Inter-procedural Analysis of Concolic Execution

A Thesis Submitted

in Partial Fulfilment of the Requirements

for the Degree of

MASTER OF TECHNOLOGY

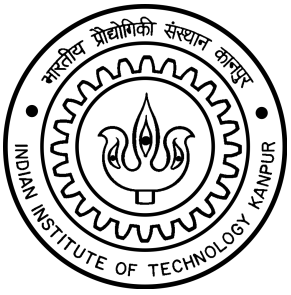
by

Ashwini Kshitij

supervised by

Dr. Subhajit Roy

Dr. Amey Karkare



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY KANPUR

June 2015

CERTIFICATE

This is to certify that the work contained in this thesis entitled ***"Inter-procedural analysis on Concolic Execution"***, by **Ashwini Kshitij (Roll No. 10327165)**, has been carried out under my supervision and this work has not been submitted elsewhere for a degree.

Dr. Subhajit Roy
Assistant Professor,
Department of CSE,
IIT Kanpur.

Dr. Amey Karkare
Assistant Professor,
Department of CSE,
IIT Kanpur.

ABSTRACT

Interprocedural analysis is the cornerstone of determining precise program behavioral information. Using this technique we can avoid making overly conversative assumptions about the effects of procedures and the state at call sites. It aims at gathering informations across multiple procedures.

In this thesis, to extend the concept of concolic execution to interprocedural calls we modify the function call site. Just before the function call we set up the environment for a procedure call. It enables it to guide the concolic execution of input parameters from caller to callee and back. After the callee executes the calling environment is restored with a modified symbolic state. A tool called CIL. has been used for any instrumentation purpose.

Modifying the symbolic execution engine to collect the interprocedural analysis information can have widepread applications in software verification and testing. The aim was to improve the coverage of test suites that are automatically generated by intraprocedural concolic testing tools. Experimenting this approach on a intraprocedural concolic testing tool showed improved coverage on some programs. On TCAS benchmark it increased the MCDC coverage by 30% with reduced number of test cases.

Acknowledgements

I acknowledge, with gratitude, my debt of thanks to Professor Subhajit Roy for his advise and encouragement and to Professor Amey Karkare for his aid and foresight. They presented me with the opportunity to tackle interesting problems in field of Software Testing. Their patient but firm guidance was critical to successful completion of my research.

I appreciate the support of my friends and wingmates who always provided me with the confidence and courage to tackle even the most challenging problems. Without their help and councel, the completion this work would have been immeasurably more difficult.

I also want to express my sincere gratitude to ***BRNS*** for encouraging our research work. Their encouragement motivated me to follow through this project.

-Ashwini Kshitij

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Symbolic execution has been a recipient of significant attention during the past few years. It is now considered an effective technique in generation of high coverage test suites. The idea has been invented more three decades ago[**?**] but it was after significant improvements the potential of the idea was harnessed. One such important improvement was symbolic execution alongside keeping track of concrete values (concolic execution)[**?**]. The main advantage of this technique is that whenever constraint solving complications (like timeouts) occur during classical symbolic execution, it is alleviated using the concrete values.

All practical programs involve procedure calls. The symbolic execution is relatively simple if there are no function calls involved (*intraprocedural analysis*). A function call tranfers the control from the caller to the callee. That function may very well modify the symbolic state of the variables. If we transition the symbolic state correctly through the function call, the symbolic execution will run more accurately with the modified symbolic state and collect the precise alteration made to symbolic state within the procedure.

1.1 Motivation

Intraprocedural Analysis is performed on one procedure at a time. It is simple and conservative. But almost all real world programs involve procedural calls. This poses a problem for techniques such as symbolic execution since they dont know how the procedure is going to modify the symbolic state. This brings us to the need of extending this analysis techniques to programs with functions and procedures. Although tools like DART[**?**] and

CREST[?] already have support for interprocedural analysis, we have developed our own methodology to do that.

Listing 1.1: Motivational example

```
1    #include <stdio.h>
2
3    int dbl(int x) {
4       return 2*x;
5    }
6
7    int main(void) {
8       int a;
9       scanf("%d",&a);
10      if (dbl(a) + 3 == 9) {
11         return 0;
12      } else {
13         return 1;
14      }
15    }
```

Listing 1.1 is a modified program in `/test/function.c` from CREST[?] tool directory. In this program **a** is treated as an input for which test cases are to be generated. When intraprocedural concolic testing is done on `main()` procedure of this program only fifty percent MCDC[?] coverage is generated. This implies only one branch is covered in this program. The first test case is a randomly generated value of **a** that covers one branch irrespective of its value. The next test case is generated using condition `dbl(a) + 3 == 9` in its path constraint, which is equivalent to condition `2a + 3 == 9`. But intraprocedural analyser fails to infer that `dbl(a) = 2a` because it cannot analyse the operations performed by `dbl()`. It substitutes the symbolic value of `dbl(a)` by its concrete value. Now the path constraint is `constant + 3 == 9`. Without any variable in this equation, no test input is generated in the next iteration. It fails to cover the other branch to produce 100% coverage. Thus a need arises to implement support for interprocedural analysis in concolic testing.

1.2 Problem Statement

This thesis aims to provide a method that can capture the transformation of the symbolic state of the inputs when the program goes through a function call. Our implementation uses concolic execution engine[**?**] for the purpose. But the engine cannot handle a procedure call. At the call site, the symbolic value is set to concrete value instead of modifying it as per the operations performed on it by the callee function. Merging our approach with this concolic engine will increase the precision of test case generation.

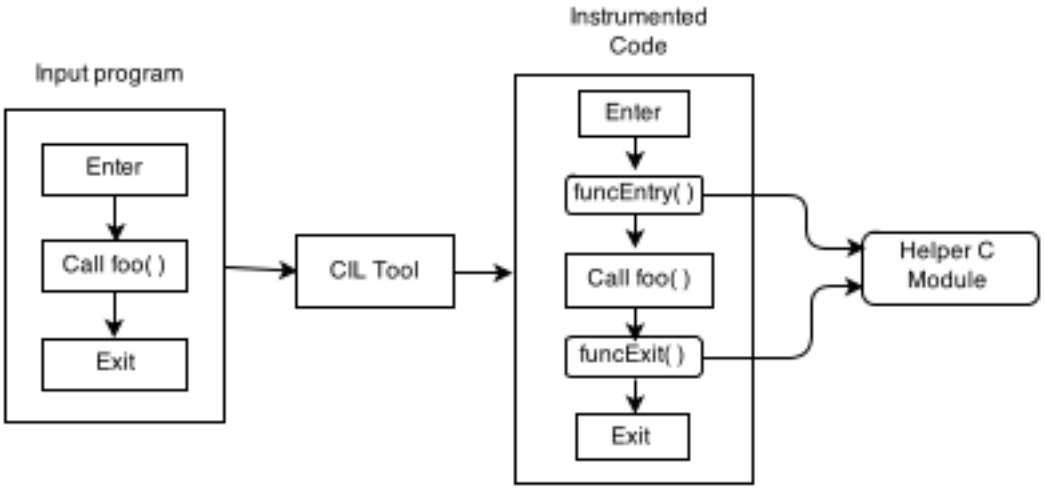


Figure 1.1: Basic Module

Figure 1.1 shown above gives a high-level idea of what our implementation does. We are given a C program with a concolic execution engine integrated with it. Problem arises with execution when it comes across a procedure call instruction. Concrete return value of function is used instead of symbolic return value as the program does not have the ability to follow through symbolic execution from caller to callee and then back to the caller.

We aim at solving this problem by instrumenting the call site. For such purposes in our approach we have used a tool called CIL[**?**]. We first set up an environment before commencing the callee procedure execution (in this case *foo()*). The operations for such tasks is carried out by *funcEntry()* function which is instrumented just before *foo()* is called. *funcEntry()* handles the mapping of symbolic and concolic values of actual parameters to formal parameters, the creation of variable stack (later explained in *Chapter 3*).

Similarly we instrument *funcExit()* function just after *foo()* call. The task of *funcExit()* is to map the symbolic state of "return variable" to appropriate variable at the call site (if

any) and clear the symbolic values that were generated during call and execution of *foo()*. It also manages the variable stack appropriately. The implementation of *funcEntry()* and *funcExit()* is done in C language and is defined in included C modules.

This approach ensures that symbolic execution is carried through correctly to and from the callee. The modified state of symbolic variables is in accordance with operations of concerned callee function (in our case *foo()*). This improves the accuracy and correctness of the test suite generation in our tool.

1.3 Contributions

In this thesis we have accomplished the following:

- Designed an approach to apply inter-procedural analysis to concolic execution.
- Developed unit tests for verifying the features and functionality of test case generation tools.
- Dynamic support to handle recursive procedure calls.

1.4 Organisation of thesis

A brief summary of the contents of other chapters of this thesis are as follows.

Chapter 2 deals with the concepts which are necessary to understand the thesis. Concolic execution is introduced and its advantages over classical symbolic execution and random testing. An overview of approaches in interprocedural analysis like Function Inlining and Call String have also been discussed in this chapter. We have also briefly mentioned the related work that has been done in this field.

Chapter 3 discusses our implementation in a detailed manner. Some terminology has been introduced to explain techniques employed in our work. But mostly the chapter pertains to the relevant algorithms designed to accomplish the aim of this thesis and how it should improve the scalability and accuracy of concolic IP analysis.

Chapter 4 contains the experimental results and their analysis after we run our modified tool on a set of standard benchmark programs. Analysis in done in terms of coverage improvement and runtime of the tool.

Chapter 5 sheds light on future work and modifications that can be done to improve the accuracy and performance of the approach discussed in this thesis.

Chapter 2

Background

2.1 Interprocedural Data-flow Analysis

This is a technique which is broadly defined as gathering of information across multiple procedures (typically over the entire program). Procedure calls pose a barrier to program analysis. Its aim is to avoid making conservative assumpions about the effect of procedures and the state at call sites.

Interprocedural analysis[**?**] is more demanding and challenging than intraprocedural analysis. Here we need to take into account the call-return and parameter passing mechanisms, local variable of the function and function recursion (can be unbounded). In some cases the called procedure is known only in run-time like with function pointers or with virtual functions.

Application : Interprocedural analysis (IPA) enables the compiler to optimize the code across different files (whole-program analysis), and can result in significant performance improvements of the program by removing spurious data dependencies. Integrating IPA with intraprocedural concolic execution will help compute the precise behavior of function calls on the symbolic state of the variables.

2.1.1 Function Inlining

One of the approaches is *function inlining*[**?**] which is the simplest and most widely used approach for accurate interprocedural symbolic execution. We simply use the copy of procedure's Control Flow Graph (CFG) at each call site. This leads to function being

re-analyzed at every call site. This is can be avoided using *function summaries*. They are implemented by merging all states at the function exit after computing an intraprocedural path constraint in terms of function input.

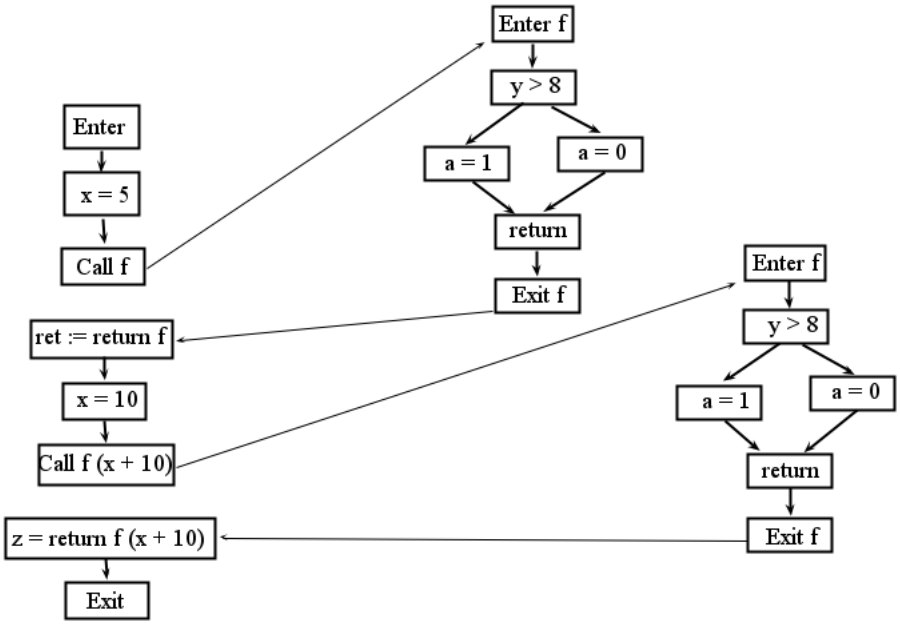


Figure 2.1: Function Inlined

This technique does not require function calling overhead. It also saves the overhead for manipulating function stack. In increases locality of reference by utilizing instruction cache.

But there are significant *drawbacks* with this approach. Firstly, the performance overhead will increase if we increase the size of the code that is to be inlined. The caller function may not fit on the cache causing high cache miss rate. Similarly if there are too many function calls involved, inlining maybe expensive since it will cause an exponential increase in the size of the CFG.

Secondly it is also going to create problems in the recursive procedures. Same can be said more generally for any scenario where there are cycles in the call graph. So basically function inlining has scalability issues pertaining to code size. Lastly, procedure inlining is only possible if the target of the call is known. Hence it will not be possible if call is via a pointer or is "virtual".

2.1.2 Call String Approach

In another IP analysis we observe the CFG's of all the procedures. In this technique,

- When we encouter a function call we interpret it as a goto from the call intruction to the first instruction of the procedure.
- Interpreting every return statement like a goto to the instruction following each call site that invoked that procedure (*Non-Deterministic*).

Using this method non-deterministically will allow non feasible paths in our analysis which will cause loss of accuracy. In a better approach called "Call String"[?] we keep track of the where we came from, that is, the context of the call and where to return. To do this we maintain a "string" that simulates a call stack. A feasible path is a control flow path that is generated in accordance with the stack regime. A perfect solution is keep record of the whole stack. This concept is easy and intuitive according to which every return jumps to the instruction that's immediately after the call site which corresponds to that particular function call. By adding the context of the call to the information in the state we can overcome the problem of passing through invalid paths.

```
int main(){
    int x ;
    c1: x = foo(2);
    ret:[x → 4]
    c2: x = foo(3);
    ret:[x → 6]
    return 0;
}
```

```
int foo(int p){
    c1:[p → 2]
    c1:[p → 3]
    return 2*p;
    c1:[p → 2, $$ → 4]
    c1:[p → 3, $$ → 6]
}
```

Figure 2.2: Calling Context

Every procedure records their state information when they are invoked. Let us consider the program in *Figure 2.2*. Here the labels *c1* and *c2* are saved in the state information of *foo()*. This is so that *p* gets different values every time function *foo()* is called. Another reason to save the labels is because they help in figuring out the respective return values of call *c1* ($\$\$ \rightarrow 4$) and *c2*($\$\$ \rightarrow 6$). Here $$ marks the return values.

This ensures context sensitivity of this approach. Call String method records the call history and information collected is propogated back to the correct point. Call String at any program point is the sequence of unfinished procedure calls (in the call stack) that

leads to that point. Implementing this algorithm is efficient for small strings. Problem arises when call string generated is large since we can keep track of only limited number of strings. This poses a limitation on the depth of function calls.

2.2 Concolic Testing

One popular approach for automated software testing is *random testing*[**?**] which involves subjecting the program to be tested to a stream of random data. It is fast but it can find only basic bugs like program crashes, code assertions or memory leaks. It is not always possible to employ this technique specially when working with binaries since it is very difficult to figure out the expected inputs. Random Testing fails to cover corner cases.

A more directed approach can be used by merging symbolic execution with concrete execution (*concolic execution*[**?**]) and then using a an SMT-solver (like Z3[**?**]) to generate test inputs. This method is called *Concolic testing*. It is much more versatile than symbolic execution. To understand why let us assume that our costraint solver cannot handle non-linear constraint. Then a path constraint with some *non-linear* function involved cannot be solved and symbolic execution will be stuck. Similar problem arises when we encounter a closed third party library function say *increment()*. Symbolic execution alogorithm is unable to determine how to modify the symbolic state according to the behavior of that function. For such cases the test cases cannot be generated by classical symbolic execution[**?**]. Concolic Tesing addresses these limitations and resolves situations like above by replacing the symbolic values by their concrete values so that the resulting constraint can be solved by the constraint solvers.

The algorithm involves initialting the inputs with randomly generated values. The program is executed and during the execution, on every conditional branch statement program collects symbolic path constraints on inputs. Symbolic constriants are essentially a set of logical constriants on input data. New program path is directed and executed by *negating/flipping* the last condition of the path constraint. This is done until all the feasible program paths are explored. To get the intuition, let us consider the program in *Figure 2.2* for concolic testing.

```
1 #include <stdio.h>
2
3 int main(){
4
5   int x,y;
6   scanf("%d",&x);
7   scanf("%d",&y);
8
9   int z = 3*x;
10
11  if(y == z){
12
13    if(y == x + 10){
14        assert(0);
15        /*error*/
16    }
17
18  }
19  return 0;
20 }
```

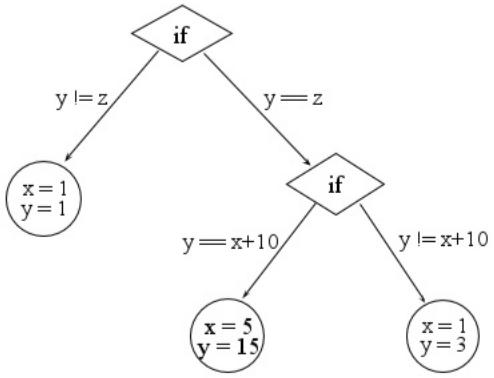


Figure 2.3: Concolic Testing

The program input variables are assigned random concrete values, say $x = 1$ and $y = 1$. The program runs and line 11 is executed. Path condition generated is :

$$\neg(y == 3x) \tag{2.1}$$

Now we negate the last condition in path constraint (PC has only one condition in this case) to execute an alternate path. So the new path constraint generated is :

$$y == 3x \tag{2.2}$$

which is examined by a SMT solver to generate the new input data. Out of many different possible values for equation *2.2* solver picks say $x = 1$ and $y = 3$ which should explore a different path than before. The program runs on these new inputs and generates a new path condition :

$$(y == 3x) \wedge \neg(y == x + 10) \tag{2.3}$$

Like previously we negate the last condition of the path constraint to generate condition for new path exploration. So we pass on the following constraint to solver to generate new inputs

$$(y == 3x) \wedge (y == x + 10) \tag{2.4}$$

Let the new inputs generated be $x = 5$ and $y = 15$ we will lead to the execution of line 13 and consequently line 14 which will hit the error or any other unexpected behavior. Since there are no more paths left to explore, the algorithm will terminate generating a set of input data for complete path coverage of this program.

2.3 Related work

2.3.1 Summary-based Analysis

New techniques have been invented that compute procedure summaries for performing an interprocedural analysis of programs. In summary-based context-sensitive analysis we create "summary"[**?**] which is succinct description of the observable behavior of each procedure. The purpose of this approach is to prevent renanalysing the behavior of same procedure when there are invoked at each call site.

The representation of every procedure has a single entry point. The analysis is divided into following two phases:

- In fist phase we summarize the effects of a procedure and a transfer function is computed in a bottom-up manner.
- In second phase we propogate the caller information to compute callee result in a top-down manner.

2.3.2 DART

DART is an abbreviation for "*Directed Automated Random Testing*"[**?**] which is a tool for automated test case generation developed by *Patrice Godefroid, Nils Klarlund* and *Koushik Sen*. It utilizes the concept of concolic execution and is comprised of the following techniques :

- *Automated* extraction of program interface from source code.
- *Random testing* the program interface by generating a test driver.
- Dynamic generation of test cases to *direct* alternate program execution path.

2.3.3 CUTE

CUTE stands for "A Concolic Unit Testing Engine for C"[**?**] which also addresses automatic test case generation with memory graphs as inputs. This tool is developed by *Koushik Sen*, *Darko Marinov* and *Gul Agha*. It is similar to DART, thereby employs concolic testing technique. It resolves some of the limitations of DART and aims at testing real-world examples of C code.

It provides a method for representing and solving approximate pointer contraints to generate test inputs. The symbolic model being used and the theorm solver both are more powerful and are built to be efficient in this system. As opposed to what DART does, CUTE does not automatically extract program interface but lets user decide relation among functions and their preconditions. The work also shows exactly how it made approximations and trade off between speed vs. correctness and scenarios where CUTE will not work correctly.

Chapter 3

Methodology

The algorithm that has been used to implement the interprocedural analysis in the concolic execution engine has been described in detail in this chapter.

3.1 Definitions

We are given a program **P** which has a concolic execution engine **C** integrated with it. Our aim is to combine interprocedural analysis with **C** so that the test case generation can also take into account the effect of function calls. To implement this algorithm we will instrument the code of P with our auxiliary code which will not alter the outcome of the program **P**. Let

- ***f*** be the callee function used in call site.
- *funcEntry()* be the instrumented method that initiates concolic environment for function call.
- *funcExit()* be the instrumented method that handles cleanup of symbolic structure and return from a function call.
- ***S*** denote Variable Stack explained in *Section 3.3.3*.

3.2 Static Analyser

We have used a tool called **CIL**[**?**] (**C** **I**ntermediate **L**anguage) to perform static analysis of program P and source-to-source transformations on it. We do this by traversing the

AST (Abstract Syntax Tree) which is the in-memory data-structure which represents the parsed program P.

3.2.1 Handle Expressions

This module accomplishes two tasks. It modifies the call site by simplifying the arguments passed to the calle function. It also simplifies the return expression in the callee function. At the call site the arguments of the function are examined one by one. If an argument is not a variable or literal then it must be an expression. In case of expression, we store the argument in a temporary local variable and pass that new variable as the parameter instead. Similar approach applies to return values. If the return statement consists of a variable, it is left unchanged otherwise it is transformed.

Listing 3.1: Expressions simplified

```
int func(int x,int y){
{
  if (y < = 1 )
     return x;
  else
     return x * func(x-1,y);
}
```

```
int func(int x,int y){
   int tmp;
   int cil_tmp4, cil_tmp4, cil_tmp4;
   if (y < = 1 )
   {
       cil_tmp4 = x;
       return cil_tmp4;
    } else {
       cil_tmp5 = x - 1;
       tmp = func(cil_tmp5, y);
       cil_tmp6 = x * tmp;
       return cil_tmp6;
    }
}
```

We can see in *Listing 3.1* the call sites and return statements have been transformed such that in further stages we only have to deal with variables and literals. We will not have to deal with complicated expressions. To create new local temporary variable we use CIL API. It ensures to generate unique variable names to avoid any ambiguity. This phase will have no effect on function calls with no arguments or functions which don't return any value (void).

3.2.2 Variable Renaming

The concolic testing engine used has symbolic table structure that is globally defined. The entries in the table are manipulated using variable names of program P as the key. In intraprocedural symbolic execution duplicate variable names are not an issue.

But when dealing with multiple procedures all of which are capable of executing symbolically, two different procedures may very well have variables with the same name. When these variables are used as the key to manipulate the values in symbol table, it leads to undefined behavior. Therefore we need to make the variables of all the procedures unique. This is done by merging the variable name with the name of the procedure is whose scope the variable belongs. *Figure 3.1* demonstrates the renaming of the variables to resolve the conflict of symbol table keys.

```
int mult(int x, int y){
        int z;
        z = x * y;
        return z;
}
```

```
int mult(int mult_x , int mult_y )
{
  int mult_z ;

  {
  mult_z = mult_x * mult_y;
  return (mult_z);
}
}
```

Figure 3.1: Variable Renaming

This method will work for non-recursive procedure calls. To handle recursion we have developed an extended version of approach which will be discussed in *section 3.4*.

3.2.3 Extend Concolic Execution Engine

The current engine uses approaches such as handling set (or assignment) instructions symbolically and unrolling of loops to carry out the symbolic execution. But these aproaches are only applied on the procedure that the user demands to be tested. The remaining procedures are unchanged.

Now on function call we want the symbolic execution to be able to continue in the callee *f*. Hence we need to instrument all the procedures of program P such that it enables them to be executed symbolically. So instead of applying the loop unrolling and symbolic assignment only on user input procedure, they are applied on every procedure.

3.2.4 Call Site Transformation

Without interprocedural call handling whenever the concolic execution encountered a function call, it used the concrete value of the function result. There interprocedural symbolic analysis could not be done. This problem is addressed by setting up a function call environment and function return environment, that will facilitate the transfer of symbolic state from caller to callee and back. The pseudo-code of the algorithm is mentioned ahead.

Algorithm 1 Transform Call Site

Input : Program P
Output : P with modified call sites

```
 1: procedure TRANSFORM
 2:     for each procedure f ∈ P do
 3:         instList ←instructionList(f)
 4:         newList ← []
 5:         for each instruction C ∈ P do
 6:             if C is a call instruction then
 7:                 argsList ← getArgumentsAsString(C.fname)
 8:                 localsList ← getLocalsAsString(C.fname)
 9:                 funcEntryInst ← makeCallInstruction(C.fname)
10:                 funcExitInst ← makeCallInstruction( )
11:                 newList ← newList::[funcEntryInst]::[C]::[funcExitInst]
12:             else
13:                 newList ← newList::[C]
14:             end if
15:             instList ← newList
16:         end for
17:     end for
18: end procedure
```

Algorithm 1 Description : The input to this algorithm is program P. Procedures in program P are inspected for call sites where we need to do transformations such that when control flows from caller to callee it has the symbolic information needed to execute symbolically. (At line 7,8 and 9 `C.fname` is the name of the function called in the call instruction *C*. At line 11 and 13 Operator `::` denotes concatenation of two lists.)
If c is a call site which invokes a function f then parameters that are passed to the function f have certain symbolic and concrete values associated with them. Before f starts executing, its actual parameters are analysed and their concolic values are mapped to its formal parameters.

The function *getArgumentsAsString* in *line 7* has two tasks. Firstly, it extracts the information about actual parameters of f like if its a literal or variable (its name). Secondly, it analyses the formal parameters of f's prototype and gathers information about their name and type. Then this information is merged into a string, one parameter at a time.

This string is passed as a paramter to *funcEntry()* along with another string containing the names of local variables (created by *getLocalsAsString()* in *line 8*). The function instrumented after call site c is *funcExit()* that does the task of cleaning up the intermediate data-structures created by *funcEntry()* and were necessary to execute f symbolically.

```
1 int main(){
2         int a,b=3,c;
3         scanf("%d",&a);
4         a = fact(a,4);
5         if(a == 20){
6                 c = 1;
7         }
8         else{
9                 c = -1;
10        }
11        printf("%d\n",c);
12        return 0;
13 }
```

Figure 3.2: sample program

In the example shown in *Figure 3.2* there is a call instruction at *line 4* invoking function *fact* with two paramters. The information about both actual and formal paramters in analysed and sent to *funcEntry()* as string argument at *line 12* in *Figure 3.3*. The second argument to *funcEntry()* is names of local variables of function *fact*. Using these arguments, environment for *fact* to execute symbolically is set up by *funcEntry()*. Similarly, *funcExit()* at *line 15 in Figure 3.3* handles the task of cleanup of this environment after *fact* has finished executing.

At *line 16* in *Figure 3.3* we also have function *add_entryToSTable* which is used to map the returned concolic values of function *fact* to appropriate variable, which is in our case *a*. This will be discussed in *Section 3.3.4*.

```
1 int main(void)
2 {
3   int a ;
4   int b ;
5   int c ;
6
7   {
8   b = 3;
9   scanf((char const   * __restrict  )"%d", & a);
10
11  //---------Transformed call site-------------------------
12  funcEntry("(int,x,variable,a) (int,y,constant,4)",
13            "tmp", "fact");
14  a = fact(a, 4);
15  funcExit();
16  add_entryToSTable("a", ret_SymValue, ret_ConValue, & a, 1);
17  //-------------------------------------------------------
18
19  if (a == 20) {
20    c = 1;
21  } else {
22    c = -1;
23  }
24  printf((char const   * __restrict  )"%d\n", c);
25  return (0);
26 }
```

Figure 3.3: Transformed Call Site

3.3 Set up Environment For Procedure Call

This section involves the runtime analysis of the arguments that are passed to *funcEntry()*. We will explain their utilization to create entries in global symbol table, to enable callee function f to execute symbolically.

3.3.1 Organising Argument Details

Using the information of actual and formals parameters passed to the *funcEntry()*, we construct a data structure *Argument* that has the following attributes:

- *funcName* : name of the function to which the argument is passed.
- *vname* : name of the associated formal parameter.
- *type* : 1 for int and 2 for real.
- *apname* : if actual argument is a variable, then its name.
- *val* : if actual argument is a literal, then its value.

Note that actual parameter can either be a variable or a literal, implying that only one among *apname* or *val* can have a valid value. The other is going to be *null* so we have to use them accordingly. We have passed this parameter information along with local

variable names of *f* to *funcEntry()* as discussed in *Section 3.2.4*. Using string tokenizing and parsing within *funcEntry()*, we get a list of *Arguments* type structures.
This step is demonstrated in *Figure 3.4* in reference to program shown in *Figure 3.3*.

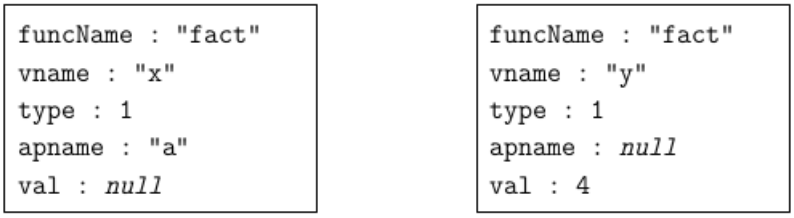


Figure 3.4: Two arguments at *line 12* in *figure 3.3*

3.3.2 Populate Symbol Table

After we created a systematic model of arguments, we need to populate the global symbol table accordingly. Besides the arguments, we also need to handle local variable in *f*. For locals, we create empty entries in symbol table. The algorithm used in this implementation is described below

Algorithm 2 Populate Symbol Table

Input : Argument list A, locals list L, symbol table S
Output : Modified symbol table S' with entries for A and L

```
1: procedure POPULATESTABLE(A, L, S)
2:     for each argument a ∈ A do
3:         if a.apname is a literal then
4:             sym ← Constant
5:             con ← a.val                          ▷ a.val is the value of a
6:         else
7:             sym ← findSymbolicValue(a.apname)    ▷ a.apname is actual parameter
8:             con ← findConcreteValue(a.apname)
9:         end if
10:        addEntryToSTable(a.vname, sym, con)
11:    end for
12:    for each local variable l ∈ L do
13:        createEmptyEntryInSTable(l.name)
14:    end for
15: end procedure
```

Algorithm 2 description : This algorithm maps the symbolic and concrete values of actuals parameters at call site to the formal parameters of the function *f*. This is done before *f* starts executing. For local variables, initially empty entries are created in the symbol table. Functions *findSymbolicValue* and *findConcreteValue* at *line 8 and 9* search the symbolic and concrete values for a particular variable in the table. They are defined in the C modules of concolic testing engine.

3.3.3 Variable Stack

In previous section, before the execution of *f* we have to populate symbol table with respective variables of *f*. But it is important to do systematic clean up after *f* is done executing. This includes deletion of entries from the table that were temporarily required for symbolic execution of *f*. It is required so that spurious values may not only cause inconsistencies but also caused the symbol table to overflow. For this purpose, we maintain a stack S called *variable stack*. Each element of the stack stores the information regarding variable entries in the symbol table for indiviual function. The contents of stack element are as follows :

- *funcName* : name of the function for which this element is created.
- *args* : string array containing the names of formal paramters of *funcName*.
- *locals* : string array containing names of local variable in the scope of *funcName*.
- *occurence* : used to indicate the instance of *funcName* in the call stack. This is explained in detail and how it is used to resolve recursive calls in *Section 3.4.2*.

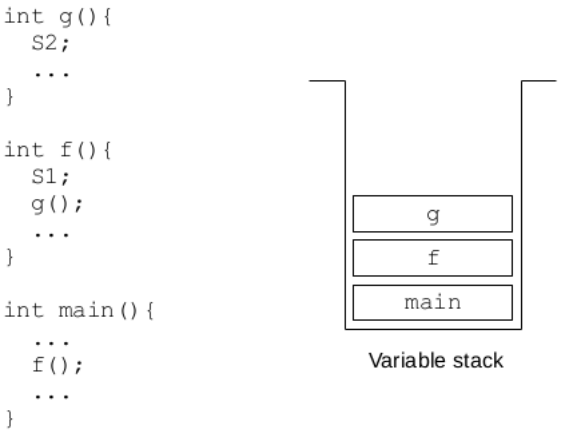


Figure 3.5: Variable Stack

The behavior of variable stack is similar to function call stack. Whenever a function call is encountered the variables (both arguments and locals) of the callee function are pushed onto the stack S.. For example, in *Figure 3.5* when f is called in *main*, its variable details are pushed onto S. At the time statement S2 executes in g, stack has three elements as shown in the figure. When we return from a function call, we pop an element from the stack and delete the entries in the symbol table by referring the popped element.
The task of pushing the element on encountering a function call is done by a method *funcEntry()* and popping the element after returning from that function is done by *funcExit()*, both of which are defined in C modules.

3.3.4 Return Handling

After the environment is set up for f, it executes symbolically covering a specific path in the control flow graph of f. It collects a summary that tranforms the symbolic values of parameters to new symbolic return values. We need to map the concolic return values to appropriate variable at call site.
For this, just before the function f returns, we copy the concolic return value to a global variable. Then at call site, we assign the lvalue of call instruction to this global concolic variable. Thus, the concolic execution flows to and from the callee correctly.

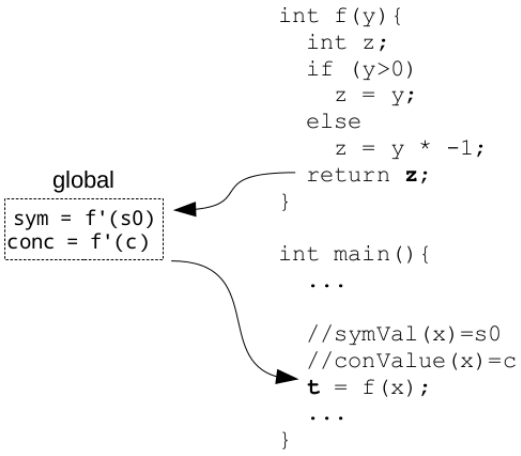


Figure 3.6: Concolic Return

Here in *Figure 3.6*, function f is called with parameter x. Its symbolic and concrete value is modified according to function f', which is a transition function along that path in f. Variable z attains this new concolic value and is returned. To ensure transfer of modified

symbolic information on function return, we map the concolic value of z to global variable and from them to t in *main*. Note that even if a function does not return a value it may alter the global values.

3.4 Recursion

3.4.1 Limitation of previous approach

The approach that we have discussed until now can only handle non-recursive procedure calls. The reason being that our algorithm did not have support for multiple versions of same variable of the same function in the symbolic table. Our variable stack can handle atmost one instance of a function at any given time. To explain this, given is an example of recursive code.

```
1 int fac(int fac_n)
2 {
3   int fac_m;
4   if (fac_n == 0)
5      fac_m=1;
6   else
7      fac_m = fac_n * fac(fac_n-1);
8   return fac_m;
9 }
10
11 int main (void)
12 {
13   int i = 5;
14   int j = fac(i);
15   return 0;
16 }
```

Figure 3.7: Program to compute factorial

In this example, at *line 14* a function call has been made to *fac*. Before the call instruction executes, our algorithm maps concolic values from i to *fac_n* and creates an entry in the symbol table using *fac_n* as the key. Now when the *fac* executes, it again encounters a call instruction at *line 7*. Our algorithm should map the values from `fac_n-1` to "new" *fac_n*. But it has no mechanism implemented to distinguish between different versions of same variable of the same procedure. This leads to overwriting of previously present entry of *fac_n* in the symbol table (refer to *table 3.1*) . More important problem that will arise is that after *fact* at *line 7* executes and returns, it cannot restore the previous symbolic state (i.e. before call was made). The concolic values of previous versions of the

key	symVal	conVal
	:	
	:	
fac_n	s0	c0
	:	

(a) call at line 14

key	symVal	conVal
	:	
fac_n	s1	c1
~~fac_n~~	~~s0~~	~~c0~~
	:	

(b) call at line 7

Table 3.1: Symbol Table

recurring function in the stack will be lost.

Therefore, we need to modify the key of symbol table such that we can keep track of multiple instances of same function in the variable stack, allowing recursive call analysis.

3.4.2 Need for Versions

Versions of function variables have to be maintained according to number of instances they have in the variable stack. It is made sure that the entries in the symbol table dont get deleted or overwritten unless they are not needed anymore. One way to handle this is to modify the key of symbol table such that they have function version associated with them.

Occurence of a procedure at any time is defined as the number of instances of that same procedure present in the call stack.

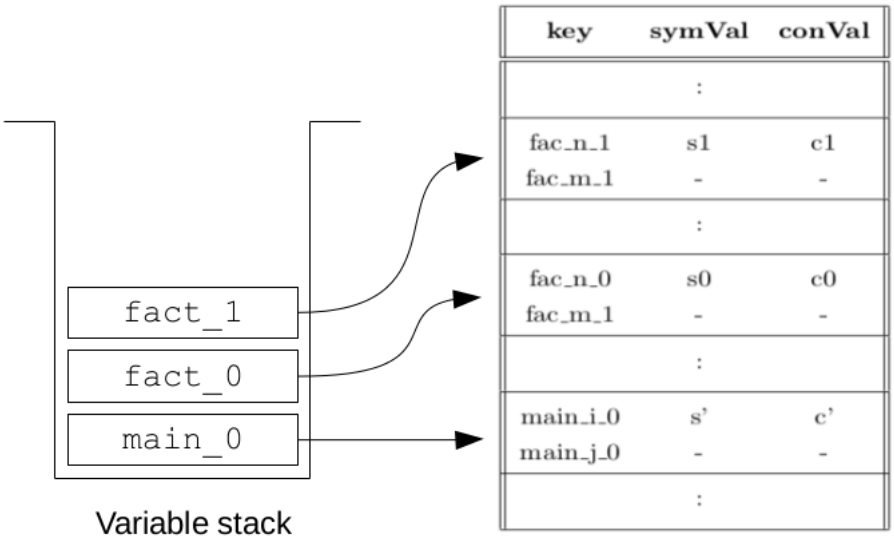


Figure 3.8: Function Variable Versions in Symbol Table

Therefore the key which was earlier defined in *Section 3.2.2* is remodifed to keep track the associated variable, procedure to which that variable belongs and that procedure's version (*occurence*) it the call stack. The new key is:

$$K' = f'(V, P, O_P) \tag{3.1}$$

where O_P is the occurence of P in function call stack. *Figure 3.8* shows the version handling of multiple instances of same functions in the symbol table. When another instance of function is pushed in the stack (recursively), no value in the symbol table is overwritten. Similarly, when a function instance is popped from the stack, the concolic values of previous version should be accessed.

3.4.3 Variable Hash-Map

The issue that arises with this approach is that we cannot rename variables in conjunction with their versions (like we did in *Section 3.2.2* using static analysis). This is because the versions are created dynamically at runtime when function call stack is manipulated. Therefore we create a hashmap M that maps the variable name to key in the symbol table.

Whenever concolic execution needs to refer the symbol table using variable V, we look up the *value* corresponding to V in M and return the *key*. But we still have to maintain the hashmap so that it always give the correct version of the key (the state of the variable stack is dynamic and so are the versions of function variables). This can be explained by considering a call instruction in the program that invokes function *f*. The operations before and after the execution of f are as follows.

Before Function Execution

In *Section 3.2.2* we discussed the algorithm to populate the symbol table with specific values before executing a call instruction. We modify that algorithm slightly, instead of using variable name as key in *addEntryToSTable*, we use our new key K'. Then we add the new mapping (variable name, K') to *M*. This keeps track of key K' and its concolic values, for when we need to do a query on the symbol table.

After Function Execution

When a function is done executing, the entries in the symbol table for that particular version of function are deleted. To do this, we again use the hashmap to get the current keys (containing current version of variables also) of symbol table that are no longer required.

3.4.4 Bound on Recursion

To keep the size of symbol table in check and prevent the variable stack from overflowing, we have to put an upper limit on the maximum number of instances that the variable stack can have of a given function. This puts a cap on the precision of symbolic execution. If the maximum number of versions allowed for a particular function be X, then the symbolic execution will only continue uptil stack depth of X, after that the function executes using only concrete values. The symbolic execution resumes when the stack depth with respect to that function becomes less than or equal to X.

Chapter 4

Experiments

This chapter states the results generated by the tool on a certain set of programs. These results are then compared with intraprocedural concolic testing.

4.1 System Specifications

Operating System	Ubuntu 14.04 LTS
OS Type	64-bit
Processor	Intel Core i7-4710HQ CPU @ 2.50 GHz × 8
Memory	7.7 GiB, 1600 MHz

4.2 Results

Interprocedural approach shows improvements where there are functions calls involved. When a function summaries its effect on the inputs variables, we get more precise path conditions. This is because instead of functions returning just concrete values they return symbolic values as well. These path conditions act as prepositional formula for test case generation.

The benchmarks program are judged based on the procedures tested. Procedures having no branch conditions are not selected becuse the coverage reported is MCDC coverage[**?**] and it is calculated using the number of branches in a procedure. The programs are timed using `GNU-time`[**?**] that measures total number of CPU-seconds that the process spent in

user mode. Table 4.1 shows the results generated on some benchmark programs. The time recorded includes the time taken to complete the instrumentation and runtime of test case generation for a program. Instrumentating the code takes maximum of 250 milliseconds among the test programs. The interprocedural analysis causes a slowdown of maximum 10 milliseconds in the test case generation. The maximum time recorded to run the tool on a program is just under 1.5 seconds.

S NO	**Program**	**Intraprocedural Tesing**			**Interprocedural Testing**		
		Coverage	**#TestCases**	**Time**	**Coverage**	**#TestCases**	**Time**
1	TCAS (Inhibit_Biased_Climb)	50	1	1.109 sec	50	2	1.235 sec
2	TCAS (Non_Crossing_Biased_Descend)	14	11	1.009 sec	85	9	1.240 sec
3	TCAS (Non_Crossing_Biased_Climb)	33	8	0.998 sec	83	7	1.289 sec
4	TCAS (alt_sep_test)	52	4	1.092 sec	52	8	1.354 sec
5	primemod (main)	66	2	1.010 sec	100	4	1.219 sec
6	heapSort (heapSort)	100	2	0.900 sec	100	2	1.143 sec
7	crest1 (main)	50	1	0.915 sec	100	2	1.108 sec
8	crest2 (main)	50	1	1.048 sec	50	1	1.191 sec
9	spectral-norm (eval_At_times_u)	75	5	1.304 sec	75	5	1.412 sec
10	spectral-norm (eval_A_times_u)	75	5	1.295 sec	75	5	1.398 sec
11	faculty (main)	50	3	0.853 sec	100	2	1.160 sec
12	linpack (dgesl)	84	43	1.185	52*	6*	1.395 sec*
13	linpack (daxpy_ur)	36	17	1.129	36	17	1.235 sec
14	linpack (dscal_r)	50	10	1.014	50	10	1.217 sec
15	linpack (ddot_r)	42	2	1.037	42	2	1.267 sec
16	prime (prime)	83	3	1.029 sec	50	2	1.243 sec

Table 4.1: Benchmark Results

In *Table 4.1* the names of the procedures tested are below the name of the program in parenthesis.

4.3 Analysis

Interprocedural implementation shows good improvements in TCAS benchmark and some others. In shows coverage improvement from **33%** to **83%** in TCAS-2. It achieves this with a reduced number of test cases.

All the programs have function calls but not all of them show improvement in coverage. The reason is that to influence the coverage, the functions should effect the path conditions. To do this, functions should alter the symbolic state of the variables of the program. That is why programs like **crest2** have several function calls but shows no improvement in coverage whereas program discussed in *Section 1.1* shows 100% coverage.

A WCET benchmark program **faculty** highlights the handling of recursive features. This program contains a functions that calculates factorial recusively. The intraprocedural testing gives 50% coverage in 3 test cases whereas interprocedural testing improves it to 100 % in 2 test cases.

In concolic execution test cases are generated according to the path constraint obtained in the consecutive runs. The last condition of path constraint is flipped and passed to constraint solver to obtain new test case that explores a different path. But concolic engine we are using also aims at minimizing the number of test cases. Therefore flipping of conditions is done on the basis of levels. The conditions on the same level are flipped at the same time.

```
if (C1)
  statement1;
if (C2)
  statement2;
```

If there are two conditions *C1* and *C2* that are on the same level then they will be flipped together to get the new test case. If execution of *C2* block depends on the execution of *statement1* then the coverage achieved may not be maximum. This is what happens in WCET program **prime** where the intraprocedural testing gives 83% coverage while interprocedural testing gives 50% coverage.

Chapter 5

Conclusion

In this thesis, we have designed a method to intergrate interprocedural analysis with intraprocedural concolic execution engine. This not only improves the coverage of a program but in some cases does it with lesser number of test cases. Only in a few cases the coverage of intraprocedural testing gives better results than interprocedural testing.

We have handled recursive procedural. Recursion is nested only until a certain bound after which only concrete execution takes place.

A small C unit test suite have also been a contribution which can be used to verify and compare the features of different test case generation tools like CREST[**?**], PathCrawler[**?**] and more.