UNIVERSITATEA

ALEXANDRU IOAN CUZA IAŞI

FACULTATEA DE

INFORMATICĂ

LUCRARE DE LICENŢĂ

# Code Refractor

Compilation and optimization of

Stack Virtual Machines

Proposed by

**Khlud Ciprian**

Sesiunea: iunie, 2014

Scientific coordinator

Professor, PhD **Ferucio Laurentiu Țiplea**

This page is intentionally left blank

# Claim about originality and respecting intelectual property and copyright

With present paper I submit that the Barchelor's paper, “Code Refractor - Compilation and optimization of

Stack Virtual Machines” is written entirely by me and it was never presented to any other faculty or other institution for higher level of education in Romania or any other country. I also claim, that all claims, even the ones taken from internet, are clearly described in this work, with respecting rules of avoiding the plagiarism:

- all the fragments of text are exactly reproduced, even if they are translated from another language, they are written in quotes and they keep an exact reference to the source

- paraphrasing in my own worlds of written texts to another authors keep an exact reference to the source;

- source code, source images, etc.taken from open-source projects or another sources are used in accord with the copyright laws and it keeps direct references

- the summary of ideas of other authors keeps exact reference to original text

Iassy, <today>

Graduating student Khlud Ciprian

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(original signature)

# Declaration of acceptance

I claim that I agree that the Bachelor’s paper with title: “Code Refractor: Compilation and optimization of stack-based virtual machines“, the source code of the programs and the other third party content (graphs, multimedia, input data, etc.) that come with this paper to be used by the Faculty of Informatics.

Also, I agree that Faculty of Informatics from University “Alexandru Ioan Cuza” Iassy to use, change, copy and distribute under noncommercial ways the programs, executables and sources, written by me as part of this Bachelor’s paper.

Iassy, <today>

Graduating student Khlud Ciprian

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(original signature)

# Agreement about rights of ownership

Faculty of Informatics agrees that the copyright of computer programs, executable or source code to belong to the author of the present paper, Khlud Ciprian.

This agreement is important for the following reasons:

- the Code Refractor compiler front-end, initial middle-end design and the backend (code generator) is under GPL version 2+, a license which is owned by the Free Software Foundation. As author of CodeRefractor, I (Khlud Ciprian) allow the version on the time of submitting of the Bachelor’s paper to be submitted under a license that fits the University's needs, but the Code Refractor as is remains Free Software under Free Software Definition of GPL2 or any newer versions;

- Code Refractor's runtime library is under MIT/X11 license, a liberal license which allows public modification. Because Code Refractor's is public domain (even I'm the author of it), the public version at the time of publishing the Bachelor’s paper can be used for University's needs, but Code Refractor's runtime library cannot be bind under any proprietary license that does not allow public domain changing, including but not limited to, commercial needs

Iași, <today>

Dean,

Khlud Ciprian

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Table of Contents

# 0. Introduction

## 0.1. Abstract

Virtual machines implement various techniques like interpretation, Just-in-time (JIT) compilation or Ahead-Of-Time (AOT) compilation. This project proposes a method of AOT compilation using the host CIL virtual machine (also known as .Net ™ virtual machine) and using a standard C++ compiler (C++ 11) to give the same semantics and in the meantime to also improve performance offering a full compilation step, and in the very same time to remove the dependency to the host virtual machine.

A similar technique could be implemented in a JVM (Java Virtual Machine), but CIL is in more ways advanced, because the CIL opcodes do not specify types, and the CIL Virtual Machine (VM) will have to infer based on context.

## 0.2. Introduction

An advanced (optimizing) compiler is split into following steps:

- scanning (sometimes named tokenization);

- parsing (which generates an AST tree);

- semantic analysis;

- transformation of AST tree to intermediate representation (IR) which defines all operations of the source program into a small steps, that can be easily understood what they do;

- optimization steps which consist into various visitors of the IR and they rewrite the IR into an equivalent, but a bit more efficient more form of the defined operations;

- IR is visited and written into a “low level representation” like Assembly language, or binary form. A critical part of this part, is to find a good way to use the minimum resources (mostly CPU registers).

This project will show a way to write an optimizing compiler against virtual machines, and focusing for correctness, performance and simplicity of understanding of the code. Also, in this introduction, the reader is informed how a virtual machine works, and how to map most operations into a low level implementation.

A virtual machine like Java (JVM = Java Virtual Machine) or .Net executes “on demand” an intermediate form (look earlier for IR), named “bytecode”. This bytecode describes the original semantic of the original Java written language (in Java world) or C# (Vb.Net, Boo, F#, etc. in .Net technologies).

If any user wants to get a better performing application, would likely want to get an efficient compiler to evaluate its program. Historically there are many approaches in literature:

- Java’s HotSpot: is a JIT compiler which will compile “hot” parts of the program. This is great for some programs which benefit from this dynamic evaluation, but sometimes the startup time is a direct consequence of this approach, as some parts of the code are interpreted and profiled before knowing which parts are hot

- another approach is done by Excelsior Jet[[1]](#footnote-2), which (even is not fully documented how it works, being a proprietary product), but it looks that it compiles every bytecodes into an intermediate representation, and based on your feature choice based on the bought edition, you will have enabled more or less optimizations. It also looks that their static compiler is written in Java (the controls look like written in Swing, and at least once in the evaluation version, I've got Java exception in compilation)

- RoboVM[[2]](#footnote-3) and GCJ[[3]](#footnote-4) both read the Java bytecodes and rewrite it directly into an optimizing compiler backend (RoboVM into LLVM bytecodes, GCJ into GCC's GIMPLE bytecodes)

- Mono[[4]](#footnote-5) has two Ahead-of-time compilers, one is their made LinearIR[[5]](#footnote-6) (their “optimizing framework”) which is fast, and another one slow, which writes into LLVM bytecodes

For practical reasons, if a person wants to implement a full virtual machine specification, will have to implement tens to hundreds of bytecode instructions that are defined by the virtual machine, to make possible to execute average or a bit bigger than “Hello world” but also it shouldn't implement all (but as not all instructions are implemented it will make it a non-conformant virtual machine).

This thesis describes the implementation of a nonconforming Ahead-of-time compiler for CIL instruction set virtual machine of instruction set, and will try to use practical approaches to make it easy to be studied, extended and to make possible to give both a good performance of the final code and a fairly decent compatibility, regardless of virtual machine versions.

# 1. AOT description and evaluation

There are many parts in implementing a virtual machine, but before starting this, is better to describe what a virtual machine does in between the bytecode and the final executing code.

## 1.1. Stack based Virtual Machines Overview

### 1.1.1. Execution model

A Java and .Net VMs would do the following:

- would start reading your entry method;

- right after that will read the bytecode as operations;

- depending of the design, will start executing that method up-to when will hit another method,

- if this method was never compiled before, this method will be read too and executed recursively

- if the method is compiled before, the method is executed normally, and the virtual machine will put guards around areas that can give exceptions or execute another method (named trampolines)

### 1.1.2. Memory model

Virtual machines do implement a garbage collector, which means, through other things that memory is cleared (collected) periodically if is not reachable from other pointers.

### 1.1.3. Compilation model

All stack based virtual machines, if they compile the method, they will compile not the bytecode directly, but they will make more optimizations on the intermediate reprezentation (IR). Also this form is later converted into binary object code, and this code is executed. These optimizations are limited in scope because the compilation happens interactively with running application. This mostly means, that the Register Allocation step is not optimal (for example in HotSpt Client version is using Linear Scan Allocator)

### 1.1.4. “Stack Based” versus “Register based” instruction set

This means that operations between various entities are defined using an “evaluator stack” which means that instructions to be executed, they have to to read the last states (or to write them) before being able to define them.

In a stack based code, even for simplest operations like:

int a = 5;

int b = 3;

int c = a + b;

The code at the bytecode level is:

1. push int 5

2. pop value from stack into “a”

3. push int 3

4. pop value from stack into “b”

5. push 'a'

6. push 'b'

7. add top two values on stack

8. pop value from stack into “c”

A register based virtual machine has a already allocated number or local variables, named “registers” and the assignments of the operations, will not use stack as evaluator state, but the instructions will use the register index as source of data.

In a “register based” VM the code would be like following:

1. set reg\_0 5

2. set into 'a' reg\_0

3. set reg\_1 3

4. set 'b from' reg\_1

5. set reg\_2 from 'a'

6. set reg\_3 from 'b'

7. add reg\_4 reg\_2 reg\_3

8. set 'c' from reg\_4

Even the number of instructions is the same in this simple problem, a precompiler can optimize the register usage of the register based, and rewrite as following:

1. set reg\_0 5

2. set reg\_1 3

3. add reg\_2 reg\_0 reg\_1

4. set 'c' from reg\_2

which will remove the usages of “a” and “b” as not being necessary, and even instruction 3 can be removed (and evaluate reg\_2 as value 8).

This is an important part as performance and optimization steps are concerned, if the stack VM can be rewritten as a Register based virtual machine, the optimizations can be applied more easily, because the instruction set contains more information to track the data.

### 1.1.5. Specified instruction set

Both Java and .Net (CIL) have publicly defined instruction set. They can be understood from public made documentation. This is great for implementers and makes a case of making easier to test virtual machine compatibility[[6]](#footnote-7). This also can simplifies the testing of virtual machine’s semantics as writing a C#/Java program that gives the asked bytecode means at the end, if a bug arise, it is easier to check if the semantics of the bytecode reflect correctly the program (so it can be a language compiler bug).

Also, it is good to know that both JVM and CIL try to fit (in most cases) in one byte, so many operations are compact. So writing a virtual machine is easier starting from given high level source programs that use any specific instructions, as it is easier to write human readable code than binary bytecode.

## 1.2. Evaluation and implementation directions

### 1.2.1. Evaluating the execution model

Both Java and Net create a method graph of Callers and Callees, and all the logic can be extracted with reflection (a Java or .Net API to scan the runtime information) or with bytecode manipulation libraries.

So the Code Refractor as first compile step, will get the assembly file which needs to be scanned, and will read the bytecodes especially the calls ones. This “linker” will scan bytecodes, and every time will find a new call instruction, will track and if finds a new function, will scan it too, up-to the moment will create the closure of all functions.

A note-worthy implementation detail is that the “new object” instruction includes an (implicit) constructor call (if this constructor is with no parameters), and this code is also scanned.

### 1.2.2. Evaluating the memory model

An advanced (and fast final code) compilation and memory model is a very hard problem, garbage collectors are as of today improved. Similarly, just a good register allocation, as the Register Allocation (RA) is an NP-hard problem, is not solvable and can be redefined as a coloring problem (which is again NP-hard).

The garbage collector can be implemented with C++'s “smart-pointer” class. This class defines a number to count all references to an instance of an object. When the references are zero, the object instance is automatically deleted.

### 1.2.3 Evaluating the compilation model

It is easier and practical to reuse solutions already existent. The very often used solution for AOT compilers for virtual machines is LLVM or another compiler intermediate representation (like in case of GCJ, using GIMPLE form), but they also require to implement a full runtime, which again is hard to be made practical.

As a solution to approximate the original virtual source, a C++ compiler can reflect fairly well the original bytecode (look to a later chapter where is defined how the C++ language reflects some of the semantics of the virtual machine, without introducing a memory or performance overhead. C++ has also another good side effect, C++ is supported in all GCC and major LLVM platforms, or even in platforms like Windows RT, so it is possible (given enough changes and platform support) to get access to all platforms which offer a C++ compiler, which includes embedded platforms, cars, etc.

### 1.2.4 Evaluating the Stack Based VM model

All operations in the .Net machine as they work against a stack which stores the states, it has to be implemented with main stack operations:

Push

Pop

Top / Peek

Introducing for a simple addition a full stack class, would make performance horrendous, but C++ (and C) pushes all (simple) variables to the CPU stack (which is very fast, as is implemented in the hardware of the CPU), so the solution is that every time we have a push of a value, we define a variable on stack to keep this variable, which is pushed on the process' stack, and when a Pop operation is called, this variable is read and will reset the counter to keep the correctness of the implementation.

The another advantage with this implementation that it will change the virtual machine to be a Register-based VM, which at the end gives a lot of optimization opportunities.

### 1.2.5. Optimization overview

First of all, I will want to make a small comment about the word “optimization” which in many ways is considered to give an “optimum” of a starting program. In fact optimizations will never give the optimum program because:

- optimum can be defined by more criteria, as most compilers (including Code Refractor one) optimize for runtime performance, will have some tradeoffs. One of them, is that it occupies more CPU stack (a memory tradeoff)

- the algorithm in the first place, that is defined in the user's code may be optimal, and the compiler has to guarantee that will create the code with the same result

- other factors, like but not limited to: some optimization passes in literature are not implemented as they are too complex to be made

Based on this, I would define an

An **optimization step** is an algorithm which starts with the intermediate representation (IR) and will rewrite it to another IR (which is equivalent with the original IR's side effects and results) in case some conditions are met for that specific algorithm, which is in a form more advantageous as resources or strides to enable another optimization steps.

Definition: the **compiler optimizations** is the application of all optimization steps until no optimization step is possible to an IR.

Optimizations can be thought in some categories:

- simplifications (like math simplifications)

- propagation (like transitivity properties)

- analysis steps (which would try to prove that some properties keep hold). If so, they do make possible that another optimization step to use these properties.

- data flow optimizations (this will be a large section of the optimizations)

- etc.

# 2. Code Refractor – implementation of an AOT virtual machines

## 2.1. Small overview of CodeRefractor

CodeRefractor does the following:

- reads the entry point of an assembly

- a pre-scan of the method call tree is created

- a second scan will take all methods and will transform their CIL bytecodes into intermediate representation

- the compiler optimizations are performed for every method

- the intermediate representation is written into C++

## 2.2. Intermediary representation of Code Refractor

As the logic of most code is method based, CR does store the intermediate representation per every method as two parts:

- variable parts which are themselves split into:

\* local variables

\* virtual registers

\* arguments (method's parameters)

- local operations which will use either the local variables or will call another methods

The local operations, are themselves split into:

- assignments

- operators:

\* binary operators

\* unary operators

- conditional operators

- creation of arrays and objects

- accessing array elements or fields

- call of a separate methods

- branch operations

- labels

These operations are in fact a more compact form of the original CIL operations, and as a difference, the IR operation will keep unify the bytecodes that do a common semantic as a single operation kind, instead of having an individual bytecode for every operation.

This means that if you make an **optimization step** that works against operators, it is easier to specify first a group of operations to target, and after that, the operations are inspected to see the specific operator you want to target.

Describing a simple optimization step with a code sample:

internal class MergeConsecutiveLabels : ResultingOptimizationPass

{

public override void OptimizeOperations(MetaMidRepresentation intermediateCode)

{

var operations = intermediateCode.LocalOperations;

var found = operations.Any(operation => operation.Kind == LocalOperation.Kinds.Label);

if (!found)

return;

for (var i = 0; i < operations.Count - 2; i++)

{

var operation = operations[i];

if (operation.Kind != LocalOperation.Kinds.Label)

continue;

var operation2 = operations[i + 1];

if (operation2.Kind != LocalOperation.Kinds.Label)

continue;

var jumpId = (int) operation.Value;

var jumpId2 = (int) operation2.Value;

OptimizeConsecutiveLabels(operations, jumpId, jumpId2);

operations.RemoveAt(i + 1);

Result = true;

}

}

private static void OptimizeConsecutiveLabels(List<LocalOperation> operations, int jumpId, int jumpId2)

{

for (var i = 0; i < operations.Count - 2; i++)

{

var operation = operations[i];

if (operation.IsBranchOperation())

continue;

switch (operation.Kind)

{

case LocalOperation.Kinds.AlwaysBranch:

var jumpTo = (int) operation.Value;

if (jumpId2 == jumpTo)

operation.Value = jumpId;

break;

case LocalOperation.Kinds.BranchOperator:

var destAssignment = (BranchOperator) operation.Value;

if (destAssignment.JumpTo == jumpId2)

destAssignment.JumpTo = jumpId;

break;

}

}

}

}

This *optimization step* will find if there are two consecutive instructions defined as label (target of jump instructions), and if they are found, the labels are merged into one, and all jumps that are targeting the deleted label, they will be redirected to the adjacent label.

Before starting optimizations, it is really great to know which optimization steps do make sense, as not all optimizations do have the same effect, but in short some rules can be easily discovered:

- when a program executes less instructions, is more likely it will run faster

- when the code that is not accessible in any execution path and is removed, this is also powerful as it possibly enable another optimizations

- when you replace a variable with a constant value (when possible) is faster as execution time

- removing calls, and replacing it with the equivalent code will save at least the performance of the final code

- removing pointer assignments, this is crucial as an assignment requires more smart pointer updating which is expensive

Some performance improvements (yet they are not optimization steps, as they don't work against the IR) can be done without any semantic analysis of the final code:

- platform native libraries which are imported, can be merged as data structures as a compilation step. This will reduce the impact of reloading every pointer to function and reloading/searching for a DLL if is loaded already

- duplicate array data initialization (that sometimes is initialized only with zeroes) can be merged, and the same can be done in all strings over the application

- string type can contain in a linear data set the length and the string data. This improves the memory accessing pattern, as there is no required L1/L2 cache misses if the string length is separated from the data

# 3. Optimization and optimization steps in the CR

## 3.1 Optimization short overview

Optimization passes are logically split in between:

- local optimizations (the ones that go just over a basic block)

- global optimizations (optimization passes that go over more than a basic block)

- program-wide optimizations (optimizations that do work over functions)

Even it doesn’t always appear to be in this way, optimizations that are the most effective are local optimizations, right after the global optimizations and at last the program-wide optimizations. The reason is that the .Net states defined by its specified stack based VM require to define many stack operations that after these states are transformed into a Register-Based VM (as it was described later), the impact of the higher order optimizations appear to be less important.

Let’s take a typical code which computes a sequence of prime numbers:

<code of prime numbers>

## 3.2.Local optimizations (block based optimizations)

References

1. CIL infrastructure, ECMA standard - <http://www.ecma-international.org/publications/standards/Ecma-335.htm>
2. Alex Aiken’s page: <http://theory.stanford.edu/~aiken/>
3. Alex Aiken’s Compiler Course on Coursera (2012): <https://class.coursera.org/compilers-2012-002>
4. Paper: **Virtual Machine Showdown: Stack Versus Registers** (Yunhe Shi, David Gregg, Andrew Beatty): *We found that a register architecture requires an average of 47% fewer executed VM instructions, and that the resulting register code is 25% larger than the correpsonding stack code. The increased cost of fetching more VM code due to larger code size involves only 1.07% extra real machine loads per VM instruction eliminated. On a Pentium 4 machine, the register machine required 32.3% less time to execute standard benchmarks if dispatch is performed using a C switch statement. Even if more efficient threaded dispatch is available (which requires labels as first class values), the reduction in running time is still around 26.5% for the register architecture.*
5. Paul Biggar – „Compiling and Optimizing Scripting Languages” - <http://youtube.com/watch?v=kKySEUrP7LA>

Slides: <http://www.slideshare.net/LittleBIGRuby/compiling-and-optimizing-scripting-languages>

1. Wikipedia’s listing of JVM bytecodes: <http://en.wikipedia.org/wiki/Java_bytecode_instruction_listings>
2. Eric Brumer – Native Code Performance and Memory: The Elephant in the CPU <http://channel9.msdn.com/Events/Build/2013/4-329>

1. <http://www.excelsior-usa.com/jet.html> [↑](#footnote-ref-2)
2. <http://www.robovm.org/> [↑](#footnote-ref-3)
3. <http://gcc.gnu.org/java/> [↑](#footnote-ref-4)
4. <http://www.mono-project.com/Main_Page> [↑](#footnote-ref-5)
5. <http://mono-project.com/Linear_IR> [↑](#footnote-ref-6)
6. Look to references 1 and 6 [↑](#footnote-ref-7)