

# Building a Compiler on Java for a mini-C Programming Language

J. Agustín Barrachina  
IEEE Student Member  
École Polytechnique  
Université Saclay-Paris

Philémon Poux  
École Polytechnique  
Université Saclay-Paris

## CONTENTS

|            |   |   |
|------------|---|---|
| <b>I</b>   | <b>Introduction</b>                           | 1 |
| I-A        | Mini C . . . . .                              | 1 |
|            | I-A1 What it does . . . . .                   | 1 |
|            | I-A2 What it doesn't . . . . .                | 2 |
|            | I-A3 Known bugs and work to be done . . . . . | 2 |
| I-B        | Structure of a Compiler . . . . .             | 2 |
| <b>II</b>  | <b>Lexical Analyzer</b>                       | 2 |
| II-A       | Regular Expressions . . . . .                 | 2 |
| II-B       | Finite Automata . . . . .                     | 3 |
| II-C       | Implementation . . . . .                      | 3 |
| <b>III</b> | <b>Syntax Analyzer</b>                        | 3 |
| III-A      | Implementation . . . . .                      | 3 |
| <b>IV</b>  | <b>Semantic Analyzer</b>                      | 4 |
| IV-A       | Implementation . . . . .                      | 4 |
| IV-B       | Results . . . . .                             | 4 |
| <b>V</b>   | <b>Code Generation</b>                        | 4 |
| V-A        | Register Transfer Language (RTL) . . .        | 4 |
|            | V-A1 Implementation . . . . .                 | 4 |
| V-B        | Explicit Transfer Language (ERTL) . .         | 5 |
|            | V-B1 Implementation . . . . .                 | 5 |
| V-C        | Location Transfer Language (LTL) . .          | 6 |
|            | V-C1 Life Duration Analysis . . .             | 6 |
|            | V-C2 Interference Graph . . . . .             | 6 |
|            | V-C3 Coloring . . . . .                       | 6 |
|            | V-C4 Implementation . . . . .                 | 6 |
| V-D        | x86-64 Assembly Code . . . . .                | 7 |
|            | V-D1 Implementation . . . . .                 | 7 |
| <b>VI</b>  | <b>Conclusion</b>                             | 7 |
|            | <b>References</b>                             | 8 |

**Abstract**—In this project, a compiler was created to generate a x86-64 assembler code from a C fragment called mini-C. This is a 100% C-compatible fragment, in the sense that any Mini C program is also a C program.

## I. INTRODUCTION

*"Optimizing compilers are so difficult to get right that we dare say that no optimizing compiler is error-free! Thus, the most important objective in writing a compiler is that it is correct" [1]*

The reader is suppose to have some basic knowledge of C and for that reason almost no explanation regarding that language will be treated in this report. For further information about C language refer to [2].

Simply stated, a Compiler is a program that can read a code written in a specific programming language and translate it into an equivalent code of another language. A fairly good analogy can be made by a translator between two different languages like Spanish and French for example.

The objective of this project is to create a compiler for a fragment of C denominated *Mini C* towards a reasonably effective code x86-64. In this project, a ".s" file will be created containing the translation between the C file to assembly. After the file is generated, the use of another compiler will be needed to produce the final output file that can be run by the computer.

General knowledge of x86-64 assembly and the functioning of processors will also be required. The concept of the stack and registers like the callee saved and the caller saved must be clear for the reader.

### A. Mini C

1) *What it does:* *Mini C* is a fragment of the language C which contains integers and pointers to structures. *Mini C* is 100% compatible with C in the sense that every *Mini C* program is also a C program. This will enable to use a C compiler such us **gcc** to use as reference.

*Mini C* can deal both with integers and structures. It can even support pointers inside structures to other structures or to the same structure in order to create lists for example. *Mini* is a Turing Complete language thus, despite its small structure, every single code in C can be reduced to *Mini C*.

2) *What it doesn't*: *Mini C* can deal both with integers and structures. But it doesn't work with any other types of variables such as floating point numbers or characters directly.

The function *sizeof* is not yet implemented. The compiler will read it correctly and will know the program is correctly written but it will not know how to translate *sizeof* resulting in an incomplete and broken assembly code.

*Mini C* doesn't implement a for loop. Even though is almost the same as implementing a while loop. It was not believed to be relevant the implementation of it as an academic objective as the implementation of it will only require time but will not help in the better understanding of a compiler.

The "++" and "--" commands to increase or decrease a variables value were not implemented as well.

3) *Known bugs and work to be done*: There was no real communication between the typer and the RTL (section V-A) generation code. For which reason, the toRLT function had to create in some cases some information already treated. This was not only working twice but also made the syntax class longer and more difficult to read. It is work for the future to create a better communication between both parts.

In the program, global structures were not yet implemented, although the code is prepared for it. In the section of code which it should be done, a throw error is prompt saying that the global structures are not yet implemented.

Although the compiler computes "5 + 4" directly as 9, it will fail to do so when more complex integer operations are done like "4\*9+(4&&0)". It is to be done more optimization on this side. Jumps are also always done by the program itself and the compiler will not compute a "if(1)" as a direct jump to the section of code inside the if loop.

The Appel & George algorithm can be applied for coloring the interference graph (section V-C3) in order to achieve better registers assignments for a better optimized assembly code.

## B. Structure of a Compiler

A compiler can be divided into two parts. The *analysis* (front end) and the *synthesis* (back end)

The *analysis* brakes the source program into constituent pieces and imposes a grammatical structure of them in order to create a intermediate representation of the source program. During this part, syntactical formation and semantical unsound is checked. The analysis also collects information about the source program and stores it in a data structure called a *symbol table* which will be used by the *synthesis* part.

The *synthesis* part makes use of the *symbol table* and the intermediate representation constructed by the analysis part and creates the target program.

A more detailed diagram of the structure can be seen in figure 1. Where the last part (Code Generation) correspond to the *synthesis* phase and the rest are all from the *analysis* phase. The diagram is longer that the one displayed, having also a converser from the assembler to the machine language and from there to the executable code. But in this project, those stages are not treated.

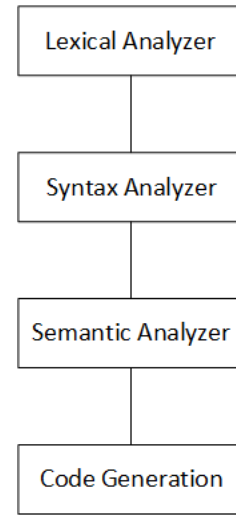


Fig. 1. Structure of a Compiler

## II. LEXICAL ANALYZER

A lexical analyzer (*Lexer*) is the first front-end step in compilers, matching keywords, comments, operators, etc, and generating a token stream for parsers called *lexemes* consisting of a *token name* and the *attribute value*. The *token name* is an abstract symbol that will be used during syntax analysis, while the *attribute value* is an entry in the *symbol table* (discussed on I-B) which will be used during the semantic analysis and the code generation.

The Lexer reads input from the programming language to compile (mini c in our case) and matches it against regular expressions and runs a corresponding action if such an expression is matched.

To make the Lexer are going to use:

- Regular Expressions: To describe the lexemes
- Finite Automata: To recognize the expressions

### A. Regular Expressions

The concept of regular expression arose in the 1950's when the American mathematician Stephen Cole Kleene formalized the description of a regular language.

A regular expression is a sequence of characters that define a search pattern. In other words, there are a conjunction of letters and digits that follow a certain rule.

Let us define *letter* as any letter in the Latin alphabet and *digit* any number [0-9]. Then we can define rules as follow:

$$0|[1-9]( < digit > * [ ] ) \quad (1)$$

Last equation 1 is a declaration of a decimal digit. The "||" is a logic or, it means, either the digit is 0 or it will be another thing. If it is not only 0, the number cannot start by 0 in C syntax, so it must start with a digit different from 0, which is range from 1 to 9 (encoded as [0-9]). Secondly, this digit can be followed by either nothing (represented by: [ ]) or by any digit for as many digits are they must be. The format <rule>\* means the repetition of a rule for as many times as necessary, or no repetition at all.

### B. Finite Automata

A *finite automata* is basically a binary graph which just say "yes" or "no" by means of a *recognizer* to each possible string.

There are two different classes of automatas:

- 1) *Nondeterministic Finite Automata* (NFA)
- 2) *Deterministic Finite Automata* (DFA)

The first class (NFA) have no restrictions on the labels of their edges. A symbol can label several edges out of the same state. The DFA on the other hand have for each state and symbol exactly one edge with that symbol leaving that state.

### C. Implementation

For the lexical analyzer, a flex library was used [3]. A .flex file was created and then, by means of jflex, converted to the final java class.

Jflex lexers are based on a DFA automata. For more information about jflex library please refer to [4].

The work of the lexer was only to read input from the file and create new symbols (*lexemes*) containing the information as explained in II.

## III. SYNTAX ANALYZER

The syntax of a programming language describes the proper form of its programs.

The *syntax analyzer* or *parser* uses the first component of the *lexemes*, the *tokens*. The *parser* creates some kind of tree representation that depicts the grammatical structure of the token stream called the *syntax tree*. In this tree, each node represent an operation and the children of the node represent the arguments of that operation.

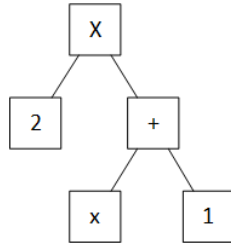


Fig. 2. Syntax Tree Example

On figure 2 an example of a tree representation can be seen. Many syntaxes can be the cause of that tree, for example, both 3 and 2 can be used to generate such tree.

$$2 * (x + 1) \quad (2)$$

$$(2 * ((x) + 1)) \quad (3)$$

### A. Implementation

The Syntax Analyzer was done using CUP, a parser generator for java [5].

In that file, precedence where applied to each operator using the following table:

| Operator      | Associativity |
|---------------|---------------|
| =             | right         |
|               | left          |
| &&            | left          |
| == !=         | left          |
| < <= > >=     | left          |
| + -           | left          |
| * /           | left          |
| ! -(negative) | right         |
| ->            | left          |

TABLE I  
PRECEDENCE

The higher the operator is in the table, the last will it be applied. For example, the operator "=" will be applied after all the other operators have been applied. There is a very simple way to apply this table in the .cup file by simply writing the following command: precedence {associativity} {token name} . Where associativity is either left or right and the token name is the token from the *lexem* created by the Lexer.

After the precedence, the following grammar was implemented (table II). The details on how to implement it where omitted. In order to see how to implement the following grammar in cup please refer to the user manual [6].

|           |   |
|-----------|---|
| < file >  | < decl >* EOF   |
| < decl >  | < var >   < type >   < funct >  |
| < var >   | int < ident >+ ;<br>struct < ident > ( * < ident >+ , ) ;   |
| < type >  | struct < ident > < var >* ;   |
| < funct > | int < ident > ( < param >* , ) < bloc ><br>struct < ident > * < ident > ( < param >* , ) < bloc >   |
| < param > | int < ident >   struct < ident > * < ident >  |
| < expr >  | < integer ><br>< ident ><br>< expr > -> < ident ><br>< ident > ( < expr >* , )<br>! < expr >   -< expr ><br>< expr > < op > < expr ><br>sizeof ( struct < ident > )<br>( < expr > ) |
| < op >    | =   ==   !=   +   -   *   /   &&   <   <=   >   >=  |
| < instr > | ;<br>< expr >;<br>if ( < expr > ) < instr ><br>if ( < expr > ) < instr > else < instr ><br>while ( < expr > ) < instr ><br>< bloc ><br>return < expr >;                             |
| < instr > | < var >* < instr >* *   |

TABLE II  
GRAMMAR

Where '|' is either one or the other is found. The '\*' is as many as necessary (including none at all), while on the other hand, '+' means as many as necessary but at least one. If under any of those symbols there is a comma (',') it means there are comma indented.

As an example on how to read the table, wwe can see that within a file, one will find a list of declarations. These declarations can be either a variable (global variables) or a structure declaration or functions. The declarations of functions have

it's own parameters (or none) and a bloc in which it will be the code. The bloc is composed by the declarations of the variables followed by instructions. The instructions are loops such as if or while or simply expressions. Expressions can be all type of C & C++ expressions such as integers, pointers, negation, call to functions, etc.

#### IV. SEMANTIC ANALYZER

*"Well typed programs do not go wrong"*

The semantics of a programming language defines what each program does when executing.

A *Semantic Analyzer* uses the *syntax tree* and the information in the *symbol table* to check the source program for semantic consistency with the language definition.

An important part of the *semantic analysis* is the *type checking* where it gathers type information and checks that each operator has matching operands. An example of the type checking will be to make sure the index which whom an array is accessed is an integer and not any other incompatible type. In an equation like  $8.0 + 4$ , the type checking will make sure to convert the integer "4" into a floating point before making the operation.

The *type checking* will make sure that the variables of an equation like  $e1 + e2$  are from the same type and reject the incoherent programs. There are some languages that use **dynamic types**, which means they check the type of the variables dynamically. Such languages are for example PHP, Python or Lisp. On the other hand, there are also **static types** languages which is the compiler the one in charge on checking the types. For example OCaml, Java and C (which will be our case).

##### A. Implementation

The bigger and more critic class of the entire program was created in *Syntax.java* class. An object of the class *File*, created by the parser (which uses the lexer) implements a method called *Typer* which takes care of what was explained in this section (IV).

##### B. Results

Our *Typer* types correctly all the good tests given beforehand. However there are still some functions that are not well implemented in the *typer* as far as lifespan of the variables is concerned. Indeed some tests provided such as "typing/bad/testfile-redef-4.c" are compiled but should fail. This is due to the lifespan of the local variables within the blocks. A solution could presumably be easily implemented but we had no success in trying and decided to move on forward in order to complete the project

#### V. CODE GENERATION

In this section we will actually generate the assembly code itself. It takes as input an intermediate representation of the source program and maps it into the target language. It is too difficult to be able to do this part in only one step. So it will actually be divided into 3 stages:

- 1) *Register Transfer Language* (RTL)
- 2) *Explicit Register Transfer Language* (ERTL)
- 3) *Location Transfer Language* (LTL)

Each stage will be explained in their corresponding sub section.

##### A. Register Transfer Language (RTL)

For this stage we will use the *syntax tree* created on III in order to create what will be called as *RTL tree*. We suppose that the local and global variables are already differenced and that the type of each variable recognized as it has been done in the last section.

The main objective (more precisely the first part) of the RTL is to create a set of instructions x86-64 from the operations of C.

The second phase is to create a *Register Transfer Language*. Here a *Control Flow Graph* (CFG) is created that will facilitate the ulterior phases and that will eliminate the distinction between statements and expressions. This RTL will create the so called pseudo registers, which are an infinite number of intermediate registers to realize operations. This registers will be converted into actual x86-64 registers in the future.

1) *Implementation*: Each file contains a list of declarations of functions (as can be seen on table II) that contains a bloc statement. Each function, contains a list of parameters and the list of declarations of variables. A RTL graph is created for each function in a recursive way. As seen in table II, a function has a list of statements that will be converted into one or more assembler commands. For each command, a label will be created and saved into the RTL graph, the RTL graph will contain each assembly operation with a label of reference to it.

Each *statement* and *expression* class (declared in the *syntax.java* file) will have a "toRTL" method that will save its label into the graph. The function "toRTL" will have as arguments:

- **Register** Register to be used.
- **Exit Label** The label to which the program will go after making the current instruction.
- **The RTL graph which will be added** In case it is necessary to add two instructions or more which is almost always the case.

Later on, there were added some more information to the function to be able to work with variables and structures.

- **Variables** Is a Map that links a string with its register. This is used to know in which pseudo-register each variable is stored in order access it. If the variable is not there, it is supposed to be global. The *Typer* section will make sure the global really exists.
- **Struct Definition** Is a Map between a string (the structure name) and a list of strings which are all the variables inside the structure. This is used for knowing what number to add as an offset when doing something like  $p - > a$

- **Struct Declaration** This is a Map between all structure pointers and the structure they are actually pointing. To actually compute  $p \rightarrow a$  first it will be necessary to use the structure declaration to know which structure is it pointing and then use the structure definition to know the offset. Because of the typer, this Maps will always contain what is is looking for, but an extra level of security was add to display an error if some variable is not in the Map.

The RTL function will return it's own label to be given to the next toRTL call in order to line up every statement and expression.

To make **condition branches**, another function will be created called "toRTLC" which will receive two label, one to be done if the expression is true, and another in order to be done in the other case. The structure will be represented as toRTLC(e, s1Label, s2Label) where 'e' will be the expression with a true or false value. s1Label will be the label to go if 'e' is true and s2Label in the other case.

In order to realize the && expression, for example with the case: "if e1 && e2 do s1 else s2" The following conversion will be done:

$$\text{toRTLC}(e1 \&\& e2, s1Label, s2Label) \rightarrow$$

$$\text{toRTLC}(e1, \text{toRTLC}(e2, s1Label, s2Label), s2Label)$$

Using a similar logic, the expression: "if e1 || e2 do s1 else s2" will be converted:

$$\text{toRTLC}(e1 || e2, s1Label, s2Label) \rightarrow$$

$$\text{toRTLC}(e1, s1Label, \text{toRTLC}(e2, s1Label, s2Label))$$

In order to make the **negative sign**, for example  $-2$ , the compiler actually does the operation  $0 - 2$ .

In order to make the **not operation**, for example !a. The compiler does an if statement such as: "if a then 0 else 1". In which case, making something like !!41 will return 1 as a result. Which is what actually happens in C code.

Some simplifications were done when doing binary operations. The compiler will compute the line:

$$x = 4 + 5$$

as

$$x = 9$$

## B. Explicit Transfer Language (ERTL)

The *Explicit Register Transfer Language* is in charge of the function call conventions. Here, the first parameters are sent to a function are stored in the registers which convention dictates will be used (%rdi, %rsi, %rdx, %rcx, %r8, %r9) and the rest of the parameters will be stocked at in the stack. Also, it will return the result of the function in the register %rax. It will also make sure the registers known as *callee saved* are correctly saved by the function before returning to the previous function.

During this stage, some pseudo-registers will be converted to real registers.

1) *Implementation:* Inside a class ERTLfile was created. The class implements a method called "createERTL" that receives an RTL class and creates an ERTLfile class from it. Each RTL has an implementation of a method called "toERTL" that returns an ERTL class. This method is used by "createERTL" to create the ERTL file in question.

Basically, the RTL graph is maintained untouched (copied into ERTL almost as it was) except for the call to a function. There is also some extra code added to the beginning and the end of the function bloc.

For calling the function, the following was added:

- 1) Move the arguments to the input registers and to the stack if necessary. The registers used as parameters of a function were declared as "Register.parameters". The code was done so that if this is to be changed later, adding or removing registers from the parameters list, nothing must be changed outside the Registers class.
- 2) Make the call to the function.
- 3) Save the result from register %rax (Again, a definition Register.result was made that applies to %rax). Should the return register be changed, only the class Register must be changed.
- 4) Recover the parameters on the stack with a manipulation of %rsp.

After entering a function. Some instructions where added.

- 1) Allocate the frame stack
  - 2) Save all the *callee saved* registers defined in "Register.callee\_saved"
  - 3) Recover all the parameters from "Registers.parameters"
- Before exiting the function, some other things are done:

- 1) Copy the result to the result register (%rax).
- 2) Recover the *callee saved* registers.
- 3) Deallocate the frame stack.
- 4) finish with the "ret" command.

Once applying the changes discussed, the output would be as follows. A RTL function will be printed as:

```
#10 main[]
entry   : L15
exit    : L11
locals  : [#8]
        L15: mov $42 #11 —> L14
        L14: #10 <- call fact[#11] —> L13
        L13: Mmov #10 #8 —> L12
        L12: Mmov #8 #9 —> L11
```

While the ERTL extension will now be extended to:

```
main(0)
entry   : L32
locals  : [#8]
        L32: alloc_frame —> L31
        L31: Mmov %r12 #18 —> L30
        L30: Mmov %rbx #17 —> L15
        L15: mov $42 #11 —> L14
        L14: Mmov #11 %rdi —> L29
```

```

L29: call fact(1) —> L28
L28: Mmov %rax #10 —> L13
L13: Mmov #10 #8 —> L12
L12: Mmov #8 #9 —> L27
L27: Mmov #9 %rax —> L11
L11: Mmov #18 %r12 —> L35
L35: Mmov #17 %rbx —> L34
L34: delete_frame —> L33
L33: return

```

Note that the return statement makes it unnecessary to have an extra field to save the exit label. There is also no need to tell the function who is been called, on which register is each parameter stored as by knowing only the number of parameters it will be enough for the function to know where they are. It can be seen also that they are many calls to registers that are not needed. All label L28, L13, L12 and L27 can be removed for example. Hopefully the following parts of the code will take care of reducing that to a more effective code.

### C. Location Transfer Language (LTL)

The Location Transfer Language is the final stage before the actual production of the assembly code which will assign all the other pseudo-register to actual available registers and add, when needed, the rest of the register to the stack.

In the Location Transfer part, all the register are re-written using allocatable ones and if there is no need to do the actual operation, like for example in a case like "mov r1 r1", it will be replaced with a "goto L", where L is the label of the next operation.

To make the LTL graph, previous stages are needed. Which will be explained by what follows.

1) *Life Duration Analysis*: The definition of an *alive variable* is as follows:

**Definition V.1.** Alive Variable A variable is said to be alive at a given breakpoint of the program if the value it contains is susceptible to be used in a future point of the execution code.

Some other definitions will be needed to make:

**Definition V.2.** def(I) Defined as the set of variables defined by the instruction 'I'

**Definition V.3.** use(I) Defined as the set of variables used by the instruction 'I'

Here there are some examples of "def" and "use" registers for some simple commands:

| Instruction    | def         | use         |
|----------------|-------------|-------------|
| mov n r        | {r}         | $\emptyset$ |
| mov r x        | $\emptyset$ | {r}         |
| load n(r1) r2  | {r2}        | {r2}        |
| store r1 n(r2) | $\emptyset$ | {r1, r2}    |

TABLE III  
DEF & USE

Having defined that, we are going to make two more definitions:

**Definition V.4.** in(I) Defined as the set of variables alive at the moment of the call of instruction 'I'

**Definition V.5.** in(I) Defined as the set of variables alive at the moment of exit from the instruction 'I'

We can generate two equations for these new definitions as follows:

$$\begin{cases} in(I) &= use(I) \cup (out(I) \setminus def(I)) \\ out(I) &= \cup_{s \in succ(I)} in(s) \end{cases}$$

The algorithm is applied by recursion and iterates until the smallest fixed point. We actually implemented the Kildall algorithm that is not as costly as this naive algorithm that, in the worst cases, is of complexity  $\mathcal{O}(N^4)$  where  $N$  is the number of vertices. The Kildall algorithm only recomputes values that have changed and do not iterate over the whole table each time. We used a stack to store the vertices that had changed and had to be recomputed and in order not to add them twice we implemented a HashMap *labelsInStack*.

2) *Interference Graph*: An interference graph is created using the information obtained by the *Life Duration Analysis*.

**Definition V.6.** Interference Two variables are said to interfere if they cannot use the same register or memory location.

The interference graph is it's a non oriented one that contains all the variables with two types of connections, *interference* or *preference*.

If we go back to the example given on V-B1 which contained the ERTL result of a simple function. In this case, all registers %rax, r10, r8 and r9 should connect with themselves in order to be able to remove all the "mov" commands.

3) *Coloring*: In this stage, the actual allocatable registers are mapped with the pseudo-registers so that all registers left are actually the ones being used in the assembly code. This section could be seen as actually coloring the interference graph, where each color is an actual allocatable register.

One of the best's algorithms to implement this stage is the one from George and Appel [7]. However this code was not applied in the code, but a simpler one was made.

4) *Implementation*: Each section described (Life Duration, Interference Graph and Coloring) was implemented with different classes each in the java files Arcs.java, Liveness.java and Coloring.java.

For the **Life Duration Analysis**, each ERTL class will implement both "use" and "def" functions which will give a set of *Registers* defined for each ERTL command. The Kildall algorithm was then applied.

To create the **interference graph** the following logic was applied. For every instruction that defines a variable  $v$  which the alive variables in "out" are  $w_1, \dots, w_n$ :

- For the instruction mov  $w \ v$ , we add interference between  $v$  and all  $w_i$  different from  $w$ . And we add the preference between  $v$  and  $w$ .
- For every other instruction we add the interference between  $v$  and all  $w_i$  with  $i \in [1, \dots, n]$

For **Coloring** the interference graph, first a "possible" registers list was made for each pseudo-register with all the possible allocatable registers that could be used for that pseudo-register. After the list was made, some iterations were made to assign all the pseudo-registers by the following functions (and in that order).

- 1) *onlyOnePossibility* In this function, every register with one and only one possibility was assigned to that register only if that same register appeared as a preferred one.
- 2) *onlyOnePossibility2* In this function, every register with one and only one possibility was assigned to that register regardless if that same register appeared as a preferred one.
- 3) *preferredOne* In this function, every register was assigned to a possible register if that register appeared as a preferred one.
- 4) *preferredPseudo* In this function, if the preferred register is a pseudo-register (not listed in the possible register list), it checks if that register (the preferred one) has already been assigned to an allocatable register and, if that allocatable register is not in the interference, it assigns it to himself.
- 5) *assignTheRest* In here the registers were assigned to whatever register possible left.
- 6) *spillRegisters* If there are register left without possibilities just stack them.

The reason why 1 and 2 are not made together was because if two registers say r1 and r2 have only one option and they interfere with each other, then one of those will be forced to be stored in the stack. If one of those register however, has the option as preferred while the other doesn't, then that register will not be chosen as the one stored on the stack.

Each ERTL subclass applies "toLTL" method. Equivalent as the RTL application of "toERTL", the LTL graph is created using the color graph created. In this section, every register is directly replaced with the register contained in the color graph. If the replacement of it makes the operation meaningless, then the operation is replaced by a "goto" command. The allocate and deallocate stack frame can now be replaced by the actual number or by a "goto" if that number is empty.

IN the following code we can see how the example shown in V-B1 is presented in the LTL graph.

```
main()
entry : L32
      L32: goto L31
      L31: goto L30
      L30: goto L15
      L15: mov $42 %rdi —> L14
      L14: goto L29
      L29: call fact —> L28
      L28: goto L13
      L13: goto L12
      L12: goto L27
      L27: goto L11
      L11: goto L35
```

```
L35: goto L34
L34: goto L33
L33: return
```

We can see that almost all the "mov" functions as well as the allocate and deallocate frame were replaced by the "goto" function. As stated before, all registers %rax, r10, r8 and r9 were mapped to the same colors so that all the "mov" operations on labels L28, L13, L12 and L27 were replaced by "goto".

#### D. x86-64 Assembly Code

1) *Implementation:* Finally, the final assembly code was generated from the LTL graph. This was done by something called *Linearization* which visits the LTL graph creating the code. Labels are created when visiting each for a second time. That means that a "goto L" will not generate the L label by it's own. The code on the label L will be put where the operation "goto L" was placed and L will be added to a visited list. If another operation "goto L" appears, then there it will be replaced by a "jmp L" only in that case. Two functions mutually recursive are created to make the *Linearization*.

- 1) **lin** Giving a certain label, lin created the code if the label was not yet visited. If not it creates a jmp to that label, marking that label as needed (so that it is actually added to the file).
- 2) **instr** Done by the visitor, this generates the code by a given label without checking any condition.

An example on how the recursion is done will be as follows.

```
instr (L1 : mov n d —> L) :=
      produce L1 : mov n d
      call lin (L)
```

For branching cases, special care will be taken to choose in which order to apply lin for both labels to make sure that, if possible, only one jump is needed. Both the conditional jump and the forced jump implementation will try to be avoided.

## VI. CONCLUSION

The implementation of a compiler proved to be very time consuming and very difficult to debug. Following the lectures and faithfully as possible proved to be the most efficient tool for debugging and making the correct implementation of the code.

It is a personal pride having been able to implement a working compiler although with bugs and incomplete in some aspects. The general knowledge of how to do it and the theory of how the compiler work were correctly assimilated.

## REFERENCES

- [1] A. V. A. M. S. L. R. S. J. D. Ullman, *Compilers. Principles, Techniques & Tools*, 2nd ed. Addison Wesley, 2006.
- [2] B. K. . D. Richie, *The C Programming Language*, 2nd ed. Prentice Hall, mar 1988.
- [3] J. Team. (2015, mar) Jflex - the fast scanner generator for java. [Online]. Available: <http://jflex.de/>
- [4] G. K. S. R. R. Dcamps, *JFlex User's Manual*, apr 2015.
- [5] S. Hudson. Cup parser generator for java. [Online]. Available: <http://www.cs.princeton.edu/appel/modern/java/CUP/>
- [6] S. E. Hudson, *CUP User's Manual*, jul 1999.
- [7] A. W. George, Lal & Appel, "Iterated register coalescing," *ACM Transactions on Programming Languages and Systems*, jul 1995.