

## 4.4 Other Instructions

In this section, we informally discuss how our approach deals with CIL instructions that are slightly different from other instructions in JB. We decided to handle these instructions informally since their translation is mostly straightforward. It is intended for readers that are expert of JB, CIL and more advanced C# features, such as generic type erasure in JB, or delegates in C#.

**Numerical and Reference Comparison** CIL compares numerical or reference values in two ways: through conditional branches (*e.g.*, `beq` branches when the topmost two values on the stack are equals) and comparisons (*e.g.*, `ceq` pushes 1 iff the topmost two values on the stack are equals, and 0 otherwise). As usual, these instructions are type independent and apply to numerical (int, float, long, ...) as well as reference values. JB uses a different approach, since its instructions are type dependent. If the topmost two values on the stack are integers, it uses a conditional branch instruction (`if_icmpeq`) similar to that of CIL (`beq`). However, JB has no comparison instruction on integers and we need to simulate it through a sequence of JB instructions relying on constants and branch. For instance, a `ceq` statement on integers is simulated as in Fig. 8a. Instead, if the topmost two values on the stack are long, JB uses a comparison statement `lcmp` that pushes to the stack 1, 0, or -1 iff the first value is less than, equal to, or greater than the

[6] Appendix A reports the detailed formal proof.

```
if_icmpeq 4
iconst_0
goto 5
iconst_1
nop
```

**(a) On integers**

```
lcmp
iconst_0
if_icmpeq 6
iconst_0
goto 7
iconst_1
nop
```

**(b) On longs**

**Figure 8: Translation of `ceq`.**

```
ldloc.0
ldc.i4.0
callvirt !0 List<A>::get(int32)
stloc.1
```

**(a) In CIL**

```
aload_0
iconst_0
invoke List.get:(I)LObject;
checkcast A
astore_1
```

**(b) In JB**

**Figure 9: Getting an element from a list.**

second, respectively. Hence, we simulate CIL conditional branch and comparison instructions through `lcmp`, integer constants and a conditional branch on integers. For instance, `beq` is translated into the sequence `lcmp; ifne #i;` where `i` is the target JB instruction of `beq`. The comparisons over long is similar to int. Namely, `ceq` is translated into the code in Fig. 8b. Conditional branch and comparison work also on references. Equality and inequality statements are treated as for integers, since JB defines an `if_acmpeq` statement. Other CIL operators (*e.g.*, `bgt`) can be applied to arbitrary references, as long as one of them is `null`.

**Generic Types** CIL keeps information about generic types, while JB erases it into `Object`. For instance, imagine that we have a local variable `list` of type `List⟨A⟩`. At source code level, a method call like `A a = list.get(0)` in Java or `A a = list[0]` in C# is legal since the elements of the list have type `A` in both languages. At bytecode level, getting an element from the list effectively returns an object of type `A` in CIL (see Fig. 9a), while it returns an object whose static type is `Object` in JB and casts it dynamically to `A` through a `checkcast` (Fig. 9b). Hence, our translation of a CIL method call with generic return type `T` adds a `checkcast` instruction to `T` after the call. Primitive types (*e.g.*, `int` and `long`) can be passed as generic types in CIL but not in JB. Hence, when using a primitive type for the generic parameter or return value of a CIL method call, we box and unbox the primitive value into a Java wrapper class such as `java.lang.Integer`.

**Delegates** Lambda expressions have only been introduced in Java 8, while C# has been using *delegates* since its very beginning. C# implements delegates through CIL instructions that load a pointer to a method (`ldftn`) and execute it, sometime by using inner classes. Namely, C# accesses a pointer to the method through `ldftn` and calls the `Invoke` method of the delegate class. Consider for instance Fig. 10. The C# code in Fig. 10a uses a delegate to call a method. In Fig. 10b, this is compiled into a `ldftn` statement at line 4 followed by a call to `Invoke` at line 7. We translate this by using reflection and string constants. Namely, the signature of the method pointed by `ldftn` is represented by a string, passed to an instrumentation library call in class `Reflection`, that calls this method by reflection (Fig. 10c). However, many static analyzers (including Julia) are unsound for reflection. Hence, our translation marks all signatures accessed in this way as entry points (that is, methods that might be directly called from outside the application and therefore are analyzed under the most generic assumptions). This might cause a loss of precision, since contextual information on delegates is lost, but preserves soundness.

| Library | # met. | # fail | % fail | Tr. t. | Mem. |
|---|---|---|---|---|---|
| mscorlib | 28,344 | 870 | 3.07% | 23" | 158 |
| Sys.Core | 6,988 | 47 | 0.68% | 4" | 96 |
| Sys.Design | 13,509 | 4 | 0.03% | 20" | 180 |
| Sys | 17,851 | 242 | 1.36% | 21" | 142 |
| Sys.Runtime.Serial | 5,624 | 74 | 1.32% | 5" | 86 |
| Sys.ServiceModel | 34,603 | 80 | 0.23% | 34" | 156 |
| Sys.Web | 28,249 | 38 | 0.13% | 37" | 216 |
| Sys.Web.Extensions | 4,245 | 0 | 0.00% | 4" | 109 |
| Sys.Windows.Forms | 28,319 | 53 | 0.19% | 42" | 189 |
| Sys.XML | 12,727 | 171 | 1.34% | 23" | 146 |
| Total | 180,460 | 1,579 | 0.87% | 3'33" | |

**Table 2: Experimental results on libraries.**

**Async and Await** In C#, an `async` method returns a `Task` object that allows the caller to execute the code of the method asynchronously. On the other hand, statement `await` waits until the execution of the asynchronous method ends and extracts the results of the computation. This pattern is compiled into method pointers and reflection at CIL bytecode level, in the same way delegates are treated. Therefore, we apply the same solution for delegates we described in Sec. 4.4.

## 5 EXPERIMENTAL RESULTS

We implemented our translation from CIL to JB through (i) a C# program that translates a CIL program to an intermediate XML representation (representing Java bytecode), and (ii) a Java program that produces a `jar` file from an XML representation. We had to split the implementation in this way since the library to read CIL bytecode (`Mono.Cecil`) is written in .NET, while the library writing `jar` bytecode (`BCEL`) is written in Java. The first part of the translation (CIL to JB) runs in parallel on different classes through the System.Threading.Tasks library (part of the standard .Net framework). We used an Intel Core i5-6600 CPU at 3.30GHz machine with 16 GB of RAM, 64-bit Windows 7 Professional, and Java SE Runtime Environment v.1.8.0_111-b14.

As a first experiment to assess the efficiency and precision of our approach, we translated and analyzed the five most popular GitHub repositories (as on February 27th, 2017) written in C# and tagged as C# repositories[7]. Tab. 1 reports (i) the number of C# LOC of each projects (Column **LOC**)[8] (our benchmarks range between 17 and 120 KLOC, hence they are real world applications); (ii) the number of stars of the GitHub repository as on February 27th, 2017 (**GH** *); (iii) the total number of methods (**# meth.**), and for how many of them the translation failed because of unsafe code (**# fail**); (iv) the time (**Tr. t.**) and memory (**Mem.**, in MB) consumed by the translation from CIL to JB; (v) the time of Julia's analyses (**Analysis t.**), and (vi) the number of alarms (**Al.**), of false alarms because of loss of information introduced by the translation (**False al.**), and the precision (ratio of false alarms *w.r.t.* the total number of alarms, column **Precision**) of Julia's analysis.

[7]We consider the number of watchers as measure of popularity of a repository. We discarded some projects *tagged* as C# that actually mostly contain native code (corefx, coreclr, mono), that did not compile in Visual Studio (roslyn, powershell), that have been dismissed (shadowsocks), or that are particularly small (wavefunction, below 1KLOC).

[8]LOC are computed with LocMetrics http://www.locmetrics.com/

```
delegate void Del(string message);
void DelegateMethod(string message) {...}
void go() {
        Del handler = DelegateMethod;
        handler("Hello World");
}
```

(a) C# code

```
void go () {
        ldarg.0
        ldftn A::DelegateMethod(string)
        newobj A/Del::.ctor(object, int)
        ldstr "Hello World"
        call void A/Del::Invoke(string)
}
```

(b) CIL

```
void go() {
        ldc "DelegateMethod(LString;)V"
        invokestatic
Reflection.GetMethod:(LString;)LMethod;
        ldc "Hello World"
        invokevirtual A/Del.Invoke:(LString;)
}
```

(c) JB

**Figure 10: An example of CIL delegate.**

| Program | LOC | GH * | # met. | # fail | Tr. t. | Mem. | Analysis t. | Al. | False al. | Precision |
|---|---|---|---|---|---|---|---|---|---|---|
| CodeHub | 32,510 | 7,718 | 4,887 | 0 | 0'07" | 115 | 0'43" | 9 | 1 | 89% |
| SignalR | 71,207 | 6,285 | 6,610 | 3 | 0'07" | 131 | 0'50" | 8 | 1 | 88% |
| Dapper | 22,513 | 5,815 | 1,058 | 0 | 0'07" | 77 | 0'29" | 13 | 3 | 77% |
| ShareX | 171,580 | 5,208 | 11,568 | 14 | 0'58" | 193 | 2'08" | 57 | 0 | 100% |
| Nancy | 109,139 | 4,969 | 8,817 | 0 | 0'07" | 136 | 1'25" | 18 | 1 | 94% |
| Total | 406,949 | | 32,940 | 17 | 1'26" | | 4'35" | 105 | 6 | 94% |

**Table 1: Experimental results on the 5 most starred Github C# projects.**

In order to assess the efficiency and library coverage of our approach, we also analyzed the 10 largest (based on the size of the .dll files) system libraries of the Microsoft .Net framework version 4.0.30319. They contain unsafe code (such as cryptographic code in mscorlib.dll) and might not be compiled from C#, but possibly from VB.Net. Tab. 2 reports the number of methods of the library (**# met.**), the number and percentage of methods where the translation fails because of unsafe code (**# fail** and **% fail**), and time (**Tr. t.**) and memory (**Mem.**, in MB) for the translation.

**Research Question 1: Efficiency.**
In 4 out of 5 top Github projects, our translation took 7" (Tab. 1); it took almost 1 minute for ShareX. These times are much shorter (overall, less than a third) than the analysis time. The memory consumed by the translation is small (below 200MB). The results for .Net framework libraries (Tab. 2) show a similar trend: we translated about 180K methods in about 3'30" (that is, a bit more than 1 msec per method) consuming at most 189MB of memory. So, our system deals with industrial-size software with a translation time comparable to the analysis time.

**Research Question 2: Precision**
We manually checked only the 105 high severity alarms issued by Julia on the top 5 Github projects, over a total of several thousands. Tab. 1 reports their number (column **Al.**) and the number of false alarms (**False al.**) due to our translation. The static analysis might generate false alarms as well, for instance because of disjunctive constraints not tracked by Julia; we do not count these as false alarms, since we want to evaluate the imprecision due to the translation, and not that inherent to Julia. In particular, 6 alarms out of 105 (about 6%) are false because of imprecision introduced by the translation. This shows that our approach satisfies Research Question 2. The origins of this imprecision are (i) `async` and `await` statements (in particular in Dapper), and (ii) `try-catch-finally` blocks (*e.g.*, in SignalR). These would require to modify Julia to recognize these features more precisely (through automatic annotations produced by the translation).

**Research Question 3: Libraries**
We manually checked that all methods of the 5 top Github C# projects where our translation fails are actually unsafe. Column **# fail** in Tab. 1 shows that there is no failure for CodeHub, Dapper and Nancy. Instead, there are 3 failures for SignalR, due to unsafe methods in class `Infrastructure.SipHashBasedString-EqualityComparer`, and 14 failures for ShareX, due to unsafe methods in two classes: (i) `GreenshotPlugin.Core` has methods setting or getting colors in fast implementations of bitmaps (`UnsafeBitmap` and subclasses); (ii) `ShareX.ImageHelpers` uses unsafe classes (such as `UnsafeBitmap`). This shows that our approach fails only for unsafe code with unsafe pointer manipulation (storing pointers in fields, returning them from methods, performing pointer arithmetic). Tab. 2 shows that the translation succeeds for 99.13% of the methods, with a worst case of 96.93%. Hence, our approach satisfies Research Question 3.

## 6 CONCLUSION

This article introduced, formalized and proved correct a translation from CIL to JB, for static analysis. To assess its feasibility and interest, it has been implemented and connected to the Julia analyzer. Experiments show positive results for efficiency, precision, and libraries' coverage. As future work, we plan to (i) improve Julia precision in the corner cases highlighted by our experiments, (ii) investigate new .Net properties of interest, and (iii) rely on `invokedynamic` when translating delegates.

## REFERENCES

[1] Checkstyle. http://checkstyle.sourceforge.net.
[2] Clr to jvm. http://xmlvm.org/clr2jvm/.
[3] CodeSonar – Static Analysis SAST Software. https://www.grammatech.com/products/codesonar.
[4] Findbugs$^{TM}$ – Find Bugs in Java Programs. http://findbugs.sourceforge.net/.
[5] Ndepend. http://www.ndepend.com.
[6] .Net Reflector Add-Ins. https://www.microsoft.com/en-us/research/project/spec.
[7] PMD. https://pmd.github.io/.
[8] WALA. http://wala.sourceforge.net/wiki/index.php/Main_Page.
[9] M. Ameri and C. A. Furia. Why Just Boogie? Translating between Intermediate Verification Languages. In *Proceedings of IFM '16*, LNCS. Springer, 2016.
[10] R. Atkey and D. Sannella. ThreadSafe: Static Analysis for Java Concurrency. *Electronic Comm. of the European Association of Software Science and Technology*, 72, 2015.
[11] M. Barnett, K. Leino, and W. Schulte. The Spec# Programming System: An Overview. In *Proceedings of CASSIS '04*, 2004.
[12] M. Bebenita, F. Brandner, M. Fähndrich, F. Logozzo, W. Schulte, N. Tillmann, and H. Venter. SPUR: a Trace-based JIT Compiler for CIL. In W. R. Cook, S. Clarke, and M. C. Rinard, editors, *Proceedings of OOPSLA '10*. ACM, 2010.

[13] P. Cousot and R. Cousot. Abstract interpretation: a unified lattice model for static analysis of programs by construction or approximation of fixpoints. In *Proceedings of POPL '77*. ACM Press, 1977.

[14] P. Cousot and R. Cousot. Abstract Interpretation: A Unified Lattice Model for Static Analysis of Programs by Construction or Approximation of Fixpoints. In *Proc. of Principles of Programming Languages (POPL'77)*, pages 238–252, 1977.

[15] Coverity. Coverity Prevent$^{TM}$. http://www.coverity.com/library/pdf/coverity_prevent.pdf.

[16] ECMA. *Standard ECMA-335: Common Language Infrastructure (CLI).* 2012.

[17] T. A. S. Foundation. Apache Commons BCEL. https://commons.apache.org/proper/commons-bcel, checked on June 24, 2016.

[18] JetBrains. Resharper. https://www.jetbrains.com/resharper.

[19] T. Lindholm, F. Yellin, G. Bracha, and A. Buckley. *The Java Virtual Machine Specification, Java SE 7 Edition.* Addison-Wesley Professional, 1st edition, 2013.

[20] F. Logozzo. Cibai: An Abstract Interpretation-Based Static Analyzer for Modular Analysis and Verification of Java Classes. In *Proceedings of VMCAI '07*, LNCS. Springer, 2007.

[21] F. Logozzo and M. Fähndrich. Static Contract Checking with Abstract Interpretation. In *Proceedings of FoVeOOS '10*, LNCS. Springer, 2010.

[22] Microsoft. FxCop. https://msdn.microsoft.com/en-us/library/bb429476(v=vs.80).aspx.

[23] F. Spoto. The Julia Static Analyzer for Java. In *Proceedings of SAS '16*, LNCS. Springer, 2016.

[24] Wikipedia. List of Tools for Static Code Analysis. https://en.wikipedia.org/wiki/List_of_tools_for_static_code_analysis#Java.

## A PROOF OF CORRECTNESS

We present here the main steps of the correctness proof of our approach.

Lemma A.1. *The function convertLocals is an identity embedding.*

Proof. It is sufficient to observe that, by construction, the function concatenates the two arrays by shifting indexes when a 64 bits value occurs, hence preserving the values and the ordering of the elements' indexes. □

Lemma A.2. *The translation of* `ldloc` *is correct.*

Proof. Let us consider the case of integer arguments (the other case can be treated analogously). Let us prove that the translation of `ldloc` for integer values is correct, *i.e.*, that $\forall \sigma_{\mathsf{CIL}} \in \Sigma_{\mathsf{CIL}}$, if $\langle \mathtt{ldloc}\ \mathtt{i}, \sigma_{\mathsf{CIL}} \rangle \rightarrow_{\mathsf{CIL}} \sigma'_{\mathsf{CIL}}$ and
$\langle \mathbb{T}[\![\mathtt{ldloc}\ \mathtt{i}, (\bar{\mathsf{s}}, \bar{\mathsf{l}}, \bar{\mathsf{a}}, \bar{\mathsf{w}})]\!], \mathbb{T}_\sigma[\![\sigma_{\mathsf{CIL}}]\!] \rangle \rightarrow_{\mathsf{JB}} \sigma'_{\mathsf{JB}}$ then $\mathbb{T}_\sigma[\![\sigma'_{\mathsf{CIL}}]\!] = \sigma'_{\mathsf{JB}}$.
Let $\sigma_{\mathsf{CIL}} = (s, l, a, h)$ be arbitrary. By the `ldloc`−CIL rule we have $\sigma'_{\mathsf{CIL}} = (\mathsf{s} :: \mathtt{l[i]}, \mathtt{l}, \mathsf{a}, \mathsf{h})$. By rule (3) of Fig. 7, $\mathbb{T}[\![\mathtt{ldloc}\ \mathtt{i}, \bar{\mathsf{s}}, \bar{\mathsf{l}}, \bar{\mathsf{a}}, \bar{\mathsf{w}}]\!] = \mathtt{iload}\ j$ where $j = |\bar{\mathsf{a}}| + 64^{|\bar{\mathsf{a}}|}_{\bar{\mathsf{a}}} + i + 64^{i}_{\bar{\mathsf{l}}}$. By definition of $\mathbb{T}_\sigma[\![]\!]$:

$$\begin{array}{rcl} \mathbb{T}_\sigma[\![\sigma_{\mathsf{CIL}}]\!] & = & \mathbb{T}_\sigma[\![(\mathsf{s}, \mathtt{l}, \mathsf{a}, \mathsf{h})]\!] \\ & = & (\mathsf{s}', convertLocals(\mathtt{l}, \mathsf{a}), \mathsf{h}'). \end{array}$$

By the `iload`−JB rule:

$\langle \mathtt{iload}\ \mathtt{j}, (\mathsf{s}, convertLocals(\mathtt{l}, \mathsf{a}), \mathsf{h}) \rangle \rightarrow_{\mathsf{JB}}$
$(\mathsf{s}' :: convertLocals(\mathtt{l}, \mathsf{a})[j], convertLocals(\mathtt{l}, \mathsf{a}), \mathsf{h}') = \sigma'_{\mathsf{JB}}$.

By definition of $\mathbb{T}_\sigma[\![]\!]$ we have

$$\begin{array}{rcl} \mathbb{T}_\sigma[\![\sigma'_{\mathsf{CIL}}]\!] & = & \mathbb{T}_\sigma[\![(\mathsf{s} :: \mathtt{l[i]}, \mathtt{l}, \mathsf{a}, \mathsf{h})]\!] \\ & = & (\mathsf{s} :: \mathtt{l[i]}, convertLocals(\mathtt{l}, \mathsf{a}), \mathsf{h}). \end{array}$$

By definition of *convertLocals* we have

$$convertLocals(\mathtt{l}, \mathsf{a})[j] = \mathtt{l}[i],$$

which implies $\mathbb{T}_\sigma[\![\sigma'_{\mathsf{CIL}}]\!] = \sigma'_{\mathsf{JB}}$, and thus $\mathbb{T}_\sigma[\![\sigma'_{\mathsf{CIL}}]\!] =_\bullet \sigma'_{\mathsf{JB}}$. □

Lemma A.3. *The translation of* `call` *is correct.*

Proof. Let us consider the case of static call (the other case can be treated analogously). Moreover, let us denote by $\ell_{[i,j]}$ the sequence $\ell_i, \ldots, \ell_j$, and by $\mathsf{s}_{(\mathtt{i},\mathtt{j})}$ the sequence of types $[\mathsf{k}_1, \ldots, \mathsf{k}_\mathtt{i}, \mathsf{t}_1, \ldots, \mathsf{t}_\mathtt{j}]$.
We show that $\forall \sigma_{\mathsf{CIL}} = ([u_{[1,n]}, v_{[1,i]}], \mathsf{l}, \mathsf{a}, \mathsf{h})$, if
$\langle \mathtt{call}\ \mathsf{m}(\mathsf{arg}_{[1,\mathtt{i}]}), \sigma_{\mathsf{CIL}} \rangle \rightarrow_{\mathsf{CIL}} \sigma'_{\mathsf{CIL}} = ([u_{[1,n]}], \mathsf{l}, \mathsf{a}, \mathsf{h}')$ and
$\langle \mathbb{T}[\![\mathtt{call}\ \mathsf{m}(\mathsf{arg}_{[1,\mathtt{i}]}), (\mathsf{s}_{(\mathtt{n},\mathtt{i})}, \bar{\mathsf{l}}, \bar{\mathsf{a}}, \bar{\mathsf{w}}_{[1,\mathtt{i}]})]\!], \mathbb{T}_\sigma[\![\sigma_{\mathsf{CIL}}]\!] \rangle \rightarrow_{\mathsf{JB}} \sigma'_{\mathsf{JB}}$,
then $\mathbb{T}_\sigma[\![\sigma'_{\mathsf{CIL}}]\!] =_\bullet \sigma'_{\mathsf{JB}}$.
Rule $\langle \mathtt{call}\ \mathsf{m}(\mathsf{arg}_{[1,\mathtt{i}]}), \sigma_{\mathsf{CIL}} \rangle \rightarrow_{\mathsf{CIL}} \sigma'_{\mathsf{CIL}}$ requires, to be applied, that the condition

$$\begin{array}{c} \langle body(\mathsf{m}(\mathsf{arg}_{[0,\mathtt{i}]}), (v_{[1,i]})), ([\,], \emptyset, [j-1 \mapsto v_j : j \in [1..i]], \mathsf{h}) \rangle \\ \downarrow_{\mathsf{CIL}} \\ (\mathsf{s}', \mathsf{l}', \mathsf{a}', \mathsf{h}') \end{array}$$

is satisfied.
Observe that for $\mathsf{a} = [j{-}1 \mapsto v_j : j \in [1..i]]$, we get $\mathbb{T}_\sigma[\![([\,], \emptyset, \mathsf{a}, \mathsf{h})]\!] = ([\,], \hat{\mathsf{l}}, \mathsf{h})$, where for each $j \in [1..i]$, $\hat{\mathsf{l}}[j + 64^{j}_{\mathsf{a}}] = v_j$. Moreover, $\mathbb{T}[\![body(\mathsf{m}(\mathsf{arg}_{[0,\mathtt{i}]})), (\mathsf{t}_1, \ldots, \mathsf{t}_\mathtt{i}), \emptyset, \mathsf{a}, \mathsf{h}]\!] = body(\mathsf{m}(\mathsf{arg}_{[0,\mathtt{i}]}))$. Therefore, by inductive hypothesis, we get that $\langle body(\mathsf{m}(\mathsf{arg}_{[0,\mathtt{i}]})(v_{[1,i]})),$
$([], [j{-}1 \mapsto v_j : j \in [1..i]], \mathsf{h}) \rangle \rightarrow_{\mathsf{JB}} (\hat{\mathsf{s}}, \hat{\mathsf{l}}, \hat{\mathsf{h}})$ is such that $\mathbb{T}_\sigma[\![(\mathsf{s}', \mathsf{l}', \mathsf{a}', \mathsf{h}')]\!] =_\bullet (\hat{\mathsf{s}}, \hat{\mathsf{l}}, \hat{\mathsf{h}})$, and in particular $\mathsf{h}' = \hat{\mathsf{h}}$. It is sufficient now to recall that by Fig. 7 (static call) $\mathbb{T}[\![\mathtt{call}\ \mathsf{m}(\mathsf{arg}_{[1,\mathtt{i}]}), \mathsf{s}_{(\mathtt{n},\mathtt{i})}, \bar{\mathsf{l}}, \bar{\mathsf{a}}, \bar{\mathsf{w}}_{[\mathtt{n},\mathtt{i}]}]\!]$ is obtained by applying `invokestatic` $\mathsf{m}(\mathsf{arg}_{[1,\mathtt{i}]})$ followed by the update of all the local variables passed by reference to the called method, and that by the `invokestatic` rule of Fig. 6, $\sigma'_{\mathsf{JB}} = ([\mathsf{u}_{[1,\mathsf{n}]}], \hat{\mathtt{1}}, \mathsf{h})$. Finally, by the definition of $\mathbb{T}_\sigma[\![\,]\!]$, we get $\mathbb{T}_\sigma[\![\sigma'_{\mathsf{CIL}}]\!] =_\bullet \sigma'_{\mathsf{JB}}$. □

Lemma A.4. *The translation of* `bgt` *is correct.*

Proof. Consider $\mathbb{T}[\![\mathtt{bgt}\ \mathsf{k}, \bar{\mathsf{s}} :: \mathsf{t}_1 :: \mathsf{t}_2, \bar{\mathsf{l}}, \bar{\mathsf{a}}, \bar{\mathsf{w}}]\!]$ in the case $\mathsf{t}_1 = \mathsf{t}_2 = \mathtt{int}$, and assume $\sigma_{\mathsf{CIL}} = ([s_{[1,n]} :: v_1, v_2], \mathsf{l}, \mathsf{a}, \mathsf{h})$ with $v_2 > v_2$ (the other case is similar), yielding to $\langle \mathtt{bgt}\ \mathtt{l}, \sigma_{\mathsf{CIL}} \rangle \rightarrow_{\mathsf{CIL}} \sigma'_{\mathsf{CIL}} = \langle \mathtt{l}, \sigma_{\mathsf{CIL}} \rangle$.
We show that if $\langle \mathbb{T}[\![\mathtt{bgt}\ \mathsf{k}, \bar{\mathsf{s}} :: \mathsf{t}_1 :: \mathsf{t}_2, \bar{\mathsf{l}}, \bar{\mathsf{a}}, \bar{\mathsf{w}}]\!], \mathbb{T}_\sigma[\![\sigma_{\mathsf{CIL}}]\!] \rangle \rightarrow_{\mathsf{JB}} \sigma'_{\mathsf{JB}}$, then $\mathbb{T}_\sigma[\![\sigma'_{\mathsf{CIL}}]\!] =_\bullet \sigma'_{\mathsf{JB}}$. By the corresponding rule in Fig. 7, $\mathbb{T}[\![\mathtt{bgt}\ \mathsf{k}, \bar{\mathsf{s}} :: \mathsf{t}_1 :: \mathsf{t}_2, \bar{\mathsf{l}}, \bar{\mathsf{a}}, \bar{\mathsf{w}}]\!] = \mathtt{if\_icmpgt}\ k'$ where
$k' = statementIdx(getBody(\mathtt{bgt}\ \mathsf{k})(k))$.
By the semantics of `if_icmpgt` we have that $\langle \mathtt{if\_icmpgt}\ k', ([s_{[1,n]} :: v_1 :: v_2], convertLocals(\mathtt{l}, \mathsf{a}), \mathsf{h}) \rangle \rightarrow_{\mathsf{JB}} \langle k', (s_{[1,n]}, convertLocals(\mathtt{l}, \mathsf{a}), \mathsf{h}) \rangle$.
As $\mathbb{T}_\sigma[\![\sigma_{\mathsf{CIL}}]\!] = ([s_{[1,n]} :: v_1 :: v_2], convertLocals(\mathtt{l}, \mathsf{a}), \mathsf{h})$, and by definition of $statementIdx()$, we get that $\mathbb{T}_\sigma[\![\langle \mathtt{l}, \sigma_{\mathsf{CIL}} \rangle]\!] = \langle k', \mathbb{T}_\sigma[\![\sigma_{\mathsf{CIL}}]\!] \rangle = \sigma'_{\mathsf{JB}}$, and thus $\mathbb{T}_\sigma[\![\langle \mathtt{l}, \sigma_{\mathsf{CIL}} \rangle]\!] =_\bullet \sigma'_{\mathsf{JB}}$ □

Lemma A.5. *The translation of* `newobj` *is correct.*

Proof. Assume $\sigma_{\mathsf{CIL}} = ([s_{[1,n]} :: v_{[1,i]}], \mathsf{l}, \mathsf{a}, \mathsf{h})$. By definition, $\langle \mathtt{newobj} \mathsf{T}(\mathsf{arg}_{[1,\mathtt{i}]}), \sigma_{\mathsf{CIL}} \rangle \rightarrow_{\mathsf{CIL}} \sigma'_{\mathsf{CIL}} = (s :: r, (l, a, h'))$, where $r$ and $h'$ satisfy the constraints of the corresponding rule of Fig. 5. In particular, *fresh* $\mathsf{T}\ \mathsf{h} = (r, \mathsf{h}')$ allocates the memory for an object of type $\mathsf{T}$ on heap $\mathsf{h}$ and returns (i) the reference $r$ of the freshly allocated object, and (ii) the heap $\mathsf{h}'$ resulting from the allocation of memory on $\mathsf{h}$.
Let $\sigma'_{\mathsf{JB}} = \langle \mathbb{T}[\![\mathtt{newobj}\ \mathsf{T}(\mathsf{a}_{[1,\mathtt{i}]}), \bar{\mathsf{s}} :: \mathsf{t}_{[1,\mathtt{i}]}, \bar{\mathsf{l}}, \bar{\mathsf{a}}, \bar{\mathsf{w}}]\!], \mathbb{T}_\sigma[\![\sigma_{\mathsf{CIL}}]\!] \rangle$, and compare $\mathbb{T}_\sigma[\![\sigma'_{\mathsf{CIL}}]\!]$ and $\sigma'_{\mathsf{JB}}$ componentwise. We may observe that in both cases the store is equal to $s :: r$, as the new elements added to the store during the translation in order to implement the object initialization are finally removed when applying the `invokevirtual` call, whose correctness is granted by structural inductive hypothesis. Moreover, the single array in JB for both local variables and method arguments is updated properly by storing and loading the values of the constructor arguments in the expected ordering. Finally, the heap $\mathsf{h}'$ results in both cases from the allocation of corresponding memory on $\mathsf{h}$. □

Lemma A.6. *The translation of* `stind` *is correct.*

Proof. By Fig. 7, we have that
$\mathbb{T}[\![\mathtt{stind}, \bar{\mathsf{s}}, \bar{\mathsf{l}}, \bar{\mathsf{a}}, \bar{\mathsf{w}}]\!] = \mathbb{T}[\![\mathtt{stfld}\ \mathtt{value}, \bar{\mathsf{s}}, \bar{\mathsf{l}}, \bar{\mathsf{a}}, \bar{\mathsf{w}}]\!] = \mathtt{putfield}\ \mathtt{value}$
By Fig. 5, we have that $\langle \mathtt{stind}, (s :: r_i :: v, \mathsf{l}, \mathsf{a}, \mathsf{h}) \rangle \rightarrow_{\mathsf{CIL}} (s, \mathsf{l}, \mathsf{a}, \mathsf{h}[r_i \mapsto v])$. By definition of the translation of concrete states (assuming that the only direct reference in the stack is $r_i$), we have that $\mathbb{T}_\sigma[\![(\mathsf{s} :: \mathsf{r}_\mathtt{i} :: \mathsf{v}, \mathsf{l}, \mathsf{a}, \mathsf{h})]\!] = (s :: r' :: v, cnvrtLoc(\mathsf{l}, \mathsf{a}), \mathsf{h}[r' \mapsto [\mathtt{value} \mapsto$

$l(i)]])$ where $r'$ is a freshly allocated references pointing to a wrapper. Then by Fig. 6, we have that $\langle \texttt{putfield value}, (s :: r' :: v, \mathit{cnvrtLoc}(\mathsf{l}, \mathsf{a}), \mathsf{h}[r \mapsto [\texttt{value} \mapsto l(i)]])\rangle \rightarrow_{\mathsf{JB}} (s, \mathit{cnvrtLoc}(\mathsf{l}, \mathsf{a}), \mathsf{h}[r' \mapsto [\texttt{value} \mapsto v]])$. Finally, we obtain that $\mathbb{T}_\sigma[\![(\mathsf{s}, \mathsf{l}, \mathsf{a}, \mathsf{h}[\mathtt{r_i} \mapsto \mathsf{v}])]\!] =_\bullet (s, \mathit{cnvrtLoc}(\mathsf{l}, \mathsf{a}), \mathsf{h}[r' \mapsto [\texttt{value} \mapsto v]])$ proving the soundness of the translation of `stind`. □

Lemma A.7. *The translation of* `ldind` *is correct.*

Proof. By Figure 7, we have that

$\mathbb{T}[\![\texttt{ldind}, \bar{\mathsf{s}}, \bar{\mathsf{l}}, \bar{\mathsf{a}}, \overline{\mathsf{w}}]\!] = \mathbb{T}[\![\texttt{ldfld value}, \bar{\mathsf{s}}, \bar{\mathsf{l}}, \bar{\mathsf{a}}, \overline{\mathsf{w}}]\!] = \texttt{getfield value}$

By Figure 5, we have that $\langle \texttt{ldind}, (s :: r_i, \mathsf{l}, \mathsf{a}, \mathsf{h})\rangle \rightarrow_{\mathsf{CIL}} (s :: \mathsf{h}(r_i), \mathsf{l}, \mathsf{a}, \mathsf{h})$. By definition of the translation of concrete states (assuming that the only direct reference in the stack is $r_i$), we have that $\mathbb{T}_\sigma[\![(\mathsf{s} :: \mathtt{r_i}, \mathsf{l}, \mathsf{a}, \mathsf{h})]\!] = (s :: r', \mathit{cnvrtLoc}(\mathsf{l}, \mathsf{a}), \mathsf{h}[r' \mapsto [\texttt{value} \mapsto l(i)]])$ where $r'$ is a freshly allocated references pointing to a wrapper. Then by Fig. 6, we have that $\langle \texttt{getfield value}, (s :: r', \mathit{cnvrtLoc}(\mathsf{l}, \mathsf{a}), \mathsf{h}[r' \mapsto [\texttt{value} \mapsto l(i)]])\rangle \rightarrow_{\mathsf{JB}} (s :: l(i)]], \mathit{cnvrtLoc}(\mathsf{l}, \mathsf{a}), \mathsf{h}[r' \mapsto [\texttt{value} \mapsto v]])$. Finally, we obtain that $\mathbb{T}_\sigma[\![(\mathsf{s} :: \mathtt{l(i)}, \mathsf{l}, \mathsf{a}, \mathsf{h})]\!] =_\bullet (s, \mathit{cnvrtLoc}(\mathsf{l}, \mathsf{a}), \mathsf{h}[r' \mapsto [\texttt{value} \mapsto l(i)]])$ proving the soundness of the translation of `ldind`. □

Lemma A.8. *The translation of* `ldloca` *is correct.*

Proof. By Fig. 5, we have that $\langle \texttt{ldloca i}, (s, \mathsf{l}, \mathsf{a}, \mathsf{h})\rangle \rightarrow_{\mathsf{CIL}} (s :: r_i, \mathsf{l}, \mathsf{a}, \mathsf{h})$ where $r_i$ is the direct reference pointing to the i-th local variable. By Fig. 7, we have that `ldloca i` is translated into a sequence of statements that (i) creates a wrapper object containing the value of the i-th local variable, (ii) stores its reference into an instrumentation variable, and (iii) leaves its reference on the operand stack as well. Then, by definition of the concrete semantics of JB, we obtain a final state $\sigma'_{\mathsf{JB}}$ appending to the initial stack a reference to the wrapper object whose value is the one of the i-th local variable. Therefore, $\sigma'_{\mathsf{JB}} =_\bullet \mathbb{T}_\sigma[\![(\mathsf{s} :: \mathtt{r_i}, \mathsf{l}, \mathsf{a}, \mathsf{h})]\!]$ since $=_\bullet$ ignores the instrumentation variables. This proves the soundness of the translation of `ldloca`. □

Theorem A.9. *The translation of a* CIL *program into* JB *code is correct.*

Proof. We prove that the translation of each statement from CIL to JB depicted in Fig. 7 satisfies the correctness property introduced in Sec. 4.3. In fact, by Lemma A.2, the translation of `ldloc` is correct, and a similar proof can be provided for the `add`, `stloc`, `ldarg`, `ldfld`, and `stfld` statements. The correctness of a new object creation is proved by Lemma A.5. The correctness proof of static and dynamic calls translation has been given in Lemma A.3, and that of `stind`, `ldind`, `ldloca` was proved by Lemmas A.6, A.7 and A.8, respectively. Finally, the correctness of the comparison statements translation is shown in Lemma A.4. □