FINE-GRAINED ASPECTS OF AUTOMATIC REFACTORIZING IN C2EIFFEL

MASTER THESIS

Adrian Friedli
ETH Zurich
adrianfriedli@student.ethz.ch

April 1, 2011 - September 30, 2011

Supervised by:
Marco Trudel
Prof. Bertrand Meyer
Abstract

The C2Eiffel framework [1] translates C source code to Eiffel source code. It aims to generate code that looks natural to a programmer such that it is readable and maintainable. While this is already the case for most parts of common C programs, there are areas for improvements.

External C functions are currently wrapped in the generated Eiffel code since the used libraries are not translated. In this thesis I will replace some standard and often used C functions by the corresponding Eiffel equivalents. In the attempt to replace standard C string functions I will also try to replace the current translation of C strings to integer arrays with native Eiffel strings.

I will also remove jump statements like break that are often used in C programs but are not available in Eiffel. This will increase the readability of the code compared to the current implementation that emulates the jump mechanism in Eiffel and produces hard to read code.
# Contents

1 Introduction ........................................... 7
  1.1 Reference Projects .................................. 8

2 Translation of printf statements to Eiffel's counterpart ....... 9
  2.1 The printf statement in C ............................. 9
  2.2 Translatable printf statements ......................... 10
  2.3 Translation ............................................ 10
  2.4 Translation in real examples ............................ 12

3 Translation of C strings to Eiffel Strings .................... 13
  3.1 C Strings .............................................. 13
  3.2 Eiffel Strings .......................................... 14
  3.3 Standard C string functions ............................. 14
  3.4 Translatable string variables ............................ 15
  3.5 Translation of standard C string functions to Eiffel features ... 19
  3.6 Implementation of the analysis .......................... 19
  3.7 Statistics .............................................. 22

4 Translation of C structs to Eiffel classes .................... 25
  4.1 Structs in C ............................................. 25
  4.2 Translatable structs .................................... 26
    Pointer to structs ....................................... 27
    Inheritance between structs .............................. 27
  4.3 Analysis Implementation ................................ 27
  4.4 Statistics .............................................. 28

5 Removing break, continue and return statements ............. 31
  5.1 Code transformation .................................... 32
  5.2 Optimization .......................................... 35
  5.3 Statistics .............................................. 37

6 Conclusions ............................................. 41
  6.1 Conclusions ............................................. 41
  6.2 Future Work ............................................ 41

A Additional translations of printf calls ............................... 43
Chapter 1

Introduction

The C2Eiffel framework [1] developed by Marco Trudel at ETH Zurich translates C source code to Eiffel source code. One of the goals of C2Eiffel is that the translated code looks natural to a programmer, that means the differences between the translated Code and manually written code should be as small as possible. The advantages of natural looking code is that is is more readable and easier to maintain. For most parts of common C programs C2Eiffel generates already such code.

The goal of this thesis is to improve the generated code even further. To measure the improvements we will test the generated code with the ten real world projects described in section 1.1. There are three main areas for improvements:

The first area is the translation of function calls to external C libraries. Since those libraries are not translated to Eiffel, the generated Eiffel code still calls those C functions directly. In this thesis I will replace some of those function calls with the corresponding Eiffel features. In chapter 2 we translate as many printf statements as possible to the Eiffel's counterpart Io.put_string. We will restrict us to the features already present in the Eiffel base library. Even with this restriction we can translate more than two thirds of the printf calls in the reference projects. In chapter 3 we translate some of the standard C string functions to the corresponding Eiffel features.

The second area is the translation of C strings and structs to Eiffel. At the moment C strings are translated to arrays of integers. In chapter 3 we try to replace them by Eiffel strings but we will see that this attempt will not be very successful. The current translation of structs to Eiffel classes uses embedded C code. The attempt to remove this C code in chapter 4 will unfortunately not be successful too.

The third area is the translation of jump statements like break or return that are not available in Eiffel. Currently those statements are handled in a generic way that is hard to read and maintain (see section 5). To improve this code I will remove break, continue and return statements by refactoring the C source code in chapter 5. This is done amongst others by adding if statements that allow to skip code when a jump statement was executed. Those code changes
have only a small impact on the readability of the code compared to the original C source code as the analysis with the reference projects shows. Compared to the currently generated code it considerably improves readability.

1.1 Reference Projects

We will evaluate our efforts on the following real world projects:

- **Hello World**
  A simple hello world program that just prints *hello world* to the command line with the `printf` function.

- **Micro_httpd [3]**
  It’s a small HTTP server with basic features and about 200 lines of code.

- **Xeyes [4]**
  A graphical program that shows two eyes which follow the mouse cursor.

- **Less [2]**
  A Unix program that shows the content of a file page by page in the terminal.

- **Wget [6]**
  A downloader that is part of the GNU Project.

- **Links [5]**
  A text based web browser.

- **Vim [10]**
  A feature rich text editor.

- **Curl [7]**
  A multiprotocol file transfer library.

- **Gmp [8]**
  The GNU multiple precision arithmetic library.

- **Gsl [9]**
  The GNU scientific library that provides a wide range of numerical routines.
Chapter 2

Translation of printf statements to Eiffel’s counterpart

In this chapter we look at C’s printf statements and which ones can be translated to Eiffel’s Io.put_string feature and how.

2.1 The printf statement in C

The method signature of printf is the following:

```c
int printf ( const char * format, ... );
```

The format string is the text to write to stdout and can contain format tags. Those tags get replaced by the additional arguments of the printf statement and formatted according to the formatting arguments inside the tags. A format tag has the following structure:

```
%[flags][width][.precision][length]specifier
```

where the parts in [] are optional. A format tag always starts with a % followed by a sequence of arguments:

specifier The specifier is the only required argument. It defines the type of the value that will replace this tag. The most common specifiers are ‘c’ for characters, ‘d’ or ‘i’ for integers, ‘f’ for floating-point values and ‘s’ for strings.

flags The flags control the justification (‘-’), signs before the value (‘+’ and space) and padding with zeros (‘0’).
Table 2.1: Supported printf specifiers and format tags

<table>
<thead>
<tr>
<th>specifier</th>
<th>format arguments</th>
</tr>
</thead>
<tbody>
<tr>
<td>c</td>
<td>precision (has no effect)</td>
</tr>
<tr>
<td>d, i</td>
<td>width (incl. *)</td>
</tr>
<tr>
<td>f</td>
<td>width, precision (incl. *)</td>
</tr>
<tr>
<td>s, u</td>
<td>-</td>
</tr>
</tbody>
</table>

**width** The width specifies the minimal number of characters to be printed. If the representation of the value is shorter it will be padded with blanks.

**precision** The precision specifies the minimal number of digits to be printed for integer specifiers and the exact number of decimal digits for floating-point specifiers.

**length** The length specifies how the argument value is interpreted (as short int, long int, ...).

The width and precision argument can be '*', that means that the value is given as an additional argument before the argument that will be formatted. If the printf statement is executed successfully the number of written characters, otherwise a negative number is returned.

### 2.2 Translatable printf statements

Since the return type of the Eiffel feature `Io.put_string` is Void, no printf statements can be translated where it is used as an expression, for example as the right hand side of an assignment. To make sure the translation works in all possible places, the replacement has to be a single statement too. If we restrict us to built in features of Eiffel the translatable printf statements only contain the specifiers and format arguments listed in table 2.1.

In addition to those printf statements we can support fprintf statements where the destination file stream is either stdout or stderr. In the case of stdout the fprintf and printf statements are equal and in the stderr case the output is redirected to the error stream.

### 2.3 Translation

In the following we look at five examples and how they are translated.

The first example (line 2 in listing 2.1) is a printf statement with a character specifier. In C2Eiffel characters are stored in an INTEGER8 and can be converted to a character with the feature `to_character_8`. The translated print statement is shown in listing 2.2 on line 2.
Listing 2.1: C: Five exemplary printf statements

```c
// c is of type char
printf("A character: %c\n", c);

// i is of type int
printf("An integer: %d\n", i);

// width is of type int
printf("An integer with variable width: %*i\n", width, i);

// d is of type double and precision is of type int
printf("A double with variable width and precision: %*.*f\n", width, precision, d);

// s is of type char
printf("A string: %s\n", s);
```

Listing 2.2: Eiffel: The five translated printf statements

```eiffel
-- c is of type INTEGER_8
Io.put_string("A character: " + (c).to_character_8.out + "%N")

-- i is of type INTEGER_32
Io.put_string("An integer: " + i.out + "%N")

-- width is of type INTEGER_32
Io.put_string("An integer with variable width: " + (create {FORMAT_INTEGER}.make (width)).formatted (i) + "%N")

-- d is of type REAL_64 and precision is of type INTEGER_32
Io.put_string("A double with variable width and precision: " + (create {FORMAT_DOUBLE}.make (width.max (precision)), precision)).formatted (d) + "%N")

-- s is of type CE_POINTER [INTEGER_8]
Io.put_string("A string: " + eif_string (s) + "%N")
```
The second example (line 5 in both listings) contains an integer specifier. The string representation of the integer is concatenated with the preceding string.

In the third example on line 8 we have the same with a variable width argument. In this case we use the FORMAT_INTEGER class which pads the integer with blanks if its representation is shorter than the specified width. The class offers more options to format the integer but for this we would need multiple instructions, which is not possible because the translation should be a single statement as discussed before.

On line 11 of the two listings we have a print statement that formats a floating-point value with variable width and precision. In the Eiffel code the FORMAT_DOUBLE class is used analogously to the integer formatting. For the width we have to use the maximum of width and precision since the make feature requires that the width is greater or equal to the precision.

In the last example (line 14) we have a print statement with a string specifier. In the Eiffel code the array of characters (CE_POINTER [INTEGER 8]) has to be converted to a STRING. This is done by the helper feature eil_string.

The translation of a fprintf statement with the error stream as first argument is the same as for the printf statement except that Io.put_string is replaced by Io.error.put_string.

For more translations see Appendix A with a list of 26 printf translations and the output of the executed print statements.

2.4 Translation in real examples

In order to see how many printf statements can be translated, we look at the reference projects listed in section 1.1 (Table 2.2). In those projects more than two-thirds of the printf statements can be translated in the above way to Eiffel.

Table 2.2: Number of translated printf statements

<table>
<thead>
<tr>
<th>Project</th>
<th>Total printf statements</th>
<th>Replaced printf statements</th>
</tr>
</thead>
<tbody>
<tr>
<td>hello world</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>micro_httpd</td>
<td>16</td>
<td>13</td>
</tr>
<tr>
<td>xeyes</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>less</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>wget</td>
<td>63</td>
<td>21</td>
</tr>
<tr>
<td>links</td>
<td>29</td>
<td>27</td>
</tr>
<tr>
<td>vim</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>libcurl</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>libgmp</td>
<td>16</td>
<td>10</td>
</tr>
<tr>
<td>libgsl</td>
<td>82</td>
<td>66</td>
</tr>
<tr>
<td>Total</td>
<td>219</td>
<td>148</td>
</tr>
</tbody>
</table>
Chapter 3

Translation of C strings to Eiffel Strings

In this chapter we look at C strings and how they can be translated to Eiffel strings. We will also see how the standard string function calls (from the string.h header file) can be replaced.

3.1 C Strings

In C a string is stored as an array of characters. Like for all arrays there are two possibilities to declare and allocate space for a string variable.

The first way is to declare the variable as character pointer as shown in listing 3.1 on line 1. With this declaration only space for the pointer is allocated, not for the content of the string. This is done with one of the memory allocation functions like malloc.

The second way to declare a string variable is as character array as shown in listing 3.1 on line 2. In this case also memory for $N$ characters is allocated.

Besides the difference in memory allocation, there is no difference in the usage of the string variables.

Listing 3.1: C: Declaration of string variables

<table>
<thead>
<tr>
<th>Line</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>char *char_pointer;</td>
</tr>
<tr>
<td>2</td>
<td>char char_array[N];</td>
</tr>
</tbody>
</table>
3.2 Eiffel Strings

In Eiffel strings are also character arrays. Those strings have, in contrast to C strings, no terminating null character at the end. The length of the Eiffel string is stored within the STRING class. This difference has consequences when accessing individual characters of the string. We will discuss those later in this chapter.

3.3 Standard C string functions

The standard C string functions are defined in the string.h header file. The following functions are there defined:

- `char * strcat(char *str1, const char *str2);`
  Appends `str2` to `str1`, where the terminating null character of `str1` is overwritten. The function returns a pointer to `str1`.

- `char * strncat(char *str1, const char *str2, size_t n);`
  Same as `strcat` but appends at most `n` characters.

- `char * strchr(const char *str, int c);`
  Searches for the first occurrence of `c` in the string `str` and returns a pointer to that location.

- `char * strrchr(const char *str, int c);`
  Same as `strchr` but searches the last occurrence of `c`.

- `int strcmp(const char *str1, const char *str2);`
  Compares the two strings and returns zero if they are equal, less than zero if the first string is less than the second one and greater zero otherwise.

- `int strncmp(const char *str1, const char *str2, size_t n);`
  Same as `strcmp` but compares only the first `n` characters.

- `char * strcoll(const char *str1, const char *str2);`
  Same as `strcmp` but takes the LC_COLLATE setting into account.

- `char * strcpy(char *str1, const char *str2);`
  Copies `str2` to `str1` and returns a pointer to `str1`.

- `char * strncpy(char *str1, const char *str2, size_t n);`
  Same as `strcpy` but copies only `n` character.

- `size_t strspn(const char *str1, const char *str2);`
  Searches for the first sequence of characters in `str1` that does contain only characters from `str2` and returns the length of this sequence.

- `size_t strcspn(const char *str1, const char *str2);`
  Same as `strspn` but searches for a sequence that does not contain any characters from `str2`. 
Translation of C strings - Translatable string variables

- `char * strerror(int errnum);`
  Returns a pointer to the corresponding error message.

- `size_t strlen(const char *str);`
  Returns the length of the string.

- `char * strpbrk(const char *str1, const char *str2);`
  Returns a pointer to the first character in `str1` that is also present in `str2`.

- `char * strstr(const char *str1, const char *str2);`
  Searches for the first occurrence of the entire string `str2` in `str1` and returns a pointer to the beginning.

- `char * strtok(char *str1, const char *str2);`

- `size_t strxfrm(char *str1, const char *str2, size_t n);`

3.4 Translatable string variables

The following types of C strings can potentially be translated to Eiffel strings:

- Global variables
- Function parameters and return types of internal functions
- Local variables

Supported operations on the string variables are:

- Assignment, where the right hand side of the assignment can be a string constant, another string variable or the return value of a function. This introduces a dependency that both sides of the assignment can be translated to Eiffel strings or none.

- Function calls to internal functions with a string variable as parameter. This introduces a dependency that both the declared and the actual parameter can be translated or none.

- Function calls to some external functions like `strcpy`, `strlen`, `printf` and others. The whole list of supported external functions is given in section 3.5.

- Equality or inequality check on two string variables. This introduces also a dependency between the two variables like an assignment.

As soon as one of the following operation is applied to a string variable, this string variable can not be translated to an Eiffel string. Also all variables that have a dependency to this one introduced through an assignment or a function call can not be translated.

- The string is declared as a struct field.
• An array of strings is not supported.

• The string is used inside the signature of a function pointer or as parameter in a call of one.

• The string variable is used in an unsupported external function call (see section 3.5).

• The string variable is used with pointer arithmetic like \( sp + 1 \).

• The string variable is casted to another type. A check if the pointer is zero (includes a cast to unsigned integer) is supported.

• Accessing an individual character of a string variable (see end of this section for the discussion of the problem).

• Dereferencing a string variable (equal to accessing the first character of the string).

Now we look at a small meaningless C program with string variables (listing 3.2) and the translation to Eiffel (listing 3.3). The Eiffel code is slightly refactored to have everything in one class and to fit on one page.

Most of the translation is straight forward. The types \texttt{char *} and \texttt{char []} are replaced by \texttt{STRING}, assignments remain normal assignments and \texttt{printf} calls are replaced where possible like described in chapter 2. Interesting is the replacement of the standard C string function \texttt{strcpy} (line 24 in the C code). It is replaced by creating an identical copy of the string by calling \texttt{twin} and an assignment (line 47 in the Eiffel code). Also noticeable is the \texttt{printf} call that can’t be translated (line 25 in the C code and line 48 in the Eiffel code). The string \texttt{s1} can still be translated to an Eiffel string but has to be converted before the call. In section 3.5 we will see more translations of standard string functions.

Array indexing problem

Now we discuss why accessing individual characters of a string can not be supported.

In C we allocate a given number of bytes for a string either by calling a memory allocation function or by declaring an array of characters with a given size. Then accessing any of those bytes with \texttt{s[n]} is valid even if \( n \) is greater or equal than the length of the string (the size returned by \texttt{strlen}). An example usage is when multiple strings are stored consecutively in one variable. In Eiffel you must not access any index greater than the length of the string. Since we can not decide on compile time if an access is within the bounds it can not be supported.
Listing 3.2: C: Program with string variables

```c
#include <stdio.h>
#include <string.h>

char *cp;
static char string[12] = "Hello World";

char* select(char *s, int i) {
    char *tmp;
    if (i != 0)
        tmp = s;
    else
        tmp = cp;
    return tmp;
}

int main()
{
    cp = "Test String";
    printf("%s\n", cp);
    char s1[100];
    strcpy (s1, cp);
    printf("%4s\n", s1);
    char *s2 = select(string, 1);
    printf("%s\n", s2);
    return 0;
}
```
Listing 3.3: Eiffel: The translation of the C program in listing 3.2

class P_TEST_STRING_TRANSFORMER
 inherit S_STDIO
 create
  default_create

feature (NONE) -- Initialization
  default_create
  do
    string := "Hello World"
  end

feature (ANY) -- attributes
  cp: STRING_8 assign set_cp
  set_cp (a_cp: STRING_8)
    do
      cp := a_cp
    end

  string: STRING_8 assign set_string
  set_string (a_string: STRING_8)
    do
      string := a_string
    end

feature (ANY) -- routines
  select2 (a_s: STRING_8; a_i: INTEGER_32): STRING_8
 local
  l_tmp: STRING_8
  do
    if a_i /= 0 then
      l_tmp := a_s
    else
      l_tmp := cp
    end
    Result := l_tmp
  end

main: INTEGER_32
 local
  l_s1: STRING_8
  l_s2: STRING_8
  do
    cp := "Test String"
    Io.put_string (cp + "%N")
    l_s1 := cp.twin
    printf ([ce_string ("%%.4s%N"), ce_string (l_s1)]).do_nothing
    l_s2 := (select2 (string, 1))
    Io.put_string (l_s2 + "%N")
    Result := 0
  end
 end
3.5 Translation of standard C string functions to Eiffel features

From the standard C string functions listed in section 3.3, six can be replaced by Eiffel features. The functions and their replacements are listed in Table 3.1.

Table 3.1: Standard C string functions and their Eiffel replacement

<table>
<thead>
<tr>
<th>C string function</th>
<th>Eiffel code</th>
</tr>
</thead>
<tbody>
<tr>
<td>strcat(s1, s2)</td>
<td>s1.append(s2)</td>
</tr>
<tr>
<td>strncat(s1, s2, n)</td>
<td>s1.append(s2.substring(1, s2.count.min(n)))</td>
</tr>
<tr>
<td>strcmp(s1, s2)</td>
<td>s1.three_way_comparation(s2)</td>
</tr>
<tr>
<td>strncmp(s1, s2, n)</td>
<td>s1.substring(1, s1.count.min(n)).three_way_comparation(s2.substring(1, s2.count.min(n)))</td>
</tr>
<tr>
<td>strcpy(s1, s2)</td>
<td>s1 := s2.twin or s1 := s2 (if s2 is a string literal)</td>
</tr>
<tr>
<td>strlen(s)</td>
<td>s.count</td>
</tr>
</tbody>
</table>

The functions marked with a * store the string result in the first argument and also return a pointer to it. This is not possible in Eiffel so those functions can only be replaced if they are used as statements and not as expression. Unfortunately, there is no feature in Eiffel to get a prefix of a string with at most $n$ characters. So we have to use substring and we have to calculate the end position since substring returns an empty string if $n$ is greater than the length of the string.

The replacement of the functions in Table 3.1 is possible when at least the first string argument can be transformed to an Eiffel string. If the second string argument can’t be transformed, then it can be converted on runtime with eif_string to an Eiffel string since it is only read.

In addition to the above replacements, it is for most functions listed in section 3.3 possible to have one or more arguments as Eiffel string even if the function cannot be replaced by a Eiffel feature. Those Eiffel strings are converted with ce_string before the call. This conversion is possible when the string is only read and no pointer to it is kept or returned by the function. Table 3.2 lists all standard string functions listed in section 3.3 and which arguments can be Eiffel strings.

3.6 Implementation of the analysis

The translation is done in two phases. In the first phase all string variables are collected and their usage in the code is analyzed. In the second phase, the types are replaced by a new AST node called AST_EIFFEL_STRING_TYPE, the external functions are replaced like described in section 3.5 and some other adjustments are done.

In this section, we will look at the first phase. The second phase is straight
Table 3.2: Eiffel strings as arguments to standard C string functions

<table>
<thead>
<tr>
<th>C string function signature</th>
<th>Eiffel string as 1st arg.</th>
<th>Eiffel string as 2nd arg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>char * strcat(char *, const char *)</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>char * strncat(char *, const char *, size_t)</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>char * strchr(const char *, int)</td>
<td>no</td>
<td>-</td>
</tr>
<tr>
<td>char * strrchr(const char *, int)</td>
<td>no</td>
<td>-</td>
</tr>
<tr>
<td>int strncpy(const char *, const char *, size_t)</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>char * strcoll(const char *, const char *)</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>char * strcpy(char *, const char *)</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>char * strncpy(char *, const char *, size_t)</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>size_t strspn(const char *, const char *)</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>size_t strspn(const char *, const char *, size_t)</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>char * strerror(int)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>size_t strlen(const char *)</td>
<td>yes</td>
<td>-</td>
</tr>
<tr>
<td>char * strpbrk(const char *, const char *)</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>char * strstr(const char *, const char *)</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>char * strtok(char *, const char *)</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>size_t strxfrm(char *, const char *, size_t)</td>
<td>no</td>
<td>yes</td>
</tr>
</tbody>
</table>

forward and is not discussed here.

For the analysis we use a dependency graph to store the dependencies between the string variables introduced through assignments or function calls (see section 3.4). Figure 3.1 shows the dependency graph for the C code in listing 3.2. The numbers next to the arrows indicate the line in the code that introduces the dependency. For example the assignment on line 11 introduces a dependency between the variables tmp and s.

![Dependency Graph](image)

Figure 3.1: The dependency graph for the C code in Listing 3.2. The numbers next to the arrows indicate the line in the code where the dependency is introduced. The variable s1 has no dependency to other variables.

While collecting those dependencies we also check the code for unsupported operations on the string variables (see section 3.4). In such situations the affected variables are marked as bad in the dependency graph. At the end of this first phase all variables that have a dependency to a bad one (possibly through other variables) are also marked bad. After that all variables that are
not marked as bad can be translated to Eiffel strings.

To get a better idea why a C string can not be translated to an Eiffel string, we also set a flag while marking a variable as bad. The used flags and in what situations they are set is described below.

Bad_ref flag
This flag is set while updating the dependency graph when a node has a dependency to a bad one and no other flag was set before (so it was not marked as bad before). Those variables could be translated to Eiffel strings if their dependencies wouldn’t have been marked as bad.

Extern_function flag
This flag is set when a variable is used as argument of an external function that can’t be replaced.

Struct_member flag
This flag is set when the variable is a struct field.

Pointer flag
This flag is set if the pointer of the string variable is used, for example in pointer arithmetic like \texttt{cp++}.

Cast flag
This flag is set when a string variable is casted to another type. Most of the time this is a cast to or from \texttt{void *}.

Function_pointer flag
This flag is set when a string variable is used inside the signature of a function pointer.

Assignment flag
This flag is set when a variable is part of an unsupported assignment like \texttt{array[0] = a\_string}.

Initializer flag
This flag is set when a variable is initialized with something else than a string literal. In those cases the variable is mostly used as an array for small integers.

Array flag
This flag is set when an individual character is accessed. See the array indexing problem at the end of section 3.4.

Other flag
This flag is set when a variable is part of an \texttt{AST\_EIFFEL\_EXPRESSION}. This happens when another transformation step has already introduced Eiffel code into the \texttt{AST}. 
3.7 Statistics

Table 3.3 shows how many strings in the reference projects (see section 1.1) can be translated to Eiffel strings. In average over all projects only 1.9 percent of the strings can be translated.

Table 3.3: Number of translatable strings in real projects

<table>
<thead>
<tr>
<th>Project</th>
<th>Total strings</th>
<th>Translatable strings</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) hello world</td>
<td>0</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>(2) micro_httpd</td>
<td>31</td>
<td>6</td>
<td>19.4%</td>
</tr>
<tr>
<td>(3) xeyes</td>
<td>0</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>(4) less</td>
<td>508</td>
<td>6</td>
<td>1.2%</td>
</tr>
<tr>
<td>(5) wget</td>
<td>2274</td>
<td>10</td>
<td>0.4%</td>
</tr>
<tr>
<td>(6) links</td>
<td>246</td>
<td>2</td>
<td>0.8%</td>
</tr>
<tr>
<td>(7) vim</td>
<td>2315</td>
<td>27</td>
<td>1.2%</td>
</tr>
<tr>
<td>(8) libcurl</td>
<td>1036</td>
<td>2</td>
<td>0.2%</td>
</tr>
<tr>
<td>(9) libgmp</td>
<td>143</td>
<td>3</td>
<td>2.1%</td>
</tr>
<tr>
<td>(10) libgs1</td>
<td>263</td>
<td>74</td>
<td>28.1%</td>
</tr>
<tr>
<td>Total</td>
<td>6818</td>
<td>130</td>
<td>1.9%</td>
</tr>
</tbody>
</table>

Table 3.4 shows in more detail why the strings can not be translated. It lists for all flags introduces in section 3.6 for how many strings this flag was set. The projects hello world and xeyes are omitted, since they have no strings. Table 3.5 summarizes those numbers.

Table 3.4: Statistics for individual flags and projects

<table>
<thead>
<tr>
<th>Flags</th>
<th>(2)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
<th>(8)</th>
<th>(9)</th>
<th>(10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Total strings)</td>
<td>31</td>
<td>508</td>
<td>2274</td>
<td>246</td>
<td>2315</td>
<td>1036</td>
<td>143</td>
<td>263</td>
</tr>
<tr>
<td>Bad_ref</td>
<td>8</td>
<td>196</td>
<td>713</td>
<td>58</td>
<td>376</td>
<td>300</td>
<td>26</td>
<td>61</td>
</tr>
<tr>
<td>Extern_function</td>
<td>15</td>
<td>96</td>
<td>785</td>
<td>31</td>
<td>399</td>
<td>506</td>
<td>56</td>
<td>26</td>
</tr>
<tr>
<td>Struct_member</td>
<td>0</td>
<td>14</td>
<td>113</td>
<td>6</td>
<td>35</td>
<td>136</td>
<td>5</td>
<td>21</td>
</tr>
<tr>
<td>Pointer</td>
<td>6</td>
<td>125</td>
<td>338</td>
<td>18</td>
<td>51</td>
<td>177</td>
<td>32</td>
<td>7</td>
</tr>
<tr>
<td>Cast</td>
<td>0</td>
<td>80</td>
<td>377</td>
<td>105</td>
<td>1612</td>
<td>236</td>
<td>36</td>
<td>8</td>
</tr>
<tr>
<td>Function_pointer</td>
<td>0</td>
<td>14</td>
<td>54</td>
<td>0</td>
<td>2</td>
<td>6</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>Assignment</td>
<td>2</td>
<td>24</td>
<td>37</td>
<td>1</td>
<td>13</td>
<td>89</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>Initializer</td>
<td>0</td>
<td>0</td>
<td>32</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Array</td>
<td>6</td>
<td>166</td>
<td>620</td>
<td>80</td>
<td>86</td>
<td>235</td>
<td>54</td>
<td>87</td>
</tr>
<tr>
<td>Other</td>
<td>0</td>
<td>18</td>
<td>122</td>
<td>4</td>
<td>8</td>
<td>47</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td>Bad marks</td>
<td>70</td>
<td>1278</td>
<td>5874</td>
<td>480</td>
<td>3348</td>
<td>2980</td>
<td>383</td>
<td>416</td>
</tr>
<tr>
<td>Average bad marks</td>
<td>2.3</td>
<td>2.5</td>
<td>2.6</td>
<td>2.0</td>
<td>1.4</td>
<td>2.9</td>
<td>2.7</td>
<td>1.6</td>
</tr>
</tbody>
</table>

The three most often set flags are the cast flag, the extern_function flag and the array flag. If we ignore the array indexing problem, more string variables could be translated (41 more in the examples (1) to (10) in total) but we can’t guarantee that the generated code doesn’t rise a precondition violation. But in most of the programs this wouldn’t be the case.
Table 3.5: Summarized statistics for individual flags

<table>
<thead>
<tr>
<th>Flags</th>
<th>Total</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Total strings)</td>
<td>6816</td>
<td></td>
</tr>
<tr>
<td>Bad_ref</td>
<td>1738</td>
<td>25.5%</td>
</tr>
<tr>
<td>Extern_function</td>
<td>1914</td>
<td>28.1%</td>
</tr>
<tr>
<td>Struct_member</td>
<td>330</td>
<td>4.8%</td>
</tr>
<tr>
<td>Pointer</td>
<td>754</td>
<td>11.1%</td>
</tr>
<tr>
<td>Cast</td>
<td>2454</td>
<td>36.0%</td>
</tr>
<tr>
<td>Function_pointer</td>
<td>87</td>
<td>1.3%</td>
</tr>
<tr>
<td>Assignment</td>
<td>175</td>
<td>2.6%</td>
</tr>
<tr>
<td>Initializer</td>
<td>34</td>
<td>0.5%</td>
</tr>
<tr>
<td>Array</td>
<td>1334</td>
<td>19.6%</td>
</tr>
<tr>
<td>Other</td>
<td>212</td>
<td>3.1%</td>
</tr>
<tr>
<td>Bad marks</td>
<td>14'829</td>
<td></td>
</tr>
<tr>
<td>Average bad marks</td>
<td>2.2</td>
<td></td>
</tr>
</tbody>
</table>
Chapter 4

Translation of C structs to Eiffel classes

The goal of this chapter is to replace C structs with simple Eiffel classes that only contain the attributes and setter features for them and no interfacing with C code.

4.1 Structs in C

In C a struct is used to group multiple variables together in a collection. It is a light version of a class like they are known from C++ or Eiffel, since structs can only have fields and no functions.

The only operations allowed with structs are copying, assigning to another struct (this includes passing structs as arguments to functions), accessing the address with the & Operator or accessing the fields with the . or -> Operators.

Listing 4.1 shows how structs can be declared.

Listing 4.1: C: Declaration of structs

```
struct fpoint_2 {
    float x;
    float y;
};
struct fpoint_2 fp2;
struct fpoint_3 {
    float x;
    float y;
    float z;
} fp3;
```
On the lines 1 to 4 a struct with the name \texttt{fpoint\_2} is declared. Then line 5 declares a variable with the name \texttt{fp2} of this struct type. The declaration of the struct and variables of this type can be combined. This is done on line 7 to 11 where a struct with the name \texttt{fpoint\_3} and a variable with the name \texttt{fp3} is declared.

### 4.2 Translatable structs

Listing 4.2 show the translation of the struct \texttt{fpoint\_2} from listing 4.1 to a simple Eiffel class that only has the attributes and the setter features.

Structs can be translated as long as no variable of this struct type is used in one of the following situations:

- An array of structs is not supported.
- The struct type is used inside the signature of a function pointer or as parameter in a call of one.
- A struct variable is used in an external function call.
- A struct variable is casted to another type. Casts between struct types can be supported if one can inherit from the other (see later in this section).
- The struct variable is used in embedded assembler code.
Listing 4.3: C: Inheritance for structs

```c
struct list {
    void *next ;
    void *prev ;
};

struct sub_list {
    sub_list *next ;
    sub_list *prev ;
    char *s;
};
```

**Pointer to structs**

In Eiffel and other object oriented languages variables are references to objects. When assigning one variable to the other only the reference is copied, the object itself is not. This is different for structs in C. They behave like simple data types, that means that in assignments and function calls the whole struct is copied. When we translate this to Eiffel we have to copy the whole class too. When the variable is declared as a pointer to a struct then the behavior is the same as with references in Eiffel.

In both cases we can replace the variable by the same Eiffel class but we have to handle assignments and function calls accordingly.

**Inheritance between structs**

When a struct extends another struct we can introduce an inheritance relationship. The necessary condition for such an inheritance is that the extending struct has all struct fields of the base struct with the same types or possibly subtypes. The names don’t have to be the same. They can be renamed with the `rename` clause in Eiffel. In addition to these fields the extending struct can have more fields with any type. casts between two structs that have an inheritance relation are supported in the translation to simple Eiffel classes. Listing 4.3 shows two structs, where the second one extends the first one.

### 4.3 Analysis Implementation

The translation works the same way as the translation of C strings to Eiffel strings described in section 3.6. But since the number of translatable structs (see section 4.4) is really small, the actual replacement of the structs is not implemented. For the analysis we use also a dependency graph like we did in the analysis for the translation of C string to Eiffel strings.

This time we don’t translate the type of individual variables, now we want to replace the whole type implementation. This means that a struct type is
replaced in the whole program or nowhere. As a consequence assignments or function calls don’t introduce dependencies anymore. The only situation where a dependency is introduced is when a struct type is used inside the declaration of another struct. This dependency is directed from the inner to the outer struct type, means that when the outer struct type is marked as bad so is the inner one.

In addition to the flags introduced in section 3.6 we need two more flags for the analysis of the structs:

Union flag

In C2Eiffel unions and structs are treated equally, only a boolean value indicates the difference. So this flag is set for all unions because unions cannot be translated to simple Eiffel classes.

Assembler flag

This flag is set whenever a struct type is used inside embedded assembler code.

4.4 Statistics

Table 4.1 shows how many structs in the reference projects can be translated to simple Eiffel classes. The numbers marked with a * are upper limits. Since the translation is not implemented, the translation of those structs is not tested. So the actual number could be less if not all unsupported operations are recognized. In average over all projects only 6.5% of the structs can be translated (at most).

<table>
<thead>
<tr>
<th>Project</th>
<th>Total structs</th>
<th>Translatable structs</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) hello world</td>
<td>0</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>(2) micro_httpd</td>
<td>6</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>(3) xeyes</td>
<td>61</td>
<td>1</td>
<td>1.6%</td>
</tr>
<tr>
<td>(4) less</td>
<td>28</td>
<td>1</td>
<td>3.6%</td>
</tr>
<tr>
<td>(5) wget</td>
<td>86</td>
<td>1</td>
<td>1.2%</td>
</tr>
<tr>
<td>(6) links</td>
<td>149</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>(7) vim</td>
<td>463</td>
<td>7*</td>
<td>1.5%</td>
</tr>
<tr>
<td>(8) libcurl</td>
<td>133</td>
<td>1</td>
<td>0.8%</td>
</tr>
<tr>
<td>(9) libgmp</td>
<td>36</td>
<td>5*</td>
<td>13.9%</td>
</tr>
<tr>
<td>(10) libgsl</td>
<td>303</td>
<td>66*</td>
<td>21.8%</td>
</tr>
<tr>
<td>Total</td>
<td>1265</td>
<td>82*</td>
<td>6.5%</td>
</tr>
</tbody>
</table>

The table 4.2 shows in more detail why the structs can’t be translated. It lists for all flags from section 3.6 and 4.3 how many structs those flag have. Additionally it shows how often the variables were marked as bas in total and on average. The project hello world is omitted, since it has no structs. Table 4.3 summarizes those numbers.
### Table 4.2: Statistics for individual flags and projects

<table>
<thead>
<tr>
<th>Flags</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
<th>(8)</th>
<th>(9)</th>
<th>(10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Total structs)</td>
<td>6</td>
<td>61</td>
<td>28</td>
<td>86</td>
<td>149</td>
<td>463</td>
<td>133</td>
<td>36</td>
<td>303</td>
</tr>
<tr>
<td>Bad ref</td>
<td>1</td>
<td>38</td>
<td>4</td>
<td>9</td>
<td>102</td>
<td>25</td>
<td>2</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Extern_func.</td>
<td>3</td>
<td>11</td>
<td>7</td>
<td>17</td>
<td>89</td>
<td>13</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Union</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>28</td>
<td>17</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Pointer</td>
<td>3</td>
<td>6</td>
<td>17</td>
<td>47</td>
<td>78</td>
<td>174</td>
<td>69</td>
<td>15</td>
<td>43</td>
</tr>
<tr>
<td>Cast</td>
<td>0</td>
<td>3</td>
<td>10</td>
<td>54</td>
<td>98</td>
<td>190</td>
<td>55</td>
<td>16</td>
<td>200</td>
</tr>
<tr>
<td>Function_ptr.</td>
<td>1</td>
<td>8</td>
<td>0</td>
<td>2</td>
<td>20</td>
<td>52</td>
<td>6</td>
<td>3</td>
<td>13</td>
</tr>
<tr>
<td>Array</td>
<td>1</td>
<td>5</td>
<td>7</td>
<td>17</td>
<td>31</td>
<td>88</td>
<td>12</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>Assembler</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Other</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>14</td>
<td>20</td>
<td>27</td>
<td>9</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Bad marks</td>
<td>27</td>
<td>185</td>
<td>257</td>
<td>1030</td>
<td>5507</td>
<td>9848</td>
<td>2386</td>
<td>844</td>
<td>2295</td>
</tr>
<tr>
<td>Av. bad marks</td>
<td>4.5</td>
<td>3.0</td>
<td>9.2</td>
<td>12.0</td>
<td>37</td>
<td>21.3</td>
<td>17.9</td>
<td>23.4</td>
<td>7.6</td>
</tr>
</tbody>
</table>

### Table 4.3: Summarized statistics for individual flags

<table>
<thead>
<tr>
<th>Flags</th>
<th>Total</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Total structs)</td>
<td>1265</td>
<td></td>
</tr>
<tr>
<td>Bad_ref</td>
<td>189</td>
<td>14.9%</td>
</tr>
<tr>
<td>Extern_func</td>
<td>160</td>
<td>12.6%</td>
</tr>
<tr>
<td>Union</td>
<td>64</td>
<td>5.1%</td>
</tr>
<tr>
<td>Pointer</td>
<td>452</td>
<td>35.7%</td>
</tr>
<tr>
<td>Cast</td>
<td>626</td>
<td>49.5%</td>
</tr>
<tr>
<td>Function_ptr</td>
<td>105</td>
<td>8.3%</td>
</tr>
<tr>
<td>Array</td>
<td>178</td>
<td>14.1%</td>
</tr>
<tr>
<td>Assembler</td>
<td>6</td>
<td>0.5%</td>
</tr>
<tr>
<td>Other</td>
<td>78</td>
<td>6.1%</td>
</tr>
</tbody>
</table>

| Bad marks    | 22379 |            |
| Average bad marks | 17.7 |         |
Chapter 5

Removing break, continue and return statements

The C programming language allows to jump from one place in the code to another with jump instructions like goto, break and others. The purpose of this chapter is to remove some of those jump instructions because Eiffel doesn’t support them. In particular the following jump instructions are removed:

\textit{break} (only those belonging to a loop)

When a \textit{break} statement is executed, the execution immediately exits the current loop.

\textit{continue}

When a \textit{continue} statement is executed, the execution skips the rest of the loop. In a \textit{for} loop the execution continues with the increment block. For all other loops the execution continues with the checking of the loop condition.

\textit{return}

When a \textit{return} statement is executed, the function terminates immediately. No further code is executed in this function.

\textit{Goto}’s are not removed by this transformation, they are handled by a later transformation step.

Existing translation

The existing handling of \textit{break}, \textit{continue} and \textit{return} emulates jumps (also \textit{goto}) in Eiffel. This is achieved with an inspect instruction inside a loop and an additional variable (\texttt{lpos}) that stores the jump destination. The different inspect values mark the different jump destinations and a special value (−1) is used as an exit condition for the loop. After the inspect statement the \texttt{lpos} variable is incremented. So when no jump statement has changed the variable, execution
Removing jump statements - Code transformation

Listing 5.1: C: Gcd Algorithm

```c
int gcd(int a, int b) {
    if (a == 0) {
        return b;
    }
    while (1) {
        if (b == 0) {
            break;
        }
        if (a > b) {
            a -= b;
        } else {
            b -= a;
        }
    }
    return a;
}
```

just continues with the directly following instructions. As an example listing 5.2 shows the transformation of the `gcd` algorithm from listing 5.1. This simple example shows also that the generated Eiffel code is not easy readable.

The advantage of this approach is that all jump statements can be handled at the same time. In the next section we will see an approach that is more readable but can’t handle `break`’s in `switch` statements and `goto`’s. For the handling of those jump statements the existing approach is kept. The advantage of having both approaches at the same time is that the number of jump destinations is reduced and for many functions the emulation of jumps isn’t needed anymore. Therefore the readability is improved.

### 5.1 Code transformation

The jump instructions that we remove in this chapter (`break`, `continue` and `return`) don’t jump to programmer defined places in the code. They basically just skip the following instructions inside a loop or a function. This allows us to use `if` statements to implement those jumps by moving the following statements into a new `if` statement with an appropriate condition [11]. We will call those `if` statements `skip blocks`.

For this we introduce a new class `JUMP_STATE` (see listing 5.3) that tracks which jump statement was called and has features for the usage in the condition of the skip blocks. Every function that has a `break`, `continue` or `return` gets an additional local variable called `js` of type `JUMP_STATE`.

Then every `break` statement is replaced with the feature call `js.break` that sets the value of `break_called` to `True`. Analogously `continue` and `return` statements...
Listing 5.2: *Eiffel*: Gcd Algorithm from listing 4.1

```eiffel
gcd (a_a : INTEGER_32; a_b : INTEGER_32): INTEGER_32
do
from
l_pos := 0
until
l_pos = -1 -- -1: return
loop
  inspect l_pos
  -- <function body>
when 0 then -- execution start:
  l_a := a_a
  l_b := a_b
  if l_a = 0 then
      Result := l_b
      l_pos := -1 + -1 -- return
  end
when 1 then
  if False then
      l_pos := -1 + 4 -- skip loop
  end
when 2 then -- loop body:
  if l_b = 0 then
    l_pos := -1 + 4 -- break
  end
when 3 then
  if l_a > l_b then
    l_a := l_a - l_b
  else
    l_b := l_b - l_a
  end
  l_pos := -1 + 1 -- back to loop condition
when 4 then -- loop done
  Result := l_a
  l_pos := -1 + -1 -- return
end
-- </function body>
end
```

---

Removing jump statements - Code transformation
Listing 5.3: Eiffel: Jump State class

```
expanded class
   JUMP_STATE

feature {ANY} -- return
   return_called: BOOLEAN
   -- "return" called

   return
   do
      return_called := True
   end

feature {ANY} -- loops
   break_called: BOOLEAN
   -- "break" called

   break
   do
      break_called := True
   end

   continue_called: BOOLEAN
   -- "continue" called

   continue
   do
      continue_called := True
   end

feature {ANY} -- queries
   loop_jumped: BOOLEAN
   -- "break", "continue" or "return" called

   do
      Result := break_called or continue_called or return_called
   end

   loop_left: BOOLEAN
   -- "break" or "return" called

   do
      Result := break_called or return_called
   end

feature {ANY}
   reset
   do
      break_called := False
      continue_called := False
   end
```

end
Removing jump statements - Optimization

are replaced with the corresponding feature calls. The call \texttt{js.return} is omitted when the execution of the function ends anyway after the \texttt{return} statement. The other features of the class \texttt{JUMP\_STATE} are:

- The feature \texttt{loop\_jumped} indicates if the remaining statements of a loop should be skipped. This is the case if any of the three jump statements was executed. It is used for the conditions of the new skip blocks.

- The feature \texttt{loop\_left} indicates if the execution should leave the current loop. This is the case if either \texttt{break} or \texttt{return} was called. It is used as additional condition for the loops.

- The feature \texttt{reset} resets the values of \texttt{break\_called} and \texttt{continue\_called} to \texttt{False}. It is used as first instruction of a loop to revert the setting of those values in the previous loop iteration. It is also used after every loop since those values are only valid within the loop they are called.

Listing 5.4 shows the transformation of the \texttt{gcd} algorithm from listing 5.1.

Nested loops

At first one would think that the value of \texttt{break\_called} and \texttt{continue\_called} could be changed by a nested loop and therefore one would need to have separate values for every nested loop. But whenever the execution enters a nested loop \texttt{break\_called} and \texttt{continue\_called} are \texttt{False}. Otherwise we would have skipped the loop in the first place.

There is one exceptional case: When the increment block of a \texttt{for} loop contains a loop. Then the \texttt{continue\_called} can be \texttt{True}. But at this point we don’t need the value anymore since the \texttt{continue} is already done and it doesn’t influence the nested loop since the value is reset at the beginning anyway.

Also jumps with \texttt{goto} into or out of a loop are no problem. \texttt{continue\_called} and \texttt{break\_called} are always \texttt{False} when a \texttt{goto} is executed. This can be verified in the same way as above.

5.2 Optimization

Through the insertion of additional skip blocks the code indentation increases. This can affect readability when multiple skip blocks are nested and therefore the indentation level increases a lot. To counteract this incrementation two optimizations are implemented.

The first optimization tries to flatten the nesting of skip blocks. If a skip block is the last statement in another skip block and they have the same skip condition, it can be removed from the other skip block and inserted right after it. This reduces the indentation level of that skip block and all statements inside by one. Listings 5.5 and 5.6 illustrate this optimization.
Listing 5.4: Eiffel: Gcd Algorithm from listing 4.1 (second version)

```eiffel
1 gcd (a_a: INTEGER_32; a_b: INTEGER_32): INTEGER_32
2 local
3  l_a: INTEGER_32
4  l_b: INTEGER_32
5  js: JUMP_STATE
6  do
7    l_a := a_a
8    l_b := a_b
9 if l_a = 0 then
10   Result := l_b
11   js.return
12 end
13 if not js.return_called then
14  from
15  until js.loop_left
16  loop
17    js.reset
18    if l_b = 0 then
19      js.break
20    end
21 if not js.loop_jumped then
22    if l_a > l_b then
23      l_a := l_a - l_b
24    else
25      l_b := l_b - l_a
26    end
27 end
28 end
29 js.reset
30 Result := l_a
31 end
32 end
```

Listing 5.5: Eiffel: Nested skip blocks

```eiffel
1 if not js.loop_left then
2  if foo then
3    js.break
4  end
5 if not js.loop_left then
6  bar
7 end
8 end
```
Removing jump statements - Statistics

Listing 5.6: Eiffel: Nested skip blocks after optimization

```eiffel
1 if not js.loop_left then
2   if foo then
3     js.break
4   end
5 end
6 if not js.loop_left then
7   bar
8 end
```

Listing 5.7: Eiffel: Merged skip block and if statement

```eiffel
1 if not js.loop_left and then foo then
2   js.break
3   end
4 if not js.loop_left then
5   bar
6 end
```

The second optimization combines a skip block with an if statement inside under the condition that the if statement is the only statement inside the skip block and it has no else part. The combination is done by combining the condition of the skip block with the condition of the if statement with an and then. This ensures that the second condition is only evaluated when it would also be evaluated in the original code. This optimization also reduces the indentation level of the if statement and all statements inside by one. Listing 5.7 shows the result of this optimization for the code from listing 5.6.

5.3 Statistics

In the following we will see how much the transformation described in this chapter changes the code of the reference projects (see section 1.1).

At first we look at how many jump statements could be removed. The remaining jump statements are break statements that belong to a switch or goto statements. Table 5.1 shows how many jump statements are in the projects before and after the refactoring. In the last column is the number of added skip blocks. Over all projects more than 80 percent of the jump statements could be removed.

Functions that have no jump statement after the transformation don’t need the emulation of jumps described in the beginning of this chapter. Table 5.2 lists how many functions have jump statements before and after the transformation. The number of such functions decreases clearly.

As discussed in section 5.2 the transformation changes the indentation level of the code. Table 5.3 shows the maximal and average code indentation before
Table 5.1: Removed jump statements

<table>
<thead>
<tr>
<th></th>
<th>before</th>
<th>after</th>
<th>removed</th>
<th>removed [%]</th>
<th>skip blocks</th>
</tr>
</thead>
<tbody>
<tr>
<td>micro_httpd</td>
<td>38</td>
<td>1</td>
<td>37</td>
<td>97.4</td>
<td>22</td>
</tr>
<tr>
<td>xeyes</td>
<td>21</td>
<td>4</td>
<td>17</td>
<td>81.0</td>
<td>0</td>
</tr>
<tr>
<td>less</td>
<td>1426</td>
<td>376</td>
<td>1050</td>
<td>73.6</td>
<td>602</td>
</tr>
<tr>
<td>wget</td>
<td>2636</td>
<td>566</td>
<td>2070</td>
<td>78.5</td>
<td>885</td>
</tr>
<tr>
<td>links</td>
<td>3650</td>
<td>1072</td>
<td>2578</td>
<td>70.6</td>
<td>1372</td>
</tr>
<tr>
<td>vim</td>
<td>18085</td>
<td>5023</td>
<td>13062</td>
<td>72.2</td>
<td>7597</td>
</tr>
<tr>
<td>libcurl</td>
<td>3369</td>
<td>974</td>
<td>2395</td>
<td>71.1</td>
<td>1499</td>
</tr>
<tr>
<td>libgmp</td>
<td>5483</td>
<td>367</td>
<td>5116</td>
<td>93.3</td>
<td>733</td>
</tr>
<tr>
<td>libgsl</td>
<td>19404</td>
<td>1796</td>
<td>17608</td>
<td>90.7</td>
<td>10118</td>
</tr>
<tr>
<td>Total</td>
<td>54112</td>
<td>10179</td>
<td>43933</td>
<td>81.2</td>
<td>22828</td>
</tr>
</tbody>
</table>

Table 5.2: Functions with jump statements

<table>
<thead>
<tr>
<th></th>
<th>before</th>
<th>after</th>
<th>less</th>
<th>less [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>micro_httpd</td>
<td>7</td>
<td>1</td>
<td>6</td>
<td>85.7</td>
</tr>
<tr>
<td>xeyes</td>
<td>17</td>
<td>2</td>
<td>15</td>
<td>88.2</td>
</tr>
<tr>
<td>less</td>
<td>375</td>
<td>70</td>
<td>305</td>
<td>81.3</td>
</tr>
<tr>
<td>wget</td>
<td>599</td>
<td>100</td>
<td>499</td>
<td>83.3</td>
</tr>
<tr>
<td>links</td>
<td>994</td>
<td>199</td>
<td>795</td>
<td>80.0</td>
</tr>
<tr>
<td>vim</td>
<td>4113</td>
<td>875</td>
<td>3238</td>
<td>78.7</td>
</tr>
<tr>
<td>libcurl</td>
<td>640</td>
<td>137</td>
<td>503</td>
<td>78.6</td>
</tr>
<tr>
<td>libgmp</td>
<td>580</td>
<td>91</td>
<td>489</td>
<td>84.3</td>
</tr>
<tr>
<td>libgsl</td>
<td>5073</td>
<td>515</td>
<td>4558</td>
<td>89.8</td>
</tr>
<tr>
<td>Total</td>
<td>12398</td>
<td>1990</td>
<td>10408</td>
<td>83.9</td>
</tr>
</tbody>
</table>

and after the transformation. It shows that the change is acceptably small and therefore doesn’t have a negative influence on the readability.

The transformation also introduces additional statements. Table 5.4 shows that the number of statements increases on average by 15 percent. Since the jump statements have to be modeled in Eiffel somehow, additional statements can not be avoided.
Table 5.3: Change in code indentation

<table>
<thead>
<tr>
<th></th>
<th>maximal nesting</th>
<th>average nesting</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>before</td>
<td>after</td>
<td>before</td>
<td>after</td>
</tr>
<tr>
<td>micro_httpd</td>
<td>5</td>
<td>5</td>
<td>0.90</td>
<td>1.10</td>
</tr>
<tr>
<td>xeyes</td>
<td>5</td>
<td>5</td>
<td>0.46</td>
<td>0.45</td>
</tr>
<tr>
<td>less</td>
<td>13</td>
<td>13</td>
<td>0.75</td>
<td>1.01</td>
</tr>
<tr>
<td>wget</td>
<td>18</td>
<td>18</td>
<td>0.92</td>
<td>1.14</td>
</tr>
<tr>
<td>links</td>
<td>22</td>
<td>25</td>
<td>0.86</td>
<td>1.11</td>
</tr>
<tr>
<td>vim</td>
<td>96</td>
<td>96</td>
<td>1.32</td>
<td>1.59</td>
</tr>
<tr>
<td>libcurl</td>
<td>20</td>
<td>26</td>
<td>1.02</td>
<td>1.33</td>
</tr>
<tr>
<td>libgmp</td>
<td>17</td>
<td>18</td>
<td>1.14</td>
<td>1.37</td>
</tr>
<tr>
<td>libgsl</td>
<td>19</td>
<td>20</td>
<td>1.00</td>
<td>1.25</td>
</tr>
</tbody>
</table>

Table 5.4: Number of statements

<table>
<thead>
<tr>
<th></th>
<th>before</th>
<th>after</th>
<th>difference</th>
<th>difference [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>micro_httpd</td>
<td>234</td>
<td>284</td>
<td>50</td>
<td>21.37</td>
</tr>
<tr>
<td>xeyes</td>
<td>246</td>
<td>249</td>
<td>3</td>
<td>1.22</td>
</tr>
<tr>
<td>less</td>
<td>7'019</td>
<td>8'263</td>
<td>1'244</td>
<td>17.72</td>
</tr>
<tr>
<td>wget</td>
<td>17'971</td>
<td>20'382</td>
<td>2'411</td>
<td>13.42</td>
</tr>
<tr>
<td>links</td>
<td>26'957</td>
<td>29'396</td>
<td>2'439</td>
<td>9.05</td>
</tr>
<tr>
<td>vim</td>
<td>138'274</td>
<td>152'477</td>
<td>14'203</td>
<td>10.27</td>
</tr>
<tr>
<td>libcurl</td>
<td>19'371</td>
<td>22'765</td>
<td>3'394</td>
<td>17.52</td>
</tr>
<tr>
<td>libgmp</td>
<td>30'670</td>
<td>35'937</td>
<td>5'267</td>
<td>17.17</td>
</tr>
<tr>
<td>libgsl</td>
<td>123'059</td>
<td>148'884</td>
<td>25'825</td>
<td>20.99</td>
</tr>
<tr>
<td>Total</td>
<td>363'801</td>
<td>418'637</td>
<td>54'836</td>
<td>15.07</td>
</tr>
</tbody>
</table>
Chapter 6

Conclusions

6.1 Conclusions

The C and Eiffel programming languages are quite different and therefore translating C strings or structs to the Eiffel counterpart is not very successful. The main reasons for this is the usage of pointers and external libraries in most C programs. The number of translatable strings is for most of the projects too small to justify the usage of two different strings representations. Those would raise difficulties when the code is changed later and string variables with a different representations are used together for example in an assignment.

In contrast to the translation of strings and structs the translation of printf statements to the Eiffel feature Io.put_string is possible for the common usages. The refactoring of the code to remove jump statements like break and continue always works and is a considerable improvement over the emulation of jumps like it is described in chapter 5. It enhances the readability with an acceptable increase in the code size.

6.2 Future Work

Possible future topics are the translation of C arrays to Eiffel arrays and the iteration over those from pointers to Eiffel iterators. Also the file handling could be translated to Eiffel’s own file handling. But after the results of the string and struct translations the success of those is uncertain.
Appendix A

Additional translations of printf calls

Listing A.1: C: test_printf_transformer class

```c
#include <stdio.h>

/* Different printf statements to test the PRINTF_TRANSFORMER */
int main(int argc, char* argv[])
{
    // supported printf statements
    char c = 'A';
    int code = 66;
    printf("A character: %c\n", c);
    printf("A character from ASCII code: %c\n", code);
    printf("A constant character: %c\n", 'C');
    printf("A constant character from ASCII code: %c\n", 68);

    int i = -123;
    long int l = 1234567l;
    long long int ll = 123456789000ll;
    unsigned int ui = 123;
    unsigned short us = 123;
    int width = 6;
    printf("An integer: %d\n", i);
    printf("An integer with width: %5i\n", i);
    printf("An integer with variable width: %*i\n", width, i);
    printf("A long integer: %d\n", l);
    printf("A long long integer: %d\n", ll); // will be truncated
    printf("An integer from unsigned int: %i\n", ui);
    printf("An integer from unsigned short: %i\n", us);
    printf("A constant integer: %d\n", 1234);

    float f = 123.45678; // will be truncated
    double d = 123.45678;
```
int small_width = 1;
int big_width = 12;
int precision = 3;
printf("A float: %f\n", f);
printf("A float with precision: %.1f\n", f);
printf("A float with variable precision: %.3f\n", precision, f);
printf("A float with (to) small width: %5f\n", f);
printf("A float with variable width: %*f\n", big_width, f);
printf("A double with width and precision: %1.3f\n", d);
printf("A double with variable width and precision: %*.*f\n", small_width, precision, d);
printf("A double with variable width and fix precision: %*.3f\n", big_width, d);
printf("A double with fix width and variable precision: %12.*f\n", precision, d);
printf("A constant float: %f", 1.234);

char* s = "A variable String";
printf("A string: %s\n", s);
printf("A constant string: %s\n", "Hello World");
printf("An unsigned integer: %u\n", ui);
printf("An unsigned integer from int: %u\n", i); // to test conversion

return 0;

Listing A.2: Output from the code in listing A.1

1 A character: A
2 A character from ASCII code: B
3 A constant character: C
4 A constant character from ASCII code: D
5 An integer: -123
6 An integer with width: -123
7 An integer with variable width: -123
8 A long integer: 1234567
9 A long long integer: -1097262584
10 An integer from unsigned int: 123
11 An integer from unsigned short: 123
12 A constant integer: 1234
13 A float: 123.456779
14 A float with precision: 123.5
15 A float with variable precision: 123.457
16 A float with (to) small width: 123.456779
17 A float with variable width: 123.456779
18 A double with width and precision: 123.457
19 A double with variable width and precision: 123.457
20 A double with variable width and fix precision: 123.457
21 A double with fix width and variable precision: 123.457
22 A constant float: 1.234000
Additional translations of `printf` calls

A string: A variable String
A constant string: Hello World
An unsigned integer: 123
An unsigned integer from int: 4294967173

Listing A.3: Eiffel: Translated test printf transformer class

class
TEST_PRINTF_TRANSFORMER

inherit
TEST_PRINTF_TRANSFORMER_DATA

feature (ANY) — routines
main (argc: INTEGER_32; argv: CE_POINTER [CE_POINTER [INTEGER_8]]) : INTEGER_32

local
c: INTEGER_8
code: INTEGER_32
i: INTEGER_32
l: INTEGER_32
ll: INTEGER_64
ui: NATURAL_32
us: NATURAL_16
width: INTEGER_32
f: REAL_32
d: REAL_64
small_width: INTEGER_32
big_width: INTEGER_32
precision: INTEGER_32
s: CE_POINTER [INTEGER_8]

do
create s.make

c := ('A').code).to_integer_8
code := 66
Io.put_string("A character: " + (c).to_character_8.
out + "%N")
Io.put_string("A character from ASCII code: " + (code
).to_character_8.out + "%N")
Io.put_string("A constant character: C%N")
Io.put_string("A constant character from ASCII code:
D%N")
i := -123
l := 1234567
ll := {INTEGER_64} 123456789000
ui := {NATURAL_32} 123
us := (123).to_natural_16
width := 6
Io.put_string("An integer: " + i.out + "%N")
Io.put_string("An integer with width: " + (create
 {FORMAT_INTEGER}.make (5)).formatted (i) + "%N")
Io.put_string("An integer with variable width: " + (create
 {FORMAT_INTEGER}.make (width)).formatted (i)
43. Io.put_string("A long integer: " + l.out + "%N")
44. Io.put_string("A long long integer: " + ll.to_integer_32.out + "%N")
45. Io.put_string("An integer from unsigned int: " + ui.to_integer_32.out + "%N")
46. Io.put_string("An integer from unsigned short: " + us.out + "%N")
47. Io.put_string("A constant integer: 1234%N")
48. f := (123.45677999999999).truncated_to_real
49. d := 123.45677999999999
50. small_width := 1
51. big_width := 12
52. precision := 3
53. Io.put_string("A float: " + (create {FORMAT_DOUBLE}.make(6, 6)).formatted(f) + "%N")
54. Io.put_string("A float with precision: " + (create {FORMAT_DOUBLE}.make(1, 1)).formatted(f) + "%N")
55. Io.put_string("A float with variable precision: " + (create {FORMAT_DOUBLE}.make((1).max(precision), precision)).formatted(f) + "%N")
56. Io.put_string("A float with (to) small width: " + (create {FORMAT_DOUBLE}.make(6, 6)).formatted(f) + "%N")
57. Io.put_string("A float with variable width: " + (create {FORMAT_DOUBLE}.make(big_width.max(6), 6)).formatted(f) + "%N")
58. Io.put_string("A double with width and precision: " + (create {FORMAT_DOUBLE}.make(3, 3)).formatted(d) + "%N")
59. Io.put_string("A double with variable width and precision: " + (create {FORMAT_DOUBLE}.make(small_width.max(precision), precision)).formatted(d) + "%N")
60. Io.put_string("A double with variable width and fix precision: " + (create {FORMAT_DOUBLE}.make(big_width.max(3), 3)).formatted(d) + "%N")
61. Io.put_string("A double with fix width and variable precision: " + (create {FORMAT_DOUBLE}.make((12).max(precision), precision)).formatted(d) + "%N")
62. Io.put_string("A constant float: " + (create {FORMAT_DOUBLE}.make(6, 6)).formatted(1.234) + "%N")
63. s.ce_assign(ce_string("A variable String"))
64. Io.put_string("A string: " + eif_string(s) + "%N")
65. Io.put_string("A constant string: Hello World%N")
66. Io.put_string("An unsigned integer: " + ui.out + "%N")
67. Io.put_string("An unsigned integer from int: " + i.to_integer_32.out + "%N")
Result := (0)
end
List of Tables

2.1 Supported printf specifiers and format tags .................. 10
2.2 Number of translated printf statements ..................... 12

3.1 Standard C string functions and their Eiffel replacement .... 19
3.2 Eiffel strings as arguments to standard C string functions .... 20
3.3 Number of translatable strings in real projects .............. 22
3.4 Statistics for individual flags and projects .................. 22
3.5 Summarized statistics for individual flags .................... 23

4.1 Number of translatable structs in real projects .............. 28
4.2 Statistics for individual flags and projects .................. 29
4.3 Summarized statistics for individual flags .................... 29

5.1 Removed jump statements ..................................... 38
5.2 Functions with jump statements ............................... 38
5.3 Change in code indentation ................................... 39
5.4 Number of statements ......................................... 39
List of Figures

3.1 Dependency graph ............................................. 20
## Listings

<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td><em>C</em>: Five exemplary <em>printf</em> statements</td>
<td>11</td>
</tr>
<tr>
<td>2.2</td>
<td><em>Eiffel</em>: The five translated <em>printf</em> statements</td>
<td>11</td>
</tr>
<tr>
<td>3.1</td>
<td><em>C</em>: Declaration of string variables</td>
<td>13</td>
</tr>
<tr>
<td>3.2</td>
<td><em>Eiffel</em>: The translation of the <em>C</em> program in listing 3.2</td>
<td>18</td>
</tr>
<tr>
<td>4.1</td>
<td><em>C</em>: Declaration of structs</td>
<td>25</td>
</tr>
<tr>
<td>4.2</td>
<td><em>Eiffel</em>: The translation of the struct <em>fpoint2</em> from listing 4.1</td>
<td>26</td>
</tr>
<tr>
<td>4.3</td>
<td><em>C</em>: Inheritance for structs</td>
<td>27</td>
</tr>
<tr>
<td>5.1</td>
<td><em>C</em>: Gcd Algorithm</td>
<td>32</td>
</tr>
<tr>
<td>5.2</td>
<td><em>Eiffel</em>: Gcd Algorithm from listing 4.1</td>
<td>33</td>
</tr>
<tr>
<td>5.3</td>
<td><em>Eiffel</em>: Jump State class</td>
<td>34</td>
</tr>
<tr>
<td>5.4</td>
<td><em>Eiffel</em>: Gcd Algorithm from listing 4.1 (second version)</td>
<td>36</td>
</tr>
<tr>
<td>5.5</td>
<td><em>Eiffel</em>: Nested skip blocks</td>
<td>36</td>
</tr>
<tr>
<td>5.6</td>
<td><em>Eiffel</em>: Nested skip blocks after optimization</td>
<td>37</td>
</tr>
<tr>
<td>5.7</td>
<td><em>Eiffel</em>: Merged skip block and <em>if</em> statement</td>
<td>37</td>
</tr>
<tr>
<td>A.1</td>
<td><em>C</em>: test_printf_transformer class</td>
<td>43</td>
</tr>
<tr>
<td>A.2</td>
<td>Output from the code in listing A.1</td>
<td>44</td>
</tr>
<tr>
<td>A.3</td>
<td><em>Eiffel</em>: Translated test_printf_transformer class</td>
<td>45</td>
</tr>
</tbody>
</table>
Bibliography


    http://jaftec.origo.ethz.ch.