How Much Relaxation Is Too Much?

Determining the performance vs. accuracy tradeoff for S2E's relaxed consistency models

Mariana D'Angelo
Department of Electrical and Computer Engineering
University of Toronto
Toronto, Canada
mariana.dangelo@utoronto.ca

Dhaval Miyani
Department of Electrical and Computer Engineering
University of Toronto
Toronto, Canada
dhaval.miyani@utoronto.ca

I. Introduction

S2E is a symbolic execution platform that operates directly on application binaries for analyzing the properties and behaviour of software systems. It analyzes programs in-vivo (the whole environment) within a real software stack (user program, libraries, kernel, drivers, etc.) rather than using abstract models of these layers. It is commonly used for performance profiling, reverse engineering of proprietary software, and finding bugs in user-mode and kernel-mode binaries [1].

The goal of this experiment is to evaluate two consistency models in the S2E system which do not have predictable accuracy or performance; the Relaxed Consistency Overapproximate (RC-OC) and Relaxed Consistency CFG Consistency (RC-CC) models. These two models are relaxed in the sense that they admit infeasible paths (which other consistency models from S2E do not). This provides a performance advantage by limiting the analysis to only certain parts of the system, thus allowing the target code to be reached sooner [1]. However, as these models allow locally infeasible paths (i.e., the state of the unit can be inconsistent), the analyses will be prone to false positives as these paths cannot be produced by concrete runs [1]. The main difference between the two models is that in RC-OC only real environmental behaviour is admitted whereas in RC-CC unrealistic environmental behaviour is possible. These models are of particular interest because the effect of admitting unrealistic environmental behaviour is unclear and as such the accuracy of each model is not easily predicted.

Our primary objective is to determine the accuracy of these models for different types of bugs. A secondary objective is to observe the performance of these consistency models. These two objectives will allow us to determine if there is a tradeoff between performance and accuracy when using these two models. For example, does running S2E with a given consistency model for an extra hour improve accuracy by 1% or 10%? Ideally, we want to be able to determine what program attributes (e.g., program length, type of bug, location of bug, etc.) make these models more effective (if any). For example, it is well known that symbolic execution suffers from exponential performance and resource usage, implying that bugs located later in the code after several branches are less likely to be found within a reasonable amount of time. It is our belief that by admitting more infeasible paths it is possible that the RC-CC model will suffer this exponential effect more than RC-OC, but have better accuracy within the program unit.

We believe these could be compelling results for software systems that are only exposed to a limited set of bug types, given that one consistency model may have higher accuracy for such bugs. For example, a banking application which performs batch processing of reports would be less susceptible to buffer overflow attacks as there is no external input (assuming that the reports are generated by trusted tools). Additionally, the accuracy vs. performance tradeoff would aid users in selecting how much time to allot S2E to run, given that a bit more time could increase the accuracy dramatically if program characteristics are a factor.

The two consistency models we will analyze are relaxed in the sense that they admit infeasible paths (which other consistency models from S2E do not). This provides a performance advantage by limiting the analysis to only certain parts of the system, thus allowing the target code to be reached sooner [1]. However, as these models allow locally infeasible paths (i.e., the state of the unit can be inconsistent), the analyses will be prone to false positives as these paths cannot be produced by concrete runs [1]. We expect that the Relaxed Consistency CFG Consistency model will have a higher false positive rate and lower false negative rate as it functions similar to static analysis by admitting the infeasible paths. Conversely the Relaxed Consistency Overapproximate model would have poorer performance (as it uses constraints to only admit feasible paths), but the false positive rate should be quite low and the false negative rate is expected to be higher with this model.

II. Background

There is much work in symbolic execution for testing [2], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [3]. A typical approach for testing is fuzzing [13], which involves generating random inputs for a program in an attempt to exercise all paths in a program (and hopefully hit any buggy code). This is known as concrete execution which involves running the program with deterministic values. Fuzz testing has several limitations, in particular coverage of paths in a program. For example, Figure 1 shows a if statement which is only taken when x is 10. There are 2^{32} possible values which x could take when an input is randomly generated and as such the probability of actually taking this path is 2^{-32}. This example demonstrates that the likelihood of some paths being taken when fuzzing is extremely low, which is the reason for the poor test coverage of concrete execution.

```
if (x == 10) {
// path 1
} else {
// path 2
}
```

Fig. 1. Code example where fuzzing generally has poor coverage

Symbolic execution, on the other hand, runs the application with symbolic inputs [2] which are initially unconstrained by design. The program executes with these symbolic values and replaces concrete operations with ones that can manipulate symbolic values. The symbolic execution engine "follows" both paths whenever it encounters a branch on a symbolic value and adds the path condition to a set of constraints known as the "path constraint". When a bug is encountered in the code a test case (i.e., concrete inputs) can be generated from the path constraint. There are two main limitations of symbolic execution: (1) "the path explosion problem" [1], which is due to the exponential growth of paths in a program caused by conditional branches, and (2) "the environment problem" [2], which is due to interactions with the surrounding environment (e.g, operating system, network, etc.).

In order to deal with the environment problem, one can use a mixture of symbolic and concrete execution (also known as concolic execution [3]). Concolic execution gathers constraints using symbolic execution, but then generates concrete values using a constraint solver in order to address the environment problem. Once these concrete values are generated the program can interact with its environment in a normal fashion without the overheads and problems associated with symbolically executing the environment.

S2E is a symbolic execution engine which can perform concrete execution, symbolic execution, and concolic execution depending on the consistency model chosen. There are six different consistency models in S2E, described below. Figure 2 provides an overview of the models showing how they transition from highly strict to highly relaxed models with some details about what consistency requirements are relaxed for each model. A model is consistent if there exist a globally feasible path through the system for every path explored in the unit. A unit is the block one wishes to analyze and the environment is the rest of the system.

- *Strictly Consistent Concrete Execution (SC-CE):* No symbolic execution in unit or environment.
- *Strictly Consistent Unit-level Execution (SC-UE):* Unit is symbolically executed while the environment is executed concretely.
- *Strictly Consistent System-level Execution (SC-SE):* Unit and environment are executed symbolically - this is the only model that executes the environment symbolically.
- *Local Consistency (LC):* Similar to SC-UE, but it adheres to constraints that the environment/unit API contracts impose on return values.
- *Relaxed Consistency Overapproximate Consistency (RC-OC):* Similar to LC, but it ignores the constraints from the environment/unit API contracts.
- *Relaxed Consistency CFG Consistency (RC-CC):* Similar to SC-UE, but like static analysis it can explore any path in the unit's inter-procedural control flow graph (even infeasible ones).

Our experiment will focus on evaluating the two latter models: RC-OC and RC-CC. These models are of particular interest because we cannot accurately predict the accuracy or performance of these two models in contrast to the other four models [1], as described in the Introduction section.

There are several differences between S2E and other similar symbolic execution engines. For instance, KLEE uses file system models [2] to avoid symbolically executing the actual filesystem. S2E does not take this approach as writing models is a labour intensive and error-prone process. Cute [3] executes the environment concretely (i.e., without modelling) with a consistency model similar to S2E's SC-SE, but it is limited to code-based selection and one consistency model. S2E does not use compositional symbolic execution [11], a performance optimization which saves results for parts of the program (e.g., a function) and reuses them when that part is called in a different context. With concolic execution everything runs concretely for full systems, however, when the execution crosses program boundaries it may result in lost paths [1]. Due to this, KLEE and CUTE cannot track path conditions in the environment and hence are unable to re-execute calls to enable overconstrained but feasible paths (e.g., malloc does not execute deterministically). DART [12], CUTE [3] and EXE [7] use mix-mode execution (concretely executing some parts and symbolically executing others) to increase efficiency, however, they do not use automatic symbolic-concrete bidirectional data conversion (i.e., automatic conversion between concrete and symbolic values at program boundaries such as the unit and the environment) unlike S2E, which is key to S2E's scalability and low programmer effort.

Static analysis is performed without executing the program in question and has known limitations such as a high false positive and negative rate (due to infeasible paths) [13] and a lack of runtime environment analysis (as only the application source code is parsed). To reduce the number of false positives tools such as Saturn [15] and bddbddb [16] use a path-sensitive analysis engine. Saturn aims to detect logic programming language bugs by summarizing functions. bddbddb finds buggy patterns in a database where programs are stored as relations. As can be inferred, these tools are language specific and require different implementations for different programming languages. Additionally, using these tools requires learning a new programming language. As a dynamic analysis framework, S2E addresses the limitations of static analysis tools. For example, they directly operate on and analyze binaries whereas static analysis would require disassembly and decompilation. This could involve converting an x86 binary to the LLVM format and running it through an engine like KLEE. Disassembly and decompilation [14] are classically undecidable problems (e.g., disambiguating code from data) [1].

III. Approach

Here we discuss the research methodology used in preparation for the evaluation, separated into four parts.

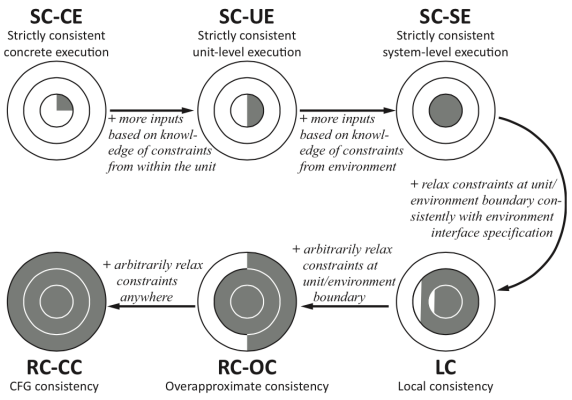

Fig. 2. S2E Consistency Models [1]

A. Sanity Check

First, we will be ensuring that these S2E models function as we would expect and that they are capable of detecting the different types of bugs we will use in our experiments. We will write a few short ($<$ 100 LOC) dummy programs with different bugs (e.g., null pointer dereference, divide by zero, buffer overflow, etc.) to determine coverage of the consistency models before beginning. If one of the consistency models cannot find any of these bugs these results will be recorded as a lack of accuracy for that model which allotting more execution time would not address, meaning that the accuracy vs. performance tradeoff for this type of bug and model would be nonexistent.

B. Dataset

We will select a diverse set of programs to allow us to determine whether different program characteristics (e.g., code length, interactions with the environment, etc.) give accuracy or performance gains for either model. The three programs will be open source and have a bug reporting system such that we can use these resources when establishing a ground truth (see next subsection). Three programs will be selected based on different levels of interaction with the environment.

1) A web server (e.g., Apache), which we expect to have a lot of network usage (and thus system calls)
2) A database (e.g., SQLite), which we expect to have a lot of file i/o (and thus system calls)
3) A computation heavy application (e.g., Fourier Transform calculator), which we expect to have minimal interaction with the environment (not many system calls are expected)

C. Ground Truth

We will look at bug reports for current (or older, if necessary) versions of the selected programs to determine what bugs we expect S2E to find and what (if any) bug types we may need to inject. This will enable us to establish a ground truth for the false negatives when measuring the accuracy of the models. Of course, S2E may find yet more bugs than those identified, so access to the source code will be necessary to verify any other bugs. This analysis will allow us to determine the false positive rate or approximate it depending on how many bugs S2E finds in a given application. This approximation may be necessary due to the manual effort required to label the bugs as valid or invalid.

D. Measurement

We will create a script to automate the different runs of our experiment. In particular, this script will record the time that a given run started and cut off the execution after different amounts of time. A unit of time, for example, could be one hour, and testing across multiples of said unit (e.g., two hours, three hours, etc.) will permit us to determine the accuracy vs. performance tradeoff discussed earlier (if it exists). We may also determine how much time it took the different models to find a given bug within a unit by exploiting any logging functionality in S2E if necessary

IV. Status

As this is an evaluation, there are two main stages for the evaluation of the consistency models for S2E: the preparation phase and the experimentation phase. We are currently in the preparation phase:

- We have set up S2E and used it on a sample application.
- We are also gathering applications with bugs in them for our data set. We are looking at the bug reports for each application to ensure that it is suitable (i.e., we can establish a ground truth as described earlier).
- Furthermore, we have created some sample programs injected with different types of bugs for sanity testing. This ensures that each of the two consistency models are able to catch the bugs from these sample programs (as described in the approach).

V. Evaluation

This section describes the experiments that will be performed to test S2E's consistency models, specifically RC-CC and RC-OC. For each consistency model, the following experiments will carried out to get the empirical amount of time it takes to catch a bug and the accuracy rate:

- Using the script described in the Approach section we will run each of the three applications for different "time units" and record the results S2E generates. The time unit could be half an hour, one hour, or several - we will fine tune this during the experiment based on preliminary results. This will determine the number of bugs S2E catches for a given time period (for measuring accuracy).
- Depending on the bug caught by S2E, a manual analysis of the source code will be performed to determine whether there was an actual bug (based on the ground truth determined earlier). This will provide the accuracy rates: false positives and false negatives. If S2E identified a bug, which manual analysis proves not to be a bug, then the false positive rate will increase. Conversely, the false negative rate will be determined using the bug reports to see if S2E misses a bug that we are certain exists.

- We will create systematic performance-accuracy trade-off graphs (per application and per consistency model). The x-axis of these graphs will be the different "time units", whereas the y-axis will be the number (or percentage) of false positives and false negatives (two data lines).

REFERENCES

[1] CHIPOUNOV, V., GEORGESCU, V., ZAMFIR, C., AND CANDEA, G Selective symbolic execution. In Workshop on Hot Topics in Dependable Systems, 2009.

[2] CADAR, C., DUNBAR, AND ENGLER, D.R. KLEE: Unassisted and automatic generation of high-coverage tests for complex systems programs. In Symp. on Operating Systems Design and Implementation, 2008.

[3] SEN, K., MARINOV, D., AND AGHA, G. CUTE: A concolic unit testing engine for C. In In 5th joint meeting of the European Software Engineering Conference and ACM Symposium on the Foundations of Software Engineering (ESEC/FSE 2005).

[4] BOONSTOPPEL, P., CADAR, C., AND ENGLER, D. RWset: Attacking path explosion in constraint-based test generation. In Proceedings of Tools and Algorithms for the Construction and Analysis of Systems (TACAS 2008).

[5] BRUMLEY, D., NEWSOME, J., SONG, D., WANG, H., AND JHA, S. Towards automatic generation of vulnerability-based sig- natures. In Proceedings of the 2006 IEEE Symposium on Security and Privacy (IEEE S&P 2006).

[6] CADAR, C., AND ENGLER, D. Execution generated test cases: How to make systems code crash itself. In Proceedings of the 12th International SPIN Workshop on Model Checking of Soft- ware (SPIN 2005).

[7] CADAR, C., GANESH, V., PAWLOWSKI, P., DILL, D., AND ENGLER, D. EXE: Automatically generating inputs of death. In Proceedings of the 13th ACM Conference on Computer and Communications Security (CCS 2006).

[8] COSTA, M., CASTRO, M., ZHOU, L., ZHANG, L., AND PEINADO, M. Bouncer: Securing software by blocking bad input. In Proceedings of the 21th ACM Symposium on Operating Systems Principles (SOSP 2007).

[9] COSTA, M., CROWCROFT, J., CASTRO, M., ROWSTRON, A., ZHOU, L., ZHANG, L., AND BARHAM, P. Vigilante: end-to-end containment of Internet worms. In Proceedings of the 20th ACM Symposium on Operating Systems Principles (SOSP 2005).

[10] EMMI, M., MAJUMDAR, R., AND SEN, K. Dynamic test input generation for database applications. In International Symposium on Software Testing and Analysis (ISSTA 2007).

[11] GODEFROID, P. Compositional dynamic test generation. In Proceedings of the 34th Symposium on Principles of Programming Languages (POPL 2007).

[12] GODEFROID, P., KLARLUND, N., AND SEN, K. DART: Directed automated random testing. In Proceedings of the Conference on Programming Language Design and Implementation (PLDI 2005).

[13] GODEFROID, P., LEVIN, M. Y., AND MOLNAR, D. Automated whitebox fuzz testing. In Proceedings of Network and Distributed Systems Security (NDSS 2008).

[14] SCHWARZ, B., DEBRAY, S., AND ANDREWS, G. Disassembly of executable code revisited. In Working Conf. on Reverse Engineering, 2002.

[15] DIILLIG, I., DILLIG, T., AND ALKEN, A. Sound, complete and scalable path- sensitive analysis. In Conf. on Programming Language Design and Implementation, 2008.

[16] LAM, M. S., WHALEY, J., LIVSHITS, V. B., MARTIN, M. C., AVOTS, D., CARBIN, M., AND UNKEL, C. Context-sensitive program analysis as database queries. In Symp. on Principles of Database Systems, 2005.