



RAPPORT DE STAGE D'OPTION SCIENTIFIQUE

Titre

NON CONFIDENTIEL

Option :	INFORMATIQUE
Champ de l'option :	Math-Informatique
Directeur de l'option :	Olivier Bournez
Directeur de stage :	Olivier Bournez
Dates du stage :	7 avril - 22 août 2014
Nom et adresse de l'organisme :	SRI International Computer Science Laboratory (CSL) 333 Ravenswood Avenue Menlo Park, CA 94025-3493 United States

4 août 2014

Abstract

PVS (standing for Prototype Verification System), is an Open Source project developed by CSL at SRI International and aiming to be both a semi-automated theorem prover providing formal support for conceptualization and debugging in the early stages of the design of hardware or software systems and a programming language. The evaluation of PVS expressions relies so far on a build-in PVS interpreter based on Common Lisp, called "Ground Evaluator". In order to allow the integration of PVS code as well as its fast execution for debugging and testing purposes, we describe here a translator of a subset of PVS to the language C.

The update of aggregate data structure, such as arrays, are frequent in functional programs and requires copying before being updated which is a significant source of space/time inefficiencies. However the execution of updates by copying is often redundant and could be safely implemented by means of destructive, inplace updates in an imperative program. We describe a method for analyzing and replacing the safe updates in an imperative program by destructive, inplace update. This method has been implemented to optimize the PVS translator.

Acknowledgements

I would like to thank my supervisor, Natarajan Shankar, for his help, explanations and suggestions as well as for the many enlightening discussions we had during this internship. I also thank Sam Owre for his explanations of the PVS API and Common Lisp in general, Robin Larrieu, from Polytechnique who shared an office and a lot of good ideas with me. I thank all of my teachers from LIX who made this internship possible, with a special mention to Stéphane Graham-Lengrand and Benjamin Doerr who recommended me. Finally, I thank all the people at the CSL, for their welcome, the interesting discussions I had with them, and for creating an exciting and inspiring environment for work. A special thank Lori Truitt for all the help she provided with administrative paperwork.

Contents

1	Introduction	4
1.1	PVS Overview	4
1.2	Why translate PVS ?	4
2	Translating PVS	5
2.1	Translator's architecture	5
2.2	Translating PVS type	6
2.2.1	PVS type system	6
2.2.2	Translating rules	6
2.3	A few translation rules	7
2.4	Translating PVS syntax	8
2.5	Using a representation of the C language	8
3	Update expressions	8
3.1	Pointer counting	8
3.1.1	How to use it	9
3.1.2	Pros and cons	10
3.2	Using a different data structure	12
3.3	Flow analysis on the PVS code	12
3.4	Analysis of the C code	12
3.4.1	Algorithm	13
3.4.2	Algorithm	16
3.5	Combination of solutions	17
4	Static analysis of the intermediate language	20
4.1	Operational semantic	20
4.2	Sets of variables	23
4.3	Analysis of update context	24
4.3.1	Analysis	25
4.3.2	Proof of bisimulation	26
4.3.3	Analysis in term of references	27
4.4	Difference with the previous analysis of PVS	27
4.5	Draft	28
5	Conclusion	29
5.1	Difficulties and successes	29
5.1.1	Working with new languages and tools	29
5.1.2	Integrating the GMP library	29
5.2	What's left to be done ?	29
5.3	My stay at SRI	30
A	PVS Syntax and CLOS representation	31
B	PVS type system and CLOS representation	33
C	Intermediate languages	34
D	Rules	36
E	Examples	37

1 Introduction

1.1 PVS Overview

PVS (Prototype Verification System) is an environment for specification and proving. The main purpose of PVS is to provide formal support for conceptualization and debugging in the early stages of the lifecycle of the design of hardware or software systems. In these stages, both the requirements and designs are expressed in abstract terms that are not necessarily executable. The best way to analyze such an abstract specification is by attempting proofs of desirable consequences of the specification. Subtle errors revealed by trying to prove the properties are costly to detect and correct at later stages of the design lifecycle.

The specification language of PVS is built on higher-order logic (functions can be treated like primitive types: functions can take functions as arguments and return them as values, quantification can be applied to function variables. Specifications can be constructed using definitions and axioms

1.2 Why translate PVS ?

Translating a specification language into an

The HACMS Project

The DARPA-funded project High Assurance Cyber-Military System (HACMS) aims to produce systems and software with proved reliability and security. It uses the abstraction of nodes and topics to describe the different components and communication channels. A PVS model of these nodes and topics allows to prove security properties. However the project requires an implementation and integration of the proven algorithms into the system. For that purpose, a translator PVS to C would be very helpful since most of the ROS systems use the C language.

Other translators

Translating PVS was already done for different purposes.

- PVS come with a native PVS to Common Lisp translator. This is used to run PVS programs for testing and debugging purposes. Since PVS's API relies heavily on the Common Lisp language, it is very easy to run PVS code within Emacs. The "Ground Evaluator" uses the generated Common Lisp code to evaluate PVS expressions. It provides a more user-friendly interface for the PVS translator by being integrated into Emacs and translating simple Common Lisp expression back to PVS.
- PVS expressions can also be translated to Yices's specification language syntax. Yices is an efficient SMT solver developed at SRI by Bruno Dutertre. This translation is a way to "plug" the Yices solver into PVS, allowing automated proof of theorems and TCCs.
- PVS was translated to the Clean language.
- Lately PVS was also translated to the SMT-LIB language (work in progress) for standardization purposes. This would allow every SMT solver able to process SMT-LIB standard inputs such as Yices 2, to be plugged into PVS for further proof automation.

2 Translating PVS

A translator is a program taking the source code of program P written in the programming language \mathcal{L}_A as an entry and generating the code of an other program Q in the target language \mathcal{L}_B as an output. In our case, \mathcal{L}_A is a subset of PVS and \mathcal{L}_B is the language C. It is expected from a translator to:

- Never fail if the entry is the code of a valid program in the input language, \mathcal{L}_A . However the translator might declare being only able to translate a fragment of the language and restrict its input language. In our case, we expect input programs to be written using the syntax in Figure 12, to be parsed and typechecked using PVS without error and finally, we expect that TCCs generated by PVS can be proven.
- Generate a valid program in the target language \mathcal{L}_B .
- For all entry x , $P(x)$ and $Q(x)$ return the same result (correctness).

2.1 Translator's architecture

The translation from PVS [Owr+98] to C follow five main steps:

- Typechecking: The PVS typechecker [Owr+99b] perform a type analysis on the PVS code to associate a PVS type to each expression. This might generates some proof obligations (TCC). The user of the translator has to make sure that the PVS code can be correctly typechecked and that all TCC can be proven.
- Lexical and syntactic analysis: The PVS parser transforms PVS [Owr+99a] code into a CLOS internal representation.

In Figure 12, we describe the syntax of the subset of PVS we are currently able to translate to C. In Figure 13, we describe the Common Lisp Object System architecture used by PVS to represent them in Common Lisp. Some classes and some slots in the classes are voluntarily omitted. For a full description of PVS parser representation, refer to [SO03].

- Translation: The translator flattens PVS expressions to generates an intermediate language which heavily relies on the use of intermediate variables to allow a simpler static analysis. The syntax of this language is described Figure 16.
- Static analysis: The intermediate language is analyzed and stripped from some of its unnecessary copies and non destructive updates using flow analysis and an enriched type system. This analysis inspired from Shankar [Sha02] and Pavol Cerny's [ČS06] previous analysis of PVS is described with more detail in Section 4.

Typically, an expression e is translated into a tuple of four elements (t, e, i, d) , where t represents a C type used to describe the expression, e is a simple expression, i is a list of instructions to be executed prior to using e , the initialization of the expression. Finally d is a list of instructions to be executed when n isn't needed anymore, the destruction of e .

- Optimizations: Several simple analysis are performed, for instance to determine where to declare and free variables as well as the most adapted C types to use. The code generated from that step can be described by the syntax in Figure 17.
- Code generation: C code is generated (.c and .h files) and can be compiled using gcc and executed when linked with the garbage collector and the GMP library. The C syntax is described in [Hus04] and the GMP library reference can be found at <https://gmplib.org/manual/>.

2.2 Translating PVS type

2.2.1 PVS type system

A PVS theory needs to be typechecked using the emacs interface `M-x typecheck` or calling the Lisp function `(tc name-theory)`. This runs the PVS parser on the code and generates CLOS objects to represent it. Then, the PVS typechecker is run on this internal representation of the theory and tries to give a type to all expressions generating TCC when needed.

The (simplified) syntax for PVS types is described in Figure 14. PVS allow a base types such as boolean or numbers. Then more complex types can be defined such as tuples, record types or functions with range and values over other types. Types can also be restricted into subtypes using predicates. For example, integers are defined as a subtype of rationals with the predicate `integer_pred`. This way, when an expression is passed to a function ranging over integers, the typechecker generates the TCC "argument must verify the predicate `integer_pred`". That TCC must then be proven (it is often proven automatically for such simple TCCs).

The Figure 15 describes how PVS type system is represented in CLOS.

2.2.2 Translating rules

PVS types: boolean, number, number_field, real, rational, integer, $A \rightarrow B$, restricted types below(10) := $\{x : \text{int} | 0 \leq x < 10\}$) enum datatype

This requires a type analysis to decide on the type of a PVS expression. For example the PVS `int` type can be represented by the `int`, `unsigned long` or `mpz_t` C types. In that case, we study the range of the expression to decide which types are allowed to represent it. Then we take the context in which the expression appears to decide. For instance in

```
incr(x:below(10)):int = x+1
```

the `x` expression, result of the function `incr` can always be represented by an `int` or `unsigned long` in C but we choose here to represent it using a `mpz_t`.

Intermediate type system : C-type with a flag : mutable (meaning that the expression it describes only has one pointer pointing to it).

```
int a = 2;      a : int[mutable]
int* a = malloc( 10 * sizeof(int*) );
```

destructive addition:

```
d_add(*mpz_t res, mpz_t[mutable] a, long b) {
  mpz_add(a, a, b);
  (*res) = a;
}
```

Rq : `d_add` is given a mutable `mpz_t` , meaning that it can modify it and is responsible for freeing it. It is also responsible for allocating memory for the result. Here it uses the memory to assign `res`.

Use an intermediate language :

```
( expr, C-type[mutable] )
```

Conversions and copies create mutables types (at a cost) : `a[mutable]_from_b`

[Fil14]

C types:[Hus04]

```
// integer and floating point types
[unsigned] char, int, long, double
type* //arrays
char* // strings
struct types // structures with fields
```

```
enum types
short int, float, union, size_t // etc...
```

Listing 1: C types

Translation rules :

subrange(a, b)	<code>int</code> // if small enough <code>unsigned long</code> // if too big or needed for function call <code>mpz_t</code> // else
int	<code>mpz_t</code>
rat	<code>mpq_t</code>
[below(a) -> Type]	<code>(Ctype)*</code>
$T : \text{TYPE} = [\# x_i : t_i \#]$	<pre>struct CT { ... Ct_i x_i; ... }; // These types must be declared</pre>
[Range -> Domain]	C closure parameterized by the Domain return type.

Figure 1: Translation rules for PVS types

We can only translate a subset of all PVS types. What's missing ?

2.3 A few translation rules

We first define a function T to translate an expression e .

$$T(e) = (T^t(e) , T^n(e) , T^i(e) , T^d(e))$$

$$\begin{aligned}
 T(2) &= (\text{int}, "2", [], []) \\
 T(4294967296) &= (\text{mpz_t}, ?, \\
 &\quad [\text{mpz_init}(?); | \\
 &\quad \text{mpz_set_str}(?, "4294967296");], \\
 &\quad [\text{mpz_clear}(?);]) \\
 T(\text{lambda}(x:\text{below}(10)):x) &= (\text{int}^*, ?, \\
 &\quad [? = \text{malloc}(10 * \text{sizeof}(\text{int})); | \\
 &\quad \text{int } i; | \\
 &\quad \text{for}(i = 0; i < 10; i++) \\
 &\quad \quad ?[i] = i;] \\
 &\quad [\text{free}(?);])
 \end{aligned}$$

Figure 2: Translation examples: number expressions

It may occur that $T^n(e) = ?$. In that case, the symbol $?$ appearing in $T^i(e)$ and $T^d(e)$ needs to be replaced by a proper variable name.

We then define two other operators:

- R wich take an expression and a type and may add an extra conversion in the instructions to make sure its result has the expected type. Also the result of this function has a proper name.

- S which take an expression, a type and a name. It makes sure that the given variable (type + name) is set to a value representing the expression.

```
number-expr "2"
(C-int, "2", [], [])

number-expr "12315468453213"
(C-mpz, nil,
 [mpz_t ~a; | mpz_init("12315468453213"); ],
 [mpz_clear ~a;])

application "f(e1, e2)"
(C-mpz, nil,
 [ instr(e1) | instr(e2)
   | mpz(~a); | f(~a, e1, e2) ]
 [mpz_clear(~a);])
```

2.4 Translating PVS syntax

We can only translate a subset of PVS syntax. What's missing ?

2.5 Using a representation of the C language

Figure 17

3 Update expressions

It is a complicated problem to decide while compiling a functional language whether an update expression should be translated into a destructive or non destructive update in the target imperative language.

Update expressions are represented by PVS as `update-expr` objects.

$$E := t \text{ with } [e1 := e2]$$

Problem : t is an expression typed as a function. Therefore it might be represented in C as an array (if domain type is `below(n)`). We want to know if we can update t in place to obtain a C object representing E or if we have to make a copy of t .

We consider a few solutions to this problem.

3.1 Pointer counting

Several systems rely on a reference counting garbage collectors. This family of garbage collectors has many advantages [JL]. Along with its simplicity and the instantaneity of garbage identification, the one we are interested in is the possibility to determine when a local variable is the only pointer to a complex data structure. In that case, at the cost of a simple test, we can avoid copies and perform destructive updates.

The idea is to keep track of the number of pointers pointing to an array or a struct. If an array is referenced in several portions of the code (nested reference in other data structure, local variable in calling function, ...) then we must be able, using the pointer counter, to perform all updates non destructively to avoid inconsistency.

We implement a very simple "Reference Counting Garbage Collector" as described in [JL] and integrate it to the C code generated.

The GC consists on a hashtable of pointer counters that we maintain during the execution of the code. Each pointer to data allocated on the heap is a key in the hashtable to which we associate an int counter as value. We then make sure that all memory allocations in the code make a call to the GC to "declare" the new memory.

<code>T* a = malloc(10 * sizeof(int));</code>	<code>T* a = (T*) GC_malloc(10 * sizeof(int));</code> All memory allocation are handled by the GC to make sure every new reference on the heap is in the reference table and has a pointer counter associated to it.
<code>free(a);</code>	<code>GC_free(a);</code> This will decrement the reference counter on <code>a</code> and might free it if this counter is now 0.
<code>T* a = b;</code>	<code>T* a = (T*) GC(b);</code> The reference count on <code>b</code> is incremented to represent that the local variable <code>a</code> now also points to the structure <code>b</code> points to.
<code>t[0] = b;</code>	<code>GC_free(t[0]);</code> <code>t[0] = (T*) GC(b);</code> This time, we also make sure the reference counter of <code>t[0]</code> is decremented and <code>t[0]</code> has a chance to be freed if nothing else points to it.

This requires to build our own C garbage collector [3](#).

3.1.1 How to use it

The garbage collector must be used for every manipulation of pointers to memory allocated on the heap. This occurs typically when representing PVS arrays or data structure. These arrays are created in the code.

When `A` points to an array (or `struct`) we want to update destructively, we first check if the pointer counter on `A` is 1. If so, we can update in place because only the local variable `A` points to the array.

However, we need to be carefull.

```
g(A:Array) : int = f(A, A WITH [(0) := 3] )
```

should not be translated to

```
g(int* A) {
  A[0] = 3;
  return f(A, A);
}
```

for (at least) two reasons:

- The variable `A` is updated destructively but it is later used as a reference to the previous value of the array.
- `f` is given twice a pointer to the same data structure. Its reference counter should be incremented.

```

struct entry_s {
    void*   pointer;
    int     counter;
    struct entry_s *tl;
};
typedef struct entry_s* entry;

struct hashtable_s {
    int     size;
    entry*  table;
};
typedef struct hashtable_s* hashtable;

hashtable ht_create ( int size );
int        ht_hashfunc( hashtable hashtable, void* pointer );
entry      ht_newentry( void* pointer );

hashtable GC_hashtable;
void       GC_start();
void       GC_quit();
entry      GC_get_entry( void* pointer );
void       GC_add_entry( entry e);
void       GC_new( void* pointer );
void*      GC( void* pointer );
int        GC_count( void* pointer );
void*      GC_malloc( int length, int size );
int        GC_free(void* pointer);

```

Figure 3: Garbage collector C header file: GC.h

Instead we could flatten the expression 4.

But again, we are lucky here that *A* is the first argument of *f*. If the updated *A* were the first arguments, the update would have been done destructively. This is why the GC alone is not enough. We need an analysis of the C code to determine whether a variable is going to be used later in the code or not. cf [3.4 Analysis of the C code](#).

3.1.2 Pros and cons

The use of a garbage collector integrated in the C code seems like a good idea when translating a functional language to C. Using a pointer counting GC allows to

We need an analysis of the C code for two reasons:

- To `GC_free` variable as soon as they are not needed anymore. Otherwise copies that could be avoided are performed because an other (useless) pointer still points to the structure we're interested in.

```

int* B = GC( A );
update(B, 0, 1); // Can't be done destructively because A also points to
GC_free(A);      // the same data as B
f( GC(B) );      // f is given a variable with a reference counter of 2.
GC_free( B );    // It might not be able to perform some update destructively

```

Should be

```

int* B = A;
update(B, 0, 1); // Can be done destructively
f( B );          // f is given a variable with a reference counter of 1.

```

```

void main() {
    GC_start();

    int* A = GC_malloc(10, sizeof(int) ); // Pointer counter of A = 1
    int i;
    for(i = 0; i < 10; i++) // Initialisation of A
        A[i] = i;          // Here A = lambda(x):x
    int* B = g( GC(A) );    // We need A further, we make sure that g knows
    int* C = A;             // main still has a pointer to A
    printf("Pointers to C = %d", GC_count(C) ); // equal to 2
    GC_free(B); // Frees B
    GC_free(C); // Only decrement the counter of C
    GC_free(A); // Frees A (and C)
    GC_quit();
}

g(int* A) {
    int* arg1 = GC(A);      // A and arg1 now both point to the array
    int* arg2;
    if (GC(A) == 1)         // This is false
        arg2 = GC( A );
    else {                  // The update must be done non destructively
        arg2 = GC_malloc( 10, sizeof(int) );
        int i;
        for(i =0; i < 10; i++)
            arg2[i] = A[i];
    }
    arg2[0] = 3;
    GC_free(A);             // A is never used afterwards, we free it here
                           // (this requires an analysis of the C code)
    int* result = f(arg1, arg2); // A function is responsible for freeing its arguments
                           // (this is why we don't free arg1 and arg2)
    return result;
}

```

Figure 4: Example of the use of the GC

•

Every update require now tests and calls to hashtable functions. This is a small cost compared to the copying it may allow to avoid but no so small compared to a single in place update that could be decided by a code analysis.

Besides, the code gets much bigger since every update or copy requires the code to both destructive and non destructive operation and the if statement to decide which one to use.

Passing argument to function :

```

int* f(int* arg) {
    int* result;
    if ( GC_count(arg) == 1)
        result = GC( arg );
    else {
        result = GC_malloc(10, sizeof(int));
        int i;
        for(i = 0; i < 10; i++)
            result[i] = GC( arg[i] );
    }
    GC_free(arg);
    result[0] = 3;
    return result;
}

```

```
}
```

This add quite some code compared to the simple :

```
int* f(int* arg) {
  arg[0] = 3;
  return arg;
}
```

3.2 Using a different data structure

The Lisp code generated by PVS and used for example by the ground evaluator to compute PVS expressions represents PVS arrays with a more complex data structure than a simple array. It basically consists in an array and a replacement list. Every time an update on $(A, 1)$ is performed, the result is a pointer to the same array A and a replacement list with an extra term: $(A, (0:=0) :: 1)$. When the length of the list becomes too big, we create a new array A' by applying the replacement terms to a copy of A and we return (A', nil) .

We could represent C data structure with a similar C structure. For example :

```
struct r_list {
  int key;
  int value;
  r_list* tl;
};
struct array_int {
  int *data;
  r_list* replacement_list;
};
```

Each structure represent the array `data` with the modifications contained in the linked list `r_list_int`.

As in the previous solution, we have the following issues:

- This adds some extra code
- This adds some extra computation. We need runtime tests for updates, and access to an element not requires reading the whole replacement list.
- This relies on lot on the GC.

3.3 Flow analysis on the PVS code

An other optimization would be to perform a analysis on the PVS variables to make sure an update Pavol [ČS06] suggests three analysis...

3.4 Analysis of the C code

This solution consist in performing an analysis on the C code internal representation before generating the actual output C code.

We use flags and two different version of the translated functions to translate update expressions (or dangerous function calls) into a destructive update as often as possible.

We define three flags:

- **mutable** means that the variable is the only pointer to the structure or array it points to. For instance if we have `f(A:Arr):Arr = A WITH [(0) := 0]` then when `f` is called in

```
let A = lambda(x:int):x in let B = f(A) in B(0)
```

we know that `f` can update `A` in place. We call the following version of `f`.

```
int* f(int* A) {
    A[0] = 0;
    return A;
}
```

- **safe** means that an occurrence of a variable is the last occurrence of that variable in the code. We need this flag to avoid updating destructively variables that appears later in the code. In the previous example, if we encounter

```
let A = lambda(x:int):x in let B = f(A) in B(0) + A(0)
```

we know we can't update `A` destructively and we call instead a non-destructive version of `f`:

```
int* f(int* A) {
    int* res = malloc(...);
    for( i ...) res[i] = A[i];
    res[0] = 0;
    return res;
}
```

- **duplicated** means that this expression may find itself nested in the result of the current function. For instance the identity function, `id(A:Arr):Arr = A`, has its argument flagged **duplicated**. Therefore when `id` is called we know that the result contains a pointer to its argument.

```
...
int* A = malloc(...);
[ init A somehow ]
int* B = id(A);
\\ From now on B and A point to the same array
\\ For instance, A should probably not be modified in place
...
```

We want to ensure the following properties:

3.4.1 Algorithm

Each PVS function is translated into two different C functions:

- A "cautious" non destructive version whose arguments are never **mutable** and therefore never modifies the arguments in place, always making copies when necessary. This doesn't mean this function can't make destructive update. For instance locally created arrays (using `init_array`) will be flagged **mutable** and might be destructively updated, should the conditions be met.
- A destructive version which requires as many arguments as possible to be **mutable** and tries to do destructive updates as often as possible. This function only requires **mutable** arguments if it uses it destructively though.

In destructive versions of all functions : Flag all array arguments to "mutable". Then for each of these arguments : - If it never occurs destructively, then remove flag (function just read the arg) - If it occurs destructively, it can never occur at all AFTER. - Need to define the order of evaluation of expression (easy rules on simple expressions) - Need to be able to detect occurrences of a name-expr - Otherwise, unflag the arg

A variable `V` of type array is created in these cases:

- Only function declarations and variables with type struct or array can be flagged **mutable** .
- Only a single occurrence of a variable may be flagged **safe** .
- Only expressions and arguments can be flagged **duplicated** .
- The last and only the last occurrence of a variable is flagged **safe** .
- Arguments of a non destructive function are never flagged **mutable** .
- A function is flagged **mutable** iff its return variable is flagged **mutable** .
- A variable may be flagged bang if it is created with a `copy`, `init_array`, `init_record` or is the result of a call to a function flagged **mutable** .
It may not be flagged **mutable** if it is the result of a call to a function not flagged **mutable** .
- A call to a destructive function `f_d(a_i, b_j, c_k)` (where a_i are flagged **mutable** and b_j are flagged **duplicated** and c_k are not flagged) may only occurs if the following conditions on the arguments passed (A_i, B_j, C_k) are met:
 - All A_i are either calls to functions flagged **mutable** or variables flagged **mutable** and **safe** .
 - All B_j are either calls to functions or variables flagged **safe** or not flagged **mutable** .
 - If the function call is flagged **duplicated** , then all B_j are also flagged **duplicated** .
- If a variable is once flagged **duplicated** , then if it is an argument, this argument is also flagged **duplicated** .

Figure 5: Propeties of the flags

```

f(int* A, int* B) {           // A and B are both flagged duplicated
    if (A[0] == 0) {
        return B;
    } else {
        int* arg1 = copy(B); // arg1 is flagged mutable and duplicated
        arg1[0] = arg1[0] - 1;
        f(arg1, A); // Both these occurrences of arg1 and A are flagged safe
    }
}

f_d(int* A, int* B) { // A and B are both flagged mutable and duplicated
    if (A[0] == 0) {
        return B;
    } else {
        int* arg1 = B; // No need to copy since B is mutable
                        // and never occurs afterwards
        arg1[0] = arg1[0] - 1;
        f_d(arg1, A); // we can call f_d since the requirements are met:
                      // both arg1 and A are flagged mutable
    }
}

```

Figure 6: Example of the two different versions of a C function generated (stripped from GC instructions)

- Create the two versions of a function
- Flag all arguments **mutable** in destructive version
- Perform several passes and move flags to make sure the properties Figure 5 are verified.
- Modify the code if the flags allow it according to the rules defined in the Annex D.
- Redo the two previous steps until stabilization.

Figure 7: Algorithm

- $V = \lambda x. e(x) : V$ has bang type
- $\text{update}(V, T, \text{key}, \text{value}) : V$ has bang type because this is basically a copy and a destructive update.
- $f(V, \dots) = \dots : \text{type of } V \text{ depends on } f.$

In these case, it has always bang type. Or it can be set to an other referenced object.

- $V = T \rightarrow V$ (should have bang type iff T has !type too and never occurs afterwards). Happens in
- $V = T[i] \rightarrow$ depends on the target type of T .
- $V = T.\text{field} \rightarrow$ depends on the type of the field.

At first all updates are non destructive.

First pass : All array variables (actuals and local variables) found in the code are flagged. Local variables are flagged according to the previous rules and actuals are flagged **mutable** in destructive version and **not mutable** in non-destructive versions. In functions returning an array (or record type), the variable result is also flagged.

Other passes : Reading the code backwards, for every occurrence T of a variable flagged **mutable**:

- If it is found in a $V = T$ instruction, then we give the bang type to V and remove bang type from T so that previous occurrences of T won't assume the uniqueness of the reference. This adds a new variable to the set of bang variables, hence the need to make several passes.
- If it is used in a $V = \text{copy}(T)$ instruction, then we replace it with a $V = T$ instruction and do as previous.
- If it is found in an $\text{update}(V, T, i, e)$, then turn that into $V = T; \text{destr_update}(V, i, e)$.
- If it is a function call
- If we reach the declaration of a variable that is marked **mutable**, this means this variable is never read. In that case, we actually don't need it (unflag it I guess...).

At the end, when we have reached the transitive closure of this definition, if we reach the beginning of the function and some arguments are still bang, this means their bangness is never used, put the flag on that argument to **non mutable** and remove the instructions freeing that variable (reminder : mutable arguments of a functions are freed inside the function or are used in a mutable way and appear somewhere in the result (trapped in closures) or are freed in other function calls.

3.4.2 Algorithm

All variables have three flags: M (mutable), D (duplicated) and T (treated).

Init: All arguments of a destructive function are flagged ($M = \text{true}$, $D = \text{false}$, $T = \text{true}$). All arguments of a non destructive function are flagged ($M = \text{false}$, $D = \text{false}$, $T = \text{true}$). All other variables are flagged ($M = \text{false}$, $D = \text{false}$, $T = \text{false}$).

Rules: When M is changed, T is set to **true**. When T is **true**, the flag M can only be set to **false**. This prevent infinite change of the flag M . The flag D can only be set to **true**.

Initialization:

We initialize a set M of mutable variables to all array arguments of a function f . We also initialize a set F of variables to free to M since f has the responsibility to free all variables flagged as mutable arguments.

We read the code backwards. T_i refer to variables that are in the set M . S_i refer to variables that are not in the set M .

$S = T$	$M \leftarrow M \cup \{S\} - \{T\}$ $F \leftarrow F \cup \{S\} - \{T\}$
$S = \text{update}(S_2, \text{key}, \text{value})$	$M \leftarrow M \cup \{S\}$ $F \leftarrow F \cup \{S\}$
$S = \text{update}(T, \text{key}, \text{value})$	$M \leftarrow M \cup \{S\} - \{T\}$ $F \leftarrow F \cup \{S\}$
$S = g(T_i, S_i)$	If the arguments of g don't allow g to be called destructively: $M \leftarrow M - \{T_i\}$ $M \leftarrow M \cup \{S\}$ if return type of g is mutable $F \leftarrow F \cup \{S\}$
$S = g(T_i, S_i)$	Otherwise: $\rightarrow S = g_d(T_i, S_i)$ $M \leftarrow M \cup \{S\} - \{T\}$ $F \leftarrow F \cup \{S\}$
$S = S_2[i]$	$M \leftarrow M \cup \{S\} - \{T\}$ $F \leftarrow F \cup \{S\}$

All arguments of the function are flagged **mutable**

What is a destructive occurrence :

$$E := f(t \text{ with } [e1 := e2] , t(0))$$

order of eval : $e1$ and $e2$ (t can occur non destr) t (expression of an update : destr) $t(0)$ (occurrence of t (even non destr))

$f(x:\text{Arr}):\text{int} = g(h(t), t)$ is destructively translated to

```

1  int f_d(int* t) {    // t has type ! since this is destructive f
2    int* arg1 = h(t); // h can't be called destructively because
3                      // even though t is !, it appears later (line 4)
4    int* arg2 = t;    // t is ! and never appears later => arg2 is !
5    return g( arg1, arg2); // arg2 is ! but g can only be called
6  }                  // destructively if arg1 is

```

Listing 2: Example

if g has type $[\text{Array!} \rightarrow ?]$ then t can't be destructive

if g has type $[\text{Array} \rightarrow ?]$ then t can be destructive

First algorithm:

Need multiple passes as the flags disappear

3.5 Combination of solutions

We use the C code analysis to write some updates as destructive. However a few updates remain non destructive. For example:

If a function is called but requires its two argument to be **mutable** and only the first is **mutable** . Then the non-destructive version is called and the first argument gets copied even though it was **mutable** .

If we perform an update on $T[i]$, our analysis doesn't tell if $T[i]$ is **mutable** or **non-mutable** .

update(A, key, value)	A[key] = value;	A must be mutable
set(A, <i>expr</i>)	A = <i>expr</i> ;	
declare(A, <i>expr</i> (<i>i</i>))	<pre> A = malloc(1 * sizeof(T)); int i; for(i = 0; i < 1; i++) A[i] = i + 1; </pre>	
copy(A, B)	<pre> A = malloc(1 * sizeof(T)); int i; for(i = 0; i < 1; i++) A[i] = B[i]; </pre>	
init(A)	int* A;	
free(A)	free(A);	
base(str, (A, B, ...))	int aux = A[0] + B[1];	A and B are only read.
return	return result;	

Figure 8: C instructions

value(<i>cste</i>)	42	
variable(<i>type</i> , <i>name</i>)	name	
call(f, <i>exprs</i>)	f(<i>expr</i> ₁ , ..., <i>expr</i> _{<i>n</i>})	

Figure 9: C expressions

To prevent that, we also perform a GC check. An update is actually a test whether an object is **mutable** or not and the appropriate update.

4 Static analysis of the intermediate language

We describe here the static analysis of the intermediate language (which syntax is defined in Figure 16) implemented in the translator.

```

Expr      ::=  Number | Variable
              |  Variable [ Variable ]
              |  if ( Variable ) Expr  else Expr
              |  array( Variable )
              |  Variable [ ( Variable ) := Variable ]
              |  Variable [ ( Variable ) <- Variable ]
              |  lambda( Function , Number , Variables )
              |  Variable ( Variables )
              |  Function ( Variables )
              |  PrimOp ( Variables )
              |  set( Variable , Expr ); Expr

Variable   ::=  Id

Function    ::=  Id

PrimOp      ::=  + | - | * | / | %
              |  < | <= | > | >= | =
              |  not | and | or | iff

FunctionDecl ::=  Id ( Variable* ) = Expr

Program     ::=  FunctionDecl* Expr

```

Figure 10: Syntax of the intermediate language

We assume a few properties on valid programs:

1. The **set** instruction is a declaration and an assignment of a variable to a value at the same time. If an expression contains two **set** of the same variable, the second **set** will override the first definition.
2. The only free variables in the body of a function declaration are the arguments of that function.

We first define the semantics of the language using a small-steps operational semantics. Then we define a few operators on the language and exhibit some properties. Finally we describe an algorithm to replace non destructive updates with destructive updates under certain conditions and prove that there is a bisimulation between programs before and after applying this algorithm. This proves that the execution of the program is not disturbed by the replacements and thus the correctness of the algorithm.

4.1 Operational semantic

A *value* is either an integer $n \in \mathbb{N}$, a reference $r \in R$ representing an array (or pointer) or a function id $f \in F$. The metavariable v ranges over the set of all values: $V := \mathbb{N} \cup R \cup F$.

An *evaluation context* (sometimes simply called *context*) E is an expression with an occurrence of a hole \square . A context and is of one of the forms

1. \square
2. $\text{set}(x, \square); e$

3. `pop()`
4. $E_1[E_2]$, where E_1 and E_2 are evaluation contexts.

A *redex* is an expression of the following form

1. x
2. $X[y]$
3. `if (x) a else b`
4. `array(x)`
5. $X[(x) := y]$
6. $X[(x) <- y]$
7. `lambda(f, m, x1, ..., xm)`
8. $y(x_1, \dots, x_n)$
9. $f(x_1, \dots, x_n)$
10. $p(x_1, \dots, x_n)$
11. `set(x, v); e`
12. `pop(v)`

We define a *local environment*, s_i , as a function ranging over the set N of all variable names with values in V . The *stack state*, s , is a series of local environments: $s = (s_0, \dots, s_n)$. We call \square the empty function and if s_i is a local environment ranging over the variables U , we write $s_i \uplus (x \mapsto v)$ the function ranging over $U \cup \{x\}$ mapping x to v and y to $s_i(y)$ for $y \neq x$. For $s = (s_0, \dots, s_n)$ a stack state, we write $s \uplus (x \mapsto v) := (s_0 \uplus (x \mapsto v), s_1, \dots, s_n)$. We also define $s(x)$ as $s_i(x)$ where $\forall j < i, s_j(x)$ is not defined and to simplify notations, we call $s' :: s := (s', s_0, \dots, s_n)$ and even $s' :: S := (s' :: s, h)$.

The *heap state* function, h is mapping references r to arrays of values, V^* .

The *store* (or *state*) function, S , describing the state of the memory at a certain point in the execution is defined as the couple (s, h) . We define $S(x) := s(x)$ and $S(r) := h(r)$.

A program is list of function declarations followed by a closed expression. For each function with id f declared before the evaluation of the expression, we call f_i the arguments of this function (variables) and $[f]$ its body (expression). A function is associated not only an id but also a number when declared. This number is used in lambda terms to refer to a function using a value.

The meta-variable conventions are that x and y range over variables, X ranges over variables typed as arrays n ranges over numbers, p ranges over primitive function symbols, f ranges over defined function symbols, a, b and e range over expressions.

A *reduction* transforms a pair consisting of a redex and a store. The reductions corresponding to the redexes above are

1. $\langle x, S \rangle \longrightarrow \langle S(x), S \rangle$
2. $\langle x[y], S \rangle \longrightarrow \langle h(s(x))(s(y)), S \rangle$
3. $\langle \text{if } (x) \text{ a else } b, S \rangle \longrightarrow \begin{cases} \langle \text{pop}(a), \square :: S \rangle & \text{if } s(x) = 0 \\ \langle \text{pop}(b), \square :: S \rangle & \text{otherwise} \end{cases}$
4. $\langle \text{array}(x), S \rangle \longrightarrow \langle r, (s, h \uplus (r \mapsto (0)_{0 \leq i < s(x)})) \rangle$ where r is a fresh pointer.
5. $\langle X[(x) := y], S \rangle \longrightarrow \langle r, (s, h') \rangle$ where r is a fresh pointer and

$$h' = h \uplus (r \mapsto h(s(X)) \uplus (s(x) \mapsto s(y)))$$

6. $\langle X[(x) <- y], S \rangle \longrightarrow \langle X, (s, h') \rangle$ where

$$h' = h \uplus (s(X) \mapsto h(s(X)) \uplus (s(x) \mapsto s(y)))$$

7. $\langle \text{lambda}(f, m, x_1, \dots, x_m), S \rangle \longrightarrow \langle r, (s, h') \rangle$ where r is a fresh pointer and

$$h' = h \uplus (r \mapsto (f, m, s(x_1), \dots, s(x_m)))$$

8. $\langle y(x_1, \dots, x_n), S \rangle \longrightarrow \langle \text{pop}([f]), s' :: S \rangle$ where $E = h(s(y))$, $f = E(0)$, $m = E(1)$,

$$s' : \left| \begin{array}{ll} \{f_1, \dots, f_{m+n}\} & \rightarrow V \\ f_i & \mapsto E(i+1) \quad \text{for } i \leq m \\ f_{m+i} & \mapsto s(x_i) \quad \text{for } i \leq n \end{array} \right.$$

9. $\langle f(x_1, \dots, x_n), S \rangle \longrightarrow \langle \text{pop}([f]), (\biguplus_{1 \leq i \leq n} f_i \mapsto s(x_i)) :: S \rangle$

10. $\langle p(x, y), S \rangle \longrightarrow \langle p(s(x), s(y)), S \rangle$ for binary operators.

11. $\langle p(x), S \rangle \longrightarrow \langle p(s(x)), S \rangle$ for the unary operator (**not**).

12. $\langle \text{set}(x, v); e, S \rangle \longrightarrow \langle \text{pop}(e), (x \mapsto v) :: S \rangle$

13. $\langle \text{pop}(v), ((s_0, \dots, s_n), h) \rangle \longrightarrow \langle v, ((s_1, \dots, s_n), h) \rangle$

An evaluation step operates on a pair $\langle e, S \rangle$ consisting of a closed expression and a store, and is represented as $\langle e, S \rangle \longrightarrow \langle e', S' \rangle$. If e can be decomposed as a $E[a]$ for an evaluation context E and a redex a , then a step $\langle E[a], S \rangle \longrightarrow \langle E[a'], S' \rangle$ holds if $\langle a, s \rangle \longrightarrow \langle a', s' \rangle$. This is represented by the following rule.

$$\frac{\langle a, s \rangle \longrightarrow \langle a', s' \rangle}{\langle E[a], s \rangle \longrightarrow \langle E[a'], s' \rangle}$$

One of the greatest advantage of using evaluation contexts is that we define the semantics of this language using only this one small-step rule.

The reflexive-transitive closure of \longrightarrow is represented $\xrightarrow{*}$. The computation of a program is defined as the evaluation of its expression $\langle e, S_0 \rangle$ on an empty store: $S_0 := (([]), [])$. If $\langle e, S_0 \rangle \xrightarrow{*} \langle e', S' \rangle$ then we can prove that $e' \in V$ and the result of the computation is then defined as $eval_h(e')$ where $eval_h$ is defined as follow:

$$eval_h : \left| \begin{array}{ll} V & \longrightarrow E \\ n & \mapsto n \in \mathbb{N} \\ r & \mapsto (eval_h(u_i))_{0 \leq i \leq n} \text{ with } (u_i)_{0 \leq i < n} := h(r) \end{array} \right.$$

Theorem 1. For all $\langle e, S_0 \rangle \xrightarrow{*} \langle v, (s, h) \rangle$, h is defined on $R \cap (\{v\} \cup Im(s) \cup Im(h))$. All references stored in the stack or in the heap or reduced from an expression are defined in the heap state.

Proof. We proof this theorem by induction on the structure of the code.

1. Base cases: All redexes generating a fresh pointer (4, 5 and 7) modify the heap state to define it on that new pointer.
2. Induction step:
 - References defined in the store are never undefined. If $\langle e, S \rangle \longrightarrow \langle e', S' \rangle$ and h is defined on r then h' is also defined on r . This is easily proven since no redex remove a definition in the heap.
 - Whenever a value in the heap or in the store is modified, it is replaced with functions (redex 7), integers (redexes 4 and 7), values from the heap (redexes 2, 5, 6 and 8), values from the stack (redexes 1, 5, 6, 7, 8, 9, 10 and 11) or values reduced from an expression (redexes 12 and 13).

□

4.2 Sets of variables

We define the free variables, Fv , of an expression as the set of all variables that occur in that expression. For this study, we are only interested in variables that may refer to an array.

We also define the *output* variables, $Ov(e)$, of an expression. This can be understood in three ways:

- This is the set of all variables that may have their value "trapped" into the expression e .
- It corresponds to all the variables which content might get modified if e gets modified in place.
- It is the set of all variables which content the pointer corresponding to the evaluation of e may point to.

Finally, we define the *active* variables, Av , of an expression. In a similar way, this can be understood as:

- The set of all variables that may contain a reference to the expression e .
- It corresponds to all the variables that could be accessed from the value returned by the expression.
- It is the set of all variables which evaluation may be a pointer pointing to e .

If X is a variable, X^* refers to all the references X may point to. It is obvious that if $X \in Av(e)$, then $X^* \in Av(e)$ as well. In that case, we voluntarily omit to mention X^* in the definition of these set below.

Expression	Ov	Av	Fv
n	\emptyset	\emptyset	\emptyset
X	$\{X\}$	$\{X\}$	$\{X\}$
$X[x]$	$\{X^*\}$	$\{X^*\}$	$\{X\}$
if (x) a else b	$Ov(a) \cup Ov(b)$	$Av(a) \cup Av(b)$	$Fv(a) \cup Fv(b)$
array(x)	\emptyset	\emptyset	\emptyset
$X[(x) := y]$	\emptyset	$\{X^*, y\}$	$\{X, y\}$
$X[(x) <- y]$	$\{X\}$	$\{X, y\}$	$\{X, y\}$
lambda($f, m,$ x_1, \dots, x_m)	\emptyset	$\{x_1, \dots, x_m\}$	$\{x_1, \dots, x_m\}$
$y(x_1, \dots, x_n)$	$\{x_1, x_1^*, \dots, x_n, x_n^*\}$	$\{x_1, \dots, x_n\}$	$\{y, x_1, \dots, x_n\}$
$f(x_1, \dots, x_n)$	$\{x_i f_i \in Ov([f])\}$ $\cup \{x_i^* f_i^* \in Ov([f])\}$	$\{x_i f_i \in Av([f])\}$ $\cup \{x_i^* f_i^* \in Av([f])\}$	$\{x_1, \dots, x_n\}$
$p(x_1, \dots, x_n)$	\emptyset	\emptyset	\emptyset
set(x, a); e	$Ov(e) \cup Av(a) - \{x, x^*\}$ if $x^* \in Ov(e)$ $Ov(e) \cup Ov(a) - \{x\}$ if $x \in Ov(e)$ $Ov(e)$ otherwise	$Av(e) \cup Av(a) - \{x, x^*\}$ if x or $x^* \in Av(e)$ $Av(e)$ otherwise	$Fv(a) \cup Fv(e) - \{x\}$

Theorem 2. For all expression e , $Ov(e) \subset Av(e) \subset Fv(e)$.

Proof. Simple induction proof on the expression form. □

4.3 Analysis of update context

The advantage of using contexts is to be able to place critical expressions like updates into a context where the evaluation order is well defined and we can identify expression evaluated before and after the reduction of a critical redex like a function call or a destructive update. These two being the only redexes that can modify the heap store in place.

We introduce *update contexts* as an expression with a single occurrence of a hole:

1. $\{\}$
2. $\text{set}(x, U); e$
3. $\text{set}(x, a); U$
4. $\text{if } (x) U \text{ else } b$
5. $\text{if } (x) a \text{ else } U$

To define which update can be made destructively, we now build a reference graph in a context U . We define the pointer analysis of U as $PA(U)$ where $PA(U)(X)$ is the set of all variables X may point to in the context U .

This will allow us to define the variables free in this context as well as the set of variable that can be modified by a destructive update.

$$\begin{aligned}
 PA(\{\}) &: \begin{cases} X \mapsto \{X, X^*\} \\ X^* \mapsto \{X^*\} \end{cases} \\
 PA(\text{set}(x, U); e) &:= PA(U) \\
 PA(\text{set}(x, a); U)(X) &:= PA(U)(X) \cup \begin{cases} \{x, x^*\} & \text{if } X \in Ov(a). \\ \{x^*\} & \text{if } X \in Av(a). \\ Av(a) & \text{if } x^* \in PA(U)(X). \\ \emptyset & \text{otherwise.} \end{cases} \\
 PA(\text{if } (x) U \text{ else } b) &:= PA(U) \\
 PA(\text{if } (x) a \text{ else } U) &:= PA(U)
 \end{aligned}$$

For X a variable and U a context, we also consider the set of critical variables $Cv(U)(X)$ smaller set such that :

$$\forall Y, \quad PA(U)(Y) \subset Cv(U)(X) \text{ or } PA(U)(Y) \cap Cv(U)(X) = \emptyset$$

This set contain all variables that may point to X and all variable that may point to these variables and so on.

The variables live in a context are the variables that could be evaluated after the hole the in the context and the variable that these variables may point to and so on.

When we consider turning a non destructive update $X[(x) := y]$ in a context U into a destructive update, we want to make sure that critical variables are not live U .

$$Cv(U)(X) \cap Lv(U) = \emptyset$$

The set $Lv(U)$ of the variables live in a context U is defined as the smallest set containing $Lv^0(U)$

$$\begin{aligned}
 Lv^0(\{\}) &:= \emptyset \\
 Lv^0(\text{set}(x, U); e) &:= Fv(e) \cup Lv^0(U) \\
 Lv^0(\text{set}(x, a); U)(X) &:= Lv^0(U) \\
 Lv^0(\text{if } (x) U \text{ else } b) &:= Lv^0(U) \\
 Lv^0(\text{if } (x) a \text{ else } U) &:= Lv^0(U)
 \end{aligned}$$

an close under

$$PA(U)(Lv(U)) \subset Lv(U)$$

4.3.1 Analysis

We consider here a function f which declaration body e doesn't contain any destructive update before the analysis. The reason is that a safe, naive translation from PVS to this language would only perform non destructive updates.

Intuitively, if a function $f(f_1, \dots, f_n) = e$ contains a non destructive update in a context $e = U\{X[(x) := y]\}$. That update can be turned into a destructive update if none of the variables that may be aliased to X are live in the context.

$$Cv(U)(X) \cap Lv(U) = \emptyset$$

This way all this variables that may point to X are never used after the update. Since they are the only variables whose evaluation is modified by making the update destructive, we can say that this is a safe transformation.

The problem is that some of these variables possibly pointing to X might be included in the set of arguments of f : $\{f_1, \dots, f_n\}$. And we can't assume anything about these variables since we don't have any information regarding the context in which the function f is called.

We define the bang analysis $BA(f)$ as the set of all variables that are involved in a destructive update or are destructive arguments in a call to a destructive function:

$$BA(f) := \left(\bigcup_{e=U\{X[(x) <- y]\}} Cv(U)(X) \right) \cup \left(\bigcup_{e=U\{g(x_1, \dots, x_n)\}} Cv(U)(\{x_i | g_i \in BA(g)\}) \right)$$

Basically $BA(f)$ is the set of all critical variables in the evaluation of a function call to f .

We define two versions of all function declared. A non destructive version f with body e_{nd} and a destructive version f^d with body e_d . Both new definitions may only differ from the original body e of f in some very specific substitutions: non destructive updates in e may be destructive in e_{nd} and e_d and function calls to a function g may become function calls to g^d with the same arguments in e_d and e_{nd} .

These functions use two different strategies:

- The second consists in allowing destructive updates of variables that could be aliased arguments. We however keep track of these arguments and only call this function when we are sure the arguments are safe in the context of the call. This yields the definition e_d of f^d verifying the following properties:

$$\circ \forall U, X, x, y, \text{ such that } e = U\{X[(x) := y]\},$$

$$e_d = U\{X[(x) <- y]\} \iff Cv(U)(X) \cap Lv(U) = \emptyset$$

$$\circ \forall U, g^d, (x_i)_{1 \leq i \leq n}, \text{ such that } e_d = U\{g^d(x_1, \dots, x_n)\},$$

$$Cv(U)(\{x_i | g_i^d \in BA(g^d)\}) \cap Lv(U) = \emptyset$$

$$BA(g^d) \ni g_i^d \neq g_j^d \in BA(g^d) \implies PA(U)(x_i) \cap PA(U)(x_j) = \emptyset$$

- The first consists in forbidding the use of destructive updates which argument may be pointing to by arguments. This is equivalent to saying that all arguments of non destructive functions are live in all contexts (or just in the empty context). This yields the new definition e_{nd} of f verifying the following properties:

$$\circ \forall U, X, x, y, \text{ such that } e = U\{X[(x) := y]\}$$

$$e_{nd} = U\{X[(x) <- y]\} \iff Cv(U)(X) \cap (Lv(U) \cup \{f_1, \dots, f_n\}) = \emptyset$$

◦ $\forall U, g^d, (x_i)_{1 \leq i \leq n}$, such that $e_{nd} = U\{g^d(x_1, \dots, x_n)\}$

$$Cv(U)(\{x_i | g_i^d \in BA(g^d)\}) \cap (Lv(U) \cup \{f_1, \dots, f_n\}) = \emptyset$$

$$BA(g^d) \ni g_i^d \neq g_j^d \in BA(g^d) \implies PA(U)(x_i) \cap PA(U)(x_j) = \emptyset$$

Theorem 3. *For all non destructive version of a function f , $BA(f) = \emptyset$.*

4.3.2 Proof of bisimulation

We define the accessible cells, from a variable X given a store S as the set of all references (heap store entries) that can be accessed from the variable X .

$$Ac(S)(X) := R \cap S^\infty(X)$$

We extend that definition to expressions and updates

$$\begin{aligned} Ac(S)(e) &:= Ac(S)(Av(e)) \\ Ac(S)(U) &:= Ac(S)(Lv(U)) \end{aligned}$$

It correspond to the only references already created in the heap store that can still be accessed in the context or expression.

The proof of correctness relies on three main invariants:

1. The results of function calls to f and f_d have the same evaluation. This means they can't be told apart by looking only at the value of the result. However they differ in their use of the arguments.
2. The only references which value may be modified during the evaluation of the call of $f(x_i)$ with a store S is $\bigcup_{f_i \in BA(f)} Ac(S)(x_i)$.
3. The union above is always a disjoint union.

These properties are obviously true before the analysis since $f = f^d$, $BA(f) = \emptyset$ for all function and no reference ever have its value modified.

The destructive and non destructive versions of a function f are built from the original definition e by replacing function calls (updates can be considered as a function call). To prove the correctness of this algorithm we need to prove that the three invariants described above are preserved between programs before and after a single replacement.

We consider the replacement of g to g^d in a function f^d of body $e = U\{g(x_i, y_j, z_k)\}$ with x_i variables corresponding to the critical arguments $g_i \in BA(g)$, y_j corresponding to arguments that are active in the body of e and z_k arguments that are not arrays or are not in either of the previous sets.

We have the following hypothesis

$$(H1) \quad g_i^d \in BA(g^d) \implies Cv(U)(x_i) \cap Lv(U) = \emptyset \quad (\text{Analysis})$$

$$(H2) \quad BA(g^d) \ni g_i^d \neq g_j^d \in BA(g^d) \implies Cv(U)(x_i) \cap Cv(U)(x_j) = \emptyset \quad (\text{Analysis})$$

$$(H3) \quad \text{The only references which value may be modified during the evaluation of the call with a store } S \text{ is the union } \bigcup Ac(S)(x_i).$$

$$(H4) \quad \text{The evaluation of the call with a store } S \text{ return the same result than if } g \text{ was called.}$$

$$(H5) \quad f \text{ is always going to be called in an evaluation context with a store such that the sets } Ac(S)(f_i) \text{ are disjoint.}$$

Proposition 4. *For all store S with which this function call is evaluated, the sets $Ac(S)(x_i)$ are disjoint.*

Proof. (H2) and (H5). □

This analysis would have worked as well for non destructive versions of functions and for the replacement of an update rather than a function call.

4.3.3 Analysis in term of references

Theorem 5. *If $\langle e_1, (s_1, h_1) \rangle \longrightarrow \langle e_2, (s_2, h_2) \rangle$ then $\langle e_1, (s_1, h'_1) \rangle \longrightarrow \langle e_2, (s_2, h'_2) \rangle$ where $h'_i = h_i|_{Lc(S_1)(e_1)}$.*

Proof. It can easily be verified for every redex. When $a = U\{b\}$, we have $Lc(a) = Lc(U) \cup Lc(b)$ and the inductive step is proved . □

4.4 Difference with the previous analysis of PVS

Difference with Shankar [Sha02] and Pavol's [ČS06] previous analysis of PVS:

- The non destructive version of a function is also optimized and might have some destructive updates or function calls.
- The destructive function does not have all its function call destructive.

4.5 Draft

The translator from PVS to that intermediate language guarantees the following properties

- All arguments passed to a function are *different* variables.
- Before the analysis, no destructive updates are used.

For the need of the analysis, we enrich the state function with a reference counter $c : V \rightarrow \mathbb{N}$.

Theorem 6. *If $\langle a, S \rangle$ is **mutable**, and $\langle a, S \rangle \longrightarrow \langle a', S' \rangle$ then $\langle a', S' \rangle$ is **mutable**.*

A function f is called **mutable** when

$$Ov([f]) \cap \{f_1, \dots, f_n\} = \emptyset$$

This basically means that the output variables of any call to that function is the empty set.

5 Conclusion

5.1 Difficulties and successes

This project was a great challenge and an opportunity for me to conduct my own research on a subject I chose. The development of the working translator was also

5.1.1 Working with new languages and tools

To be able to translate PVS, I had to fully understand not only the syntax and semantics of the PVS language but also the structure of the PVS API written in Common Lisp. This means I also had to learn Common Lisp which I decided then to use to write the translator mostly because it made the integration of the native PVS parser and typechecker easier. Finally I had to discover the C language which I only had a basic knowledge of.

5.1.2 Integrating the GMP library

In PVS (and in other languages such as Common Lisp or Python), the `integer` type represent the whole set \mathbb{Z} of all relative numbers (and `rational` also describe \mathbb{Q}). To implement that in C, we need more than the finite types `int`, `long`, etc.

The translator uses the GMP library which introduces the types `mpz_t` and `mpq_t`. These types are pointers (technically arrays) to structures and they had to be used with caution (allocation, freeing, ...).

For example (Figure 11), a function returning a `mpz_t` should actually take a first `mpz_t` argument and set it to the return value. Its return type being `void`.

```
norm(x:int, y:int):int = x*x + y*y
void norm(mpz_t result, mpz_t x, mpz_t y) {
    mpz_t aux1;
    mpz_init(aux1);
    mpz_mul(aux1, x, x);
    mpz_clear(x);
    mpz_t aux2;
    mpz_init(aux2);
    mpz_mul(aux2, y, y);
    mpz_clear(y);
    mpz_add(result, aux1, aux2);
    mpz_clear(aux1);
    mpz_clear(aux2);
}
```

Figure 11: Example of the GMP library use

5.2 What's left to be done ?

One of my biggest regret was not having the time to finish the translator and fully implement closures. Some work is already done in that direction. We use a C structure to represent a closure:

```
struct r_list_int {
    int (*body)(void* env, void* args); // body is a function pointer
    void* env; // env contains the environment variables
    void* args; // args will contain the arguments
};
```

5.3 My stay at SRI

Besides the conception and implementation of the PVS to C compiler, my stay at SRI International was rich in interesting events.

The first weeks of my stay were the occasion to discover PVS and Coq as I started working on a translator Coq to PVS. With Robin, we also wrote as an exercise a basic linear algebra library.

I discovered Lisp the hard way while discovering the middle- and back-end of the PVS API. Among other exercises, I decided to write a Common Lisp parser to help me understand the huge architecture of the PVS API code (classes definitions, inheritances and organization, function dependances, ...)

I also have had the chance to attend to the many interesting seminars SRI hosted every week. The "Crazy Ideas" seminar hosted every other week was a ...

The SRI also organized a Summer School to which we were allowed to attend and which was very interesting.

Shankar never hesitated to include us in many project

I've been included in the HACMS project which was very interesting. With other: Correcting translator PVS to SMT-LIB

Draft

Discovering PVS : Translating Coq proofs to PVS PVS library for basic linear algebra

Robin project, HACMS Contest week-end 14-15 June Summer School Parsing Lisp code - generate HTML architecture file Correcting translator PVS to SMT-LIB [ČS06]

References

- [ČS06] Pavol Černý and N. Shankar. "Static Analyses for Guarded Optimizations of High Level Languages". SRI Internship Report. MA thesis. ENS Paris, 2006 (cit. on pp. 5, 12, 27, 30).
- [Fil14] Jean-Christophe Filliâtre. *Langages de programmation et compilation*. MPRI. <https://www.lri.fr/~filliatr/ens/compil/>. 2013-14 (cit. on p. 6).
- [Hus04] Eric Huss. "The C Library Reference Guide". In: *Webmonkeys: A Special Interest* (2004) (cit. on pp. 5 sq.).
- [JL] Richard Jones and Rafael Lins. "Garbage Collection: Algorithms for Automatic Dynamic Memory Management, 1996". In: *John Wiley & Sons Ltd., England* () (cit. on p. 8).
- [Owr+98] S. Owre et al. *User Guide for the PVS Specification and Verification System*. Three volumes: Language, System, and Prover Reference Manuals. Computer Science Laboratory, SRI International. Menlo Park, CA, Sept. 1998 (cit. on p. 5).
- [Owr+99a] S. Owre et al. *PVS Language Reference*. Computer Science Laboratory, SRI International. Menlo Park, CA, Sept. 1999 (cit. on p. 5).
- [Owr+99b] S. Owre et al. *PVS System Guide*. Computer Science Laboratory, SRI International. Menlo Park, CA, Sept. 1999 (cit. on p. 5).
- [Sha02] Natarajan Shankar. "Static analysis for safe destructive updates in a functional language". In: *Logic Based Program Synthesis and Transformation*. Springer, 2002, pp. 1–24 (cit. on pp. 5, 27).
- [SO03] N. Shankar and S. Owre. *PVS API Reference*. Computer Science Laboratory, SRI International. Menlo Park, CA, Sept. 2003 (cit. on p. 5).

A PVS Syntax and CLOS representation

<i>Expr</i>	::=	<i>Number</i> <i>Name</i> <i>Expr</i> (<i>Expr</i> ⁺) <i>Expr</i> <i>Binop</i> <i>Expr</i> <i>Unaryop</i> <i>Expr</i> <i>Expr</i> ‘ { <i>Id</i> <i>Number</i> } (<i>Expr</i> ⁺) (# <i>Assignment</i> ⁺ , #) <i>IfExpr</i> LET <i>LetBinding</i> ⁺ IN <i>Expr</i> <i>Expr</i> WHERE <i>LetBinding</i> ⁺ <i>Expr</i> WITH [<i>Assignment</i> ⁺ ,]
<i>Number</i>	::=	<i>Digit</i> ⁺
<i>Id</i>	::=	<i>Letter</i> <i>IdChar</i> ⁺
<i>IdChar</i>	::=	<i>Letter</i> <i>Digit</i>
<i>Letter</i>	::=	A ... Z
<i>Digit</i>	::=	0 ... 9
<i>IfExpr</i>	::=	IF <i>Expr</i> THEN <i>Expr</i> { ELIF <i>Expr</i> THEN <i>Expr</i> } * ELSE <i>Expr</i> ENDIF
<i>Name</i>	::=	true false integer? floor ceiling rem ndiv even? odd? cons car cdr cons? null null? restrict length member nth append reverse
<i>Binop</i>	::=	= \= OR \ / AND & /\ IMPLIES => WHEN IFF <=> + - * / < <= > >=
<i>Unaryop</i>	::=	NOT -
<i>Assignment</i>	::=	<i>AssignArg</i> ⁺ { := -> } <i>Expr</i>
<i>AssignArg</i>	::=	(<i>Expr</i> ⁺) ‘ <i>Id</i> ‘ <i>Number</i>
<i>LetBinding</i>	::=	{ <i>LetBind</i> (<i>LetBind</i> ⁺) } = <i>Expr</i>
<i>LetBind</i>	::=	<i>Id</i> [: <i>TypeExpr</i>]

Figure 12: Syntax of the PVS subset of the translator

<code>expr</code> \subset syntax [<i>abstract class</i>] <i>type</i> the type of the expression
<code>name</code> \subset syntax [<i>mixin class</i>] <i>id</i> the identifier <i>actuals</i> a list of actual parameters <i>resolutions</i> singleton This is a mixin for names, i.e., name-exprs , type-names , etc.
<code>name-expr</code> \subset name expr [<i>class</i>]
<code>number-expr</code> \subset expr [<i>class</i>]
<code>tuple-expr</code> \subset expr [<i>class</i>] <i>exprs</i> a list of expressions
<code>application</code> \subset expr [<i>class</i>] <i>operator</i> an expr <i>argument</i> an expr (maybe a tuple-expr)
<code>field-application</code> \subset expr [<i>class</i>] <i>id</i> identifier <i>argument</i> the argument A field application is the internal representation for record extraction, e.g., r'a
<code>lambda-expr</code> \subset binding-expr [<i>class</i>] This is the subclass of binding-expr used for LAMBDA expressions.
<code>if-expr</code> \subset application [<i>class</i>]
<code>record-expr</code> \subset expr [<i>class</i>] <i>assignments</i> a list of assignments
<code>update-expr</code> \subset expr [<i>class</i>] <i>expression.</i> an expr <i>assignments</i> a list of assignments An update expression of the form e WITH [x := 1, y := 2] , maps to an update-expr instance, where the expression is e , and the assignments slot is set to the list of generated assignment instances.
<code>assignment</code> \subset syntax [<i>class</i>] <i>arguments.</i> the list of arguments <i>expression</i> the value expression Assignments occur in both record-exprs and update-exprs . The arguments form is a list of lists. For example, given the assignment 'a(x, y)'1 := 0 , the arguments are ((a) (x y) (1)) and the expression is 0.

Figure 13: (Partial) CLOS representation of PVS syntax

B PVS type system and CLOS representation

$TypeExpr ::= Name$
 $\quad \quad \quad EnumerationType$
 $\quad \quad \quad Subtype$
 $\quad \quad \quad TypeApplication$
 $\quad \quad \quad FunctionType$
 $\quad \quad \quad TupleType$
 $\quad \quad \quad CotupleType$
 $\quad \quad \quad RecordType$
 $EnumerationType ::= \{ IdOps \}$
 $Subtype ::= \{ SetBindings \mid Expr \}$
 $\quad \quad \quad (Expr)$
 $TypeApplication ::= Name Arguments$
 $FunctionType ::= [FUNCTION \mid ARRAY$
 $\quad \quad \quad [- [IdOp :] TypeExpr^+ \rightarrow TypeExpr]$
 $TupleType ::= [- [IdOp :] TypeExpr^+]$
 $CotupleType ::= [- [IdOp :] TypeExpr^+]$
 $RecordType ::= [\# FieldDecls^+ \#]$
 $FieldDecls ::= Ids : TypeExpr$

Figure 14: Fragment of the PVS type system

type-expr \subset syntax	[abstract class]
.....	
type-name \subset type-expr name	[class]
adt	
.....	
subtype \subset type-expr	[class]
supertype	
predicate	
.....	
funtype \subset type-expr	[class]
domain	
range.	
.....	
tupletype \subset type-expr	[class]
types	
.....	
recordtype \subset type-expr	[class]
fields	
.....	

Figure 15: (Partial) CLOS representation of PVS types

C Intermediate languages

Expr ::= *Number* | *Variable*
| *Variable* [*Variable*]
| if (*Variable*) *Expr* else *Expr*
| array(*Variable*)
| *Variable* [(*Variable*) := *Variable*]
| *Variable* [(*Variable*) <- *Variable*]
| lambda(*Function* , *Number* , *Variables*)
| *Variable* (*Variables*)
| *Function* (*Variables*)
| *PrimOp* (*Variables*)
| set(*Variable* , *Expr*); *Expr*

Variable ::= *Id*

Function ::= *Id*

PrimOp ::= + | - | * | / | %
| < | <= | > | >= | =
| not | and | or | iff

FunctionDecl ::= *Id* (*Variable*^{*} ,) = *Expr*

Program ::= *FunctionDecl*^{*} *Expr*

Figure 16: Syntax of the intermediate language

<i>Expr</i>	::=	<i>Number</i> <i>String</i> <i>Function</i> (<i>Exprs</i>) <i>Pointer</i>
<i>Pointer</i>	::=	<i>Variable</i> <i>Variable</i> . <i>Id</i> <i>Variable</i> [<i>Expr</i>]
<i>Variable</i>	::=	<i>Id</i>
<i>Type</i>	::=	int unsigned long int mpz_t mpq_t array(<i>Type</i> , <i>Number</i>) struct(<i>Id</i>)
<i>Instruction</i>	::=	decl(<i>Variable</i>) free(<i>Variable</i>) if (<i>Expr</i>) { <i>Instructions</i> } else { <i>Instructions</i> } <i>MPZFunction</i> (<i>Variable</i> [, <i>Exprs</i>]) init_array(<i>Variable</i> , <i>Instructions</i> , <i>Expr</i>) init_record(<i>Variable</i> , <i>Instructions</i> , <i>Exprs</i>) set(<i>Pointer</i> , <i>Expr</i>)
<i>Function</i>	::=	+ * ... <i>Id</i>
<i>MPZFunction</i>	::=	mpz_set_str mpz_add ...
<i>FunctionDecl</i>	::=	<i>Id</i> (<i>Variables</i>) : <i>Type</i> = <i>Instructions</i> [return <i>Expr</i>] ;
<i>StructDecl</i>	::=	struct <i>Id</i> : <i>Types</i>
<i>Types</i>	::=	[<i>Type</i> [, <i>Types</i>]]
<i>Exprs</i>	::=	[<i>Expr</i> [, <i>Exprs</i>]]
<i>Variables</i>	::=	[<i>Variable</i> [, <i>Variables</i>]]
<i>Instructions</i>	::=	[<i>Instruction</i> ; [<i>Instructions</i>]]

Figure 17: Syntax of the representation language. Every *Expr* is typed.

D Rules

	A safe	A not safe
A mutable	Replace every occurrence of the variable B by the variable A	<pre> B = GC_malloc(...); for(i ...) { B[i] = A[i]; } </pre>
A non-mutable	<pre> if (GC_count(A) == 1) { B = A; } else { B = GC_malloc(...); for(i ...) { B[i] = A[i]; } } </pre>	<pre> B = GC_malloc(...); for(i ...) { B[i] = A[i]; } </pre>

Figure 18: Rules for `copy(B, A)`

	A safe	A not safe
A mutable	Replace every occurrence of the variable B by the variable A	<pre> B = GC_malloc(...); for(i ...) { B[i] = A[i] } </pre>
A non-mutable	Replace every occurrence of the variable B by the variable A	<pre> B = GC(A); </pre> <p>If B is flagged duplicated then A must be too.</p>

Figure 19: Rules for `set(B, A)`

	A safe	A not safe
A mutable	<code>f_d(A)</code>	<code>f(A)</code>
A not mutable	<code>f(A)</code>	<code>f(A)</code>

Figure 20: Rules for `f(A)` with A flagged **mutable** in the destructive version

	A safe	A not safe
A mutable	<code>f(A)</code>	<code>copy(B, A)</code> <code>f(B)</code>
A not mutable	<code>f(A)</code>	<code>f(A)</code>

Figure 21: Rules for `f(A)` with A flagged **deduplicated**

E Examples

PVS code	Intermediate language code	C code generated
<code>f(A:Arr):Arr = A</code>	<pre>f: ((A, int*)) -> int* set(result, A) return(result)</pre>	<pre>int* f(int* A) { return A; }</pre>
<pre>f(A:Arr):Arr = let B = A in B</pre>	<pre>f: ((A, int*)) -> int* set(B, A) set(result, B) return(result)</pre>	<pre>int* f(int* A) { return A; }</pre>
<pre>f(A:Arr):Cint = let B = A in A(0) + B(0)</pre>	<pre>f: ((A, int*)) -> int set(B, A) set(result, +(A(0), B(0))) return(result)</pre>	<pre>int* f(int* A) { int* B = (int*) GC(A); int result = A[0] + B[0]; GC_free(B); GC_free(A); return result; }</pre>
<pre>f(A:Arr):Arr = let B = A in A WITH [(0) := B(0)]</pre>	<pre>f: ((A, int*)) -> int set(B, A) set(L, 0) set(R, B(0)) update(result, A, L, R) return(result)</pre>	<pre>int* f_d(int* A) { int* B = GC_malloc(...); for(i ...) B[i] = A[i]; int L = 0; int R = B[0]; int* result = GC(A); result[L] = R; GC_free(B); GC_free(A); return result; }</pre>

Figure 22: Examples of setting variables

PVS code	Intermediate language code	C code generated
<pre>f(A:Arr):Arr = A WITH [(0) := 0]</pre>	<pre>f: ((A, int*)) -> int* set(L, 0) set(R, 0) copy(result, A) update(result, L, R) return(result)</pre>	<pre>int* f(int* A) { int L = 0, R = 0; int* result; if(GC_count(A) == 1) { result = GC(A); } else { result = GC_alloc(...); for(i ...) result[i] = A[i]; } result[L] = R; GC_free(A); return result; } int* f_d(int* A) { int L = 0, R = 0; A[L] = R; return A; }</pre>
<pre>f(A:Arr):Arr = let B = A WITH[(0):=0] in A WITH[(0) := B(0)]</pre>	<pre>f: ((A, int*)) -> int* set(L1, 0) set(R1, 0) copy(B, A) update(B, L1, R1) set(L2, 0) set(R2, B(0)) copy(result, A) update(result, L1, R1) return(result)</pre>	<pre>int* f(int* A) { int R1 = 0, L1 = 0; B = GC_alloc(...); for(i ...) B[i] = A[i]; B[L1] = R1; int R2 = 0, L2 = B[0]; result = GC_alloc(...); for(i ...) result[i] = A[i]; result[L2] = R2; GC_free(A); GC_free(B); return result; } int* f_d(int* A) { int R1 = 0, L1 = 0; B = GC_alloc(...); for(i ...) B[i] = A[i]; B[L1] = R1; int R2 = 0, L2 = B[0]; A[L2] = R2; GC_free(B); return A; }</pre>

Figure 23: Examples of copying variables