



Universidade Federal de Pernambuco

Centro de Informática

Pós-Graduação em Ciência da Computação

**Whole Program Optimizations of J2ME bytecode**

by

Tarcisio Pinto Camara

**Dissertação de Mestrado**

Recife, agosto de 2004



UNIVERSIDADE FEDERAL DE PERNAMBUCO  
CENTRO DE INFORMÁTICA

TARCISIO PINTO CAMARA

**Whole Program Optimizations of J2ME bytecode**

*Este trabalho foi submetido à Pós-Graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Mestre em Ciência da Computação*

Orientador: Prof. Dr. André L. M. Santos

Co-orientador: Prof. Dr. Geber Lisboa Ramalho

Recife, 17 de agosto de 2004



## ACKNOWLEDGMENTS

I thank God and my parents for my life and my education.

I thank a lot my teachers, André Santos and Geber Ramanho, for all dedication, attention and motivation that make this work possible.

I could not miss to thank Cleber Zanchettin and Joel da Silva, two new brothers that the life gave to me, for being my family during this journey. Thanks guys.

Finally, I would like to thank Eric Lafortune, for developing and publishing ProGuard, as well as C.E.S.A.R/Meantime for opening the source code of some J2ME applications and allowing us to publish results; and I also thank CNPq (Brazilian Research Council) for supporting this research project.



## RESUMO

Aplicações para os dispositivos móveis, como telefones celulares e *paggers*, implementadas em J2ME (*Java 2 Micro Edition*) são desenvolvidas sob severas restrições de tamanho e desempenho do código. A indústria tem adotado ferramentas de otimização, como *obfuscators* e *shrinkers*, que aplicam otimizações de programa inteiro (*Whole Program Optimizations*) considerando que o código gerado não será estendido ou usado por outras aplicações. Infelizmente, os desenvolvedores frequentemente não conhecem (ou não confiam) suficientemente nestas ferramentas e continuam sacrificando a qualidade do código na tentativa de otimizar suas aplicações. Este trabalho apresenta um estudo original identificando a efetividade das otimizações mais comuns nos *obfuscators*. Este estudo mostra também que a otimização de *Method Inlining*, conhecida pelos benefícios de desempenho, tem sido negligenciada por estas ferramentas por normalmente esperar-se que ela tenha efeito negativo sobre o tamanho de código. Assim, este trabalho contribui com uma implementação de *method inlining* entre classes e fundada no princípio de otimização de programa inteiro, capaz de melhorar tanto o tamanho do código como o desempenho da aplicação, ao remover cerca de 50% dos métodos alcançáveis. Finalmente, na tentativa de ajudar os desenvolvedores a tirar o melhor proveito destas ferramentas, o estudo inclui também um guia de boas práticas de programação considerando as otimizações implementadas pelos *obfuscators*.





## ABSTRACT

Applications for mobile devices, like cell phones and pagers, implemented in the *J2ME Platform* (Java 2 Micro Edition) are developed under strong performance and code size constraints. Industry has been adopting optimization tools, such as obfuscators and shrinkers, which apply *Whole Program Optimizations* considering that the generated code will not be extended or used by other applications. Unfortunately, developers often don't know (or don't trust in) these tools enough and keep sacrificing code quality in order to optimize their applications. This work presents an original study identifying the effectiveness of the most common optimizations in the obfuscators. This study has shown us that *Method Inlining*, an important optimization with known performance benefits, has been disregarded by these tools since it often has negative effects on code size. Thus, this work contributes with a cross-module whole-program method inlining implementation that improves both performance and application code size, while removing around 50% of the reachable methods. Finally, in order to help developers to take the best advantage of these tools, we have also included a best programming practices guide considering the optimizations implemented by obfuscators.



# TABLE OF CONTENTS

<b>1</b>	<b>INTRODUCTION.....</b>	<b>1</b>
1.1	WORK DESCRIPTION .....	1
1.2	DOCUMENT STRUCTURE.....	1
<b>2</b>	<b>JAVA 2 MICRO EDITION.....</b>	<b>1</b>
2.1	J2ME HISTORY.....	1
2.2	J2ME ARCHITECTURE .....	1
2.3	KVM/CLDC CONSTRAINTS .....	1
<b>3</b>	<b>WHOLE PROGRAM OPTIMIZATION .....</b>	<b>1</b>
3.1	WHOLE PROGRAM OPTIMIZATION IN J2ME .....	1
3.2	WHOLE PROGRAM OPTIMIZATION IN OBFUSCATORS .....	1
3.3	METHOD INLINING IN OBFUSCATORS.....	1
3.4	FINAL REMARKS.....	1
<b>4</b>	<b>METHOD INLINING AND THE PROBLEM OF CODE SIZE INCREASE .....</b>	<b>1</b>
4.1	INTERNAL STRUCTURE OF THE JAVA VIRTUAL MACHINE .....	1
4.2	BYTECODE EXPANSION WHEN COPYING .....	1
4.2.1	<i>Creation of temporary variables .....</i>	<i>1</i>
4.2.2	<i>Re-indexing variable instructions.....</i>	<i>1</i>
4.2.3	<i>Jump instructions.....</i>	<i>1</i>
4.2.4	<i>Replacing return by goto.....</i>	<i>1</i>
4.2.5	<i>Switches.....</i>	<i>1</i>
4.2.6	<i>Accesses to the constant pool .....</i>	<i>1</i>
4.3	THE DECISION ALGORITHM.....	1
4.3.1	<i>Call graph.....</i>	<i>1</i>
4.3.2	<i>Restrictions imposed by the Java Virtual Machine.....</i>	<i>1</i>
<b>5</b>	<b>PROPOSED METHOD INLINING TECHNIQUE .....</b>	<b>1</b>
5.1	IMPLEMENTATION DECISIONS.....	1
5.2	COPY AND MODIFICATION OF THE BYTECODE .....	1
5.3	DECISION ALGORITHM .....	1
<b>6</b>	<b>EXPERIMENTAL RESULTS.....</b>	<b>1</b>
6.1	EXPERIMENTS SETUP .....	1
6.2	PARAMETERIZATION .....	1
6.3	OPTIMIZATION OCCURRENCE .....	1
6.4	EXECUTION MEMORY AND PERFORMANCE .....	1

6.5	CODE SIZE REDUCTION BY OBFUSCATORS .....	1
<b>7</b>	<b>BEST PROGRAMMING PRACTICES .....</b>	<b>1</b>
7.1	SITUATIONS ALREADY RESOLVED BY OBFUSCATORS .....	1
7.1.1	<i>No identifier needs to be shorted.....</i>	<i>1</i>
7.1.2	<i>Unused features in frameworks do not need to be suppressed .....</i>	<i>1</i>
7.1.3	<i>Primitive type constant values do not need to be substituted by hand.....</i>	<i>1</i>
7.1.4	<i>Fields do not need to be made public to avoid field access methods (get and set).....</i>	<i>1</i>
7.1.5	<i>Long methods can be divided into small context methods called once.....</i>	<i>1</i>
7.2	SITUATIONS NOT RESOLVED BY OBFUSCATORS .....	1
7.2.1	<i>Constant propagation is not still available.....</i>	<i>1</i>
7.2.2	<i>Dead code elimination is not still available .....</i>	<i>1</i>
7.2.3	<i>Control flow analysis is not still available .....</i>	<i>1</i>
7.2.4	<i>Devirtualization is not still available.....</i>	<i>1</i>
7.2.5	<i>Merging of adjacent superclass is not still available .....</i>	<i>1</i>
7.2.6	<i>Call graph considers objects instantiated anywhere, not only locally .....</i>	<i>1</i>
7.3	SITUATIONS THAT JEOPARDIZE OBFUSCATORS .....	1
7.3.1	<i>Reflection API usage .....</i>	<i>1</i>
7.3.2	<i>Relative resource addressing.....</i>	<i>1</i>
7.3.3	<i>Unnecessary code.....</i>	<i>1</i>
7.3.4	<i>Throwing exceptions.....</i>	<i>1</i>
7.3.5	<i>Synchronization .....</i>	<i>1</i>
7.3.6	<i>Switches.....</i>	<i>1</i>
7.4	OTHER RECOMMENDATIONS .....	1
7.4.1	<i>Do not initialize big arrays in line.....</i>	<i>1</i>
7.4.2	<i>Types byte, short, char and boolean are usually converted to int .....</i>	<i>1</i>
7.4.3	<i>Avoid nested and anonymous classes .....</i>	<i>1</i>
7.4.4	<i>Avoid reinvent API (Application Program Interface) already available .....</i>	<i>1</i>
7.4.5	<i>Reuse objects .....</i>	<i>1</i>
<b>8</b>	<b>RELATED WORKS .....</b>	<b>1</b>
8.1	LANGUAGE-INDEPENDENT METHOD INLINING RESEARCHES.....	1
8.2	METHOD INLINING IN COMPILERS AND TOOLS .....	1
8.3	RELATED WORKS OF BEST PROGRAMMING PRACTICES .....	1
<b>9</b>	<b>CONCLUSIONS .....</b>	<b>1</b>
9.1	CONTRIBUTIONS.....	1
9.2	FUTURE WORK .....	1
	<b>REFERENCES.....</b>	<b>1</b>

# FIGURES

FIGURE 2-1: JAVA 2 EDITIONS AND THEIR TARGET MARKETS [KVMWP].....	1
FIGURE 2-2: J2ME SOFTWARE LAYER STACK [KVMWP] .....	1
FIGURE 2-3: VARIANTS OF THE JAVA PLATFORM FOR SMALL DEVICES [ORTIZ, 2002] .....	1
FIGURE 3-1: BUILD PROCESS FOR J2ME APPLICATIONS. ....	1
FIGURE 4-1: JVM RUNTIME DATA AREAS. ....	1
FIGURE 4-2: EXAMPLE OF METHOD INLINING WITH TEMPORARY VARIABLES. ....	1
FIGURE 4-3: RTA EXAMPLE. ....	1
FIGURE 5-1: MAKE ABSTRACT EXAMPLE. ....	1
FIGURE 5-2: EXAMPLE OF METHOD INLINING WITHOUT TEMPORARY VARIABLES.....	1
FIGURE 5-3: FORMULAE OF CODE SIZE INCREASE ESTIMATION.....	1
FIGURE 6-1: CODE SIZE REDUCTION BY OBFUSCATORS. ....	1

# TABLES

TABLE 3-1: OPTIMIZATIONS IMPLEMENTED BY ANALYZED OBFUSCATORS..... 1

TABLE 3-2: OPTIMIZATIONS OCCURRENCE IN OBFUSCATORS. .... 1

TABLE 3-3: OPTIMIZATION MECHANISMS USED BY OBFUSCATORS. .... 1

TABLE 6-1: BYTECODE SIZE REDUCTION BY ALGORITHM PARAMETERIZATION. .... 1

TABLE 6-2: NUMBER OF INLINED METHODS. .... 1

TABLE 6-3: NUMBER OF INLINED CALL SITES. .... 1

TABLE 6-4: REDUCTION ON TOTAL MEMORY ALLOCATED..... 1

TABLE 6-5: APPLICATION PERFORMANCE IMPROVEMENT. .... 1



# 1 INTRODUCTION

This new century is witnessing a new trend in computer science research: ubiquitous or pervasive computing. According to it, computation will be increasingly embedded in mobile devices (such as cell phones and pagers), providing to users relevant information and services anytime and anywhere. A myriad of applications from which users can daily benefit are being developed, ranging from simple e-mail systems to complex applications, such as intelligent Personal Digital Assistants, interactive multiplayer games, e-commerce location-sensitive transactions systems, and so on [Hansmann, 2003].

The main platform used by the industry for programming mobile devices is Java 2 Micro Edition (J2ME), a smaller version of the Java Standard Edition (J2SE) in order to fit it into the strong execution memory and processing constraints of these devices<sup>1</sup>. These constraints often force J2ME developers to sacrifice the object-oriented benefits and software quality recommendations, such as code legibility and ease of maintenance, in order to reduce the number of classes, methods and fields of the application and consequently its processing requirements.

A solution to face these constraints is to employ *Whole Program Optimizations* [Dean, 1996], where the application is globally transformed, considering that the optimizing tool knows the entire application code, this is the generated code will not be extended or used by other applications. The industry has been increasingly adopting this approach, using tools like obfuscators or shrinkers.

Obfuscators were originally implemented to make reverse engineering difficult, while applying some automatic transformations like reducing names of class, package and member (methods and fields). As a side effect, these optimizations also make the programs smaller. Nowadays, some of these tools, sometimes called shrinkers, have included other optimization techniques specifically aiming at reducing application size, like: removal of unused classes and members; and flattening the class hierarchy.

---

<sup>1</sup> The J2ME applications must be very small (often up to 50 kilobytes) [Knudsen, 2002b] and they have to consider processing capacity as low as that of 25 MHz processors [KVMds].



In spite of obfuscators being very popular in the J2ME community, we have observed that developers often don't know (or don't trust) these tools enough and keep sacrificing code quality in order to optimize their applications.

## ***1.1 WORK DESCRIPTION***

This work presents an original study identifying what optimizations are most common in the most popular obfuscators and where their implementations differ. It also identifies new optimizations being implemented and gives guidelines about what else could be taken into account to choose a tool.

This study highlighted that method inlining optimization has been neglected by these tools for often having as a side effect the increase of the code size. Method inlining, a well-known optimization, consists basically in choosing a certain set of call sites and replacing them by the code of the called method. This optimization presents a proved performance gain [Dean & Chambers, 1994] and it is a good solution for automatically resolving many situations the programmers try to avoid handly.

This work introduces a novel cross-module and whole-program technique for implementing method inlining optimization for the Java Virtual Machine, which both *improves the performance of the application and still tries to assure some code size reduction, while removing around 50% of the reachable methods*. This was possible by exploring, on one side, the low level characteristics of the Java Virtual Machine, in particular the way a method body is copied, and, on the other hand, considering the possibility of removing methods when all their call sites have been replaced.

For more than three years, our research team has been providing consulting services on J2ME for industrial scale applications [CESAR/Meantime]. This experience, combined with the knowledge acquired in this work concerning the internal structure of obfuscators, shown us that using the best optimizations is not enough to assure code quality of the J2ME applications. *It is essential for programmers to know the capabilities of the adopted tool in order to avoid unneeded design sacrifices and to improve optimization results*. A lot of effort is wasted avoiding situations already dealt with automatically, while other practices may confuse the optimization algorithms.

In order to reduce this problem, this work also introduces best practices for programming J2ME applications considering the optimizations implemented by obfuscators. To our knowledge, the technical literature has not covered these issues so far. The practices are organized so that we highlight situations not resolved by

obfuscators, as well as situations handled by them and practices that jeopardize their usefulness. These best practices have been used by an industrial software development team in CESAR/Meantime [CESAR/Meantime], our partner.

## *1.2 DOCUMENT STRUCTURE*

The next chapters lead the discussions in the rest of the work.

Chapter 2 presents the state of the art of the J2ME Technology, including its importance as a Java Technology, its architecture and constraints. This is important to justify extreme developers care about application size and performance.

Chapter 3 discusses the whole program optimizations and introduces our empirical study identifying which ones are most common in obfuscators and where their implementations differ. The chapter also comments why the J2ME platform benefit this kind of optimization.

Chapter 4 details the method inlining optimization and the problem of the code size expansion. This chapter contributes with a full analysis of low level factors of the Java Virtual Machine involved in the increase of code size during the optimization.

Chapter 5 introduces our proposed technique for implementing method inlining optimization, while detailing how we have combined a low level approach and the whole-program assumption to face each factor of code expansion.

Chapter 6 discusses our experiments to evaluate the impact of the proposed technique on the application code size, execution performance and memory, as well as the surprising reduction in the number of methods and calls after optimization.

Chapter 7 presents a best programming practices guide, developed considering our study about the optimizations available by obfuscators and our experience while extending one of them.

Chapter 8 lists the most important related work, including method inlining language-independent researches, and method inlining implementations in compilers and tools. The chapter also presents some works on best programming practices guides.

Finally, Chapter 9 discusses some conclusions of the work, presenting the most important contributions of this research and some future works.



## 2 JAVA 2 MICRO EDITION

The main platform used by the industry for programming mobile devices is Java 2 Micro Edition (J2ME). Some of the main benefits of using Java in these devices are:

- *Hardware-independence*: standardizing resources provided by the virtual machine allows applications to be executed in several different devices, since they share the same virtual machine.
- *Dynamic content*: new applications and application versions can be installed and configured in the devices, allowing better adequacy to user needs.
- *Developer community*: the application developing is not limited to device manufacturers. Any Java developer can, with some effort, implement applications to devices unavailable before.
- *Security*: applications security and validation can also be found in J2ME. Language constraints forbid illegal access to critical device resources, avoiding wrong or malicious instructions.

This chapter presents the state of the art of the J2ME Technology, including its importance as a Java Technology, its current architecture and possible trends.

### 2.1 J2ME HISTORY

The first version of the Java language was developed as part of the Sun's Green Project [Green] in order to create products to small computers and devices. Due to its strong portability, its potential for more complex computer environments was evident, and each new version added more and more features (Applets, AWT, RMI, JDBC, Serialization, Reflection, etc).

Soon, the number of required classes and resources made it unavailable to be implemented by simpler devices, like cell phones, *paggers*, PDAs (Personal Digital Assistance), radios, televisions, home applicants, etc. The software development to this emergent business domain was kept limited to manufacturer laboratories, often using hardware-dependent low level programming.

In the first four years of the Java platform, the language became the main choice for internet applications. Sun has introduced some other technologies addressed to small devices, but some of them did not have the expected acceptance:

- The *JavaCard* platform (1996) [JavaCard] defines a very small Java environment for *smart cards* e *Java rings* [Cameron & Day, 1998]. This platform is still well used by interested community, but it is too small to cover the needs of applications for larger devices, like cell phones.
- The *PersonalJava* platform (1997) [PersonalJava] implements an environment just a little smaller than the traditional one, used in personal computers (PC's). It addresses devices like portable computers and advanced PDA's. Unfortunately, the environment was too large to be embedded in cell phones, pagers and some simple PDA's.
- The *EmbeddedJava* platform (1998) [EmbeddedJava] was addressed towards embedded systems manufacturers, so that they could define which resources would be provided by the device's virtual machine. This flexibility made applications hardware-dependent and its development limited to the manufacturers themselves.

Recognizing that the Java architecture needed some radical reorganization [CLDC 1.0; Appendix1], Sun regrouped Java technologies in three editions, each one addressed to a specific business range:

- *Java 2 Platform Enterprise Edition (J2EE)* [J2EE]: For enterprises needing to provide Internet business server solutions.
- *Java 2 Platform Standard Edition (J2SE)* [J2SE]: For the desktop market, where applications do not need advanced features.
- *Java 2 Platform Micro Edition (J2ME)* [J2ME]: For consumer and embedded device manufacturers, as well as service providers who wish to deliver content to these devices.

Each edition defines a set of development tools, like libraries and APIs (Application Programming Interfaces), together with a Java virtual machine properly scaled for the execution environment. Figure 2-1 presents the three Java editions and their relations with available virtual machines.

Since its initial versions, the J2ME Platform was well accepted by the Java community. Actually, each new feature or API is identified as a Java Specification Request (JSR) and produced by a consortium of big industrial partners, including device manufacturers, service and content providers and software companies. In the end, the J2ME Platform has been a successful return to the initial goal of the Green Project, while making Java available for pervasive computing.

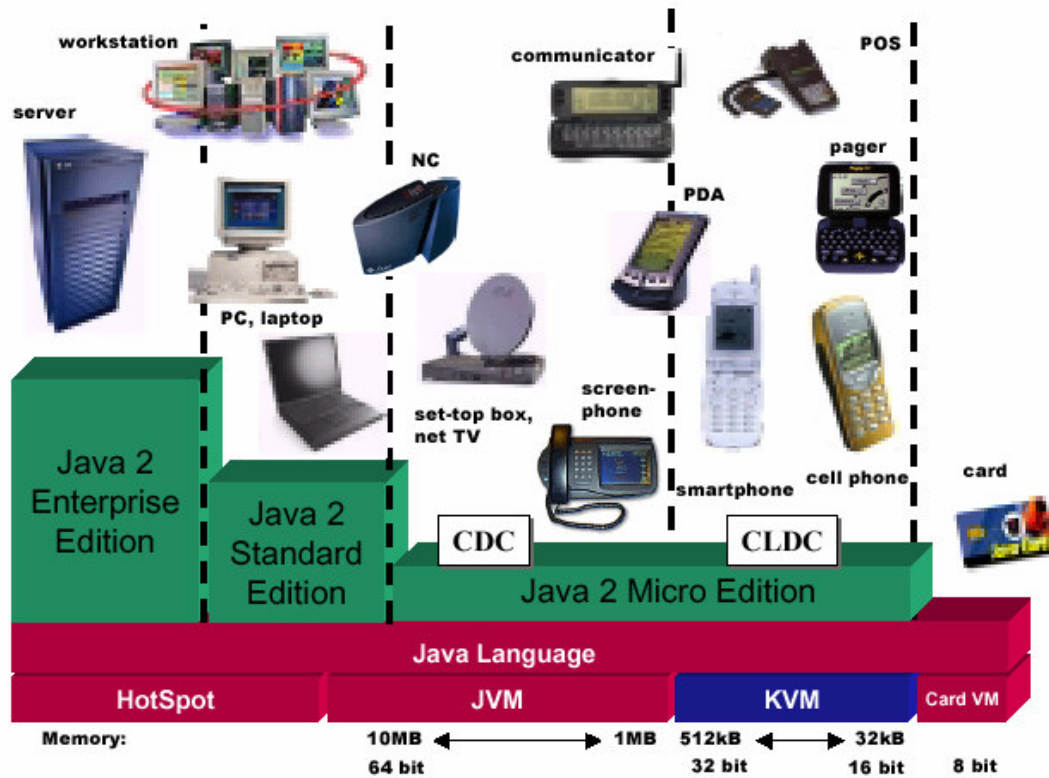


Figure 2-1: Java 2 editions and their target markets [KVMwp]

## 2.2 J2ME ARCHITECTURE

Due to the wide range of hardware and execution environments that J2ME targets, its architecture was designed to be modular and scalable, allowing flexibility while defining which features must be available for each device class. These requirements were modeled in three software layers that must be built on the Host Operating System of the device:

- *Java Virtual Machine Layer:* This layer is an implementation of a Java virtual machine that is customized for a particular device's host operating system and supports a particular J2ME configuration.
- *Configuration Layer:* The configuration layer defines the minimum set of Java virtual machine features and Java class libraries available on a category of devices and market segment. A device can support only one configuration.
- *Profile Layer:* The profile layer defines the minimum set of APIs available on a particular "family" of devices. Profiles are implemented upon a particular configuration. Applications are written for a particular profile and are thus portable to any device that supports that profile. A device can support multiple profiles.

Figure 2-2 presents a graphical representation of the J2ME architecture layers:

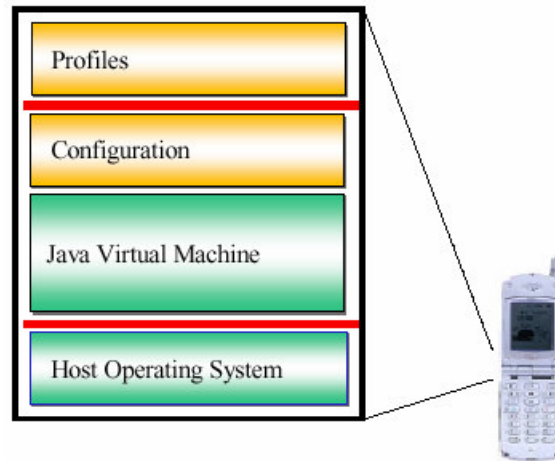


Figure 2-2: J2ME software layer stack [KVMwp]

The implementations of the configurations and virtual machines are always very closely aligned. Together they are designed to capture just the essential capabilities of each category of devices. Any differentiation into devices families must be specified in the profile layer. Nowadays, there are only two available configurations:

- *Connected Device Configuration (CDC)* [CDC 1.0]: The CDC uses a virtual machine named CVM (*Compact Virtual Machine*), with all classical resources and features, but with some constraints about memory usage. This configuration targets devices that provide at least a few megabytes of memory to the Java environment.
- *Connected Limited Device Configuration (CLDC)* [CLDC 1.0] [CLDC 1.1]: The CLDC uses a limited virtual machine named KVM (*Kilobyte Virtual Machine* or *K Virtual Machine*) and targets devices with several processing and memory constraints, making available only a few kilobytes of memory to Java.

Figure 2-1 also presents these two configurations and their relations with each virtual machine.

As the profiles address market segments, the number of profiles currently available and being developed are much wider than the number of configurations and they are continuously being reorganized. Figure 2-3 resumes the current relationship among all Java technologies for small devices, including profiles and configurations.

*Personal Profile* stack, composed by CDC [CDC 1.0], Foundation Profile [FP 1.0] Personal Basis Profile [PBP 1.0] and Personal Profile [PP 1.0], is being developed upon CDC in order to provide an environment similar to the PersonalJava Technology [PersonalJava].

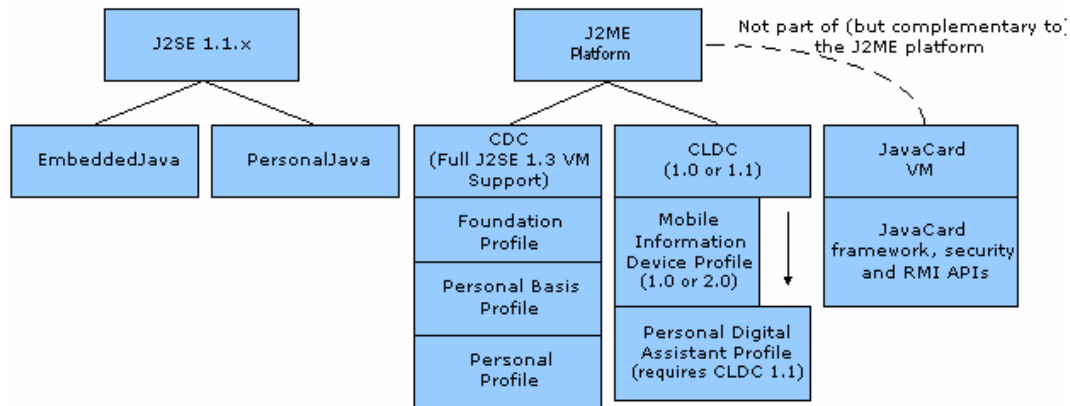


Figure 2-3: Variants of the Java Platform for small devices [Ortiz, 2002]

*Mobile Information Device Profile* (MIDP), the first and most popular profile, was developed upon CLDC addressed to *Mobile Information Devices* (MIDs), like cell phones, pagers and some simple PDAs. It is already in the second version [MIDP 2.0] and there are a number of compatible devices being sold in the world. Besides, MIDP has been used as basis for other profiles, like *Personal Digital Assistant Profile* (PDAP) that only provides some optional packages specifically for PDA applications [PDAP].

Thus, CLDC and MIDP standard are assumedly a key part of the J2ME Technology [PDAP] and they can be considered the most basic, popular and well-established J2ME stack. However, they are also one of the most limited Java execution environments, imposing several critical constraints. These constraints, which justify and require aggressive optimizations, will be discussed in the next section.

### 2.3 KVM/CLDC CONSTRAINTS

In order to allow J2ME developers to create portable applications, the profile, configuration and virtual machine specifications must require minimal resources and libraries that devices must make available to be compatible.

The high-level design goal for the KVM was to create the smallest possible “complete” Java virtual machine that would maintain all the central aspects of the Java programming language [KVMwp]. However, KVM was designed for small devices that typically contain 16-bit or 32-bit processors, clocked as low as 25 MHz, and a minimum total memory footprint of approximately 128 kilobytes [KVMds]. Regarding these requirements, KVM implementation has currently only 50-80 kilobytes of object code and needs only a few tens of kilobytes of dynamic memory to run [KVMwp]. In spite of its reduced size in memory, not much memory is left for the applications. It is easy to



find devices that reject applications larger than 50 kilobytes [Knudsen, 2002b]. Such small execution environment justifies extreme developer care about application size and performance.

The CLDC specification has currently two versions [CLDC 1.0] [CLDC 1.1]. Each one defines the subset of the Java programming language and virtual machine features that the device must provide. CLDC Specification version 1.0 defines that the supported KVM must be fully compatible with the standard *Java Virtual Machine Specification* [Lindholm & Yellin, 1999], except for the following differences:

- No floating point support
- No user-defined class loaders
- No thread groups and daemon threads
- No finalization of class instances (method `Object.finalize()`)
- Many exception and error classes are not available
- No native methods (Java Native Interface - JNI)
- No reflection (package `java.lang.reflect`)
- No weak references (class `java.lang.ref.WeakReference`)

The CLDC Specification version 1.1 was recently released and consists in an incremental release that is intended to be fully backwards compatible with CLDC Specification version 1.0 but also to address slightly bigger devices. It does not include any new major changes, just adding requirements for some features, like floating-point support and some minor library changes to make it more compatible with J2SE.

Except for these constraints, the KVM supporting each CLDC specification must be fully compatible with the standard *Java Virtual Machine Specification* [Lindholm & Yellin, 1999], including the standard *classfile* format. In fact, there is not a specific compiler for J2ME. The application is compiled on the same way, however, before being installed in the device, it is submitted to a *pre-verification* process which checks if the constraints imposed by the configuration were satisfied. That procedure still inserts some marks in the *classfile* to ease the class loading and validation tasks of the device's operating system.

Finally, both versions of the MIDP Specification [MIDP 1.0] [MIDP 2.0] were designed assuming only CLDC 1.0 features, so that they will also work on top of CLDC 1.1, and presumably any newer versions. This means that, even considering the probable MIDP stack evolution, J2ME developers must expect a severely constrained execution environment, when compared with standard Java platform. In such an environment, aggressive optimizations are unavoidable, not only to improve the application performance, but also to reduce the application size.



## 3 WHOLE PROGRAM OPTIMIZATION

*Whole Program Optimizations* consider that the optimizing tool knows the entire application code and transform it based on this assumption. They assume that the generated code will not be extended or used by other applications. This chapter discusses the state of the art of *whole program optimizations* focusing on demand of J2ME for optimization. Section 3.1 discusses why we can safely apply this kind of optimizations on most of the J2ME applications. Sections 3.2 and 3.3 present our empirical study, identifying what optimizations are most common in the most popular obfuscators and where their implementations differ. Section 3.4 discusses some final remarks about this chapter.

### 3.1 WHOLE PROGRAM OPTIMIZATION IN J2ME

*Whole program optimizations* can be safely applied on most J2ME applications because the security model defined in the CLDC Specification [CLDC 1.0] [CLDC 1.1] and the application model defined in the MIDP Specification [MIDP 1.0] [MIDP 2.0] forbid that applications interact with each other after downloaded and installed in the device. Unfortunately, the total amount of code devoted to security in Java 2 Standard Edition far exceeds the memory budget available for a Java virtual machine supporting CLDC. Therefore, some compromises and simplifications were necessary.

CLDC application-level security model uses a metaphor of a closed “sandbox” that ensures the system libraries are closed and predefined by CLDC, profiles (such as MIDP) and manufacturer-specific classes specifications. This security model still specifies that a Java application can load application classes only from its own Java Archive (JAR) file. These restrictions (about system and application classes loading) mean the application programmer can consider that no class will be dynamically loaded in execution time other those considered in design time.

The MIDP application model even allows multiple applications to be delivered in one JAR file, called *MIDlets Suite*. In these cases, each application, called *MIDlet*, can interact with each other sharing data and code, however the set of system and

classes of all applications are still predefined in the context of the JAR file, on the same way as described by CLDC security model.

Note that dynamic class loading is still available in J2ME, through the method `Class.forName(String className)`. This method allows that programmers access a class by its name (for example, to instantiate it). Even in this case, the loaded class must be inside the application JAR file or it must be one of the system classes. However, the optimizing tool is not able to automatically identify these classes as part of the application, because they are not directly referenced and the class name can be programmatically built in the execution time. In this case, optimizing tools use to make available some way to programmers indicate which classes are accessed using this mechanism. This will be better discussed in Section 7.3.1

### ***3.2 WHOLE PROGRAM OPTIMIZATION IN OBFUSCATORS***

*Obfuscators* were originally developed to make reverse engineering difficult, replacing human-readable identifiers inside the Java *classfile* with meaningless short strings, making the resulting applications more difficult to understand through *decompilation*<sup>1</sup>. As a side effect, these obfuscators also made the programs smaller [RetroGuard]. Soon, industry noticed it was possible to employ other optimizations in order to reduce even more the application size and to improve the execution time.

In this work, we use the term *obfuscator* for any tool that automatically employs *Whole Program Optimizations*, so that the submitted program is globally transformed, considering other programs will not use the generated code.

Obfuscators are often included in the build process between the compilation and pre-verification steps, acting directly over the already compiled bytecode. Thus, optimizations previously performed by the compiler are automatically kept in the final version of the application. Figure 3-1 introduces a graphic representation of the applications build cycle in J2ME.

---

<sup>1</sup> *Decompilation* is a process that generates the source code while interpreting the application bytecode.

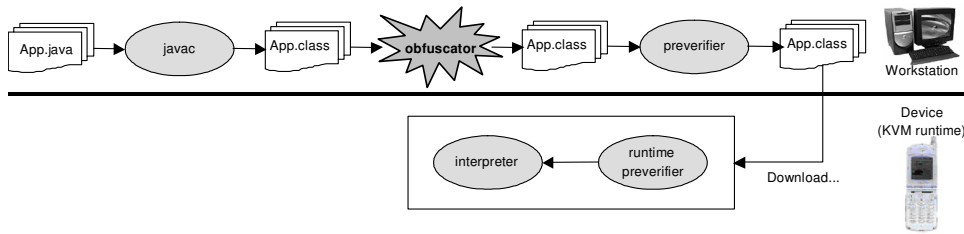


Figure 3-1: Build process for J2ME applications.

There are many available tools that can be classified as obfuscators. We studied some of the most popular tools, such as RetroGuard [RetroGuard], DashO [DashO], Jshrink [Jshrink], Jax [Jax] and ProGuard, identifying the optimizations they announced as implemented or as future work. Table 3-1 presents the name and a short description of the optimizations we found, grouped by its main goal.

Optimizations	Description
<b>Optimizations against reverse engineering</b>	
Classfile recreation	It removes unused constant pool entries and attributes used to store compiler information, such as line number of the source code and local variables names.
Class and member names compression	It replaces the class, field and method names with short names (often one letter). Overloading (methods with the same name) is used whenever possible.
Package name compression	It replaces all or part of the package name with short names (often one letter), so that the grouping of classes is kept.
Class package relocation	It moves all optimized classes to the default package, which has no name. If the moved class accesses some package or protected member in an unmoved class, these members must be made <code>public</code> to prevent access violation.
<b>Optimizations for program size reduction</b>	
Removal of unused elements	It traces and removes classes, fields and methods not referenced directly or indirectly from some start method, such as <code>startApp()</code> .
Removal of write-only fields	It traces and removes fields that are only written but never read. The instructions that wrote to the field are removed too, but the instructions that evaluated to the assigned value are kept, because they can include some side effect.
Removal of unused method body	It makes a method abstract if its body is never executed but the method cannot be removed for any reason. For example, if the method implements some interface and it is called virtually, but it does not belong to an instantiated class.
Merging adjacent superclass	It removes intermediate classes in the class hierarchy, moving all methods and fields of a class to its superclass. In order to keep the resulting objects size, either the superclass is not instantiated or the subclass has no fields.
<b>Optimizations for execution performance improvement</b>	
Devirtualization	It replaces slower virtual call instructions with faster static call instructions. In order to do that, methods are made static and private when possible.
Method inlining	It replaces some method calls with the code of the called method. It is only possible for calls that reach only one method (non-polymorphic).

Intra-procedural optimizations	Optimizations inside the method body, such as <i>constant folding</i> (evaluation of constant expressions in compile time) or <i>dead code elimination</i> (removal of write-only variables and unreachable branches) [Nullstone, 2002]. Some other optimizations, like method inlining, can create opportunities for these optimizations.
--------------------------------	--

Table 3-1: Optimizations implemented by analyzed obfuscators.

RetroGuard [RetroGuard] was developed by Retrologic as an open source project with the main goal of making applications harder to reverse engineering. It became popular because it is pre-installed in the Sun's J2ME Wireless Toolkit [J2MEWtk 1.0.4]. We used the version 1.1.9.

Jshrink [Jshrink] is a commercial tool developed by Eastridge Technology and includes a good graphical user interface for configuration and reverse engineering. It is listed by ProGuard as one of the commercial alternatives tools. We used version 2.19 and an evaluation license for the experiments.

ProGuard [Lafortune] was developed by Eric Lafortune as an open source extension of RetroGuard with the main goal of making applications smaller. It became popular because it is also pre-installed in the Sun's J2ME Wireless Toolkit [J2MEWtk 1.0.4]. We used version 1.7.2 for the experiments but ProGuard has new versions published frequently.

Jax [Jax] is a research project developed by IBM [Tip et al., 1999] [Tip & Palsberg, 2000], implementing additional complex optimizations, like *merging adjacent superclasses*. Unfortunately, its source code is not public. Nowadays the Jax project is being discontinued and integrated to the IBM development environment. We have tested the version 7.3a then available for download and free for use.

DashO [DashO] is a commercial tool developed by preEmptive Solutions that also has similar tools for .NET architecture. We used a copy of DashO Embedded Edition version 1.0 with an evaluation license for the experiments.

For each of the optimizations, we prepared an example application that was submitted to all evaluated tools, even if not mentioned in the tools' documentation. Then, we decompiled the resulting applications in order to validate the effect over the bytecode. Table 3-2 presents the list of the actual optimizations implemented in the obfuscators. We noted that there is a trend to implement optimizations for execution performance improvement. Our analysis of the generated bytecode showed that almost all optimizations have insignificant differences when implemented by different

obfuscators (marked with “X”). The only two interesting exceptions are *package name compression* and *method inlining*, which presented different results (marked with “?”).

Optimizations	RetroGuard	JShrink	ProGuard 2.1	Jax 7.3	DashO EE
Classfile recreation	X	X	X	X	X
Class and member names compression	X	X	X	X	X
Package name compression	?	?	?		
Class package relocation			X	X	X
Removal of unused elements		X	X	X	X
Removal of write-only fields				X	X
Removal of unused method body				X	
Merging adjacent superclass				X	
Devirtualization				X	
Method inlining				?	?
Intra-procedural optimizations					

Table 3-2: Optimizations occurrence in obfuscators.

*Package name compression* was implemented in different ways by the tools. RetroGuard implemented it so that only each word of the package path is compressed, keeping the number of levels of the package tree. Jshrink opted for replacing the whole package path with one letter, but keeping the groupings of classes of each package. ProGuard optionally allows all optimized classes to be moved to one user-specified package, similar to class *package relocation optimization*, but if so the groupings of classes are lost.

*Method inlining* results were even more distinct. Actually, the DashO documentation does not identify *method inlining* as having been implemented, however we found some indications of inlined method in reports generated by the tool, but only for trivial instance field access methods (non static get and set). Jax *method inlining* was mentioned briefly in some articles as being only a secondary goal and applied for methods whose only function is to set or retrieve a field’s value [Tip et al, 1999]. Understanding that method inlining is a very important optimization, we refined its analysis, as presented in the next section.



### 3.3 METHOD INLINING IN OBFUSCATORS

In order to identify the scope of the existing implementations, we opted for exploring experimentally the results of the tools (in this case DashO and Jax) when applied to carefully controlled situations.

Initially, we prepared some simple programs representing opportunities we consider promising to method inlining optimization, and we submitted them to the tools, investigating how far each tool already optimizes them.

For the experiments performed here, we used the version 7.3 of Jax and an evaluation copy of the DashO Embedded Edition. Both of them were configured so that all optimizations over method names were disabled, allowing the reverse engineering process of the resulting programs. Table 3-3 presents the sample methods, the number of times they were called and what tools dealt with them in any way, analyzing the generated bytecode.

Opportunities/Obfuscators		Number of calls	Examples	Jax 7.3	DashO EE
Instance	Trivial field access methods	3	<pre>public int getIndex () {     return this.index; }</pre>		X
	Array field access methods	3	<pre>public int getItem (int i) {     return this.items[i]; }</pre>		
	Empty methods	1	<pre>public void doNothing () { }</pre>		
	Methods called once	1	<pre>public void doSomething () {     ... }</pre>		
Static	Trivial field access methods	3	<pre>public <b>static</b> int getIndex () {     return this.index; }</pre>	X	
	Array field access methods	3	<pre>public <b>static</b> int getItem (int i) {     return this.items[i]; }</pre>	X	
	Empty methods	1	<pre>public <b>static</b> void doNothing () { }</pre>	X	
	Methods called once	1	<pre>public <b>static</b> void doSomething () {     ... }</pre>		

Table 3-3: Optimization mechanisms used by obfuscators.

While elaborating the examples, we noticed that, apparently, the tools handle **static** and **instance** methods differently. We then created additional versions of the examples, also exploring this factor.

**Trivial field access methods** is an example that recovers the value of a field or sets a value to it, being a good example of optimization opportunity without creating temporary variables, even if they have been called several times. Notice that a simple additional comparison or exception thrown can make the code non-trivial, demanding the creation of temporary variables.

**Array field access methods**, either one-dimensional or multidimensional, are good examples of small not trivial methods where generation of temporary variables is needed. However, even if the methods have been called several times, the removal of its header can compensate for the code expansion during the copying of the bytecode.

**Empty methods**, since they aren't part of interfaces implementation or a polymorphic call, they also can be removed. In this case, the optimization simply removes the method and all its call sites.

**Methods called only once** usually can be optimized and removed, regardless of their size.

In our experiments, Jax dealt many examples including non-trivial methods but we were unable to obtain any effect over instance methods, only static ones. DashO dealt only trivial field access methods and only its instance version. Anyway the obfuscators seem to have neglected method inlining while tried to apply it only where it surely does not increase the application size.

### ***3.4 FINAL REMARKS***

Our results should not be used to classify the obfuscators or identify the best one. We are only interested in (i) identifying what optimizations are more common and (ii) identifying what is the trend of new implementations. There are many other factors that must be taken into account to choose a tool, as presented above:

- *Configuration flexibility*: all obfuscator must provide some way for the user to identify which pieces of the code (classes and members) must not be optimized and the start point of the application, used to construct the call graph. Usually, this is done by configuration script for each application being optimized. However, a flexible configuration script language can allow the user to reuse the scripts in applications with the same architecture.
- *Graphical user interface*: many obfuscators provide a graphical user interface at least to help the user in the construction of the configuration scripts. However, this interface can be very powerful, including even reverse engineering tasks to allow the user to choose the unchanged code elements graphically.

- *Development environment integration:* in a software production environment, it is usual to integrate the obfuscator with the IDE (Interactive Development Environment) or some build tool, like ANT [Ant Project], so that it can be automatically executed through the development lifecycle. Some obfuscators have plug-ins to the most popular IDEs or they provide command line interfaces to easy integration.
- *Project continuity:* see if the project is really alive and its delivery of new versions and bugs fixes, as well as if there is user support available. It is always possible to choose another tool, but this can impose some work to retrain the developers and to update configuration scripts and the integration with the development environment.
- *Documentation:* check if the obfuscator has good documentation about how to use it, how to write configuration scripts and how to integrate it with the development environment. We propose the documentation should include best programming practices too, as described in Chapter 7.
- *Price/License:* of course, pay attention to the license agreement and what is needed to get new versions of the tool. Many obfuscators are GNU projects that grant free rights for use and modification. For commercial tools, they use to publish trial version free for use but often with expiration deadlines.

## 4 METHOD INLINING AND THE PROBLEM OF CODE SIZE INCREASE

Method inlining essentially consists of two parts: (i) a decision algorithm that chooses a set of method calls (call sites) to be optimized and (ii) a copy mechanism that replaces the selected call sites with the code of the method being called [Serrano, 1997].

This optimization usually brings a direct improvement on the performance of the application, since it removes the overhead for method call and return. In other words, it potentially removes all the activities related to managing the call context stack (frames). Another important method inlining benefit is to open opportunities for other intra-procedural optimizations, like *constant folding* and *dead code elimination*, since they usually can only work on continuous code blocks between calls.

Unfortunately, this technique also has a direct impact over application code size, since inlining a method replicates its code in all replaced call sites, causing code expansion. This can also degrade application performance, because it can cause “*thrashing*” on demand-paged virtual-memory systems<sup>1</sup>. In other words, if the executable size is too big, the system can spend most of its time going out to disk to fetch the next piece of code [Cline, 2003].

We believe that in order to overcome properly the problem of the increase in the application code size it is necessary to take into account the features of the underlying implementation of the programming language. In our case, this means that we should explore the runtime and bytecode features of the Java Virtual Machine. In principle, it is more powerful than the source code approach, since it is easier to trace and to control the exact impact of the changes performed by the optimization over the application size. However, intra-procedural optimizations already implemented in compilers cannot be reused. They must be implemented again on bytecode level, and performed after method inlining.

---

<sup>1</sup> It is not clear if *thrashing* is really a problem for J2ME platform, because the memory management policy is implementation-dependent and it can be specially designed for small devices [CLDC 1.0, Section 5.4.5] [Knudsen, 2002b].

On the same way, we should consider features of the applications to be optimized, like the possibility of relying on whole program analysis and optimization in J2ME. Thus, we can be much more aggressive, implementing cross-module inlining and removing methods when all their call sites have been replaced. In fact, we included the removal method benefit as a parameter of the decision algorithm.

Most results presented in the literature explore the decision algorithm with the objective of maximizing the performance of the application while trying to control code expansion. However, as far as we know, they usually do not take fully advantage of the removal method benefit as a parameter of the decision algorithm, since without the whole program assumption, only few methods, like private methods, can be removed. As we consider we know all application code, we were able to remove even public and cross-module methods if all their calls had been inlined.

In Section 4.1 we explain some of the internal structures used by the Java Virtual Machine (JVM) to keep the stack of method calls. Section 4.2 presents an extensive list of low level details of the JVM related to the size of the bytecode to be copied during inlining. We then present in Section 4.3 the factors that must be considered in any decision algorithm, discussing the influence of the call graph and the restrictions imposed by the virtual machine.

#### *4.1 INTERNAL STRUCTURE OF THE JAVA VIRTUAL MACHINE*

The specification of the Java Virtual Machine [Lindholm & Yellin, 1999] defines many data structures to manage the execution of applications. Among these structures we are particularly interested in those related to controlling method calls, notably frames, operand stack, local variables array and invoke instructions. Figure 4-1 shows a graphical representation of these JVM runtime data areas, as defined in JVM Specification [Lindholm & Yellin, 1999].

The frames are managed by each thread. They are used to store the call context of a method. Each frame contains its own array of local variables and its own operand stack. The maximum size of these two structures is defined in the method code.

The array of local variables stores the value of the parameters and local variables of the method. If it is an instance method (non static), the zero index position stores a reference to the `this` object. The positions next to it store the values of the parameters followed by the values of the local variables, in a way that variables of type `double` or `long` use two positions in the array, while the other basic types use a single position.

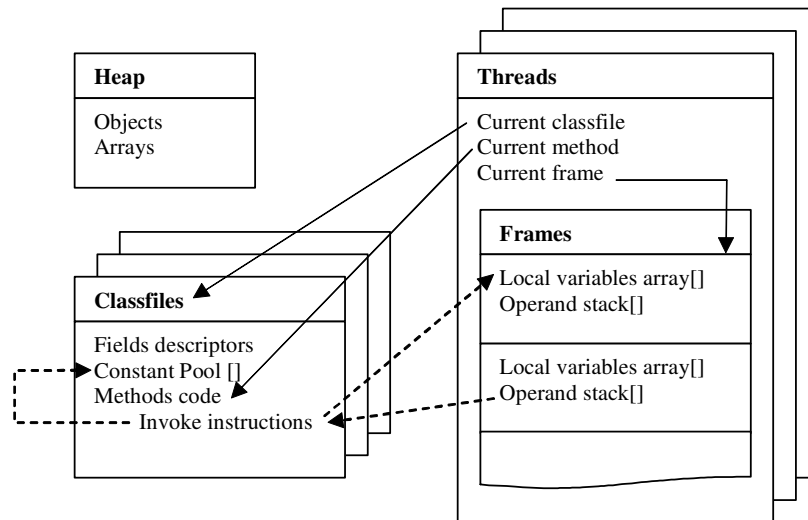


Figure 4-1: JVM Runtime data areas.

The operand stack stores the partial values resulting from the execution of the instructions. Each instruction takes its parameter(s) from the top of the stack, uses them according to the instruction semantics and eventually pushes back on the top of the stack its result. For instance, the family of `load_n` instructions is responsible for copying the value of variable  $n$  to the top of the stack. Similarly, `store_n` instructions are responsible for assigning to variable  $n$  the value on the top of the operand stack.

`Invoke` instructions are responsible for method calls. The instruction takes the method arguments from the top of the operand stack, including the reference to the called object (for non-static methods), and initializes a new frame and its array of local variables. When returning from the execution of the method, the virtual machine removes the frame and places the result returned by the method on the top of the stack.

The constant pool is also an important characteristic of the Java Virtual Machine for our work, since it is one of the structures that most contributes to application size. The constant pool is a table, present in every Java class file, containing symbolic information of all the elements accessed by the class, such as constant values (e.g. strings) and references to methods and fields from the class itself or from other classes it references. The self-sufficiency of the classfile has historical reasons: it was designed this way to ease class distribution over a network (Java applets). But this creates a significant amount of duplication of entries in the constant pools in the classes of an application. For example, if many different classes access a method from a class, each “client” class will have an entry in the constant pool referencing (naming) the method,

including its name, type descriptor and the name of the class where it is defined. Notice that calls to a same method in a given classfile share a single entry in the constant pool. The removal of methods through inlining, therefore, produces a direct impact over this structure, since it also removes entries in the constant pool related to the declaration and calls to the method.

## 4.2 BYTECODE EXPANSION WHEN COPYING

In general, the copied bytecode needs to be modified before being inserted in the code of the calling method, and many of these changes may also increase the application size.

The following sections describe these and other changes that impact on the size of the copied code, of the modified caller method code and, consequently, of the entire application. To illustrate the discussion, Figure 4-2 presents a simple example of method inlining.

Caller		Callee		Modified caller	
<b>void caller()</b>		<b>void callee(int i, int value)</b>		<b>void caller()</b>	
<b>Source code</b>	<b>Variables</b>	<b>Source code</b>	<b>Variables</b>	<b>Source code</b>	<b>Variables</b>
if ( 1 > 0 ) { callee(2, 0); }	0 this	this.f[] = value;	0 this 1 i 2 value	if ( 1 > 0 ) { Class v1 = this; int v2 = i; int v3 = value; v1.f[v2] = v3; }	0 this 1 v1 2 v2 3 v3
<b>Bytecode</b>	<b>Op.Stack</b>	<b>Bytecode</b>	<b>Op.Stack</b>	<b>Bytecode</b>	<b>Op.Stack</b>
0 iconst_1	1	0 aload_0	this	0 aload_1	1
1 iconst_0	1; 0	1 getfield #3	f	1 iconst_1	1; 0
2 if_icmple 11		4 aload_1	f; i	2 if_icmple 19	
5 aload_0	this	5 iload_2	f; i; value	5 aload_0	this
6 iconst_2	this; 2	6 iastore		6 iconst_2	this; 2
7 iconst_0	this; 2; 0	7 return		7 iconst_0	this; 2; 0
8 invokevirtual #2				8 istore_3	this; 2
11 return				9 istore_2	this
				10 istore_1	
				11 aload_1	this
				12 getfield #3	f
				15 aload_2	f; i
				16 iload_3	f; i; value
				17 iastore	
				18 return	

Figure 4-2: Example of method inlining with temporary variables.

The first column represents the caller method, the second, the callee, and the third, the modified caller after expansion due to inlining. The variables indicated in each column include, beside their names, the indexes used by variable instructions. Each column also shows the state of operand stack after each instruction.

In this example, it was necessary to create a temporary variable for each parameter, to re-index the instructions that access the variables, and to adjust the offset of the jump instructions in the modified caller method (`if_icmple` instruction). These and other transformations will be detailed in the next sections.

#### *4.2.1 Creation of temporary variables*

To create the local temporary variables, it is necessary to insert `store` instructions before the copied code in the array, in order to transfer the method arguments in the operand stack to the array of local variables. The size of each `store` instruction depends on the index of the variable it refers to, and it can use from one to four bytes [Lindholm & Yellin, 1999].

It may be necessary to generate an additional `checkcast` instruction before the `store` instruction that stores the reference to the called object, guaranteeing that the type of the created variable is compatible with the type of the object `this` expected by the callee code. This is necessary when the callee method is reachable through polymorphism. Notice that the call graph must also guarantee that only one method is reachable.

#### *4.2.2 Re-indexing variable instructions*

Once the parameters and local variables of the callee method are mapped into temporary variables in the modified caller method, all instructions accessing variables in the copied bytecode must be re-indexed to access the new variables.

Since the size of the instructions that access variables (such as `load` and `store`) depends on the index of the variable, the size of the modified bytecode may become bigger than the original bytecode. In the example in Figure 4-2 this did not take place, but since the frequency of these instructions is very high, the impact may be significant.



### 4.2.3 *Jump instructions*

The offsets of the `jump` instructions in the caller code and in the copied code must be adjusted to take into account the new inlined code.

In the example of Figure 1, the instruction `if_icmple` of the caller method had to be adjusted; now referring to a new offset (19). In theory it might be necessary to replace instructions like `goto` (3 bytes), by a bigger instruction, like `goto_w` (5 bytes), depending on the new offset. But the current version of the specification of the Java Virtual Machine imposes a restriction on the maximum size of a method bytecode [Lindholm & Yellin, 1999] that removes the need to use instructions like `goto_w`.

In short, the re-indexation of the jump instructions has no impact in code size.

### 4.2.4 *Replacing return by goto*

In case there are `return` instructions (1 byte) in the copied code, they must be replaced by `goto` instructions (3 bytes), redirecting the flow to the instruction following the copied code.

In the example of Figure 4-2 this did not happen. We even applied a small improvement on the copied code, removing the last `return` in the code, since the value returned by the method would be already in the operand stack.

### 4.2.5 *Switches*

The `tableswitch` and `lookupswitch` instructions can vary their size depending on the place where they are in the bytecode. Their offset tables must be aligned to an address that is a multiple of 4 [Lindholm & Yellin, 1999].

Thus switches in the copied code need to be modified to fit their new location in the modified caller method. In the same way, switches in the caller method may be moved if some method inlining is done before its location in the code, inserting new code. In both cases these modifications may end up increasing the size of the switches' code.

### 4.2.6 *Accesses to the constant pool*

Method inlining is often performed between methods of different classes. In this case, all entries in the constant pool accessed by the copied code must also be copied to the class that declares the caller method, if they don't exist yet.

Entries in the constant pool are one of the main factors contributing to the size of a classfile. The replication of these entries can generate a significant impact in the global size of the application. The precise measure of this impact is difficult because it's possible to share them in the scope of each classfile.

### 4.3 THE DECISION ALGORITHM

The implementation of the decision algorithm, responsible for the choice of which method calls will be inlined, is the most complex part of method inlining optimization [Serrano, 1997]. In general, if many calls are selected to be inlined, the benefit over the performance of the application is bigger, but on the other hand each inlined call has the potential of duplicating the copied code, increasing the size of the application.

Although the algorithm demands a large set of specialized information for its parameterization, the analysis uses a graph, namely the call graph, representing all the possible method calls.

In Section 4.3.1 we present the construction of the call graph and its relevance for method inlining optimization. Then the next section details the restrictions imposed by the Java Virtual Machine itself that must be taken into consideration for any decision algorithm, regardless of its objective or of the heuristics used.

#### 4.3.1 Call graph

The major difficulty in the construction of the call graph is the identification of virtual calls that can reach more than one method, through polymorphism. For example, in a call to  $e.m()$  the algorithm must decide which of the possible implementations of  $m$  may be executed from the evaluation of the expression  $e$ .

The call graph is particularly important for the method inlining optimization. Polymorphic calls cannot be directly optimized, since it is not possible to define precisely the implementation of the method that will be executed. There is also a technique, named customization, able to transform polymorphic calls in sequences of monomorphic calls, by inserting code to test the type of the expression before calling each of the possibly reachable method [Whitlock, 2000]. This technique is not explored in the context of this work since it results in an even longer code sequence.

There are many algorithms for the construction of the call graph available in the literature. Some of the most relevant ones are CHA [Dean et al., 1995], RTA [Bacon,

1997], XTA and k-FCA [Tip & Palsberg, 2000]. All of them start from the entry point of the application and traverse recursively the bytecode of the methods, analyzing the instructions that invoke methods. At each call, the method is tagged as reachable and its bytecode will be analyzed later. Methods that are never called are left tagged as unreachable.

The CHA (Class Hierarchy Analysis) is the simplest algorithm. It takes into account only the class hierarchy to determine the possible executions of a method call. However, most of the whole program tools rely on the RTA (Rapid Type Analysis) algorithm [Bacon, 1997] for this task, since its implementation is relatively straightforward and it also presents an acceptable approximation of the accesses that will occur during execution. RTA extends the CHA algorithm also taking into account, besides the class hierarchy, the class instances (new instructions) to determine reachable methods. Therefore, according to RTA, if a method is reachable and there is a call to the method  $e.m()$  in its body, then only implementations of methods with a signature compatible with  $m()$ , belonging to any instantiated subclass of the static type of the expression  $e$  are tagged as reachable.

Figure 4-3 shows a simple example where RTA can find unreachable methods. In the example, as only the class B is instantiated, RTA marks the method  $C.m()$  as unreachable, even the method  $A.m()$  have been called. In the same example, the CHA algorithm would fail and would mark  $C.m()$  as reachable.

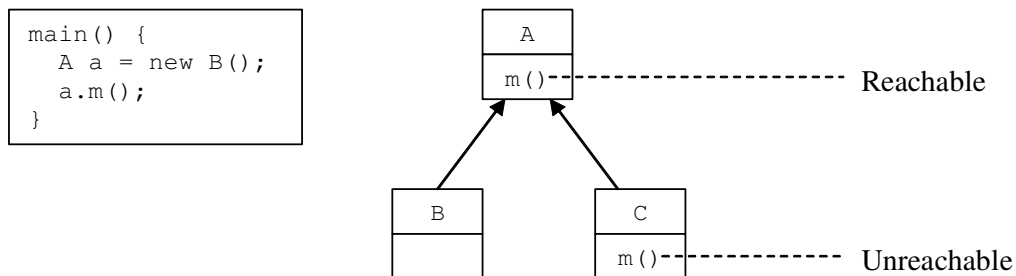


Figure 4-3: RTA example.

Other algorithms, such as XTA and k-FCA, are variations and extensions of RTA. They vary in the precision of the resulting graph, the implementation difficulty, and the amount of memory and processing time they need.

### *4.3.2 Restrictions imposed by the Java Virtual Machine*

Besides calls considered polymorphic by the call graph, the decision algorithm should reject many other situations, due to characteristics of the Java Virtual Machine itself [Lindholm & Yellin, 1999].

**Abstract** methods, by definition, cannot be optimized, since they have no code associated to them. Usually, the call graph not even identifies an abstract method as reachable. Similarly, **native** methods are rejected for not having Java bytecode, since their code is external to the virtual machine. **Synchronized** methods are also usually rejected, since they implement an implicit lock.

**Constructors** are considered in the same way as methods at the bytecode level. They would be excellent candidates for method inlining, since they can't be polymorphic. But a security mechanism of Java (the preverifier in the case of J2ME) does not allow an object to be initialized directly by the constructor of its superclass. Therefore only constructors implemented with the directive `this` may be optimized.

Methods that **catch exceptions** cannot be optimized because, when entering the catch block, the operand stack is emptied. Therefore, if the exception is caught only in the modified caller method, its behavior may be affected. On the other side, methods that **throw exceptions** can be optimized with no problem to the application execution; however they will end up changing the exception tracing info while debugging the application, since the real method that throws the exception in execution time becomes different from that written in the source code.

Methods that access special elements, such as **non-public** fields or methods declared in the same class or inherited from other classes, or even calls to methods of the superclass (using the directive **super**), must be handled specially, since it may happen that the caller method doesn't have enough permission to access these elements. In an environment that allows whole program optimization, such as in J2ME applications, some of these elements may be made public, like fields and methods defined in classes of the program. But many situations, like methods that contain calls to superclass methods or that access non-public fields or methods inherited from **libraries** can't be adequately resolved and cause the rejection of some or all of the calls to the method from the list of methods that may be inlined.

Methods that have direct or indirect **recursive** calls also need to be handled specially, since they can lead the decision algorithm to a loop.



## 5 PROPOSED METHOD INLINING TECHNIQUE

After this study of low level factors that impact method inlining optimization, we formulated our approach supported on two activities: (1) we defined some techniques and heuristics to minimize the code expansion when copying the code; and (2) we selected only the methods and call sites that, when copied, do not increase the total size of the application, considering that the method will be excluded later if all its call sites are optimized. Thus, we often manage to inline small methods, such as get and set methods, as well as methods called only once, common in many applications

Before we detail our technique, Section 5.1 presents some implementation decisions. Then, Section 5.2 defines how we handle bytecode copy problems. Section 5.3 details our proposed decision algorithm.

### 5.1 IMPLEMENTATION DECISIONS

In order to implement our solution we decided to extend ProGuard [Lafortune], one of the most popular obfuscators in the J2ME community, often cited in technical articles and development environments of Sun Microsystems [Klemm, 1999] [Knudsen, 2002a] [J2MEwtk 1.0.4]. It is also an open source project. Other tools and environments evaluated either did not provide a minimum support for a complete obfuscator [RetroGuard] [Dahm, 2002] [JikesBT] or did not publish their code [Jax] [DashO] [Jshrink].

Version 1.7.2 of ProGuard already implemented some basic optimizations like *classfile recreation*, *name compression* and *unused members elimination*. For this last optimization, ProGuard implemented a variant of the CHA (Class Hierarchy Analysis) algorithm, to trace the call graph. But it didn't build an explicit data structure for that.

The CHA algorithm is efficient enough for unused member detection, the initial goal of ProGuard, but it fails to detect non-polymorphic calls since it does not consider which classes were instantiated, generating an imprecise call graph [Bacon, 1997]. Therefore, the first change we did was the implementation of an extension of the RTA (Rapid Type Analysis) algorithm. In spite of not being the most sophisticated algorithm,

RTA is well known for being fast and very efficient for detecting non-polimorphic calls [Tip & Palsberg, 2000], a highly important feature for method inlining optimization.

Our RTA implementation is able to detect virtual calls where, even if there are several methods overriding the referenced method, only one of them belonged to an instantiated class, being marked as uniquely reachable. In this case, the referenced method can be made abstract if its body will never be executed. Figure 5-1 shows a simple example where our RTA implementation makes a method abstract. In the example, as only the class B is instantiated and it has its own implementation of `m()`, RTA marks the method `C.m()` as unreachable and makes `A.m()` abstract, so that the call `a.m()` becomes monomorphic.

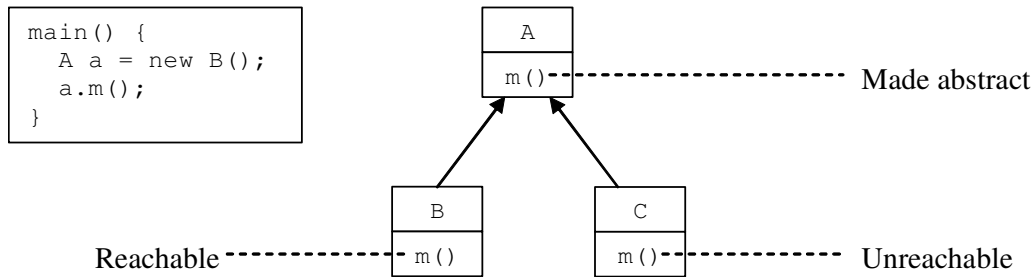


Figure 5-1: Make abstract example.

Besides the call graph construction, another auxiliary implementation needed was a bytecode handling mechanism able to modify and copy the JVM instructions. We considered to use or to integrate some bytecode toolkits already available [Dahm, 2002] [JikesBT], but in the end we decided to develop our own mechanism keeping the programming style already found in ProGuard.

After these preliminary changes, we could implement the method inlining optimization itself, presented in the next two sections.

## 5.2 COPY AND MODIFICATION OF THE BYTECODE

In order to minimize code expansion during the copy of the bytecode, initially we identified a special but very frequent situation where the first instructions of the body of the method only loads its parameters, reproducing the state of the operand stack before the call to the method.

In this case, we can copy and modify only part of the code, avoiding the creation of temporary variables and taking advantage of the previous state of the operand stack. To do that, some constraints should be satisfied, in particular parameters must not be

used again in the callee method and there must be no jumps to the region of the initial load instructions. It may seem too restrictive, but this situation handles most of the field access methods (get and set), simple functions and delegations. We named this mechanism “stack binding”, since it uses the operand stack to connect the copied code and the caller method context, against the “variable binding” mechanism, which uses temporary variables as described in the Section 4.2. Figure 5-2 shows a graphic representation of a method inlining using the stack binding mechanism.

Notice that while using the stack binding, it is often the case that the copied bytecode has the same size of the `invoke` instruction, keeping the same size of the code for each inlining operation. The variable binding is still needed for those cases where it is not possible to use the stack binding, or when it is required to include some `checkcast` instruction to match the type of the called object and the type of the parameter `this`, as described in the Section 4.2.

Caller	Callee	Modified caller																																																						
<div><div><b>void caller()</b></div><table><tr><th>Source code</th><th>Variables</th></tr><tr><td><pre>if ( 1 &gt; 0 ) {     callee(1); }</pre></td><td>0 this</td></tr></table><table><tr><th>Bytecode</th><th>Op.Stack</th></tr><tr><td>0   iconst_1</td><td>1</td></tr><tr><td>1   iconst_0</td><td>1; 0</td></tr><tr><td>2   if_icmple 10</td><td></td></tr><tr><td>5   aload_0</td><td>this</td></tr><tr><td>6   iconst_1</td><td>this; 1</td></tr><tr><td>7   invokevirtual #2</td><td></td></tr><tr><td>10  return</td><td></td></tr></table></div>	Source code	Variables	<pre>if ( 1 &gt; 0 ) {     callee(1); }</pre>	0 this	Bytecode	Op.Stack	0   iconst_1	1	1   iconst_0	1; 0	2   if_icmple 10		5   aload_0	this	6   iconst_1	this; 1	7   invokevirtual #2		10  return		<div><div><b>void callee(int value)</b></div><table><tr><th>Source code</th><th>Variables</th></tr><tr><td><pre>this.f = value;</pre></td><td>0 this 1 value</td></tr></table><table><tr><th>Bytecode</th><th>Op.Stack</th></tr><tr><td>0   aload_0</td><td>this</td></tr><tr><td>1   aload_1</td><td>this, value</td></tr><tr><td>2   putfield #4</td><td></td></tr><tr><td>4   return</td><td></td></tr></table></div>	Source code	Variables	<pre>this.f = value;</pre>	0 this 1 value	Bytecode	Op.Stack	0   aload_0	this	1   aload_1	this, value	2   putfield #4		4   return		<div><div><b>void caller()</b></div><table><tr><th>Source code</th><th>Variables</th></tr><tr><td><pre>if ( 1 &gt; 0 ) {     f = 1; }</pre></td><td>0 this</td></tr></table><table><tr><th>Bytecode</th><th>Op.Stack</th></tr><tr><td>0   iconst_1</td><td>1</td></tr><tr><td>1   iconst_0</td><td>1; 0</td></tr><tr><td>2   if_icmple 10</td><td></td></tr><tr><td>5   aload_0</td><td>this</td></tr><tr><td>6   iload_1</td><td>this; 1</td></tr><tr><td>7   putfield #4</td><td></td></tr><tr><td>10  return</td><td></td></tr></table></div>	Source code	Variables	<pre>if ( 1 &gt; 0 ) {     f = 1; }</pre>	0 this	Bytecode	Op.Stack	0   iconst_1	1	1   iconst_0	1; 0	2   if_icmple 10		5   aload_0	this	6   iload_1	this; 1	7   putfield #4		10  return	
Source code	Variables																																																							
<pre>if ( 1 &gt; 0 ) {     callee(1); }</pre>	0 this																																																							
Bytecode	Op.Stack																																																							
0   iconst_1	1																																																							
1   iconst_0	1; 0																																																							
2   if_icmple 10																																																								
5   aload_0	this																																																							
6   iconst_1	this; 1																																																							
7   invokevirtual #2																																																								
10  return																																																								
Source code	Variables																																																							
<pre>this.f = value;</pre>	0 this 1 value																																																							
Bytecode	Op.Stack																																																							
0   aload_0	this																																																							
1   aload_1	this, value																																																							
2   putfield #4																																																								
4   return																																																								
Source code	Variables																																																							
<pre>if ( 1 &gt; 0 ) {     f = 1; }</pre>	0 this																																																							
Bytecode	Op.Stack																																																							
0   iconst_1	1																																																							
1   iconst_0	1; 0																																																							
2   if_icmple 10																																																								
5   aload_0	this																																																							
6   iload_1	this; 1																																																							
7   putfield #4																																																								
10  return																																																								

Figure 5-2: Example of method inlining without temporary variables.

The impact of re-indexing variable access instructions was minimized by sharing the indexes of temporary local variables, as if they had been defined inside a block in the modified caller method. Thus, the variables’ indexes tend to be kept low even if many call sites are optimized in the same caller method. To do that, the piece of code to be copied is modified and prepared when the callee method is visited, not the caller, being attached to the respective caller site for effective insertion later in the modified caller method, when this is visited. Therefore, all callee methods are prepared for copying considering only the original variables of the caller method.



The relocation of switches in the modified caller method, in cases when a previous call site has been replaced, was handled by inserting `nop` instructions after the copied code for each of those call sites. Thus the switch instructions that follow them keep their alignment in an address that is a multiple of 4, removing the need for any correction of these instructions.

The presence of switch instructions in the copied code was not handled, causing the rejection of the method as candidate for the optimization. This decision results from the need to modify the bytecode beforehand, without effectively inserting it in the caller method. It prevents the definition of the exact location where the switch is going to be inserted in the modified code. In fact, we preferred to privilege the handling of instructions that access variables, because they are much more frequent than switch instructions.

The replacement of return instructions by `goto` instructions, the removal of the last return instruction and the handling of the constant pool entries were implemented as described in Section 4.2.

### *5.3 DECISION ALGORITHM*

The decision algorithm was designed in order to reduce the code size considering the possibility of excluding the method when all of its call sites have been optimized. Thus, in many cases, it will only be worth optimizing the calls to the method if it will be excluded after the optimization.

The algorithm visits the reachable methods of the application in an arbitrary order, deciding (a) which calls to the method will be optimized and (b) if it will be possible to remove it after the optimization. After visiting the method, and having decided to optimize part or all its call sites, this decision will never be reverted, guaranteeing that each method will be **visited up to once**. This concern with the performance of the algorithm itself is very important, with specific research work in this area [Dean & Chambers, 1994].

During the evaluation of each method, the algorithm calculates an **estimate of the application size expansion** and verifies whether the resulting value is smaller than a certain limiting factor. This factor is the main parameter of the algorithm, named `MAX_EXPANSION`. If that limiting factor is zero, it means the algorithm doesn't tolerate any size increase. Higher values allow more methods to be optimized, improving the application performance, but with a smaller percentage of code reduction,

as shown in Chapter 6. Figure 5-3 details the parameters considered in the calculation of that estimate.

<p>NCS : number of call sites to be optimized          CDL : length of the code array to be modified          CSL : length of the code array occupied by all                invoke instructions of the call sites to                be optimized          RMV : flag indicating if the method will be                removed          MDL : total size of the method, including its                header</p> <p><math display="block">CUST = (NCS * CDL) - CSL - (RMV ? MDL : 0)</math></p>
--

Figure 5-3: Formulae of code size increase estimation.

Of course, the method can only be removed (indicated by flag RMV) if all its call sites were selected to be inlined. However, it may still be worthwhile to inline only a subset of the call sites, depending on the length of the copied code array (CDL) and size occupied by replaced invoke instructions (CSL).

Additionally, notice that in the formulae of Figure 5-3 the value MDL is defined as the total size used by the method, including its header. Since the header is formed by **constant pool** entries, the exact determination of which entries will be able to be excluded is not a simple task, due to the possibility that they are being shared by other elements of the classfile. Therefore, the estimation of the code size expansion cannot be fully exact. We tried to make it as closer as possible, predefining a fixed average size to method headers.

Initially, the evaluation of each method considers an ideal situation where all calls to the method will be optimized, enabling the exclusion of the method. It also considers that it is possible to apply the stack binding mechanism, reducing the code size to be copied (CDL). **The estimate is then performed successively**, while the evaluation of the method validates each part of this initial hypothesis, in order to reject it as soon as possible. For example, if the algorithm detects that not all the calls to the method are monomorphic, or that it is not really possible to perform the stack binding mechanism, the estimation is performed again, and it could reject the optimization on the method.

In the context of the estimative of the Figure 5-3, the size of the copied code is considered the same for all the optimized calls. However, several factors presented in

the Section 4.3.2 show that this is not true. Therefore, at the end of the analysis of each method, a **last verification** is still performed considering all the real parameters, including the total size of modified bytecode for each call site. This is the last situation where the optimization of the method being visited may still be rejected.

Besides these considerations about the impact on the application size, the constraints imposed by the virtual machine, shown in the Section 4.3.2, should also be taken into account. They could influence in the number of call sites to be optimized or even reject the whole method.

Thus, **abstract**, **native**, **synchronized** and **constructor** methods are rejected, as well as methods that catch exceptions or that contains direct recursive calls. For **indirect recursive** calls, the first method where the cycle is detected is rejected, allowing the rest of the methods in the cycle to be optimized, merging them into the first method and generating direct recursive calls.

Methods that **throw exceptions** are inlined normally, assuming the behavior change over exception tracing due to the fact that the real method that throws the exception in execution time becomes different from that written in the source code.

When accessed by an inlined method, **non-public** members (fields and methods) declared in the same class or inherited from other classes are often made public. This is supported by the whole program optimization. Exceptions to this rule refer to members inherited from **library** classes, which can't be made public; and **private** methods, which are not made public because the `invokeSpecial` instruction used for private methods is faster than the one used for public methods (`invokeVirtual`) [Lindholm & Yellin, 1999].

In all situations where the accessibility problem is not resolved, that is, members that can't be made public and methods with special **super** calls, the algorithm does not reject the method immediately, but it goes ahead considering **only calls in the same class**, then it evaluates again the expansion estimative and the possibility of rejecting the inlining of the method.

In spite of these restrictions, this decision algorithm caught all example opportunities introduced in Table 3-3. It is still able to remove almost all of the optimized methods, as shown in Section 6.3.

## 6 EXPERIMENTAL RESULTS

In order to evaluate our proposed method, we compared the original ProGuard with our ProGuard version. For a given set of applications, to which the optimizations should be applied, the evaluation criteria were percentage of the code reduction and the execution performance gain.

Sections 6.1 and 6.2 present the experiment setup and the process of adjusting our algorithm's parameters, respectively. Then, Sections 6.3 shows the results concerning the number of methods and calls excluded during the optimization, whereas Section 6.4 discusses the optimization impact on performance and memory. Finally, Section 6.5 compares our method inlining code size reduction with those ones found in other obfuscators.

### *6.1 EXPERIMENTS SETUP*

Since there were no standard benchmarks for J2ME optimization, we have chosen some real applications provided by C.E.S.A.R/Meantime [CESAR/Meantime], a well established IT Brazilian company that works in J2ME applications since 2000. We have also included three J2SE (Java Standard Edition) applications in some experiments, to evaluate the proposed method results outside J2ME scope. The selected applications are:

- Eight J2ME games (BreakOut, Ship, Istari, Atlantis, SpaceInvaders, GoldHunter, Pacman and LightTenis) developed by C.E.S.A.R/Meantime, the first two using the wGEM game engine (framework) [Pessoa, 2001];
- Three J2SE applications (Ant, JDepend and the original ProGuard), freely available.

Games were chosen since this is typically a kind of application to which memory and processing power are critical resources. In order to assess the generality of our method, we have selected games with different styles, as well as the three non-J2ME applications. All selected J2ME applications suffered previous strong manual improvements, by a skilled software engineering team, in order to meet the processing

and memory restrictions of cell phones. Any kind of extra optimization in these applications is thus a good result.

All code size measurements are based on compressed JAR files, containing only the application classes, without resources (e.g., images and sounds). The impact on non-compressed code is not so important because the applications are often distributed compressed; for example, only JAR files can be installed in J2ME devices.

Concerning performance measurements, all J2ME case studies were tested on the emulator DefaultGrayPhone, supplied with the J2ME Wireless Toolkit 1.0.4 [J2MEwtk 1.0.4]. Non-J2ME applications were executed with Java 2 Standard Development Kit 1.4.1. The execution platform was a PC with an AMD Athlon 1.0 GHz processor and 256 Mb RAM memory, running Windows 2000 Professional. The applications were automatically compiled and compressed with the Java 2 Standard Development Kit 1.4.1.

## 6.2 *PARAMETERIZATION*

Table 6-1 presents the size reduction results varying the MAX\_EXPANSION parameter in our algorithm. Columns 1 and 2 show, respectively, the applications original size and its reduction percentage obtained with original ProGuard v1.7.2. Column 3 exhibits the size reduction achieved with our Extended ProGuard with no method inlining. Notice that Column 3 already improves slightly the size reduction percentage compared with original ProGuard (Column 2), due to our new implementation of the call graph, using RTA.

Columns 4 to 7 show the percentage of size reduction measured when applying method inlining with several MAX\_EXPANSION values, as indicated in parenthesis in the column headers. Column 4, when MAX\_EXPANSION is zero, indicates that the algorithm tries to reject any code size increase when inlining. Columns 5 to 7 show higher values that make the algorithm more tolerant to code size expansion. In the extreme case, Column 7 shows the code size reduction when MAX\_EXPANDED is 65536 (the maximum allowed method size), indicating the algorithm accept a great number of method inlining, regardless to the code size expansion.

The bold values highlight the best code size reduction for each application. Notice that sometimes (applications Istari and GoldHunter) the best value is not acquired by the most conservative parametrization (Column 4). That is because the

estimation of the code size reduction is not straightly exact. However, these best results are always slightly better those ones found in Column 4.

Applications		1	2	3	4	5	6	7
		Original size	ProGuard v1.7.2	Extended ProGuard (without inlining)	Extended ProGuard (MAX_EXPANSION = 0)	Extended ProGuard (MAX_EXPANSION = 100)	Extended ProGuard (MAX_EXPANSION = 200)	Extended ProGuard (MAX_EXPANSION = 65536)
J2ME	Atlantis	26.317	28,09%	28,88%	<b>31,27 %</b>	31,02%	31,02%	30,67%
	BreakOut	31.326	48,18%	48,70%	<b>51,40 %</b>	51,13%	50,75%	50,75%
	Ship	41.841	41,02%	41,33%	<b>44,07 %</b>	43,25%	42,38%	41,07%
	Istari	37.432	38,00%	38,46%	42,31%	<b>42,37 %</b>	41,52%	39,39%
	SpaceInvaders	41.723	36,72%	37,29%	<b>39,63 %</b>	39,21%	38,97%	38,62%
	GoldHunter	42.243	31,85%	32,60%	34,86%	<b>34,89 %</b>	34,46%	33,36%
	Pacman	52.326	55,41%	55,63%	<b>56,83 %</b>	56,67%	56,68%	56,09%
	Tenis	52.587	55,41%	55,79%	<b>57,73 %</b>	57,72%	57,61%	56,37%
J2SE	Ant 1.5.1	707.376	88,81%	90,74%	<b>90,95 %</b>	90,93%	90,89%	90,70%
	JDepend 2.6	84.535	59,73%	62,34%	<b>64,01 %</b>	63,79%	63,64%	62,89%
	ProGuard 1.7	188.547	49,45%	49,57%	<b>50,25 %</b>	49,92%	<u>49,44%</u>	<u>47,04%</u>
Average			<b>48,42 %</b>	<b>49,21 %</b>	<b>51,21 %</b>	<b>50,99 %</b>	<b>50,67 %</b>	<b>49,72 %</b>
		DEFAULT				AGGRESSIVE		

Table 6-1: Bytecode size reduction by algorithm parameterization.

For our surprise, the algorithm often keeps improving of application code size reduction, even when we extrapolate the value of the MAX\_EXPANSION, allowing all the possible methods to be optimized (Column 7). That is because the whole program assumption allows a great number of methods to be removed after inlining in aggressive approaches, as will be shown in Section 6.3. The underlined values in the bottom right corner of the table (Columns 6 and 7) indicate the only two values when that aggressive approach has generated some application code bigger than the original ProGuard result, shown in Column 2. Even in these cases, there was no explosion of application size.

In the remaining experiments, we worked with only two versions of our method, namely **default** and **aggressive**, respectively corresponding to columns 4 and 7 in Table 6-1.

### 6.3 OPTIMIZATION OCCURRENCE

Table 6-2 presents the number of optimized methods for each parameterization, classified in three categories: (i) *inlined and kept*, indicating the number and percentage of methods that were inlined but could not be removed; (ii) *inlined and removed*, indicating the number and percentage of methods that could be fully removed after inlining; and (iii) *not inlined*, indicating the number and percentage of methods that was not inlined at all. The average of optimized methods is grouped by application platform (J2ME or J2SE), in order to help the analysis of the optimization impact for each one of them.

			Default			Aggressive		
Application		Number of Methods	Inlined and kept	Inlined and removed	Not inlined	Inlined and kept	Inlined and removed	Not inlined
J2ME	Atlantis	122	0 0,00%	52 42,62%	70 57,38%	2 1,64%	54 44,26%	66 54,10%
	Istari	246	0 0,00%	143 58,13%	103 41,87%	3 1,22%	175 71,14%	68 27,64%
	Ship	220	0 0,00%	100 45,45%	120 54,55%	5 2,27%	113 51,36%	102 46,36%
	BreakOut	164	0 0,00%	79 48,17%	85 51,83%	4 2,44%	85 51,83%	75 45,73%
	GoldHunter	196	0 0,00%	84 42,86%	112 57,14%	1 0,51%	102 52,04%	93 47,45%
	SpaceInvasors	179	0 0,00%	70 39,11%	109 60,89%	5 2,79%	84 46,93%	90 50,28%
	Pacman	145	0 0,00%	60 41,38%	85 58,62%	6 4,14%	72 49,66%	67 46,21%
	Tenis	173	0 0,00%	89 51,45%	84 48,55%	2 1,16%	104 60,12%	67 38,73%
	Average J2ME			0,00 %	46,15 %	53,85%	2,02 %	53,42 %
J2SE	Ant 1.5.1	414	0 0,00%	108 26,09%	306 73,91%	9 2,17%	130 31,40%	275 66,43%
	JDepend 2.6	312	1 0,32%	95 30,45%	216 69,23%	15 4,81%	116 37,18%	181 58,01%
	ProGuard 1.7.2	1.163	1 0,09%	118 10,15%	1.044 89,77%	22 1,89%	211 18,14%	930 79,97%
	Average J2SE			0,14 %	22,23 %	77,64%	2,96 %	28,91 %

Table 6-2: Number of inlined methods.

For our surprise, the optimization was able to remove around 50% of methods in J2ME applications (46,15% in default parameterization and 53,42% in the aggressive one) and around 25% of methods in J2SE workbench (22,23% in default and 28,91% in aggressive).

We also highlight that only a few methods have been kept after inlining, usually only in the aggressive approach. This result indicates that, in order to assure code reduction, most of the method inlining opportunities seem only to be worth if the method can be removed after inlining.

Table 6-3 shows a similar measurement for calls selected by the algorithm, indicating the percentage of call sites that was inlined or not for each parameterization.

			Default		Aggressive	
Application	Number of Call Sites		Inlined	Not Inlined	Inlined	Not Inlined
J2ME	Atlantis	323	198 61,30%	125 38,70%	226 69,97%	97 30,03%
	Istari	755	379 50,20%	376 49,80%	520 68,87%	235 31,13%
	Ship	620	324 52,26%	296 47,74%	412 66,45%	208 33,55%
	BreakOut	387	231 59,69%	156 40,31%	268 69,25%	119 30,75%
	GoldHunter	691	279 40,38%	412 59,62%	479 69,32%	212 30,68%
	SpaceInvasors	476	232 48,74%	244 51,26%	329 69,12%	147 30,88%
	Pacman	433	179 41,34%	254 58,66%	281 64,90%	152 35,10%
	Tenis	804	481 59,83%	323 40,17%	594 73,88%	210 26,12%
Average J2ME			51,72 %	48,28%	68,97 %	31,03%
J2SE	Ant 1.5.1	884	253 28,62%	631 71,38%	352 39,82%	532 60,18%
	JDpend 2.6	721	289 40,08%	432 59,92%	383 53,12%	338 46,88%
	ProGuard 1.7.2	4.081	191 4,68%	3.890 95,32%	989 24,23%	3.092 75,77%
Average J2SE			24,46 %	75,54%	39,06 %	60,94%

Table 6-3: Number of inlined call sites.



As expected, the aggressive parameterization always optimizes more methods and calls than the default one.

These optimization occurrences were possible due to the generalization power of the algorithm, which was able to remove almost all field access methods (e.g. *get*, *set* and *is*), simple functions and delegations, methods called only once, small methods called few times, etc. All of these situations are very common in object-oriented applications.

Of course, the results also depend on the programming style and architecture of the application. For example, the ProGuard 1.7.2 (last application in Table 6-2 and Table 6-3) had a bad result since it uses excessively the *visitor design pattern* [Gamma et al, 1995], that produces many virtual and polymorphic calls that are not inlined by our technique.

## 6.4 EXECUTION MEMORY AND PERFORMANCE

We also evaluated the impact of the optimization on time and memory needed for the execution of the applications. For that, we modified the source code of some of the J2ME games, making them deterministic, i.e., simulating user input and removing random behavior, timers and threads usage. Additionally, we modified the source code of the ProGuard 1.7.2 to show the time and memory used while processing its own code. These modified applications were submitted to the original ProGuard 1.7.2, and to our Extended ProGuard with the default and aggressive parameterization. All optimizations available by each ProGuard versions were enabled, in order to reproduce real usage of the tools where the optimizations can interact each other.

Table 6-4 shows the total memory allocated by each modified J2ME application as shown by the emulator output. For the J2SE application, the ProGuard 1.7.2 itself, the presented value represents the instant memory allocated in the end of the execution.

All memory values were identical for all executions of the applications. Below each memory values, we inform the percentage of reduction compared to the values obtained with the non-optimized application.

The memory results were already expected. The method inlining optimization presented a little influence on the amount of used memory, when compared with the results already obtained by the original ProGuard 1.7.2.

Total memory allocated (bytes)		No optimization	ProGuard v1.7.2	Extended ProGuard (default)	Extended ProGuard (aggressive)
J2ME	SpaceInvaders	<b>492.412</b>	451.672 8,27%	444.844 <b>9,66%</b>	452.072 <b>8,19%</b>
	Pacman	<b>635.536</b>	626.696 1,39%	626.336 <b>1,45%</b>	634.364 <b>0,18%</b>
	Atlantis	<b>393.432</b>	363.144 7,70%	355.988 <b>9,52%</b>	359.884 <b>8,53%</b>
	Tenis	<b>1.476.828</b>	1.440.028 2,49%	1.431.244 <b>3,09%</b>	1.449.092 <b>1,88%</b>
J2SE	ProGuard 1.7.2	<b>16.415.832</b>	16.231.112 1,13%	15.754.576 <b>4,03%</b>	16.242.776 <b>1,05%</b>

Table 6-4: Reduction on total memory allocated.

Table 6-5 presents the execution time for each modified application, measured as an average of three consecutive executions. Below each execution time, we inform the percentage of reduction of the time, compared to the values obtained with the non-optimized application. Therefore, positive percentage values mean improvements on the application performance, and negative percentage values mean the resulting application is slower than the non-optimized one.

Execution time average (ms)		No optimization	ProGuard v1.7.2	Extended ProGuard (default)	Extended ProGuard (aggressive)
J2ME	SpaceInvaders	<b>14.411</b>	14.501 -0,63%	13.059 <b>9,38%</b>	12.912 <b>10,40%</b>
	Pacman	<b>13.703</b>	13.710 -0,05%	13.413 <b>2,12%</b>	13.322 <b>2,78%</b>
	Atlantis	<b>11.403</b>	11.323 0,70%	10.422 <b>8,61%</b>	10.388 <b>8,90%</b>
	Tenis	<b>17.478</b>	17.432 0,27%	16.614 <b>4,95%</b>	16.564 <b>5,23%</b>
J2SE	ProGuard 1.7.2	<b>10.018</b>	10.254 -2,36%	10.208 <b>-1,90%</b>	10.478 <b>-4,60%</b>

Table 6-5: Application performance improvement.

In our experiments, the default parameterization, shown in the third data column, always improved the application performance compared with the original ProGuard 1.7.2., shown in the second data column. That is because the application code generated by default parameterization was always smaller than original ProGuard's, as shown in Section 6.2. That avoids degrading application performance for memory cache reasons while inlining.

The negative value of the default parameterization (when the ProGuard itself is optimized) is because, in that case, the method inlining improvement was not able to compensate the performance degradation caused by previous optimizations of the original ProGuard 1.7.2. In fact, these optimizations did not target application performance.

In most of the applications, the execution time for the aggressive optimization is slightly better than for the default one. The exception to this rule was already expected: ProGuard 1.7.2 itself, where the aggressive inlining had increased the application size, degrading the application performance for memory cache reasons.

This result means that the aggressive inlining seems to be worth only if it reduces the application code size. Otherwise, if the code increases, the performance impact is probably worse or equal to the non-inlined version, due to the performance degradation by memory cache reasons.

Anyway, as the default parameterization always reduces the application size, we believe that we can apply default method inlining without degrading the application performance. In the end, the default parameterization seems to assure both a reasonable performance improvement while reducing some application code size.

## 6.5 *CODE SIZE REDUCTION BY OBFUSCATORS*

The study presented in Chapter 3 shown to us that method inlining optimization is rarely implemented, having been found only in Jax and DashO. In order to verify the overall improvement of our optimization on code size reduction, we have submitted to these obfuscators the same applications used to evaluate our solution.

For the experiments performed here, we used the version 7.3 of Jax and an evaluation copy of the DashO Embedded Edition (the same versions used in study of Section 3.2). They were configured in order to minimize the size of the applications, including all optimizations provided by each of them. The graph of Figure 6-1 presents a comparison of the code size reduction of these applications after optimized.

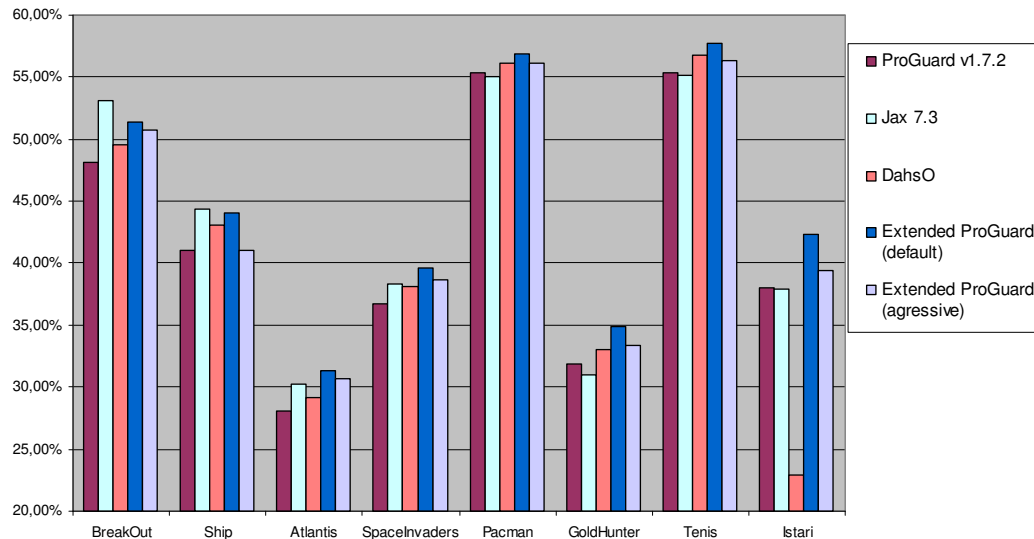


Figure 6-1: Code size reduction by obfuscators.

Despite of the additional size reduction of our optimization being apparently small that was enough to make ProGuard to stand out among the tools. The only situations in which we lost to Jax refers to applications that use frameworks (wGEM [Pessoa, 2001]) where Jax managed to apply the adjacent superclass merge optimization, still not implemented by the other tools, including ProGuard.

In fact, the elimination of unused members is surely the most effective optimization for the application size reduction, represented by the good performance of the original version of ProGuard. However the usage of other optimizations as method inlining performs an additional improvement that can be very important in restricted environments like J2ME.



## 7 BEST PROGRAMMING PRACTICES

This study on the optimizations available by obfuscators combined with our experience while extending one of them gave us valuable know-how about optimizations details. Moreover, for more than three years, our research team has been providing consulting services on J2ME for industrial scale applications [CESAR/Meantime]. This experience clearly has shown us that *it is essential for programmers to know the capabilities of the adopted tool in order to avoid unneeded design sacrifices and to improve optimization results*. A lot of effort is wasted avoiding situations already resolved automatically, and many other practices can confuse the optimization algorithms. Therefore, we have developed this best practices guide to advise the developers of possible difficulties and facilities when using obfuscators. The issues are organized in four groups:

- *Situations already resolved by the obfuscator*: practices we believe programmers implement only for optimizing the application but that are already resolved by some optimization.
- *Situations not resolved by the obfuscator*: clarifies mistaken practices e.g. practices we believe programmers can do expecting the obfuscators to fix them but that are not really resolved by any available optimization. These situations include limitations and possible future implementations of the tool, as well as special cases not handled properly.
- *Situations that jeopardize the obfuscator*: highlights undesirable practices that confuse or make some optimization inapplicable or make the obfuscator configuration difficult. These situations must be avoided when possible but can be used if needed.
- *Other recommendations*: includes all other general advices and recommendations not included in the previous cases, possibly not related to the obfuscators.

This kind of information depends on the optimizations implemented by each obfuscator. Ideally, obfuscator developers should include this information in the user documentation, since they know the details of the optimizations they implemented. Unfortunately, this is not usual. The next sections present some examples of best

programming practices considering the optimizations implemented by the obfuscator we have extended.

An expert programming staff can extend this document and produce its own best programming practices document considering the optimizations available by the adopted obfuscator and the programmers' skills. This document should be maintained and updated every time a new optimization becomes available (when the obfuscator is changed or a new version is installed) or when a new advice to the programmers seems important.

## 7.1 SITUATIONS ALREADY RESOLVED BY OBFUSCATORS

This section presents some practices we believe the programmers can try to implement only for optimizing the application but that are already resolved by some optimization. Some practices include a table presenting a source code example with *unneeded* changes; a *recommended* version of that; and the result of this version when *optimized*.

### 7.1.1 No identifier needs to be shorted

Short identifiers easily reduce code legibility. Names of classes, fields and method are already replaced with short (often one letter) ones by *class and member names compression*; names of variables are removed by *classfile recreation*; and package names are emptied by *class package relocation*.

Example:

Unneeded	<b>packabe</b> game; <b>class</b> GameObj { <b>int</b> vx; }
Recommended	<b>package</b> com.company.game; <b>class</b> GameObject { <b>int</b> xspeed; }
Optimized	<b>class</b> A { <b>int</b> a; }

### 7.1.2 Unused features in frameworks do not need to be suppressed

Frameworks often implement a lot of methods and fields to be used when needed. Many of these features are not used by one application. However it is not needed to suppress these declarations from the framework for each application, since they are automatically removed by *removal of unused elements*, *removal of unused method body*, *removal write-only fields* and *method inlining* optimizations. Some examples are fields used only by methods never called; or static methods available in utility classes.

Example:

Unneeded	<pre> <b>class</b> A { <b>void</b> m() { /* ... */ } }           // unused body <b>class</b> B <b>extends</b> A { <b>void</b> m() { ... } } <b>class</b> Util {     <b>static</b> A a; /* <b>static</b> B b; */              // write-only     /*<b>static void</b> initA() { a = <b>new</b> A(); } */ // unreachable     <b>static void</b> initB() { a = /* b = */ <b>new</b> B(); } } ... <b>public void</b> startApp() {     Util.initB();           // only class B is instantiated     Util.a.m();             // always B.m() is executed } </pre>
Recommended	<pre> <b>class</b> A { <b>void</b> m() { ... } } <b>class</b> B <b>extends</b> A { <b>void</b> m() { ... } } <b>class</b> Util {     <b>static</b> A a; <b>static</b> B b;     <b>static void</b> initA() { a = <b>new</b> A(); }     <b>static void</b> initB() { a = b = <b>new</b> B(); } } ... <b>public void</b> startApp() {     Util.initB();     Util.a.m(); } </pre>
Optimized	<pre> <b>abstract class</b> A { <b>void abstract</b> m(); } // made abstract <b>class</b> B <b>extends</b> A { <b>void</b> m() { ... } } <b>class</b> Util {     <b>static</b> A a; } ... <b>public void</b> startApp() {     Util.a = <b>new</b> B();           // initB inlined     Util.a.m(); } </pre>

### 7.1.3 Primitive type constant values do not need to be substituted by hand

Constant declarations make the code easier to understand and to maintain, however constants are implemented as common class fields on bytecode level. For primitive type values, when the constant is used, the compiler usually generates bytecode using directly the constant value, but the field is kept because some dynamic loaded class can access it later. Thus, a whole program analysis does not find any reference to those fields and *removal of unused elements* optimization removes them from the resulting application.

Example:

Unneeded	Image.createImage(10, 15);
Recommended	<pre> <b>public static final int</b> IMAGE_WIDTH  = 10; <b>public static final int</b> IMAGE_HEIGHT = 15; Image.createImage(IMAGE_WIDTH, IMAGE_HEIGHT); </pre>
Optimized	Image.createImage(10, 15);



### 7.1.4 Fields do not need to be made public to avoid field access methods (get and set)

Declaring public fields is not recommended because it does not protect the user of the class from changes in class implementation. Unfortunately, many programmers do this in order to avoid field access method declarations (get and set). Besides, method call instruction is often slower than direct field access instructions. However, you can create the proper get and set methods because *method inlining* often removes them to you and makes fields `public` if needed. Trivial methods (e.g. that only read or write a field) are always inlined. Non-trivial method inlining depends on the number of times the method is called and the size of the method.

Example:

Unneeded	<code>public int xspeed;</code>
Recommended	<code>private int xspeed;</code> <code>public int getXSpeed() { return xspeed; }</code>
Optimized	<code>public int xspeed;</code>

### 7.1.5 Long methods can be divided into small context methods called once

Long methods can make the code harder to understand. However, sometimes the programmer does not divide the method in smaller context methods to avoid the new declaration. *Method inlining* optimization often removes methods called once and replaces its unique call with the method code.

Example:

Unneeded	<code>public void update () {</code> <code>// Update map</code> <code>...</code> <code>// Update objects</code> <code>...</code> <code>};</code>
Recommended	<code>public void update () {</code> <code>updateMap();</code> <code>updateObjects();</code> <code>}</code> <code>/** Update map */</code> <code>private void updateMap() { ... }</code> <code>/** Update objects */</code> <code>private void updateObjects() { ... }</code>
Optimized	<code>public void update () { ... ... }</code>

## 7.2 SITUATIONS NOT RESOLVED BY OBFUSCATORS

This section clarifies mistaken practices e.g. practices we believe the programmers can do trusting the obfuscators but that are not really resolved by any available optimization. Some practices include a table presenting a source code example with the *unresolved* situation; and the *expected* result that can be available by some future optimization, not implemented yet<sup>1</sup>.

### 7.2.1 Constant propagation is not still available

*Constant propagation* is an intra-procedural optimization where constants assigned to a variable can be propagated and substituted at the use of the variable [Nullstone, 2002]. Compilers often implement *constant propagation* but some obfuscator optimizations, like *method inlining*, can open new opportunities to this optimization. Besides, some compilers perform constant propagation only in some cases, for example when the variable is explicitly declared as `final`.

Example:

Unresolved	<code>int x = 10; int y = x * 5;</code>
Expected	<code>int x = 10; int y = 50;</code>

### 7.2.2 Dead code elimination is not still available

*Dead code elimination* is an intra-procedural optimization where code that does not affect the program (e.g. dead stores) can be eliminated [Nullstone, 2002]. Compilers often implement *dead code elimination* but some obfuscator optimizations, like *method inlining*, can open new opportunities to this optimization.

Example:

Unresolved	<code>int i = 1;          // never used global = 1;        // dead code global = 2; return;</code>
Expected	<code>global = 2; return;</code>

### 7.2.3 Control flow analysis is not still available

*Control flow analysis* considers branch instructions, such as `if`, `while` or `switch`, to product an intra-procedural representation, the flow graph. With this graph,

---

<sup>1</sup> In fact, we know by personal communication that some of these optimizations are already included in the official ProGuard's list of future features.

it is possible to detect and remove unreachable branches. Note that, without this analysis, *removal of unused elements* optimization can fail and keep a method that actually will never be executed. Compilers often perform this analysis and some of them are able to remove unreachable branches under some conditions, like evaluation of constant boolean values.

Example:

Unresolved	<pre> <b>boolean</b>    debug_mode = <b>false</b>; <b>if</b> (debug_mode) {           // constant propagation     updateTimeCounter(); // kept but never executed } ... </pre>
Expected	<pre> <b>boolean</b>    debug_mode = <b>false</b>; ... </pre>

#### 7.2.4 Devirtualization is not still available

*Devirtualization* optimization replaces slower virtual call instructions with faster static linked call instructions. In order to do that, methods are automatically made `static`, `private` and `final` when possible. As this optimization is not still available in the analyzed obfuscator, the programmer must assure that himself.

#### 7.2.5 Merging of adjacent superclass is not still available

*Merging of adjacent superclass* optimization removes intermediate classes in the class hierarchy, moving all methods and fields of a class to its superclass. This optimization is not still available in the analyzed obfuscator. However, when it is implemented, it often does not manage to be applied due to the restrictions to keep the instantiated object size. So, the programmer must be always careful about the number of classes.

#### 7.2.6 Call graph considers objects instantiated anywhere, not only locally

*Call graph* is a data structure that indicates which methods are reachable through each call site. The major difficulty in the call graph construction is the identification of possible executions from a virtual call. The analyzed obfuscator implements an effective enough algorithm [Bacon, 1997] that verifies if the method belongs to an instantiated classes from the class hierarchy. However, once the class is instantiated, its methods are considered reachable anywhere.

Example:

Unresolved	<pre> class A { void m() { ... } } class B extends A { void m() { ... } } ... static A a = new B(); // instantiating class B public void startApp() {     a = new A(); // instantiating class A     a.m(); // A.m() body is always executed } </pre>
Expected	<pre> class A { void m() { ... } } class B extends A { } // method B.m() could be removed ... static A a = new B(); public void startApp() {     a = new A();     a.m(); } </pre>

### 7.3 SITUATIONS THAT JEOPARDIZE OBFUSCATORS

This section highlights undesirable practices that confuse or make some optimization inapplicable or make the obfuscator configuration hard. These situations must be avoided when possible but can be used if needed. Some practices include a table presenting a source code with *undesirable* practice; and a recommended approach as alternative solution.

#### 7.3.1 Reflection API usage

*Reflection* is the capability to refer some class, field or method by string statements, without knowing the exact element being accessed. Note that the name of the referenced element can be replaced by *class and member names compression*, so that the element is not found in execution time. If this feature is really needed, the user must inform the obfuscator to not change the element name.

Example:

Undesirable	<code>Class.forName("MyClass").newInstance();</code>
Recommended	<code>new MyClass();</code>

#### 7.3.2 Relative resource addressing

It is possible to refer a resource relative to the class location in the package tree. Note that the package tree can be reorganized by *class package relocation*, so that the resource is not found in execution time. If this feature is really needed, it needed to relocate the resource too or to inform the obfuscator to not change the package name.

Example:

Undesirable	<code>Class.getResourceAsStream("image.png")</code>
Recommended	<code>Class.getResourceAsStream("\\...\\image.png")</code>

### 7.3.3 Unnecessary code

Unnecessary code, especially method calls, field reading and class instantiation, must be strongly avoided. For example, *method inlining* considers the number of times the method is called to decide if it will be optimized; or if a field is read only once, it cannot be removed by *removal of write-only fields* optimization; and yet, instantiated classes automatically considers all its declared or inherited methods can be reached by virtual calls.

Example:

Undesirable	<pre> <b>class</b> A {     Object x; Object y;     A() {         x = <b>new</b> Object(); y = x;    // x cannot be write-only     } } </pre>
Recommended	<pre> <b>class</b> A {     Object x; Object y;     A() {         x = y = <b>new</b> Object();      // both can be write-only     } } </pre>

### 7.3.4 Throwing exceptions

Throw exceptions only when needed. It is slow and requires additional classfiles attributes. Moreover, methods that throw exceptions cannot be optimized by *method inlining* because it could change the program behavior. Do not use exception as control flow or to frequent user messages.

### 7.3.5 Synchronization

Synchronized methods are about 10 times slower than normal methods [Hardwick, 2003]. Moreover, they also cannot be optimized by *method inlining*, since they implement an implicit lock.

### 7.3.6 Switches

Switches are very large instructions and they require special treatment to be copied by *method inlining*, since the instruction length depends on its position in the code array. Our implementation rejected *method inlining* of calls to methods that belongs switches.

## 7.4 OTHER RECOMMENDATIONS

This section includes all other general advices and recommendations not included in the previous cases, possibly not related to the obfuscators.

### 7.4.1 Do not initialize big arrays in line

When initializing arrays in line, such as in the example, each array position initialization is compiled to an assign instruction. If some big array needs to be initialized, consider loading it from some binary resource.

Example:

Declaration	<code>int arr[] = { 0, 0, 1, 0, ... };</code>
Generated bytecode	<code>arr[0] = 0; arr[1] = 0; arr[2] = 1; ...</code>

### 7.4.2 Types byte, short, char and boolean are usually converted to int

The Java Virtual Machine Specification [Lindholm & Yellin, 1999] almost always operates `byte`, `short`, `char` and `boolean` data as `int`, inclusive when loading and storing variables, storing constants, operating math instructions or evaluating branch conditions. There are special instructions only for arrays. The effect over class fields is not imposed and it depends on the virtual machine implementation. So, the programmer usually does not need to force the use of these types only for optimization.

### 7.4.3 Avoid nested and anonymous classes

Nested and anonymous classes are inner classes, declared inside the scope of other classes. However compilers create an entire classfile for each inner class, including all internal structures. Even more, compilers still have to create some special methods and fields to allow an inner class to access its enclosing class' private information [Lindholm & Yellin, 1999]. Often, these classes can be replaced with some normal implementation.

### 7.4.4 Avoid reinvent API (Application Program Interface) already available

Try to use the API already available rather than reimplement your own one. The new classes and methods rarely manage to be faster than those already implemented by environment and they usually make the application larger.

#### *7.4.5 Reuse objects*

It takes a long time only to create an empty object (about 13 times longer than assigning a field, for example) [Hardwick, 2003], so it is often worth updating the fields of an old object and reusing it rather than creating a new one. Moreover, it also reduces the garbage collector task, since fewer objects are removed.

## 8 RELATED WORKS

Method inlining is a well-known optimization and it has been studied and implemented for decades since non-object-oriented programming languages, like Fortran and C [Allen & Johnson, 1988]. This section discusses the main method inlining researches and implementations that are related with our work somehow. Section 8.1 presents the method inlining specific researches and Section 8.2 discusses some method inlining implementations on compilers and tools. Section 8.3 discusses some related works on best programming practices.

### *8.1 LANGUAGE-INDEPENDENT METHOD INLINING RESEARCHES*

Most of the specific research on method inlining found in the literature present the problem in an abstract way, independent from language, platform or application domain. They often explore the decision algorithm in order to maximize application performance, while trying to control the code expansion somehow. Since they are generic approaches, they do not use to take full advantage of whole program environments and, as far as we know, none of them include the removal method benefit as a parameter of the decision algorithm.

Manuel Serrano [Serrano, 1997] published an article proposing a method inlining optimization that controls the code size expansion, using a 'factor' initialized experimentally, which is reduced by 1, for each nested inlining, stopping when the value becomes zero. Thus, Serrano's work provides an unconventional and interesting approach for dealing with recursive call sites. Unfortunately, the intrinsic local characteristic of its decision algorithm, while analyzing each call site isolated, makes difficult to reward by the method removal.

Jeffrey Dean and Craig Chambers [Dean & Chambers, 1994] proposed a general decision algorithm where, for each call site, the inlining is performed, its cost and benefit are calculated and, in case it is not interesting, the process is reverted. The exact function of the cost and benefit of the inlining is left out, just citing the increase of the size and the performance gain as important factors. For optimization of the algorithm, the work proposes the creation of a database with previously performed analysis (named



inlining trials) to be considered in future decisions about similar calls. Its experiments stressed a compilation time reduction due to inlining trials. Since the cost and benefit are estimated for each call site, it is difficult to take in account the global benefit from the removal of the method only if all their call sites were optimized.

Vortex [Dean et al, 1996] is a language-independent optimizing compiler that performs object-oriented-focused optimizations on a low-level intermediate language. It was developed in order to unify the effectiveness evaluation of these optimizations over the application performance in several languages. Cross-module inlining is one of the optimizations mentioned as being implemented, however the author does not detail the decision algorithm being used or the parameters taken into account. All benchmarks are substantial in size and there are almost no results about code size increasing.

## ***8.2 METHOD INLINING IN COMPILERS AND TOOLS***

Cross-module method inlining is effectively implemented only in very aggressive optimizing compilers and tools, which often have as main goal the application performance improvement. Most of these works are related to C/C++ compilers, however Java imposes some additional important language specific issues, like virtual methods by default that makes static analysis harder. Here, we discuss the most related method-inlining implementations in both languages.

Rainer Leupers published an article [Leupers & Marwedel, 1999] exposing the development of a function inlining approach for C compilers for embedded processors, imposing a global limit over final generated code size. His decision algorithm tries to find the method set that, when inlined, satisfies the limit and obtains the best execution performance. To do this, it requires several input parameters, including information about the real execution flow (profiling). Leupers' work, like ours, is very careful about the impact of method inlining on code size, however C is not an object-oriented language, which imposes many other difficulties, like polymorphism. Besides it does not consider the possibility of whole program optimizations, like including the removal method benefit as a parameter of the decision algorithm.

C++ and most ANSI C compilers allow the programmer to mark functions as a suggestion to be inlined (usually with an explicit "inline" function definition keyword) [Cline, 2003]. Unfortunately, the decision algorithm is completely unclear and the compilers can inline some, all, or none of the calls to a marked function, depending on many factors. Besides, the function can be removed only in some very restricted

situations, like declared as static and linked under special directives etc. Therefore, users have few guidelines when to mark a function to be inlined and when it will be really inlined or removed, so that it seems there is no results about optimization effectiveness over application size reduction, if any.

Sun's Java HotSpot technology [HotSpot 1.4.1], evolution of the Just In Time compilers (JIT), in principle can make inlining of frequently called methods in execution time, inclusive replacing the calls with native code. However, one of the main features of this kind of compiler is the capability to undo the optimization if it is not available or worthwhile anymore, for example due to a new class loaded dynamically or to save execution memory. Therefore, the original bytecode copy of the optimized methods can never be removed, increasing the application size in execution time. Besides, only recently Sun has published a CLDC HotSpot implementation [CLDC HotSpot] tuned to J2ME platform; however it does not implement method inlining at all.

David Whitlock's work [Whitlock, 2000] extends an academic tool, named BLOAT, implementing some inter-procedural optimizations on Java bytecode, among them method inlining. Whitlock's work presents some more details about implementation techniques and difficulties; however it has as main goal the improvement of the execution performance of the applications, not presenting enough concerns or results about the increase in the code size. In fact Nystrom [Nystrom, 1998] originally proposes the BLOAT tool, only with intra-procedural optimizations, like those found in compilers [Nullstone, 2002].

Obfuscators are also tools where method inlining can be found. We consider them the closest related work because they are also addressed to whole program optimizations. As shown in the study presented in Chapter 3, only a few obfuscators implement these optimizations (we could find it in Jax and DashO).

### ***8.3 RELATED WORKS OF BEST PROGRAMMING PRACTICES***

There are really a myriad of articles, web sites and books that address best Java programming practices [O'Hanley, 2004] [Hardwick, 2003] [Klemm, 1999], however we was not able to find any publication that considers the usage of obfuscators or the impact of automated optimizations over those practices. O'Hanley [O'Hanley, 2004] presents a long list of good programming practices for some Java technologies and constructions, such as *Servlets*, *JSPs* and *Swing*, exceptions, constructors, serialization

and so on. Hardwick's work [Hardwick, 2003] is an on-line collection of general recommendations for optimizing Java programs so that they are faster, smaller and more maintainable. Klemm [Klemm, 1999] identifies and explains the main Java performance problems sources and it presents a list of source-level guidelines for accelerating Java applications, trying to reduce the object copy and allocation tasks.

Some other documents only recommend the use of obfuscators as a good practice for J2ME [Giguere, 2002] [Larson, 2002] [J2MEwtk 1.0.4], but they do not teach how to program to take more advantage of them. Giguere [Giguere, 2002] addresses some interesting guidelines to optimize J2ME application size and it recommends the use of obfuscators to shorten the names of packages, classes, methods and data members. Larson [Larson, 2002] strongly recommends obfuscator as a great way to reduce the application size. Sun's J2ME Wireless Toolkit [J2MEwtk 1.0.4] has already suggests the use of obfuscators since version 1.0.4.

We also could not find any survey about the most common optimizations implemented by obfuscators; instead, we only found some articles that compare compiler optimizations [Nullstone, 2002] [Hardwick, 2003]. Of course, the documentation of each obfuscator indicates what optimizations it implements, but rarely presents programming facilities and difficulties.

## 9 CONCLUSIONS

In this chapter, we present the most important contributions of this research and some future works.

### *9.1 CONTRIBUTIONS*

The strong demand for programs based on platforms with high memory and processing constraints, like cell phones, is pushing the implementation and use of optimization tools, such as obfuscators and shrinkers. So far, these tools have neglected the use of method inlining due to its classical problem of increasing code size. This study presents an original implementation of cross-module and whole-program technique for method inlining that improves both performance and application code size. The experimental results show that our technique is able to optimize and exclude around 50% of the reachable methods and calls, reducing the code size more than 3%, in average, and improving the performance up to 10%. These percentages vary according to the application architecture and the algorithm's parameterization. Anyway, our technique is able both to perform a reasonable performance improvement and to reduce application code size in most cases.

The key idea behind this surprising result is to take full advantage of low-level features of the Java Virtual Machine. In fact, previous efforts failed in pointing how to overcome the problem of code size increase when using method inlining, because the solutions are too general, disregarding the languages' implementation and target environment specificities. Our results indicate that, in order to guarantee the success of some optimization methods, such as inlining, the use of these specificities is unavoidable.

Unfortunately, developers often don't know (or don't trust) these tools enough and keep sacrificing the code quality in order to optimize their applications. This study shows that obfuscators are safe but its effectiveness depends on the adopted programming practices. We noticed that to take best advantage of the obfuscators and to avoid unneeded sacrifices, it is essential that programmers (i) choose a tool that satisfies

the project requirements, considering its available optimizations and (ii) know how these optimizations can affect the programming and design decisions.

In order to help the programmer in choosing the obfuscator, this work presented an original study identifying which optimizations are most common in current obfuscators and where their implantations differs. It also identifies trends of new optimizations being implemented and gives some guidelines about what else could be taken into account to choose the tool.

Besides, in order help developers to program using obfuscators, we introduced a set of best programming practices, organized in situations not resolved by the tools; situations well resolved by the tools and situations that jeopardize the tools usefulness. Despite their obvious importance, these practices have not been enumerated or discussed in the scientific or technical literature so far.

Our Extended version of ProGuard and the presented programming practices has been tested and used by an industrial software development team in C.E.S.A.R/Meantime [CESAR/Meantime]. We submitted our extensions to ProGuard maintainers to be incorporated in the official open-source version.

This work is strongly motivated by the J2ME platform, because of its clear need for space and efficiency optimization and for allowing save whole program optimization. However, it is important to stress that none of the low level features we have explored in our inlining technique are J2ME-specific. The proposed method inlining can be directly used in any Java platform. The only restriction is the fact that the proposed technique performs whole program optimization, generating a code that will not be reused by other applications.

## ***9.2 FUTURE WORK***

We intend to study and evaluate other intra-procedural and inter-procedural optimization techniques not explored by the majority of current tools.

Method inlining opens opportunities to other intra-procedural optimizations, such as the identification and removal of unused variables, and the pre-processing of operations on constants, among others [Nullstone, 2002]. These optimizations are frequently implemented in compilers in a high level way. However, as our method inlining technique processes directly on the bytecode already compiled, we are not able to reuse those optimizations. They must be implemented again on bytecode level, and performed after method inlining. We believe that the benefit of these intra-procedural

optimizations can improve the method inlining results a lot, especially if included in the code size estimation of the decision algorithm.

Some other inter-procedural techniques, such as *devirtualization* and *adjacent merge of superclasses*, could also be examined. To take full advantage of these techniques, we intend to keep the approach of fully exploring the implementation specificities, in order to assess whether the cost-benefit ratio is worthwhile, as in the case of method inlining.

Additionally, we intend to study in depth the problem of the insertion and removal of constant pool entries. It seems to be a rich problem, since changes in the constant pool is a determinant factor for a successful code size reduction and still a risk to the aggressive approach of method inlining.

Finally, we also intend to format our best programming practices guide as a study on traditional design patterns [Gamma et al, 1995] in order to indicate how far each design pattern can benefit (or jeopardize) the most common optimizations found in obfuscators.



## REFERENCES

- [Allen & Johnson, 1988] Allen, R. Johnson, S. (1988) *Compiling C for Vectorization, Parallelization, and Inline Expansion*. In Proceedings of the SIGPLAN'88 Conference on Programming Language Design and Implementation. Atlanta, Georgia.
- [Ant Project] *The Apache Ant Project*. The Apache Software Foundation. Last change September, 2004. <http://ant.apache.org/>
- [Bacon, 1997] Bacon, D. F. (1997) *Fast and Effective Optimization of Statically Typed Object-Oriented Programs*. PhD thesis, Computer Science Division, University of California, Berkeley. December, 1997. Report No. UCB/CSD-98-1017.
- [Cameron & Day, 1998] Cameron, C. Day, B. (1998) *Knuckletop Computing: The Java Ring*. In Sun Microsystems' web site. <http://java.sun.com/features/1998/03/rings.html>
- [CDC 1.0] Sun Microsystems. *CDC - Connected Device Configuration, v1.0a*. JCP Specification, JSR 036. <http://jcp.org/aboutJava/communityprocess/final/jsr036/>
- [CESAR/Meantime] C.E.S.A.R/Meantime. *Centro de Estudos Avançados do Recife. Meantime Mobile Games*. Recife, PE. <http://www.meantime.com.br>
- [CLDC 1.0] Sun Microsystems. *CLDC - Connected, Limited Device Configuration*. JCP Specification, JSR 030. <http://jcp.org/aboutJava/communityprocess/final/jsr030/>
- [CLDC 1.1] Sun Microsystems. *CLDC - Connected, Limited Device Configuration, v1.1*. JCP Specification, JSR 139. <http://jcp.org/aboutJava/communityprocess/final/jsr139/>
- [CLDC HotSpot] Sun Microsystems. (2003) *The CLDC HotSpot Implementation Virtual Machine – White Paper*. May, 2003. [http://java.sun.com/products/cldc/wp/CLDC\\_HotSpot\\_WhitePaper.pdf](http://java.sun.com/products/cldc/wp/CLDC_HotSpot_WhitePaper.pdf)



- [Cline, 2003] Cline M. (2003) *C++ FAQ Lite: Inline functions*. Last change March, 2003. <http://burks.brighton.ac.uk/burks/language/cpp/cppfaq/inline-functions.html>
- [Dahm, 2002] Dahm, M. (2002) *BCEL Byte Code Engineering Library 4.4.1*. The Apache Jakarta Project. Last change December, 2002. <http://jakarta.apache.org/bcel>
- [DashO] PreEmptive Solutions. *DashO Embedded Edition documentation*. <http://www.preemptive.com/>
- [Dean et al, 1995] Dean, J. Grove, D. Chambers, C. (1995) *Optimization of Object-Oriented Programs Using Static Class Hierarchy Analysis*. In Proceedings of the 9th European Conference on Object-Oriented Programming (ECOOP'95). Aarhus, Denmark.
- [Dean & Chambers, 1994] Dean, J. Chambers, C. (1994) *Towards Better Inlining Decisions Using Inlining Trials*. In Proceedings of the ACM Conference on Lisp and Functional Programming Languages (LFP'94). Orlando, Florida.
- [Dean, 1996] Dean, J. (1996) *Whole-Program Optimization of Object-Oriented Languages*. Technical Report TR-96-06-02, Department of Computer Science and Engineering, University of Washington.
- [Dean et al, 1996] Dean, J., DeFouw, G., Grove, D., Litvinov, V., Chambers, C. (1996) *Vortex: An Optimizing Compiler for Object-Oriented Languages*. In Proceedings of the 11th ACM SIGPLAN conference on Object-oriented programming, systems, languages, and applications (OOPSLA'96). San Jose, California.
- [EmbeddedJava] Sun Microsystems. *EmbeddedJava Application Environment*. <http://java.sun.com/products/embeddedjava/>
- [FP 1.0] Sun Microsystems. *Foundation Profile Specification, v1.0a*. JCP Specification, JSR 046. <http://jcp.org/aboutJava/communityprocess/final/jsr046/>
- [Gamma et al, 1995] Gamma, E., Helm, R., Johnson, R., Vlissides, J. (1995) *Design Patterns, Elements of Reusable Object-Oriented Software*. Addison-Wesley, Reading, MA.
- [Giguere, 2002] Giguere, E. (2002) *Optimizing J2ME Application Size*. In the Sun Microsystems' web. Posted in February, 2002. <http://developers.sun.com/techtopics/mobility/midp/ttips/appsize/>
- [Green] Sun Microsystems. *A Brief History of the Green*

- Project*. <http://today.java.net/jag/old/green/>
- [Hansmann, 2003] Hansmann, U. Merk, L. Nicklous, M., Stober, T. (2003) *Pervasive Computing: Second Edition*. Springer Verlag.
- [Hardwick, 2003] Hardwick, J. (2003) *Java Optimization*. Last change April, 2003 – <http://www-2.cs.cmu.edu/~jch/java/optimization.html>
- [HotSpot 1.4.1] Sun Microsystems. (2002) *The Java HotSpot Virtual Machine, v1.4.1 – White Paper*. September, 2003. [http://java.sun.com/products/hotspot/docs/whitepaper/Java\\_Hotspot\\_v1.4.1/JHS\\_141\\_WP\\_d2a.pdf](http://java.sun.com/products/hotspot/docs/whitepaper/Java_Hotspot_v1.4.1/JHS_141_WP_d2a.pdf)
- [J2EE] Sun Microsystems. *Java 2 Platform, Enterprise Edition (J2EE)*. <http://java.sun.com/j2ee/>
- [J2ME] Sun Microsystems. *Java 2 Platform, Micro Edition (J2ME)*. <http://java.sun.com/j2me/>
- [J2MEwtk 1.0.4] Sun Microsystems. *J2ME Wireless Toolkit v1.0.4 documentation*. <http://java.sun.com/j2me/>
- [J2SE] Sun Microsystems. *Java 2 Platform, Standard Edition (J2SE)*. <http://java.sun.com/j2se/>
- [JavaCard] Sun Microsystems. *Java Card Technology*. <http://java.sun.com/products/javacard/>
- [Jax] IBM Research. *Jax Project*. Posted in June, 1998. <http://www.alphaworks.ibm.com/tech/JAX>
- [JikesBT] IBM AlphaWorks. *JikesBT. Jikes Bytecode Toolkit*. Posted in March, 2000. <http://www.alphaworks.ibm.com/tech/jikesbt>
- [Jshrink] Eastridge Technology. *JShrink*. Copyright 1997-2004. <http://www.e-t.com/jshrink.html>
- [Klemm, 1999] Klemm, R. (1999) *Practical Guidelines for Boosting Java Server Performance*. In Proceedings of the ACM 1999 on Java Grande Conference. San Francisco, California.
- [Knudsen, 2002a] Knudsen, J. (2002) *Obfuscating MIDlet Suites with ProGuard*, In the Sun Microsystems' web site. Posted in August, 2002. <http://developers.sun.com/techtopics/mobility/midp/tips/proguard/>
- [Knudsen, 2002b] Knudsen, J. (2002) *Understanding MIDlet Memory*. In the Sun Microsystems' web site. Posted in June, 2002. <http://developers.sun.com/techtopics/mobility/midp/tips/memory/>

- [KVMds] Sun Microsystems. *The K Virtual Machine - Data Sheet*. <http://java.sun.com/products/cldc/ds/>
- [KVMwp] Sun Microsystems. *Java 2 Platform Micro Edition (J2ME) Technology for Creating Mobile Devices – White Paper*. <http://java.sun.com/products/kvm/wp/KVMwp.pdf>
- [Lafortune] Lafortune, E. *ProGuard documentation*. Source Forge. <http://proguard.sourceforge.net/>
- [Larson, 2002] Larson, E. D. (2002) *J2ME Optimization Tips and Tools*. In Sun Microsystems' web site. Posted in November, 2002. <http://developers.sun.com/techtopics/mobility/midp/tips/optimize/>
- [Leupers & Marwedel, 1999] Leupers, R., Marwedel, P. (1999) *Function Inlining under Code Size Constraints for Embedded Processors*. In Proceedings of the International Conference on Computer-Aided Design (ICCAD). San Jose, California.
- [Lindholm & Yellin, 1999] Lindholm, T., Yellin, F. (1999) *The Java Virtual Machine Specification – Second Edition*. Sun Microsystems. <http://java.sun.com/docs/books/vmspec/>
- [MIDP 1.0] Sun Microsystems. *MIDP - Mobile Information Device Profile*. JCP Specification, JSR 037. <http://jcp.org/aboutJava/communityprocess/final/jsr037/>
- [MIDP 2.0] Sun Microsystems. *MIDP - Mobile Information Device Profile, v2.0*. JCP Specification, JSR 118. <http://jcp.org/aboutJava/communityprocess/final/jsr118/>
- [Nullstone, 2002] Nullstone Corporation. (2002). *NULLSTONE Optimization Categories*. <http://www.nullstone.com/htmls/category.htm>
- [Nystrom, 1998] Nystrom, N. J. (1998) *Bytecode-level analysis and optimization of Java classes*. Master dissertation, Department of Computer Science, Purdue University. West Lafayette, Indiana. August, 1998.
- [O'Hanley, 2004] O'Hanley, J. (2004) *Java Practices - Home*, Canada. <http://www.javapractices.com/>
- [Ortiz, 2002] Ortiz, C. E. (2002) *A Survey of J2ME Today*. In Sun Microsystems' web site. Posted in November, 2002. <http://developers.sun.com/techtopics/mobility/getstart/articles/survey/>

- [PBP 1.0] Sun Microsystems. *Personal Basis Profile Specification, v1.0*. JCP Specification, JSR 129. <http://jcp.org/aboutJava/communityprocess/final/jsr129/>
- [PDAP] Sun Microsystems. *PDA Optional Packages for the J2ME Platform*. JCP Specification, JSR 075. <http://jcp.org/aboutJava/communityprocess/final/jsr075/>
- [PersonalJava] Sun Microsystems. *PersonalJava*. <http://java.sun.com/products/personaljava/>
- [Pessoa, 2001] Pessoa, C. *wGEM: um Framework de Desenvolvimento de Jogos para Dispositivos Móveis*. Master dissertation. Centro de Informática. Universidade Federal de Pernambuco, Recife, Pernambuco. November, 2002.
- [PP 1.0] Sun Microsystems. *Personal Profile Specification*. JCP Specification, JSR 62. <http://jcp.org/aboutJava/communityprocess/final/jsr062/>
- [RetroGuard] Retrologic. *RetroGuard 1.1.9 User's Guide*. Copyright 1998-2004. <http://www.retrologic.com/>
- [Serrano, 1997] Serrano, M. (1997) *Inline expansion: when and how?*. In Proceedings of the 9th International Symposium on Programming Languages, Implementations, Logics, and Programs (PLILP'97). Southampton, New York.
- [Tip et al, 1999] Tip, F. Laffra, C. Sweeney, P. F. (1999) *Practical Experience with an Application Extractor for Java*. In Proceedings of the 14th Annual ACM SIGPLAN Conference on Object-Oriented Programming Systems, Languages, and Applications (OOPSLA'99), Denver, Colorado.
- [Tip & Palsberg, 2000] Tip, F. Palsberg, J. (2000) *Scalable propagation-based call graph construction algorithms*. In Proceedings of the 14th Annual ACM SIGPLAN Conference on Object-Oriented Programming Systems, Languages, and Applications (OOPSLA'00). Minneapolis, Minnesota.
- [Whitlock, 2000] Whitlock, D. M. (2000) *Persistence-Enabled Optimization of Java Programs*. Master dissertation, Department of Computer Science, Purdue University. West Lafayette, Indiana. May, 2000.