On Compiling JML Specifications into Bytecode

Jędrzej Fulara
Institute of Informatics
University of Warsaw
ul. Banacha 2
02-097 Warsaw, Poland
fulara@mimuw.edu.pl

Krzysztof Jakubczyk
Institute of Informatics
University of Warsaw
ul. Banacha 2
02-097 Warsaw, Poland
kjk@mimuw.edu.pl

Aleksy Schubert
Institute of Informatics
University of Warsaw
ul. Banacha 2
02-097 Warsaw, Poland
alx@mimuw.edu.pl

ABSTRACT

An idea of compiling different languages to the Java bytecode becomes more and more popular. This forces us to build one common verification platform for the bytecode level. BML (Bytecode Modelling Language), a recent specification language based on JML (Java Modelling Language), is designed to be the core of such platform. There is a need to automatically translate source code level specifications into bytecode, to be able to use them later in the verification.

In this paper we present *Jml2Bml* compiler, a tool that for a given java source file with JML annotations and corresponding class file, generates BML annotations and inserts them into the bytecode.

Categories and Subject Descriptors

D.2.4 [**Software Engineering**]: Software/Program Verification

General Terms

Reliability, Verification

Keywords

Java, byte code, JML, BML

1. INTRODUCTION

Modern, high level programming languages support creating software in a modular way. Applications can be divided into smaller parts that can be developed independently. In large systems, it is a common practice to outsource some well defined subsystems to external companies. The main problem in this approach is that the pieces of software developed by different programming teams often are not fully compatible. To avoid this incopatibility and useless code that is produced, there is a need to specify precisely the desired behaviour of the components, what they require and what can we expect from them. The solution is to define formally implementation contracts that can be then automatically verified.

Specification languages are useful in descirbing the system components behaviour. They are not only helpful in dividing the problem into smaller pieces, but also focus on *what* is expected from each part, without saying anything about *how* should it be done. The specification languages are designed to be simple enough to be understood by programmers, so they can play role of code documentation. Using specification language for documenting code has the advantage that it is possible to automatically verify that the source code implements the documented features.

One of the key aspects of modern software is security. As the end users, we usually have to trust the software we download from the internet. When the downloaded program is open source it is believed that it will not do anything inappropriate. If we download commercial software or program that is not open source, the only argument for security is a digital signature. The signature does not assure that the software is secure. There were cases when many users were deceived by respected companies eg. rootkits were installed when audio CD was inserted. The specification language can be used to describe a required security policy. In this case the verifcation stands for ensuring that the policy holds. As stated, not only the developers are interested in checking the described code property. The end user may also want to verify if the code he is running is secure. However, applications are usually distributed only in some executable form, so the specifications have also to be translated into the lower (in our case: bytecode) level. The compiled specifications are very useful in developing the Proof-Carrying Code (PCC) infrastructure [12].

The other reason, why it is important to be able to translate the source code specification into lower level language specifications, is the fact that more and more languages will be compiled to the same bytecode. Java Virtual Machine can be used with languages different than Java. Few examples:

- Jython - the Python Java implementation
- JRuby - the Ruby Java implementation
- Jacl - the Tcl Java implementation
- Rhino - the JavaScript Java implementation

- Scala - a programming language compiled to Java bytecode

At SugarCon 2008, Sun Microsystems President and CEO Jonathan Schwartz said "we are just going to take the 'J' off the 'JVM' and just make it a 'VM'". Therefore there will be a global trend with support of companies to use JVM with languages other than Java. Bytecode itself has a verification algorithm [9] but the verification is done only to ensure that the loaded bytecode will not cause the crash of the JVM. The verification algorithm includes:

- checking that all arguments on the operand stack are legal
- ensuring that all types of variables passed to methods are correct
- checking that all load and store operations have correct types

Unfortunately programming errors are not checked. Now, the Java language has JML with proper tools that can be used to verify programs - check their correctness or find errors. Most of them operates on source code. But it seems to be a better idea to develop one common verification platform at the bytecode level, then to create multiple, different for each language platforms working at the source code level. However it would be to difficult to add annotations to the compiled programs manually, so we have to provide tools translating source code level specification languages into one common bytecode specification language. The BML, proposed in [3] is a good choice for the target language. There are mainly three sets of tools needed to build the common bytecode verification platform:

- bytecode verification tools that use BML annotations
- modelling languages (such as JML for Java) for other programming languages
- compilers that compile programs to JVM bytecode along with annotation compilers

The *Jml2Bml* compiler described in this paper is designed to be a part of this scheme.

1.1 JML

The Java Modelling Language (JML) is a behavioural specification language for Java modules. It allows to write specifications according to *design-by-contract* principles. Data types and method behaviour can be precisely commented using JML annotations. They describe the input requirements (preconditions), what we can expect at the output (postconditions) and also some lower level properties of the code (i.e. loop invariants, loop variants etc). JML annotations are written in standard Java comments, so they do not affect the normal work of any Java compiler.

An important goal in the design of JML is that it should be easily understandable by Java programmers. It is achieved by staying as close as possible to the Java syntax and semantics. The tool support for JML is rich (see [2] for an overview). In particular, there are tools that check JML specification at runtime [4], in extended static checking fashion [7], and allow to perform software certification [11]. There are also tools that support annotation generation [6],[8].

1.2 BML

The Bytecode Modelling Language (BML) is a specification language for the bytecode. It was proposed by Burdy et al. [3] The design of BML directly follows the fundamental concepts of JML. It inherits most constructions and keywords from the JML syntax. As the BML is developed within the MOBIUS [1] project and the main target of the project are Java-enabled mobile devices such as mobile phones, the current version of BML assumes some simplifications of the Java bytecode which are present in the J2ME platform – the Java platform for mobile devices with restricted resources.

The class files representing bytecode with BML annotations are regular Java class files, executable by all Java tools. The annotations are stored within additional attributes. The BML related attributes start with the prefix `org.bmlspecs` and according to the specification of the Java Virtual Machine they should be ignored by the Machine, since their names are not part of the original JVM specification.

Of course, following the logical structure of class files, class specifications are stored as class attributes, method specifications, as attributes of corresponding method and specifications inserted in the code are attributes of the JVM Code attribute of the given method.

2. ANNOTATION LANGUAGE

The structure of annotations in BML and JML is very similar. We have two main types of annotations: method annotations and data type (class and interfaces) annotations.

2.1 Method Annotations

The most important type of method annotations are *method specifications* describing the input-output behaviour of the method. This are preconditions (`requires`), defining conditions that should be fulfilled before entering the method and postconditions (`ensures`) telling what we can expect after the method finishes. One can define also which fields are modified (clause `modifies`) and which exceptions might be thrown (clause `signal`).

The other type of method annotations are specifications elements appearing in the code, like:

- Assert instructions that state some facts about fields, variables etc. that should hold at this point of program execution.
- Loop specifications that describe the loop invariants (`loop_invariant`), loop variants, like `decreases` to prove the loop's termination or `modifies` that tells which fields or variables can be changed in this loop.
- Declarations of local `ghost` variables - variables that exist only in the specification. Their values can be modified only using special `set` instructions.

- `Set` instructions are similar to Java assingments, but they operate on ghost fields and variables.

2.2 Data Type Annotations

Class (and interface) specifications describe the behaviour of a class as a whole (in the `static` version) or of objects of that class (`instance`). The most important type of *class specifications* are class invariants. They describe the property that should hold for all objects of this class in all *visible* states, i.e. after all constructors and before and after all methods. For example, having a field `Object[] list`, one can write an invariant that the list is never null and its length is 10. Class invariants can be seen as additional, implicit preconditions and postcondtitions for all methods in the class.

Other important class specifications are:

- Declarations of `ghost` fields. They are similar to local ghost variables, but are visible in the whole class scope.
- Model fields – fields present only in the specifications, representing some more complicated formulas. For example one can create a model field representing the property that a collection does not contain nulls.

More details can be found in [10] and [5].

3. JML2BML COMPILER

In our work we have designed and implemented a tool called *Jml2Bml* that compiles JML specifications into BML. It takes as input a Java source file with JML annotations, the corresponding class file and outputs the class file with inserted proper BML annotations. Our compiler uses an enhanced Abstract Syntax Tree for the Java source code, taken from the *OpenJml* compiler. For different types of JML clauses, there are separate translation rules defined. At each node of the AST, all translation rules are applied. If some rule succeeds to translate this node, the result is stored in to the class file, using the *BMLlib* library [13]. This approach makes the compiler easily extensible. One can simply write new translation rule to support additional features of the JML language.

Currently the *Jml2Bml* compiler focuses on subset of the JML called JML Level 0. Due to external libraries limitations not all desired features are translated, for example the `loop_modifies` clause is not supported by the *OpenJml*.

The *Jml2Bml* is designed to be compatible with other bytecode level tools, such as the bytecode editor *Umbra*.

3.1 Architecture Description

In this section we present the overview of the architecture of Jml2Bml compiler. The *Jml2Bml* compiler uses *OpenJml* to parse the Java source code together with the JML annotations. To insert generated BML annotations, the *BMLLib* library is used. The dependencies between internal packages of *Jml2Bml* and *OpenJml* and *BMLlib* are presented in figure 3.1.

The `jml2bml.main` package provides the entry point to the application. JML annotations from given source file will be translated and inserted into corresponding `class` file. Functions to access some bytecode information are located in `jml2bml.bytecode` and helpers to *BMLlib* are collected in `jml2bml.bmllib`. The `jml2bml.rules` package contains translation rules for different aspects of JML. It should be easy to add new rules in the future. Classes for traversing the Java abstract syntax tree can be found in `jml2bml.ast`. In `jml2bml.symbols` implementation of symbol table can be found. The `jml2bml.engine` package contains the core translating mechanism.

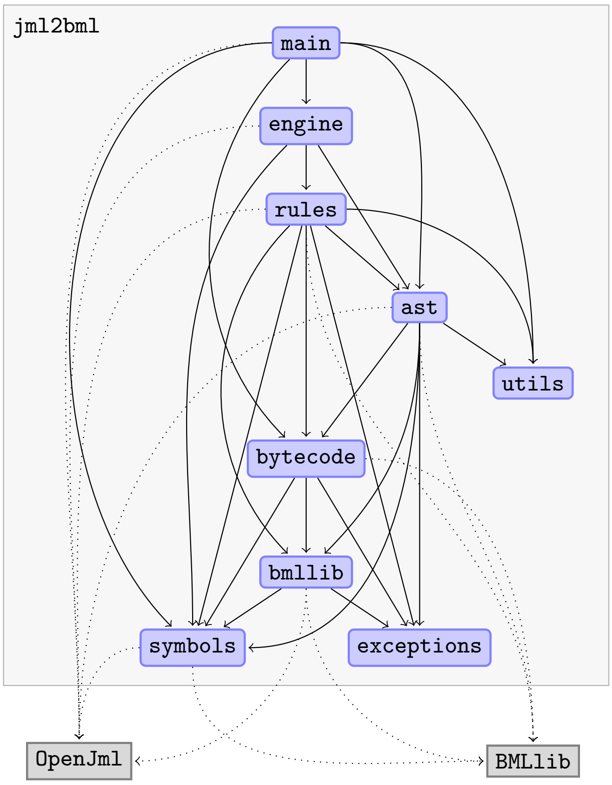

Figure 1: The dependency graph of the *Jml2Bml* packages. Dotted lines denote access to external libraries.

3.2 Translation mechanism

The full translation contains of a set of independent translation rules. Having the set of rules the AST tree of the source code with annotations is traversed. For each visited node all translation rules are applied. For most nodes translation rules do nothing - each is responsible for few node types. The translation mechanism allows to register new rules in a very simple way, what is an important issue in case of implementing new features in the future.

3.3 Translation rule

The *Jml2Bml* compiler uses a set of translation rules. The concept of translation rule is that it should be responsible for relatively small, independent piece of translation. For example we have separete rule for translating *assert* and another one for translating *loop_invariant*. Translation rule may write results of translation to the output class file (using *BMLlib*). It can hovewer only collect some translated data that may be used by other translation rules. For example both - the *assert* and *loop_invariant* annotations contain expressions. Therefore we created an expression translation rule that makes translation of an expression but it does not write anything to output file - just returns the translated expression for other translation rules.

It is relatively easy to extend translation using the translation rule concept. For example if we would like to translate an annotation that was not already translated we should create a new translation rule implementation and include it in the translation mechanism by registering the rule in the translation manager. When implementing the translation rule we should first find out which AST nodes are important and then override proper methods of the base class.

Translation rule key features are:

- the concept falls into a visitor design pattern
- translation rule is an extension of a simple abstract class
- translation process can be broken into smaller, independent pieces
- extending translation is simple

3.4 Translating expressions

To be able to translate any JML specification, one needs to translate JML expressions, so the fundamental task in writing the compiler was to write a translation rule for expressions. As BML is based on JML, the syntax of expressions is similar in both languages and includes:

- Binary arithmetic operations (+,-,*,/)
- Boolean operations
- Relational operators (<, ≤, ≥, ! =, etc)
- Logical formulas containing
  - implications
  - quantifiers (with bound variables)

As in standard Java, also in JML and BML the expressions can contain local variables, references to fields (both standard and ghost), method invocations, array access etc. There can also appear constructions specific for JML and BML, like `\old` clause. The translation of expressions is in many cases straightforward. Translation of identifiers is more complicated, because one has to distinguish between fields, ghost fields, local variables and bound variables and resolve them properly at the bytecode level.

4. AN EXAMPLE OF USING THE COMPILER

This section provides an example demonstrating the result of launching the *Jml2Bml*.

```
public class List {

  private Object[] list;

  /*@ requires list != null;
    @ ensures \result ==(\exists int i;
    @ 0 <= i && i < list.length &&
    @ \old(list[i]) == o1 && list[i] == o2);
    @*/
  public boolean replace(Object o1, Object o2){
    /*@
      @ loop_invariant i <= list.length
      @ && i >=0 && (\forall int k;0 <= k
      @ && k < i ==> list[k] != o1);
      @ decreases list.length - i;
      @*/
    for (int i = 0; i < list.length; i++) {
      if (list[i] == o1) {
        list[i] = o2;
        return true;
      }
    }
    return false;
  }
}
```

Figure 2: An example class `List.java` containing single method `replace`.

4.1 Source Code

Consider the class presented in figure 2. This is an excerpt of a class which implements a sequence of objects. We present here only one method that replaces in the `list` array the first occurence of its first parameter with the second one. True will be returned, if and only if such an element was found.

The presented code, apart from standard Java statements, contains also specifications in the JML. There is a precondition (`requires ...`) for the method `replace` defined. It requests that every time the method is invoked, the field `list` it not `null`. The next three lines (starting with `ensures` constitute the method postcondition. It states that, if the precondition was fullfiled, then the method result is true if and only if there was an element in the `list` which value has been updated from `o1` to `o2`. Note that the postcondition makes use of some JML features, like `\result`, `\old` or `\exists`. This postcondition does not describe all properties of this method. For example an implementation that replaces all elements in the `list` up to the first occurence of `o1` with `o2` will fulfill this specification.

In addititio to specification describing input-output behaviour of the method, also the loop implementing the `replace` method is annotated. The `loop_invariant` clause contains the invariant: a formula that should hold at the beginning of the loop body at each loop iteration. In this example it states that in iteration `i` there are no occurences of `o1` in `list` on positions before `i`. The annotation `decreases` describes the loop variant. It specifies an expression (in this case `list.length - i`) which value is decreased in each loop iteration by at least one.

4.2 Bytecode

In this section we describe the result of translating the source code from Figure 2. Since the binary class files are not human readable, we rely on its textual representation obtained from the BMLlib. The Figure 4.2 shows the translated `replace` method together with BML annotations inserted by our *Jml2Bml* compiler. Lines 0 and 1 correspond to initialization the `i = 0`. The loop is located between lines 2 and 33. Lines 5 - 12 represent the `if` statement, 15 - 23 correspond to lines ??? from the source code. Loading loop condition parameters is located in lines 27 - 32 and 33 performs the loop condition comparison.

The `requires-enusers` pair is translated into input-output behaviour BML specifications located just before the method code. Loops specifications are located after line 32 in the presented listing. The Jml2Bml compiler detects loops in the bytecode and inserts the annotation before the statement representing the loop condition. In this case it is the `if_icmplt` instruction comparing `i` and `list.length`. For more details about detecting loops refer to section 5. The modifies clause describes set of variables modified by the loop. Currently, because of *OpenJML* limitations it is not supported by our compiler (the default value `everything` will be inserted).

5. DETECTING LOOPS IN BYTECODE

To be able to compile the JML loop invariants, one should detect in the bytecode the corresponding loop. The created BML annotation should be associated with the bytecode instruction that represents the loop condition. Note that the loop condition is translated into multiple bytecode instructions. We are interested in the last one (comparison). A loop can be translated in one of the following ways:

In the first scenario, in the vertex a, an unconditional jump (goto) to the vertex c is done (vertex c denotes loading the condition). In d the condition is checked, and if it is fulfilled, we jump back to b. Between b and c is the loop body. The annotation should be added to the vertex d. In the second approach, the condition is tested at the beginning (a puts the condition on the stack and b checks it. If it is fulfilled, we enter the loop, otherwise we jump out). In c an unconditional jump back to a is done. The BML annotation should be associated with the instruction in the vertex a.

`Do-while` loops and loops with always true condition (i.e. `while(true){...}` or `for(;;){...}` are usually compiled in another way:

Before entering the loop (between a and c), no condition is checked. There might be some `break` inside (vertex b). In this cases, the annotation should be added to a (start of the loop). The Jml2Bml compiler covers all the cases described above. It tries to detect the first kind of loop. If it fails, tries to detect the second one. At the end checks the `do - while` case. In the first case:

- assume that the tested instruction is in vertex c. Consider all incoming edges that start in a vertex v, which is before c
- if there are no such vertices, return null (tested instruction is not the c vertex in the first kind loop)

```
/*
 * \requires true
 *  {|
 *    \precondition list != null
 *    \ensures \result ==
 *       (exists int i; 0 <= i && i < list.length
 *          && old_list[i] == o1 && list[i] == o2)
 *  |}
 */
public boolean replace(Object o1, Object o2)
0:     iconst_0
1:     istore_3
2:     goto          #27
5:     aload_0
6:     getfield      main.List.list
9:     iload_3
10:    aaload
11:    aload_1
12:    if_acmpne     #24
15:    aload_0
16:    getfield      main.List.list
19:    iload_3
20:    aload_2
21:    aastore
22:    iconst_1
23:    ireturn
24:    iinc          %3   1
27:    iload_3
28:    aload_0
29:    getfield      main.List.list
32:    arraylength
/*
 * \loop specification
 *    \modifies everything
 *    \invariant i <= list.length &&
 *        i >= 0 &&
 *        (forall int k; 0 <= k &&
 *             k < i==>list[k] != o1)
 *    \decreases list.length - i
 */
33:    if_icmplt     #5
36:    iconst_0
37:    ireturn
```

Figure 3: The method `replace` in the `List.class`

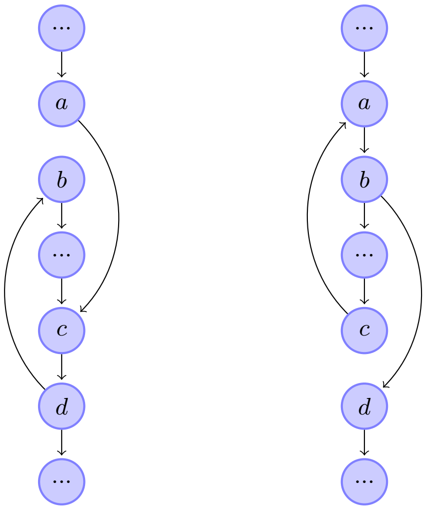

Figure 4: Two ways of compiling loops.

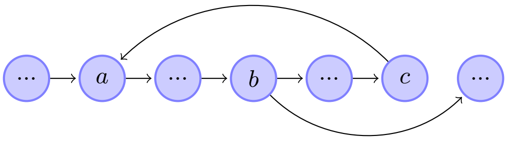

Figure 5: Compiling `do-while` loop.

- else take this v that has the longest jump to c (other jumps come from some continue instructions inside the loop). This is a from our graph,
- look at the next instruction. This is our b. Find the longest backward jump. It is our vertex d
- return d

If no loop of the first kind was detected for an instruction, try to detect the second kind:

- assume that the tested instruction is a.
- if the instruction has less than two incomming edges - return null
- find v that is after a and has the longest jump to it. This is c from our graph.
- look at the next instruction d.
- find such u that there exist an edge (u,d) and u is between a and d and there is no such u' that u' is between a and u and there is an edge (u',d). This is candidate for b
- if at u is an unconditional jump (goto), then this is a break - this is the case of loop with always true condition. Return a
- else (at u is a conditional jump) - u is really our b. Return it.

If both cases described above fail, the algorithm tries to detect the `do - while` loop. We simply check if

- there is a backward jump from the tested instruction to some a
- if yes, assuming that cases 1 and 2 failed, return a as the beginning of the loop

6. MATCHING BYTECODE LOOPS WITH SOURCE LOOPS

Let us take any source method that has loop with JML invariant and bytecode corresponding to this method. The compiler used to generate the bytecode may have used some optimizations. Unfortunately this causes some problems. Here are some exemplary loop optimizations:

- loop unwinding (loop unrolling)
  In this case invariant should be put in every copy of the loop. But the invariant may need a change.
- loop interchange
  If internal loop invariant depended on external loop variables - this invariant must also be translated.
- code-motion
  Some part of code may be moved before the loop. What if invariant depended on it?

When we want to add BML specifications to optimized bytecode, we have to know the optimizations that were used. There are two solutions of this problem:

- Include the JML to BML compiler in existing Java compiler application.
- Use non-optimizing compiler.

The first solution has the advantage that even optimized bytecode may be annotated. Unfortunately it would have to use an existing compiler infrastructure so every change in the compiler might reflect a change in translating annotations. This would be very difficult and complicated.

In second solution translations are more predictible and simplier compared to the optimizing compiler. Different compiler implementations may be used eg. Jikes or the reference one.

In both cases class level annotations (method pre-post conditions, class invariants) can be translated in the same way. Therefore when we limit to only these annotations we can use Jml2Bml compiler even with an optimizing compiler.

7. RELATED WORK

Jml2Bml compiler uses external libraries that are still in developement (*BMLlib*, *OpenJML*). They do not provide whole functionality needed to compile the JML Level 0. When they get more powerful, also our compiler should be enhanced to support more sophisticated specifications (and therefore get more adequate for the end user). The most urgent are the `modifies` clauses and `set` instructions.

8. CONCLUSION

We have presented *Jml2Bml* compiler that deals with JML annotations and translates them into BML. The resulting annotations are inserted in binary format into the class file (using the *BMLlib* library). The compiler is an important step in building a common verification platform for all languages compiled to the Java bytecode.

9. REFERENCES

[1] Mobius – mobility, ubiquity and security. enabling proof-carrying code for java on mobile devices.

[2] L. Burdy, Y. Cheon, D. Cok, M. Ernst, J. Kiniry, G. T. Leavens, K. R. M. Leino, and E. Poll. An overview of jml tools and applications. *FMICS: Eighth International Workshop on Formal Methods for Industrial Critical Systems*, 2003.

[3] L. Burdy, M. Huisman, and M. Pavlova. Preliminary design of bml: A behavioral interface specification language for java bytecode. *Fundamental Approaches to Software Engineering (FASE 2007)*, pages 215–229, 2007.

[4] Y. Cheon and G. T. Leavens. A runtime assertion checker for the java modeling language. *Proceedings of the International Conference on Software Engineering Research and Practice (SERP '02)*, 2002.

[5] J. Chrząszcz, M. Huisman, J. Kiniry, M. Pavlova, and A. Schubert. *BML Reference Manual*, 2008.

[6] M. Cielecki, J. Fulara, K. Jakubczyk, . Jancewicz, A. Schubert, J. Chrząszcz, and . Kamiński. Propagation of jml non-null annotations in java programs. *PPPJ'06: Proceedings of the 4th international symposium on Principles and Practices of Programming in Java*, pages 135–140, 2006.

[7] D. Cok and J. Kiniry. Esc/java2: Uniting esc/java and jml: Progress and issues in building and using esc/java2 and a report on a case study involving the use of esc/java2 to verify portions of an internet voting tally system. *Construction and Analysis of Safe, Secure and Interoperable Smart Devices: Proceedings of the International Workshop CASSIS 2004*, pages 108–128, 2005.

[8] M. D. Ernst, J. H. Perkins, P. J. Guo, S. McCammant, C. Pacheco, M. S. Tschantz, and C. Xiao. Daikon system for dynamic detection of likely invariants. *Science of Computer Programming*, 2007.

[9] A. Gal, M. Franz, and C. W. Probst. Java bytecode verification via static single assignment form.

[10] G. Leavens, E. Poll, C. Clifton, Y. Cheon, C. Ruby, D. Cok, and J. Kiniry. *JML Reference Manual*, 2005.

[11] C. Marche, C. Paulin-Mohring, and X. Urbain. The krakatoa tool for certification of java/javacard programs annotated in jml, 2004.

[12] G. C. Necula. Proof-carrying code. In *Conference Record of POPL '97: The 24th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, pages 106–119, Paris, France, 1997.

[13] A. Schubert, J. Chrząszcz, T. Batkiewicz, J. Paszek, and W. Wąs. Technical aspects of class specification in java byte code. *Proceedings of Bytecode'08*, 2008.