securityandtrust.lu

FR

Final Report

F. Pastore, O. Cornejo

Interdisciplinary Centre for Security, Reliability and Trust

University of Luxembourg

ITT-1-9873-ESA-FAQAS-D2

Issue 5, Rev. 3

October 13, 2021

UNIVERSITÉ DU LUXEMBOURG

Revisions

Issue Number	Date	Authors	Description
ITT-1-9873-ESA-FAQAS-D2 Issue 1 Rev. 1	June 12th, 2020	Fabrizio Pastore, Oscar Cornejo	Initial release.
ITT-1-9873-ESA-FAQAS-D2 Issue 2 Rev. 1	June 25th, 2020	Fabrizio Pastore, Oscar Cornejo	Delivered document.
ITT-1-9873-ESA-FAQAS-D2 Issue 2 Rev. 2	July 6th, 2020	Fabrizio Pastore, Oscar Cornejo	Revised document.
ITT-1-9873-ESA-FAQAS-D2 Issue 3 Rev. 3	July 8th, 2020	Fabrizio Pastore, Oscar Cornejo	Revised document after review meeting.
ITT-1-9873-ESA-FAQAS-D2 Issue 3 Rev. 1	Nov 4th, 2020	Fabrizio Pastore, Oscar Cornejo	Document delivered for TR3, Prototyping and tool chain review. With respect to previous version, the following sections had been updated: • Section 1.1.4 • Section 2.1.6 • Section **??** • Section about evaluation of code-driven mutation analysis (now moved to D4) • Section about case studies for LXS (now moved to D4) • Section about case studies for ASN (now moved to D4) Modified text is in blue.
ITT-1-9873-ESA-FAQAS-D2 Issue 4 Rev. 1	Dec 16th, 2020	Fabrizio Pastore, Oscar Cornejo	Updated empirical results for code-driven mutation testing concerning ESAIL.
ITT-1-9873-ESA-FAQAS-D2 Issue 5 Rev. 1	Sept 20th, 2021	Fabrizio Pastore, Oscar Cornejo	Updated Section **??** to include the finalized version of the data-driven mutation analysis approach. Added Section **??** to describe the code-driven test suite augmentation approach. Extended Section **??** to reflect our considerations on the fasibility of data-driven test suite augmentation. Modified text is in blue.
ITT-1-9873-ESA-FAQAS-D2 Issue 5 Rev. 2	Sept 22nd, 2021	Fabrizio Pastore, Oscar Cornejo	Updated Section **??** to include a revised version of the text, in line with what published in IEEE Transactions on Software Engineering [1].
ITT-1-9873-ESA-FAQAS-D2 Issue 5 Rev. 3	Oct 6th, 2021	Fabrizio Pastore, Oscar Cornejo	Addressing comments from ESA. In particular, we addressed typos in Figures **??** to **??**.

Delivered Items

In the following Table we provide a list of deliverable items released with this document. Each Item is identified by its path on the Alfresco system.

Deliverable	Description
Review Meeting 2/Deliverables/ASN1CC-CaseStudy/Specifications/ACN-UM-v-3-2.pdf	User manual of ASN1 Compiler
Review Meeting 2/Deliverables/ASN1CC-CaseStudy/Specifications/taste-documentation-current.pdf	Taste software documentation including ASN1 overview
Review Meeting 2/Deliverables/ASN1CC-CaseStudy/Software/asn1.container.tar.gz	Mutation testing Singularity container corresponding to the ASN1CC case study.
Review Meeting 2/Deliverables/ASN1CC-CaseStudy/Software/ASN1-container-instructions.md	Instructions to deploy the ASN1CC singularity container.
Review Meeting 2/Deliverables/GSL-CaseStudies/Specifications/gs-man-nanosoft-ms100-command-and-management-sdk-3.6.2-1-g67fe6e1.pdf	GSL software specifications
Review Meeting 2/Deliverables/LXS-CaseStudies/Specifications	Folder with LXS specifications
Review Meeting 2/Deliverables/LXS-CaseStudies/Specifications/FAQAS-LXS-MAN-001_1- SVF Software Installation and User Manual.pdf	SVF and ESAIL Software Installation and User Manual
Review Meeting 2/Deliverables/LXS-CaseStudies/Specifications/ESAIL-LXS-ICD-P-0184_2A ADCS IF SW External ICD.docx	ADCS IF SW External ICD
Review Meeting 2/Deliverables/LXS-CaseStudies/Specifications/ESAIL-LXS-SDD-P-0105_1B On-board Application Software Design Document	On-board Application Software Design Document
Review Meeting 2/Deliverables/LXS-CaseStudies/Specifications/MOC-applicable MIB egos-mcs-s2k-icd-0001-version7.0-FINAL	SCOS-2000 Database Import ICD
Review Meeting 2/Deliverables/LXS-CaseStudies/Specifications/ocp.dat	SCOS-2000 Database file containing the specifications of the nominal value ranges for ESAIL ADCS parameters.
LXS-CaseStudies/Specifications/ESAIL-VirtualMachine	Folder containing the ESAIL virtual machine. The ESAIL virtual machine contains both unit and system test cases. The location and instructions required to execute ESAIL system test cases are provided in FAQAS-LXS-MAN-001_1 Section 7. Concerning unit test cases, they are implemented using the check framework [2]. Unit test cases are stored inside the *Test* folder of the component under test. For example, for the AdcsController, they are stored in *./ApplicationLayer/AdcsController/Test/*.
Review Meeting 2/Deliverables/MLFS-CaseStudy/Software/mlfs.container.tar.gz	Mutation testing Singularity container corresponding to the MLFS case study.
Review Meeting 2/Deliverables/MLFS-CaseStudy/Software/MLFS-container-instructions.md	Instructions to deploy the MLFS singularity container.
Review Meeting 2/Deliverables/MLFS-CaseStudy/Specifications/E1356-GTD-SUM-01_I1_R2.pdf	Installation and execution instructions of the MLFS.
Review Meeting 2/Deliverables/MLFS-CaseStudy/Specifications/E1356-CS-SUM-01_I1_R5.pdf	Installation and execution instructions of the MLFS test suite.
Review Meeting 2/Deliverables/SnT-Software/FAQAS-DataDrivenMutator-Buffers.zip	Preliminary implementation of the data-driven mutation testing component
Review Meeting 2/Deliverables/SnT-Software/FAQAS-DataDrivenMutator-ASN1.tar.gz	Preliminary implementation of the data-driven mutation testing component for ASN1.
Review Meeting 2/Deliverables/SnT-Software/FAQAS-CodeDriven-Mutation.tar.gz	Preliminary implementation of the code-driven mutation testing component
Review Meeting 3/Deliverables/SnT-ScriptsAndPrograms-ForCodeDrivenMutationTesting.zip	Updated scripts for code-driven mutation testing.
Review Meeting 3/Deliverables/SnT-ScriptsAndPrograms-ForCodeDrivenMutationTesting.zip	Updated scripts for code-driven mutation testing.
Review Meeting 5/Delivered Software/ITT-1-9873-ESA-FAQAS-MASS.zip	Zipped repository for the code-driven mutation analysis toolset (MASS).
Review Meeting 5/Delivered Software/ITT-1-9873-ESA-FAQAS-SEMuS.zip	Zipped repository for code-driven mutation testing toolset (SEMuS).
Review Meeting 5/Delivered Software/ITT-1-9873-ESA-FAQAS-DAMAt.zip	Zipped repository for data-driven mutation analysis toolset (DAMAt).

Contents

Introduction

This document is the final report of the ESA activity ITT-1-9873-ESA, which concerns the development of a framework for the automated assessment and the automated improvement of test suites for space software[1].

From spacecrafts to ground stations, software has a prominent role in space systems; for this reason, the success of space missions depends on the quality of the system hardware as much on the dependability of its software. Mission failures due to insufficient software sanity checks [3] are unfortunate examples, pointing to the necessity for systematic and predictable quality assurance procedures in space software.

Existing standards for the development of space software regulate software quality assurance and emphasize its importance. The most stringent regulations are the ones that concern flight software, i.e., embedded software installed on spacecrafts, our target in this activity. In general, software testing plays a prominent role among quality assurance activities for space software, and standards put a strong emphasis on the quality of test suites. For example, the European Cooperation for Space Standardization (ECSS) provides detailed guidelines for the definition and assessment of test suites [4, 5].

Test suites assessment is typically based on code inspections performed by space authorities and independent software validation and verification (ISVV) activities, which include the verification of test procedures and data (e.g., ensure that all the requirements have been tested and that representative input partitions have been covered [6]). Though performed by specialized teams, such assessment is manual and thus error prone and time-consuming. ***Automated and effective methods to evaluate the quality of the test suites are thus necessary.*** Also, ***methods to automatically generate test cases will speed-up the improvement of test suites***.

Since one of the primary objectives of software testing is to identify the presence of software faults, an effective way to assess the quality of a test suite consists of artificially injecting faults in the software under test and verifying the extent to which the test suite can detect them. This approach is known as *mutation analysis* [7]. In mutation analysis, faults are automatically injected in the program through automated procedures referred to as mutation operators. Mutation operators enable the generation of faulty software versions that are referred to as *mutants*. Mutation analysis helps evaluate the effectiveness of a test suite, for a specific software system, based on its mutation score, which is the percentage of mutants leading to test failures.

Despite its potential, mutation analysis is not widely adopted by industry in general and space system development in particular. The main reasons include its limited scalability and the pertinence of the mutation score as an adequacy criterion [8]. Indeed, for a large software system, the number of generated mutants might prevent the execution of the test suite against all the mutated versions. Also, the generated mutants might be either semantically equivalent to the original software [9] or redundant with each other [10]. Equivalent and redundant mutants may bias the mutation score as an adequacy criterion.

The mutation analysis literature has proposed several optimizations to address problems related to scalability and mutation score pertinence. For example, scalability problems are addressed by approaches that sample mutants [11] [12] or solutions that prioritize and select the test cases to be executed for each mutant [13]. Equivalent and redundant mutants can be detected by comparing the code coverage of the original program and its mutants [14, 15, 16, 17]. However, these approaches have not been evaluated on industrial,

[1]In this report, we use the term space software to indicate software to be deployed on hardware that runs on-orbit.

embedded systems and there are no feasibility studies concerning the integration of such optimizations and their resulting, combined benefits. For example, we lack mutants sampling solutions that accurately estimate the mutation score in the presence of reduced test suites; indeed, mutant sampling, which comes with a certain degree of inaccuracy, may lead to inaccurate results when applied to reduced test suites that do not have the same effectiveness of the original test suite.

In addition, existing mutation analysis approaches cannot identify problems related to the interoperability of integrated components (integration testing). Space software, similar to software running in other types of CPSs, is often affected by problems cause by the lack of ***interoperability of integrated components*** [18, 19], mainly due to the wide variety and heterogeneity of the technologies and standards adopted. It is thus of fundamental importance to ensure the effectiveness of test suites with respect to detecting interoperability issues, for example by making sure test cases trigger the exchange of all possible data items and report failures when erroneous data is being exchanged by software components. For example, the test suite for the control software of a satellite shall identify failures due to components working with different measurement systems [20]. Unfortunately, well known, code-driven mutation operators (e.g., the sufficient set [21, 22]) simulate algorithmic faults by introducing small changes into the source code and are thus unlikely to simulate interoperability problems resulting in exchanges of erroneous data.

Finally, traditional mutation analysis approaches *cannot inject faults into black-box components* whose implementation is not tested within the development environment (e.g., because it is simulated or executed on the target hardware). For example, in a satellite system, such components include the control software of the Attitude Determination And Control System (ADCS), the GPS, and the Payload Data Handling Unit (PDHU). During testing with simulators in the loop, the results generated by such components (e.g., the GPS position) are produced by a simulator. During testing with hardware in the loop, these components are directly executed on the target hardware and cannot be mutated, either because they are off-the-shelf components or to avoid damages potentially introduced by the mutation.

Last, test generation approaches are in preliminary stages and cannot be applied in industrial space context. For example, SEMU, a state-of-the-art approach can generate test inputs only for batch programs that can be compiled with the LLVM infrastructure.

FAQAS had the objective to assess the feasibility of mutation analysis and testing for space software by identifying feasible solutions based on existing literature. FAQAS objectives were the following:

- To perform a comprehensive analysis and survey of mutation analysis/testing.
- To prototype the mutation analysis/testing process to be applied on space software.
- To develop a toolset supporting mutation analysis and testing automation.
- To empirically evaluate mutation analysis/testing by applying it to space software use cases.
- To evaluate how mutation analysis/testing can be integrated into a typical verification and validation life cycle of space software and to define a mutation analysis/testing methodology.

Overview of the contributions

FAQAS led to the following contributions:

- A survey of the literature on mutation analysis/testing, which is summarized in section 1.
- The development of a toolset that includes the following tools:
  - MASS (Mutation Analysis for Space Software), a tool that automatically executes code-driven mutation analysis. Code-driven mutation analysis consists of automatically generating mutants

    by altering the source code of the software under test. MASS implements a pipeline that makes it feasible in the context of space software.

  - DAMAt (DAta-driven Mutation Analysis with Tables), a tool that automatically executes data-driven mutation analysis. Data-driven mutation analysis is an approach newly defined within FAQAS, which, instead of mutating the implementation of the software under test, alters the data exchanged by software components. Data-driven mutation analysis enables the injection of faults that affect simulated components (e.g., sensors), which is not feasible with traditional, code-driven mutation analysis.
  - SEMuS (Symbolic Execution-based MUtant analysis for Space software), a tool that automatically generates test inputs based on code-driven mutation analysis results.
  - DAMTE, a tool-supported methodology for the generation of test inputs based on data-driven mutation analysis results.

- An extensive empirical evaluation demonstrating the feasibility, effectiveness, and scalability of the proposed approaches with space software.
- The definition of guidelines for the adoption of mutation analysis and testing strategies within ECSS activities. The proposed guidelines support both quality assurance activities described in ECSS standards and Independent Software Verification and Validation (ISVV) practices.

TODO: FAQAS.drawio.pdf

Chapter 1

State-of-the-art

This chapter briefly discusses the applicability, in the context of space software, of existing solutions related to the four contributions of FAQAS: code-driven mutation analysis, code-driven mutation testing, data-driven mutation analysis, data-driven mutation testing.

1.1 Code-driven mutation analysis

Mutation analysis can drive the generation of test cases, which is referred to as ***mutation testing*** in the literature. A detailed overview of mutation testing and analysis solutions and optimizations can be found in recent surveys [23, 24].

1.1.1 Mutation Adequacy and Mutation Score computation

A mutant is said to be killed if at least one test case in the test suite fails when exercising the mutant. Mutants that do not lead to the failure of any test case are said to be live. Three conditions should hold for a test case to kill a mutant: *reachability* (i.e, the test case should execute the mutated statement), *necessity* (i.e., the test case should reach an incorrect intermediate state after executing the mutated statement), and *sufficiency* (i.e., the final state of the mutated program should differ from that of the original program) [25].

The mutation score, i.e., the percentage of killed mutants, is a quantitative measure of the quality of a test suite. Recent studies have shown that achieving a high mutation score improves significantly the fault detection capability of a test suite [26], a result which contrasts with that of structural coverage measures [27]. However, a very high mutation score (e.g., above 75%) is required to achieve a higher fault detection rate than the one obtained with other coverage criteria, such as statement and branch coverage [27]. In other words, there exists a strong association between a high mutation score and a high fault revelation capability for test suites.

The capability of a test case to kill a mutant also depends on the observability of the program state. To overcome the limitations due to observability, different strategies to identify killed mutants can be adopted; they are known as strong, weak, firm, and flexible mutation coverage [28]. With strong mutation, to kill a mutant, there shall be an observable difference between the outputs of the original and mutated programs. With weak mutation, the state (i.e., the valuations of the program variables in scope) of the mutant shall differ from the state of the original program, after the execution of the mutated statement [29]. With firm mutation, the state of the mutant shall differ from the state of the original program at execution points between the first execution of the mutated statement and the termination of the program [30]. Flexible mutation coverage consists of checking if the mutated code leads to object corruption [31]. For space software, we suggest to rely on strong mutation because it is the only criterion that truly assesses the fault detection capability of the test suite; indeed, it relies on a mutation score that reflects the percentage of mutants leading to test failures. With the other mutation coverage criteria, a mutant is killed if the state of the mutant after execution of the mutated statement differs from the one observed with the original code, without any guarantee that either the erroneous values in state variables will propagate or the test oracles will detect them.

1.1.2 Mutation Operators

Mutation analysis introduces small syntactical changes into the code (source code or machine code) of a program through a set of mutation operators that simulate programming mistakes.

The *sufficient set of operators* is widely used for conducting empirical evaluations [32, 33, 34, 35]. The original sufficient set, defined by Offutt et al., is composed of the following operators: Absolute Value Insertion (ABS), Arithmetic Operator Replacement (AOR), Integer Constraint Replacement (ICR), Logical Connector Replacement (LCR), Relational Operator Replacement (ROR), and Unary Operator Insertion (UOI) [32]. Andrews et al. [34] have also included the *statement deletion operator* (SDL) [21], which ensures that every pointer-manipulation and field-assignment statement is tested.

The sufficient set of operators enables an accurate estimation of the mutation score of a test suite [36]; furthermore, the mutation score computed with the sufficient set is a good estimate of the fault detection rate (i.e., the portion of real faults discovered) of a test suite [34, 37].

However, empirical work has shown that, to maximize the detection of real faults, a set of operators should be used in addition to the sufficient set: Conditional Operator Replacement (COR), Literal Value Replacement (LVR), and Arithmetic Operator Deletion (AOD) [38].

The SDL operator has inspired the definition of mutation operators (e.g., *OODL operators*) that delete portions of program statements, with the objective of replacing the sufficient set with a simpler set of mutation operators. The OODL mutation operators include the delete Arithmetic (AOD), Bitwise (BOD), Logical (LOD), Relational (ROD), and Shift (SOD) operators. Empirical results show that deletion operators produce significantly fewer equivalent mutants[1] [21, 22] and, furthermore, test suites that kill mutants generated with both SDL and OODL operators kill a very high percentage of all mutants (i.e., 97%) [22].

Another alternative to the sufficient set of operators is the generation of *higher order mutants*, which result from the application of multiple mutation operators for each mutation [40, 41, 42, 43]. However, higher order mutants are easier to kill than the first order ones (i.e., less effective to assess test suites limitations) [44, 24], and there is limited empirical evidence regarding which mutation operators should be combined to resemble real faults and minimize the number of redundant mutants [24].

1.1.3 Compile-time Scalability

The potentially large size of the software under test, combined with the large number of available mutation operators, may make the compilation of all mutants infeasible.

To reduce the number of invocations to the compiler to one, *mutant schemata* include all the mutations into a single executable [45]. With mutant schemata, the mutations to be tested are selected at run-time through configuration parameters. This may lead to a compilation speed-up of 300% [46].

Another solution to address compile-time scalability issues consists of *mutating machine code* (e.g., binary code [47], assembly language [48], Java bytecode [49], and .NET bytecode [50]), thus avoiding the execution of the compilation process after creating a mutant. A common solution consists of mutating the LLVM Intermediate Representation (IR) [51], which enables the development of mutants that work with mul-

[1] For example, statement deletion can lead to equivalent mutants only if statements are redundant, which is unlikely [39].

tiple programming languages [52] and facilitates the integration of optimizations based on dynamic program analysis [53].

Unfortunately, the mutation of machine code may lead to mutants that are not representative of real faults (i.e., faults caused by human mistakes at development time) because they are impossible to generate from the source code [53]. For instance, a function invocation in the source code may lead to hundreds of machine code instructions (e.g., the function call *std::vector::push_back* leads to 200 LLVM IR instructions) and, consequently, some of the mutants derived from such instructions cannot be derived by mutating the source code. In the case of IR mutation, some of these impossible mutants can be automatically identified [53]; however, the number of generated mutants tend to be higher at the IR level than at the source code level, which may reduce scalability [52]. In addition, we have encountered three problems that prevented the application of mutation analysis tools based on LLVM IR to our case study systems. First, space software relies on compiler pipelines (e.g., RTEMS [54]) that include architecture-specific optimizations not supported by LLVM. Second, there is no guarantee that the executables generated by LLVM are equivalent to those produced by the original compiler. Third, efficient toolsets based on LLVM often perform mutations dynamically [53], which is infeasible when the software under test needs to be executed within a dedicated simulator, a common situation with space software and many other types of embedded software in cyber-physical systems.

1.1.4 Runtime Scalability

A straightforward mutation analysis process consists of executing the full test suite against every mutant; however, it may lead to scalability problems in the case of a large software under test (SUT) with expensive test executions. *Simple optimizations* that can be applied to space software consist of (S1) stopping the execution of the test suite when the mutant has been killed, (S2) executing only those test cases that cover the mutated statements [55], and (S3) rely on timeouts to automatically detect infinite loops introduced by mutation [24].

Split-stream execution consists of generating a modified version of the SUT that creates multiple processes (one for each mutant) only when the mutated code is reached [56, 57], thus saving time and resources. Unfortunately, it cannot be applied in the case of space software that needs to run with simulators because, in general, the hosting simulator cannot be forked by the hosted SUT.

Another feasible solution consists of *randomly selecting a subset of the generated mutants* [58, 12, 11]. Zhang et al. [11] empirically demonstrated that a random selection of 5% of the mutants is sufficient for estimating, with high confidence, the mutation score obtained with the complete mutants set. Further, they show that sampling mutants uniformly across different program elements (e.g., functions) leads to a more accurate mutation score prediction than sampling mutants globally in a random fashion. For large software systems that lead to thousands of mutants, random mutation analysis is the only viable solution. However, for very large systems such as the ones commonly found in industry, randomly selecting 5% of the mutants may still be too expensive.

Gopinath et al. estimate the number of mutants required for an accurate mutation score [12]. They rely on the intuition that, under the assumption of independence between mutants, mutation analysis can be seen as a Bernoulli experiment in which the outcome of the test for a single mutant is a Bernoulli trial (i.e., mutant successfully killed or not) and, consequently, the mutation score should follow a binomial distribution. They rely on Tchebysheff's inequality [59] to find a theoretical lower bound on the number of mutants required for an accurate mutation score. More precisely, they suggest that, with 1,000 mutants, the estimated mutation score differs from the real mutation score at most by 7 percentage points. However, empirical results show that the binomial distribution provides a conservative estimate of the population variance and, consequently, 1,000 mutants enable in practice a more accurate estimate ($> 97\%$) of the mutation score than expected.

In the statistics literature, the correlated binomial model [60], and related models [61, 62, 63] are used when Bernoulli trials are not independent [64]. In our work, based on the results achieved by Gopinath et al., we assume that the degree of correlation between mutants is limited and the binomial distribution can be used to accurately estimate the mutation score.

The statistics literature also provides a number of approaches for the computation of a sample size (i.e., the number of mutants, in our context) that enables estimates with a given degree of accuracy [65, 66, 67, 68]. For binomial distributions, the most recent work is that of Gonçalves et al. [69], that determines the sample size by relying on heuristics for the computation of confidence intervals for binomial proportions. A confidence interval has a probability p_c (the confidence level) of including the estimated parameter (e.g., the mutation score). Results show that the largest number of samples required to compute a 95% confidence interval is 1,568.

If used to drive the selection of mutants, both the approaches of Gopinath et al. and Gonçalves et al., which suggest sampling at least 1,000 mutants, may be impractical when mutants are tested with large system test suites.

An alternative to computing the sample size before performing an experiment is provided by sequential analysis approaches, which determine the sample size while conducting a statistical test [70]. Such approaches do not perform worst-case estimates and may thus lead to smaller sample sizes. For example, the sequential probability ratio test, which can be used to test hypotheses, has been used in mutation analysis as a condition to determine when to stop test case generation (i.e., when the mutation score is above a given threshold) [71]. In our context, we are interested in point estimation, not hypothesis testing; in this case, the sample size can be determined through a fixed-width sequential confidence interval (FSCI), i.e., by computing the confidence interval after every new sample and then stop sampling when the interval is within a desired bound [72, 73, 74]. Concerning the method used to compute the confidence interval in FSCI, the statistics literature [72] reports that the Wald method [75] minimizes the sample size but requires an accurate variance estimate. We will therefore resort to a non-parametric alternative, which is Clopper-Pearson [76]. Note that FSCI has never been applied to determine the number of mutants to consider in mutation analysis.

Other solutions to address *runtime scalability problems* in mutation analysis aim to *prioritize test cases* to maximize the likelihood of executing first those that kill the mutants [77, 78, 13]. The main goal is to save time by preventing the execution of a large subset of the test suite, for each mutant. Previous work aimed at prioritizing faster test cases [77] but this may not be adequate with system-level test suites whose test cases have similar, long execution times. Approaches that rely on data-flow analysis to identify and prioritize the test cases that likely satisfy the killing conditions [78] are prohibitively expensive and are unlikely to scale to large systems. Other work [13] combines three coverage criteria: (1) the number of times the mutated statement is exercised by the test case, (2) the proximity of the mutated statement to the end of the test case (closer ones have higher chances of satisfying the sufficiency condition) , and (3) the percentage of mutants belonging to the same class file of the mutated statement that were already killed by the test case. Criterion (3) is also used to reduce the test suite size, by only selecting the test cases above a given percentage threshold. Unfortunately, only criterion (1) seems applicable in our context; indeed, criterion (2) is ineffective with system test cases whose results are checked after long executions, while criterion (3) may be inaccurate when only a random, small subset of mutants is executed, as discussed above.

1.1.5 Detection of Equivalent Mutants

A mutant is equivalent to the original program when they both generate the same outputs for the same inputs. Although identifying equivalent mutants is an undecidable problem [9, 79], several heuristics have been developed to address it.

The simplest solution consists of relying on *trivial compiler optimisations* [80, 35, 24], i.e., compile both the mutants and the original program with compiler optimisations enabled and then determine whether their executables match. In C programs, compiler optimisations can reduce the total number of mutants by 28% [35].

Solutions that identify equivalent mutants based on *static program analysis* (e.g., concolic execution [81, 82] and bounded model checking [83]) show promising results (e.g., to automatically identify non-equivalent mutants for batch programs [82]) but they rely on static analysis solutions that cannot work with system-level test cases that execute with hardware and environment simulators in the loop. Indeed, (1) simulation results cannot be predicted by pure static analysis, (2) concolic execution tools, which rely on LLVM, cannot be run if the SUT executable should be generated with a specific compiler (see Section 1.1.3), (3) there are no solutions supporting the concolic execution of large software systems within simulation environments (state-of-the-art techniques work with small embedded software [84]), and (4) communication among components not based on direct method invocations (e.g., through network or databases) is not supported by existing toolsets.

Alternative solutions rely on *dynamic analysis* and compare data collected when testing the original software and the mutants [14, 15, 16, 17]. The most extensive empirical study on the topic shows that nonequivalent mutants can be detected by counting the number of methods (excluding the mutated method) that, for at least one test case, either (1) have statements that are executed at a different frequency with the mutant, (2) generate at least one different return value, or (3) are invoked at a different frequency [16]. To determine if a mutant is non-equivalent, it is possible to define a threshold indicating the smallest number of methods with such characteristics. A threshold of one identifies non-equivalent mutants with an average precision above 70% and an average recall above 60%. This solution outperforms more sophisticated methods relying on dynamic invariants [17]. Also, coverage frequency alone leads to results close to the ones achieved by including all three criteria above [16]. However, such approaches require some tailoring because collecting all required data (i.e., coverage frequency for every program statement, return values of every method, frequency of invocation of every method) has a computational and memory cost that may break real-time constraints.

1.1.6 Detection of Redundant Mutants

Redundant mutants are either *duplicates*, i.e., mutants that are equivalent with each other but not equivalent to the original program, or *subsumed*, i.e., mutants that are not equivalent with each other but are killed by the same test cases.

Duplicate mutants can be detected by relying on the same approaches adopted for equivalent mutants.

According to Shin et al., subsumed mutants should not be discarded but analyzed to augment the test suite with additional test cases that fail with one mutant only [10]. The augmented test suite has a higher fault detection rate than a test suite that simply satisfies mutation coverage; however, with large software systems the approach becomes infeasible because of the lack of scalable test input generation approaches.

1.1.7 Benchmark of State-of-the-art, Code-driven Mutation Testing Toolsets

This section summarizes the outcome of an experiment performed to evaluate the applicability of state-of-the-art mutation testing tools in the space context, based on the case study systems of the project.

To carry out this preliminary evaluation of mutation testing tools, we selected a set of tools presented in the literature based on the following criteria:

- **Availability of source code.** To enable optimizations, the tool under analysis should be provided along with source code.
- **Applicability to C/C++ code.** The tool under analysis should be able to process C and C++ code.
- **Licence compatible with ESA Software Community Licence Permissive (ESA SCLP).** The licence of the tool under analysis, should enable redistributing the tool itself within the FAQAS framework, which is released under ESA SCLP.

- **Age.** To avoid problems due to support for recent libraries, we should prioritize tools that are recent and actively developed.

The first three criteria mentioned above constitute mandatory requirements. Tools not meeting these requirements are not selected for evaluation in our context because they cannot be integrated into the FAQAS framework.

Table 1.1: Summary of Data-Driven Mutation Testing Benchmarks.

Reference	Approach/Tool Name	Evaluation
Hariri & Shi 2018	SRCIRor	**Source code availability.** Yes, https://github.com/TestingResearchIllinois/srciror. **Applicability to C/C++ code.** Yes. **ESA SCLP Compatible.** Yes, released under NCSA, https://opensource.org/licenses/NCSA, which allows redistribution and re-licensing. **Age.** Aged, last update in September 2018. **Outcome. The tool is applicable in space context.**
Wang et al. 2017	Accmut	**Source code availability.** Yes, https://github.com/wangbo15/accmut/ **Applicability to C/C++ code.** Yes. **ESA SCLP Compatible.** Yes, released under NCSA, https://opensource.org/licenses/NCSA, which allows redistribution and re-licensing. **Age.** Aged, last update in January 2018. **Outcome. Depends on CLANG/LLVM, which prevents compilations for some sysems.**
Phan et al. 2018	MUSIC	**Source code availability.** Yes, https://github.com/swtv-kaist/MUSIC/ **Applicability to C/C++ code.** Yes. **ESA SCLP Compatible.** No. The software is licensed with proprietary licence. In private communication via e-mail, authors have shown to be available to relicensing, however this might not fit the budget of the project. **Age.** Recent, last update in July 2019.
Denisov & Pankevich 2018	Mull	**Source code availability.** Yes, https://github.com/mull-project/Mull **Applicability to C/C++ code.** Yes. **ESA SCLP Compatible.** Yes. Apache Licence 2.0, https://opensource.org/licenses/Apache-2.0. **Age.** Ongoing, last update in June 2020. **Outcome. The tool requires compilation with CLANG/LLVM, which leads to compilation errors with systems depending on RTEMS. Also, natively, Mull performs mutations on the fly through just-in-time compilation features, which is inapplicable if the SUT is executed within a simulator.**
Delgado et al. 2018	MuCPP	**Source code availability.** No, only executables are available https://ucase.uca.es/mucpp/
Jia & Harman 2008	Milu	**Source code availability.** Yes, https://github.com/yuejia/Milu/ **Applicability to C/C++ code.** Yes. **ESA SCLP Compatible.** Yes, released under NCSA licence, https://opensource.org/licenses/NCSA. **Age.** Aged, last update in April 2018. **Outcome. The tool generates a preprocessed source code that does not compile.**
Brannstrom et al. 2015	Dextool	**Source code availability.** Yes, https://github.com/joakim- brannstrom/dextool **Applicability to C/C++ code.** Yes. **ESA SCLP Compatible.** Yes, released under Mozilla public Licence 2.0, https://opensource.org/licenses/MPL-2.0. **Age.** Ongoing, last update in June 2020. **Outcome. Depends on CLANG/LLVM, which prevents compilations for some sysems.**
Delamaro et al. 2001	Proteum	**Source code availability.** Yes, https://github.com/magsilva/proteum. **Applicability to C/C++ code.** Yes. **Age.** Aged, last update December 2015.
Shariar and Zulkernine 2008	Function Calls Mutation	**Source code availability.** No.
Dans & Hierons 2001	Floating-point Mutation	**Source code availability.** No.

Table 1.1 provides the list of selected tools along with the evaluation results. We do not evaluate all the criteria when one of the mandatory requirements is not met. For what it concerns the compatibility with the ESA Software Community Licence Permissive, we consider the licenses NCSA and Apache Licence 2.0 compatible. Indeed, both the two licences allow for redistribution of the software, a condition that is sufficient to release a mutation testing tool as component of the FAQAS framework.

For our evaluation we then selected the five most recent tools that fulfill our mandatory requirements: SRCIRor, Mull, Dextool, Accmut, and Milu. Proteum has been discarded because its latest stable version dates back to December 2015; on May 2020 a few changes had been made on Proteum GitHub repository, however, the up to date version is indicated by its developer as not usable.

To evaluate the applicability of existing mutation testing tools to space software, we evaluated each mutation testing tool considered in our study against the same case study system of the project, i.e., the System Test Suite for ESAIL provided by LXS. We selected this case study system, because (1) it is the largest case study system of FAQAS in terms of lines of code, (2) the ESAIL system test suite requires that the full software is compiled and all the required libraries linked (this may complicate the use of tools that cannot parse all the source code), (3) the software under test (SUT) is executed within a system emulator (SVF) that requires the SUT to respect its real-time constraints. ESAIL consists of 924 source files (719 files with extension ".c" and 205 with extension ".h"). In total, it consists of 74,161 LOC. ESAIL is compiled with sparc-rtems4.8-gcc, a tailored version of the gcc compiler for sparc systems, the compiler is provided by Cobham Gaisler[2].

To draw a final outcome for or evaluation (see Table 1.1), we applied each selected tool to ESAIL and verified if the mutation testing tool could successfully create mutated version of ESAIL that can be compiled and executed within the SVF. Out of all the selected tools, only SRCIRor had been successfully applied to ESAIL.

1.1.8 Summary

We aim to rely on the sufficient set of operators since it has been successfully used to generate a mutation score that accurately estimates the fault detection rate for software written in C and C++, languages commonly used in embedded software. Further, since recent results have reported on the usefulness of both LVR and OODL operators to support the generation of test suites with high fault revealing power [38], the sufficient set may be extended to include these two operators as well.

To speed up mutation analysis by reducing the number of mutants, we should consider the SDL operator alone or in combination with the OODL operators. However, such heuristic should be carefully evaluated to determine the level of confidence we can expect.

Among compile time optimizations, only mutant schemata appear to be feasible with space software. Concerning scalability, simple optimizations (i.e., S1, S2, and S3 in Section 1.1.4) are feasible. Alternative solutions are the ones relying on mutant sampling and coverage metrics. However, to be applied in a safety or mission critical context, mutant sampling approaches should provide guarantees about the level of confidence one may expect. Currently, this can only be achieved with approaches requiring a large number of sampled mutants (e.g., $1,000$). Therefore, sequential analysis based on FSCI, which minimizes the number of samples and provides accuracy guarantees, appears to be the most appropriate solution in our context. Further, test suite selection and prioritization strategies based on code coverage require some tailoring to cope with real time constraints.

Equivalent mutants can be identified through trivial compiler optimizations and the analysis of coverage differences; however, it is necessary to define and evaluate appropriate coverage metrics. The same approach can be adopted to identify duplicate mutants. The generation of test cases that distinguish subsumed mutants is out of the scope of this work.

A high-level description of a possible mutation testing pipeline was proposed in a recent survey[3] [24]. It consists of the following sequence of activities: select (sample) mutants, compile mutants, remove equivalent and redundant mutants, generate test inputs that kill mutants, execute mutants, compute mutation score, reduce test suites and prioritize test cases. Unfortunately, such pipeline does not enable the integration of many optimizations proposed above, which further motivates our work. For example, it cannot support FSCI-based sampling, which requires mutants sampling to be coupled with mutants execution. Also, it does not envision the detection of equivalent and redundant mutants based on code coverage. Moreover, it only partially addresses scalability issues since test suite reduction and prioritization are performed after mutation analysis. Further, it includes a test input generation step that is not feasible in the context of CPS. Finally, it has never been implemented and therefore its feasibility has not been evaluated.

[2]https://www.gaisler.com/index.php/products/operating-systems/rtems

[3]The main objective of such pipeline was to walk the reader through the survey, not to propose a precise and feasible solution.

1.2 Data-driven Mutation Analysis

Data-driven mutation analysis evaluates the effectiveness of a test suite in detecting ***interoperability faults***. The CPS literature reports on four different interoperability types [18]: technical (which concerns communication protocols and infrastructure), syntactic (which concerns data format), semantic (which concerns the exchanged information, that is, errors in the processing of exchanged data), and cross-domain interoperability (which concerns interaction through business process languages such as BPEL [85]). For example, a technical interoperability problem may concern two components working with two different network protocols (e.g., TCP VS UDP), a syntactic interoperability problem is that of two components using different keywords to specify a field in a json [86] data file (e.g., ”temperature” and ”temp”), a semantic interoperability problem is that of a control software that does not take appropriate actions when the voltage of the board is above nominal range, cross domain interoperability is that of a web service not following the expected flow of remote calls. Technical and syntactic interoperability are provided by off-the-shelf hardware and libraries (not tested by CPS developers) while cross-domain interoperability concerns systems integrated in online services (e.g., energy plants) but is out of scope for the type of CPSs we target in this work, which are safety-critical CPSs like flight systems, robots, and automotive systems. In this paper, we thus focus on ***semantic interoperability*** faults, that is, faults that affect CPS components integration and are triggered (i.e., lead to failures) in the presence of specific subsets of the data that might be exchanged by CPS components. We thus aim to ensure that a test suite fails when the data exchanged by CPS components is not the one specified by test cases (e.g., through simulator configurations). Related work includes mutation analysis [23, 24] and ***fault injection*** [87] techniques.

P-1

1.2.1 Mutation Analysis

Mutation analysis concerns the automated generation of faulty software versions (i.e., mutants) through automated procedures called mutation operators [23, 24]. The effectiveness of a test suite is measured by computing the mutation score, which is the percentage of mutants leading to failures when exercised by the test suite.

Mutation operators introduce syntactical changes into the code of the SUT. The ***sufficient set of operators*** is implemented by most mutation analysis toolsets [32, 33, 34, 35, 32]. Unfortunately, these operators simulate faults concerning the implementation of algorithms (e.g., a wrong logical connector), which is usually tested in unit test suites that, by definition, do not exercise the communication among components, our target in this paper. Also, as stated in the Introduction, such operators can’t be used to generate faulty data with simulated or off-the-shelf components. ***Higher-order*** mutation analysis [88], which simply combines multiple operators, has the same limitations.

Components integration is targeted by interface [89], integration [90], contract-based [91], and system-level mutation analysis [92]. The former three assess the quality of integration test suites by introducing changes that concern function invocations (e.g., switch function arguments) and inter-procedural data-flow (e.g., alter assignments to variables returned to other components); they can simulate integration faults in units integrated with API invocations but not interoperability problems concerning larger components communicating through channels (e.g., network). ***System-level mutation*** relies on operators for GUI components, which are out of our scope, and configuration files, by applying simple mutations, such as deleting a line of text, and are unlikely to lead to interoperability problems.

1.2.2 Fault injection

Fault injection techniques simulate the effect of faults by altering, at runtime, the data processed by the SUT [87]. Faults are introduced according to a fault model that describes the type of fault to inject, the timing of the injection, and the part of the system targeted by the injection. Different from data-driven mutation analysis, fault injection techniques aim to stress the robustness of the software, not assess the quality of its test suites.

Faults affecting components' communication, CPU, or memory can simulated by performing bit flips [93, 94, 95, 96]. ***Communication faults*** are simulated also by duplicating or deleting packets, altering their sequence, or introducing incorrect identifiers, checksums, or counters [97, 98]. Faults affecting ***signals*** can be simulated by shifting the signal or increasing the number of signal segments [99]. The largest set of faults affecting data exchanged through files or byte streams is simulated by Peach [100], which includes also protocol-specific fault injection procedures such as replacing host names with randomly generated ones. In general, although existing techniques may simulate a large set of faults they do not cover all the CPS interoperability faults (see Section **??**).

Approaches performing fault injections other than bit flips require a model of the data to modify. The modelling formalisms adopted for this purpose are grammars [101, 102, 103, 104], UML class diagrams [97, 98], or block models [105, 100]. Grammars are used to model textual data (e.g., XML), which is seldom exchanged by CPS components because of parsing cost. Block models enable specifying the representation to be used for consecutive blocks of bytes, which makes them applicable to a large set of systems; however, existing block model formalisms rely on the XML format, which is expensive to process and thus not usable with real-time systems [105, 100]. The ***UML class diagram*** is a formalism that enables the specification of complex data structures and data dependencies [97, 98]; however, it requires loading the data as UML class diagram instances, which is too expensive for real-time systems.

1.2.3 Summary

To summarize, the modification of the data exchanged by software components enables the simulation of communication and, therefore, semantic interoperability faults. Test suites can thus be assessed by relying on fault injection techniques to mutate data. However, existing fault injection techniques do not target mutation analysis; consequently, we lack methods for the specification of fault models and metrics for the assessment of test suites. Also, a larger set of procedures for the modification of data is needed. Finally, block models can effectively capture the structure of the data to modify but formalisms not relying on XML are needed. Our paper addresses such limitations.

1.3 Code-driven Mutation Testing

This section describes the approaches that can be adopted to automatically generate test cases that kill mutants. To kill a mutant we need test cases that (1) reach the mutation point (i.e., execute the mutated code), (2) cause corruptions in the program state right after the mutated code, and (3) manifest these corruptions into the program output (e.g., by producing an erroneous value in a state variable verified by a test assertion) thus leading to a failure [24]. These conditions are also known as the ***killing conditions*** of a mutant.

In the literature, there exist two groups of approaches for generating test cases that kill mutants: approaches based on constraint-programming, and approaches based on evolutionary computation. Below we introduces these two stream of approaches after introducing two well-known solutions for test generation driven by structural coverage, which are the basis for mutation testing solutions based on constraint programming.

1.3.1 Test Generation driven by program structure

Two alternative state-of-the-art solutions to generate test inputs that maximize structural coverage are CBMC, a bounded model checker, and KLEE [106], a symbolic execution engine. They are detailed below.

CBMC is an approach that implements ***Bounded Model Checking*** [107, 108] (BMC), an approach for purely static software verification. The idea in BMC is to represent the software together with the properties to be verified as an instance of the propositional satisfiability problem (SAT). Such a representation captures the software behavior exactly, assuming that all the loop bodies in the software are repeated at most a fixed number of times. This approach has several advantages: the logical formulation is usually very compact compared to traditional model checking, where verification is reduced to a reachability problem in a graph representing the program state space; there are several high-performance SAT solvers [109, 110] that can be used for solving the instances; and the satisfying assignments of an instance can be directly translated to meaningful counterexamples for correctness in the form of fault-inducing executions. Furthermore, it is widely recognized that BMC based approaches are particularly good at quickly finding short counterexamples when they exist.

KLEE [106] is an open source tool that implements ***Concolic Execution***, a technique that performs ***symbolic execution*** along a concrete execution path. KLEE was designed to automatically generate test cases that achieve high code coverage. The tool is implemented as a modified LLVM virtual machine that targets LLVM bytecode programs. KLEE also provides a symbolic POSIX library that enable analysis of programs that uses the system's environment. For example, KLEE can be executed with the flag `-sym-stdin N` which will make stdin symbolic with size `N`.

In FAQAS, we rely on KLEE to automatically produce the inputs that make the mutated version of the program generate a different output than the original version.

1.3.2 Test Generation based on Constraint Programming

Techniques based on constraint programming use some form of automated reasoning (e.g., Propositional Satisfiability or Constraint Solving [111]) to derive data that satisfy all the conditions necessary to kill a mutant [25].Existing approaches, differ for the strategy adopted to automatically generate these constraints from the program under test.

Offutt et al. [25], for example, automatically derive such constraints from the program by extracting the predicate expressions on the program's control flow graph. Then, such constraints are encoded to form a constraint system. In their approach, they propose three strategies for identifying infeasible constraint systems, the contradictions to such systems are the new test cases for the program under analysis.

```
 1  int midval (int x, int y, int z) {
 2    int midval;
 3
 4    midval = z;
 5    if (y < z) {
 6      if (x < y) {
 7        midval = y;
 8      }
 9      else if (x < z)
10   Δ else if (x <= z) {
11        midval = x;
12      }
13    }
14    else
15      if (x > y) {
16        midval = y;
17      } else if (x > z) {
18        midval = x;
19      }
20    return midval;
21  }
```

Listing 1.1: midval function returns the mid value between three integers.

In the following, we introduce an example of the application of Offutt's [25] approach by using the `midval` function presented in Listing 1.1. This function has a mutation on line 9, which has been mutated into line 10. According to their approach, the three killing conditions would be the following:

- Reachability C_R : $(y < z) \wedge (x \geq y)$
- Necessity C_N : $(x < z) \neq (x \leq z)$
- Sufficiency C_S : Output(P) $\neq$ Output(M)

C_R defines the condition required to reach the mutated statement, in this case the conjunction between the predicate of the first if condition and the negation of the second if condition. C_N defines the condition required to assure a different program state between the original and mutated version of the program right after the mutation point. Finally, C_S defines the condition necessary to demonstrate that both the original and the mutated program returns different values.

Holling et al. [81] proposed to use a ***symbolic execution*** approach to identify new test cases. Their idea is to first execute symbolically both the original and the mutated function, then to check if their return values are equivalent or not. Symbolic execution determines what inputs cause each part of a function to be covered during execution. To symbolically execute a function, it is necessary to replace the original inputs (i.e., concrete values) with symbolic ones. The ***symbolic values*** represent a set of possible concrete values that lead to a certain program path (i.e., path condition).

Holling's approach [81] to automatically identify equivalent mutants relies on the observation that two mutants are equivalent when there are no concrete values making the mutated function produce an output that is different from the one of the original function. If a value that makes the two functions generate distinct results can be found, the mutant is non-equivalent. To automate the generation of inputs, Holling et al. rely on KLEE [106].

We introduce an example of Holling's approach in Listing 1.2. The top part of Listing 1.2 shows the function `isPositive`, which checks if an integer number is positive or not. The bottom part of Listing 1.2 presents the mutated version of `isPositive`, where the relational operator $\geq$ has been replaced by the operator $>$. To automate the generation of inputs using KLEE, all the parameters need to be treated as symbolic values. This is achieved by function `make_symbolic` (see Listing 1.3) which converts concrete variables to symbolic ones by considering their memory address and size. In Listing 1.3, the parameter `numSymbolic` is made symbolic in Line 3. Then, the original and mutated functions are called using the symbolic arguments in Lines 5 and 6. Finally, we need to introduce an assertion that makes the symbolic execution engine ***look for inputs that make the output of the two functions different***. In Listing 1.3, this is achieved with an assertion that verifies that the return values of the two functions the same (see Line 8). Despite being counter-intuitive, this approach is effective because symbolic execution engines aim to identify inputs that falsify the assertions in the program. When the equality is falsified, then the two functions can produce a different output for a same input. The input that falsifies the equality can thus be used to improve the test suite enabling it to kill the mutant. In the example of Listing 1.3, KLEE will indicate that the return values of the original and mutated function differ when `num` is equal to zero. A new test case exercising function `isPositive` with `num=0` should thus be added to the test suite in order to kill the mutant.

```c
int isPositive(int num){
  if (num >= 0){
    return 1;
  } else {
    return 0;
  }
}

int MUT_isPositive(int num){
  if (num > 0){
    return 1;
  } else {
    return 0;
  }
}
```

Listing 1.2: isPositive and MUT_isPositive functions

```c
void generate_test_input () {
  int numSymbolic;
  make_symbolic(&numSymbolic, sizeof(numSymbolic), "numSymbolic");

  int original_ret = isPositive(numSymbolic);
  int transformed_ret = MUT_isPositive(numSymbolic);

  assert(original_ret == transformed_ret);
}
```

Listing 1.3: Holling's approach for test case generation.

```c
void generate_test_input () {
  int numSymbolic;
  make_symbolic(&numSymbolic, sizeof(numSymbolic), "numSymbolic");

  int original_ret = isPositive(numSymbolic);
  int transformed_ret = MUT_isPositive(numSymbolic);

  assert(original_ret != transformed_ret);
}
```

Listing 1.4: Test case generation with assertion that reflects the desired behaviour.

Similarly to Holling's approach, Riener et al. [83] proposed to use ***bounded model checking*** techniques to search for these counter examples. In their bounded model checking approach, the original program and the mutant are unrolled with respect to a certain maximum bound. In program unrolling, loops are re-written as a repeated sequence of similar independent statements. Then, both unrolled programs are encoded into a logic formula over the same input variables. To ensure that the mutation affects the output of the mutant, a propagation condition is encoded and added to the previous logic formula, the condition asserts that there exist at least one pair of different outputs under the assumption of equal inputs. In the last step, the formula is processed by a SMT-solver, if the solver finds a satisfying assignment, the inputs of the formula are translated into a new test case for the current program under analysis.

SEMu [112] is a recent mutantion testing framework based on dynamic symbolic execution that has been built on top of the KLEE Symbolic Virtual Machine [106]. SEMu uses a form of ***differential symbolic execution*** [113] to generate test inputs that kill mutants. The approach consists of modeling the mutant killing problem as a symbolic execution search in a scalable and cost-effective way. The SEMu framework is the building block of the test generation tool developed in FAQAS (i.e., ***SEMuS***). Different from the approaches presented above, SEMu can generate test inputs that kill mutants without the need of human intervention (e.g., to ad assertions); however, it targets only bash programs, it cannot generate unit test cases, for example.

In SEMu, the multiple program versions (i.e., the original and the mutated programs) are compiled within the same executable. Because mutants differ only slightly from the original program (minor syntactic differences), the technique performs one symbolic execution search for the unchanged source code, and then it forks the symbolic execution search every time it reaches a mutated statement.

The technique follows both the mutated and the original execution and compares the state (i.e., the value of the variables) at each step of the execution. More precisely, every time the execution reaches the statement after the mutation, the technique triggers the generation of test inputs for the forked and the original programs by comparing their symbolic states. To generate test inputs that kill the mutant, the technique encodes the three killings conditions, obtained from the symbolic execution search, into a single formula (the $kill$ formula) to be passed to a constraint solver.

To speed-up test generation, SEMu integrates a number of optimizations:

- ***Meta-mutation***: all mutants are encoded into a single program called meta-mutant (in FAQAS, we extended it at source code level), the mutant is selected through a branching statement named mutant choice statement.
- **Discarding non-infected mutant paths**: mutant paths failing to infect the program state are discarded immediately.
- **Heuristic search**: stop exploration of a path after K transitions, and solve the kill formula.
- **Infection-only strategy**: Generates test inputs by aiming only at mutant infection.

1.3.3 Test Generation based on Evolutionary Computation

Test generation approaches based on ***evolutionary computation*** typically rely on population-based meta-heuristic optimization algorithms [114]. They search for program inputs that could kill mutants under the guidance of a fitness function [114]. The main research contribution of these methods is the definition of fitness functions that capture the killing conditions of a mutant and identify test inputs that satisfy those conditions.

The ***fitness function*** captures the killing conditions of a mutant. For instance, Ayari et al. [115] proposed to use an evolutionary approach based on ***ant colony optimization*** (ACO) for automatic test input data generation on mutation testing. The ACO is an optimization algorithm inspired by the behavior of ants, it is based on the ants ability to find the shortest path between their nest and the food source. In the study by Ayari et al. [115], the approach takes an existing test case and produces a new test case by slightly modifying its inputs. The fitness function measures the distance between the mutated statement, and the statement reached by the new test case (e.g., the reachability condition). More precisely, the distance is defined as the number of basic blocks between the two statements in the program's control flow graph. Papadakis et al. [78], instead, rely on fitness functions that capture the distance between mutated statement and the statement covering the branches of the different mutations (e.g., the necessity condition).

Fraser and Arcuri [116] propose to use distinct distance metrics tailored to the specific operator used to generate the mutants. This tailoring is needed because the ***necessity killing condition*** relies on changes in the program state and the execution of a mutated statement does not guarantee that the program state had been changed (i.e., values on the stack are different at the mutation point). For example, the *deletion operator*, which removes a statement, may or may not change the program state, depending on the semantic of the removed sentence (e.g., a logging instruction does not alter the program state). In case the mutation effectively changes the program state the distance is set to 0, otherwise the given value is 1. In the case of the *insert unary operator*, which adds or subtracts 1 to a numerical value, the operator always change the program state, so the distance is set to 0 when the statement is reached.

Instead, in the case of the *replace variable operator*, which replaces a specific variable with all other variables of the same type in the program scope, the distance is set to 0 only if the values of the variables being exchanged are different before executing the statement, otherwise it is set to 1.

1.3.3.1 Generation of test oracles

The automated generation of test oracles is a research topic that goes beyond the specific needs of mutation testing [117, 118].

Fraser et al. provide an overview of existing approaches for the automated generation of test oracles that have been integrated into existing test case generation tools [119]. A common solution consists of the automated synthesis of assertions for the test case. These assertions reflect the output generated by the function under test when it is exercised with the automatically generated input. For example, Line 6 in Listing 1.5 shows the oracle that can be automatically generated for the function analyzed in Listing 1.3 (i.e., function *isPositive*). The oracle, in this case, consists of an assertion verifying that the value generated by function *isPositive* matches the value '1', which is the value observed during test generation for the input value '0'. This is the approach implemented by Riener et al. [83], who generate assertions that verify variables that present different values between the original and the mutated executions.

```
1 void test () {
2   int numSymbolic = 0;
3
4   int ret = isPositive(numSymbolic);
5
6   assert( ret == 1 );
7 }
```

Listing 1.5: Automatically generated oracle for function 'isPositive'.

Randoop [120] allows annotation of the source code to identify observer methods to be used for assertion generation. Orstra [121] generates assertions based on observed return values and object states. DiffGen [122] extends the Orstra approach to generate assertions from runs on two different program versions.

A well known limitation of automatically generated assertions that reflect the actual values observed during execution is that they need to be validated. More precisely, we need to ensure that the values expected by the assertions do not reflect a failure triggered by the test case (e.g, an erroneous value being returned). Such validation activity is typically performed manually by the engineers because it should be based on domain knowledge and system specifications. Specifications are generally written in natural language because, to reduce development costs, only few components of the system are specified using formal languages. For this reasons the automated verification of such assertions is infeasible.

Approaches that support engineers in the analysis of generated oracles exist and might be considered to speed up the process [123, 124]. For example, Staats et al. [123] identify the subset of variables to be verified by oracles in order to maximize the fault finding potential of the testing process. First, they generate a collection of mutants from the SUT. Second, the test suite (automatically generated) is run against the mutants using the original system as the oracle. Third, they select the variables to verify in test oracles by focussing on those variables that show different values in the original and the mutated version.

ZoomIn. [124] automatically identifies suspicious assertions. These are assertions that verify data that is generated by functions showing anomalies during test cases executions. Anomalies are detected by automatically deriving pre- and post-conditions of the functions of the SUT based on the data recorded during the execution of a manually implemented test suite. Anomalies consist of function executions that violate such pre- and post-conditions.

1.3.4 Summary

Among all the approaches for test input generation, SEMu is the most advanced one. However, it does not integrate solutions to automatically generate oracles. Solutions for generating oracles are at their infancy and cannot be considered ready for integration into the FAQAS toolset.

1.4 Data-driven Mutation Testing

Although the literature does not address the problem of automatically generating test cases for data mutation testing, in this section we provide references to work that can be reused for this purpose. We discuss both the generation of ***test inputs*** and ***test oracles***. Concerning the generation of test inputs, we group the applicable approaches according to the type of models used to specify the data to be generated: ***UML models***, ***grammars***, ***block models***, and ***no models***.

Automated test inputs generation aims to automatically generate data that can be altered through a mutation operator. In the context of system and integration testing, which is the target of data-driven mutation testing, test input data is provided through input interfaces. However, since data-driven mutation is applied to data exchanged different communication layers, including input interfaces and interfaces used for the communication among internal components, we may observe two possible situations. First, the data targeted by mutation testing coincide with the input data (i.e., it was not transformed internally). Second, the data targeted by data mutation is the result of a transformation of the input data. We thus need to consider both the two cases when discussing related work.

1.4.0.1 UML models

When the data exchanged on the communication layer targeted by mutation testing coincide with the input data, existing approaches based on constraint-solvers can be adopted. These approaches rely on a formal or semi-formal specification of the structure of the input data and the constraints among data fields. Some techniques target the generation of inputs whose data structure had been specified using a UML class diagram where the relations among data fields have been captured using OCL constraints. The data structure model is a UML class diagram and resembles the one reported in Figure **??**. The OCL language is instead used to capture all the constraints among data fields. Existing techniques in this category work by generating class diagram instances that satisfy a set of given OCL constraints by executing appropriate constraint solvers after having transformed the OCL constraints into other formalisms such as ***Alloy models*** [**?**], ***constraint satisfaction*** [**?**], ***SMT*** [**?**], or ***SAT*** problems [**?**].

Other approaches, instead, work with models specified in formats other than UML class diagrams: ***Java classes*** [**?**, **?**], ***constraint logic*** [**?**], ***Alloy*** [**?**], or ***Z specifications*** [**?**]. These techniques have been proven to be effective for testing software systems that process classical data structures like trees. Alloy is a modelling language for expressing complex structural constraints [**?**], which has been successfully used to generate test inputs for testing object-oriented programs [**?**]. Korat, instead, is a technique that enables the generation of data structures to test Java programs [**?**]. Given a bound to the input structures (i.e., the maximum number of instances for each class to be used), Korat exhaustively generates all the nonisomorphic structures that are valid. Some of the limitations of Korat include requiring the definition of an imperative predicate that evaluates the correctness of the generated structure, which could be complex in the case of complex data models, requiring the manual definition of an input bound for each non primitive attribute or association, which might be particularly expensive in case of complex data structure, not dealing with constraints defined over integers. A more efficient, black-box test generation approach is UDITA [**?**]. What contributes to the efficiency of UDITA is the combination of both generator methods and predicates. Generator methods are used to build instances of the data structure, while predicates are used to validate the generated instances. UDITA relies upon the Java Path Finder model checker [**?**] to generate all the instances that satisfy the given predicates. However, the implementation of these generator methods that define the complete structure of a complex data model instance and lead to realistic test inputs can be quite expensive.

A common limitation of solutions based on constraint-solvers is their scalability [**?**]. In the context of mutation testing, where existing test inputs are available, a possible solution to address the scalability problem may consist of generating new test inputs by ***regenerating only portions of existing test inputs***. For example, Di Nardo et al. [**?**] automatically generate test inputs for new requirements by adapting existing field data. This is achieved by combining model transformations with constraint solving. Despite empirical results show a huge performance gain with respect to traditional constraint-based approaches, the need for dedicated parsers to translate existing test inputs into class diagram instances may limit the applicability of the approach. Also, the available test inputs may not enable the generation of all the inputs data needed.

Other approaches address the scalability problem by relying on ***hybrid input generation approaches*** [**?**]. For example, PLEDGE [**?**] combines metaheuristic search and Satisfiability Modulo Theories (SMT [**?**]) to generate UML instance models from UML class diagrams annotated with OCL constraints. It works by using the Negation Normal Form (NNF [**?**]) to represent all the constraints derived from the UML data model. Different subformulas that build the NNF formula are then solved by combining metaheuristic search and SMT. Metaheuristic search is used to handle subformulas whose satisfaction involves structural tweaks to the instance model, i.e., additions and deletions of objects and links. SMT is used with subformulas involving only primitive attributes, i.e., attributes with primitive types.

When the data exchanged on the communication layer targeted by mutation is the result of a transformation of the input data, the only applicable solution consist in the application of approaches that generate inputs from scratch. By generating a large number of inputs that include instances for all the classes of the data model, these approaches may, in principle, lead to the generation of required data by internal components. However, without means to drive the generation of such data, the application of UML-based test input generation approaches in these situation is likely to be inefficient.

1.4.0.2 Grammars

When grammars are used to model the input, ***grammar-based test input generation*** approaches relying on the expansion of the production rules of the grammar can be adopted [**?**]. Available tools are shown in Table 1.2.

Table 1.2: List of tools for grammar-based inputs generation.

Name	Grammar	Licence	Description
GramTest [**?**]	BNF	Apache 2.0	Java-based tool that allows you to generate test cases based on BNF grammars. It covers all the production rules of the grammar.
Fuzzingbook Grammar Coverage-based Fuzzer [**?**]	BNF	MIT	Python tool that implements production rules coverage. It implements the Shortest Path Selection [**?**] optimization.
GP [**?**, **?**]	Annotated BNF	Proprietary	Tool for the automated generation of inputs from Stochastic Context Free Grammars. Implemented on top of genetic programming algorithms. https://selab.fbk.eu/kifetew/downloads/gplib-607.jar
RIDDLE [101]	BNF	Proprietary	Tool that adopts a grammar to describe the format of inputs; based on the grammar, random and boundary values are generated for tokens representing input parameters.
pFuzzer [**?**]	BNF	Tool that aims at producing valid inputs for input parsers `https://github.com/uds-se/pFuzzer`.	

Among existing approaches, ***Parser-Directed Fuzzing*** (hereafter, *pFuzzer*) aims at producing valid inputs for input parsers [**?**]. The challenge is to cover all the lexical and syntactical features of a certain language. The approach systematically produces inputs for the parser and tracks all the comparisons made; after every rejection, it satisfies the comparisons leading to rejections, effectively covering the input space. Evaluated on five subjects, from CSV files to JavaScript, the *pFuzzer* prototype covers more tokens than both lexical-based (AFL) and constraint-based approaches (KLEE).

Similarly to the case of UML-based test case generation approaches that generate inputs from scratch, these grammar-based approaches can be adopted when the communication layer targeted by mutation testing is either the input layer or an internal layer.However, they may suffer of inefficiency problems; in additions, grammar can be unlikely used to model the types of data processed by space software.

1.4.0.3 Block-models

Among the existing toolsets based on block-models, ***Peach*** is the only one which supports the automated generation of test data. However, it's implementation simply generates random data, the process is referred to as ***blind fuzzing***[4].

1.4.0.4 No models

The generation of test inputs without the need of data model is typically driven by code coverage. A representative solution is given by AFL, which has been introduced in Section **??**.

More in general, approaches that address the problem of automatically testing programs that process structured data can be adopted for this purpose [**?**, **?**, **?**]. For example, SUSHI [**?**] is a tool that aims to cover test objectives that depend on non-trivial data structure instances. It relies on symbolic execution to generate path conditions that capture the relationship between program paths and input data structures. The path conditions are then translated into fitness functions to enable testing based on meta-heuristic search. A solution for the search problem is a sequence of method invocations that instantiates the structured inputs to exercise the program paths identified by the path condition.

1.4.1 Automated Generation of Test Oracles

In the case of data-driven mutation testing, solutions for the ***automated generation of oracles*** that consist of assertions verifying the output of the software under test (see Section 1.3.3.1) might still be applied. However, considering that data-driven mutation testing is more likely adopted in the context of system-level testing, where inputs consist of complex, structured data, the generation of test oracles using techniques built for unit testing is likely infeasible in this context.

Possible solutions may consist of approaches that verify the correctness of the log files generated during the execution of the program [**?**, **?**]. Given a set of log files (or execution traces) generated during valid executions, these approaches can derive ***finite state automata*** (FSAs) that capture the sequences of events and data-flow observed in valid executions. The derived FSAs can be used to verify if new executions match the inferred models. More precisely, they enable the automated detection of invalid sequences of events and data-flows. Despite none of these approaches had been applied in the context of data-driven mutation testing, traces collected from multiple runs of a same test case might be used to derive an FSA that captures the behaviour of a single test case. By relying on multiple traces collected from a same test case we can leverage the generalization power of the inference engines used to generate the FSAs; these inference engines, for example, can automatically recognize and filter out variable elements such as timestamps. The inferred FSAs could thus then be used as oracles for newer executions of the generated test cases. Such approaches have shown to be effective in detecting both functional failures [**?**] and performance problems [**?**].

[4]https://wiki.mozilla.org/Security/Fuzzing/Peach#Creating_a_Data_Model

Chapter 2

Approach

This chapter presents the theory and the methodology behind the toolset implemented by FAQAS.

2.1 Code-driven Mutation Analysis

2.1.1 Overview

Figure 2.1 provides an overview of the mutation analysis process that we propose, *Mutation Analysis for Space Software (*MASS*)*. Its goal is to propose a comprehensive solution for making mutation analysis applicable to embedded software in industrial cyber-physical systems. The ultimate goal of *MASS* is to assess the effectiveness of test suites with respect to detecting violations of functional requirements.

Different from the mutation analysis pipeline presented in related work [24], *MASS* enables the integration of all mutation analysis optimization techniques that are feasible in our context to address scalability and pertinence problems (see Section 1.1.8). *MASS* consists of eight steps: (Step 1) Collect SUT Test Suite Data, (Step 2) Create Mutants, (Step 3) Compile Mutants, (Step 4) Remove Equivalent and Duplicate Mutants Based on Compiled Code, (Step 5) Sample Mutants, (Step 6) Execute Prioritized Subset of Test Cases, (Step 7) Identify Likely Equivalent / Duplicate mutants Based on Coverage, and (Step 8) Compute the Mutation Score. Different from related work, *MASS* enables FSCI-based sampling by iterating between mutants sampling (Step 5) and test cases execution (Step 6). Also, it integrates test suite prioritization and reduction (Step 6) before the computation of the mutation score. Finally, it includes methods to identify likely equivalent and duplicate mutants based on code coverage (Step 7). We describe each step in the following paragraphs.

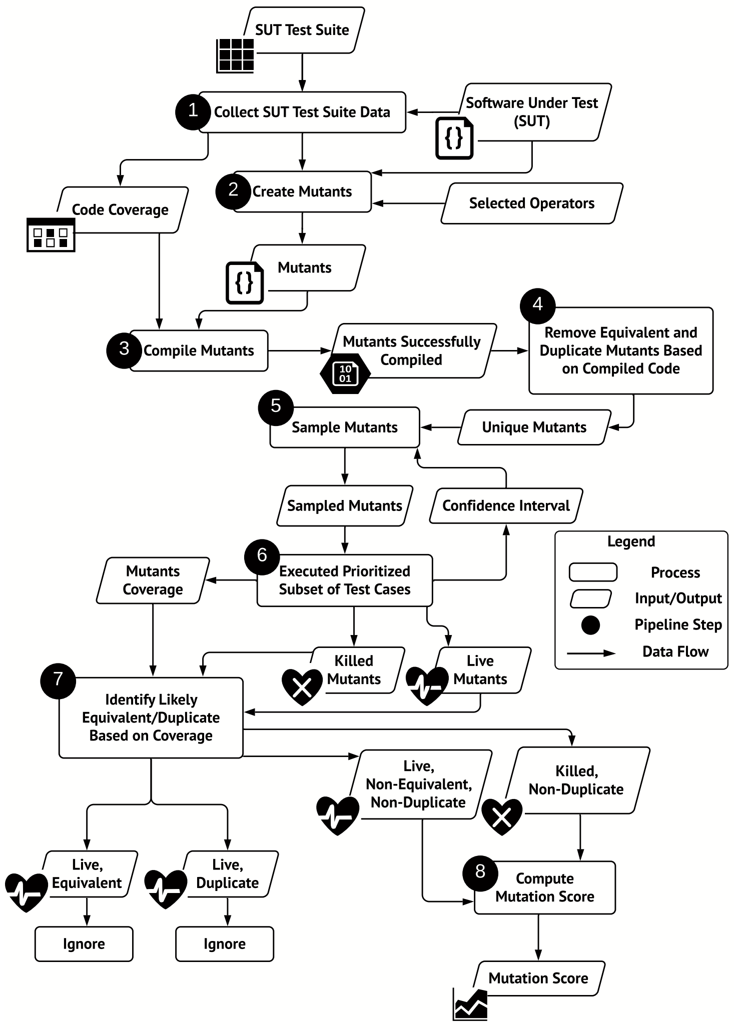

Figure 2.1: Overview of the proposed Mutation Analysis Pipeline

Table 2.1: Implemented set of mutation operators.

	Operator	Description*
Sufficient Set	ABS	$\{(v, -v)\}$
	AOR	$\{(op_1, op_2) \mid op_1, op_2 \in \{+, -, *, /, \%\} \wedge op_1 \neq op_2\}$
		$\{(op_1, op_2) \mid op_1, op_2 \in \{+=, -=, *=, /=, \%=\} \wedge op_1 \neq op_2\}$
	ICR	$\{i, x) \mid x \in \{1, -1, 0, i+1, i-1, -i\}\}$
	LCR	$\{(op_1, op_2) \mid op_1, op_2 \in \{\&\&, \Vert\} \wedge op_1 \neq op_2\}$
		$\{(op_1, op_2) \mid op_1, op_2 \in \{\&=, \vert=, \&=\} \wedge op_1 \neq op_2\}$
		$\{(op_1, op_2) \mid op_1, op_2 \in \{\&, \vert, \&\&\} \wedge op_1 \neq op_2\}$
	ROR	$\{(op_1, op_2) \mid op_1, op_2 \in \{>, >=, <, <=, ==, !=\}\}$
		$\{(e, !(e)) \mid e \in \{\texttt{if(e)}, \texttt{while(e)}\}\}$
	SDL	$\{(s, \texttt{remove}(s))\}$
	UOI	$\{(v, \text{-}v), (v, v\text{-}), (v, {+}{+}v), (v, v{+}{+})\}$
OODL	AOD	$\{((t_1\ op\ t_2), t_1), ((t_1\ op\ t_2), t_2) \mid op \in \{+, -, *, /, \%\}\}$
	LOD	$\{((t_1\ op\ t_2), t_1), ((t_1\ op\ t_2), t_2) \mid op \in \{\&\&, \Vert\}\}$
	ROD	$\{((t_1\ op\ t_2), t_1), ((t_1\ op\ t_2), t_2) \mid op \in \{>, >=, <, <=, ==, !=\}\}$
	BOD	$\{((t_1\ op\ t_2), t_1), ((t_1\ op\ t_2), t_2) \mid op \in \{\&, \vert, \wedge\}\}$
	SOD	$\{((t_1\ op\ t_2), t_1), ((t_1\ op\ t_2), t_2) \mid op \in \{», «\}\}$
Other	LVR	$\{(l_1, l_2) \mid (l_1, l_2) \in \{(0, -1), (l_1, -l_1), (l_1, 0),$ $(\mathit{true}, \mathit{false}), (\mathit{false}, \mathit{true})\}\}$

*Each pair in parenthesis shows how a program element is modified by the mutation operator on the left; we follow standard syntax [38]. Program elements are literals (l), integer literals (i), boolean expressions (e), operators (op), statements (s), variables (v), and terms (t_i, which might be either variables or literals).

2.1.2 Step 1: Collect SUT Test Data

In Step 1, the test suite is executed against the SUT and code coverage information is collected. More precisely, we rely on the combination of gcov [?] and GDB [?], enabling the collection of coverage information for embedded systems without a file system [?].

2.1.3 Step 2: Create Mutants

In Step 2, we automatically generate mutants for the SUT by relying on a set of selected mutation operators. In *MASS*, based on the considerations provided in Section 1.1.2, we rely on an extended sufficient set of mutation operators, which are listed in Table 2.1. In addition, in our experiments, we also evaluate the feasibility of relying only on the SDL operator, combined or not with OODL operators, instead of the entire sufficient set of operators.

To automatically generate mutants, we have extended SRCIRor [?] to include all the operators in Table 2.1. After mutating the original source file, our extension saves the mutated source file and keeps track of the mutation applied. Our toolset is available under the ESA Software Community Licence Permissive [?] at the following URL **https://faqas.uni.lu/**.

2.1.4 Step 3: Compile mutants

In Step 3, we compile mutants by relying on an optimized compilation procedure that leverages the build system of the SUT. To this end, we have developed a toolset that, for each mutated source file: (1) backs-up the original source file, (2) renames the mutated source file as the original source file, (3) runs the build system (e.g., executes the command `make`), (4) copies the generated executable mutant in a dedicated folder, (5) restores the original source file.

Build systems (e.g., GNU make [?] driving the GCC [?] compiler) create one object file for each source file to be compiled and then link these object files together into the final executable. After the first build, in subsequent builds, build systems recompile only the modified files and link them to the rest. For this reason, our optimized compilation procedure, which modifies at most two source files for each mutant (i.e., the mutated file and the file restored to eliminate the previous mutation), can reuse almost all the compiled object files in subsequent compilation runs, thus speeding up the compilation of multiple mutants. The experiments conducted with our subjects have shown that our optimization is sufficient to make the compilation of mutants feasible for large projects. Other state-of-the-art solutions introduce additional complexity (e.g., change the structure of the software under test [45]) that does not appear to be justified by scalability needs.

Mutants that lead to compilation errors are discarded. Concerning compilation warnings, we assume the build system of the SUT has been properly configured; more precisely, if the system should compile without warnings, the compiler is expected to be configured to treat warnings as errors otherwise mutants that lead to warning are retained.

2.1.5 Step 4: Remove equivalent and redundant mutants based on compiled code

In Step 4, we rely on trivial compiler optimizations to identify and remove equivalent and redundant mutants. We compile the original software and every mutant multiple times once for each every available optimization option (i.e., `-O0`, `-O1`, `-O2`, `-O3`, `-Os`, `-Ofast` in GCC) or a subset of them. After each execution of the compiler, we compute the SHA-512 hash summary of the generated executable.To detect equivalent mutants, *MASS* compares the hash summaries of the mutants with that of the original executable. To detect duplicate mutants but avoid combinatorial explosion, *MASS* focuses its comparison of hash summaries on pairs of mutants belonging to the same source file (restricting the scope of the comparison is common practice [35]). Hash comparison allows us to (1) determine the presence of equivalent mutants (i.e., mutants having the same hash as the original executable), and (2) identify duplicate mutants (i.e., mutants with the same hash). Equivalent and duplicate mutants are then discarded. We compare hash summaries rather than executable files because it is much faster, an important consideration when dealing with a large number of mutants. The outcome of Step 4 is a set of ***unique mutants***, i.e., mutants with compiled code that differs from the original software and any other mutant.

2.1.6 Step 5: Sample Mutants

In Step 5, *MASS* samples the mutants to be executed to compute the mutation score. *MASS* does not selectively generate mutants but samples them from the whole set of successfully compiled, nonequivalent, and nonduplicated mutants (result of Steps 2 to 4). This choice aims to avoid sampling bias which may result from the presence of such mutants; indeed, there is no guarantee that these mutants, if they were discarded after being sampled, would be uniformly distributed across program statements. Our choice does not affect the feasibility of *MASS* since Steps 2 to 4 have negligible cost.

Our pipeline supports different sampling strategies: ***proportional uniform sampling***, ***proportional method-based sampling***, ***uniform fixed-size sampling***, and ***uniform FSCI sampling***.

The strategies ***proportional uniform sampling*** and ***proportional method-based sampling*** were selected based on the results of Zhang et al. [11], who compared eight strategies for sampling mutants. The former was the best performing strategy and consists of sampling mutants evenly across all functions of the SUT, i.e., sampling $r\%$ mutants from each set of mutants generated inside the same function. The latter consists of randomly selecting $r\%$ mutants from the complete mutants set. This is included in our study because it is simpler to implement and showed to be equivalent to stratified sampling strategies, based on recent work [12].

The ***uniform fixed-size sampling*** strategy stems from the work of Gopinath et al. [12] and consists of selecting a fixed number N_M of mutants for the computation of the mutation score. Based their work, with 1,000 mutants, one can guarantee an accurate estimation of the mutation score.

In this paper, we introduce the ***uniform FSCI sampling*** strategy that determines the sample size dynamically, while exercising mutants, based on a fixed-width sequential confidence interval approach. With ***uniform FSCI sampling***, we introduce a cycle between Step 6 and Step 5, such that a new mutant is sampled only if deemed necessary. More precisely, *MASS* iteratively selects a random mutant from the set of unique mutants and exercises it using the SUT test suite. Based on related work, we assume that the mutation score computed with a sample of mutants follows a binomial distribution (see Section 1.1.4). For this reason, to compute the confidence interval for the FSCI analysis, we rely on the Clopper-Pearson method since it is reported to provide the best results (see Section 1.1.4). Mutation analysis (i.e., sampling and testing a mutant) stops when the confidence interval is below a given threshold T_{CI} (we use $T_{CI} = 0.10$ in our experiments). More formally, given a confidence interval $[L_S; U_S]$, with L_S and U_S indicating the lower and upper bound of the interval, mutation analysis stops when the following condition holds:

$$(U_S - L_S) < T_{CI}. \tag{2.1}$$

Unfortunately, the assumption about the estimated mutation score following a binomial distribution may not hold when a subset of the test suite is executed for every mutant (which could happen in Step 6). Without going into the details behind the implementation of Step 6, which is described in Section 2.1.7, we can expect that a reduced test suite may not be able to kill all the mutants killed by the entire test suite, i.e., the estimated mutation score may be affected by negative bias. Consequently, over multiple runs, the mean of the estimated mutation score may not be close to the ***actual mutation score*** (i.e., the mutation score computed with the entire test suite exercising all the mutants for the SUT) but may converge to a lower value. To compute a correct confidence interval that includes the actual mutation score of the SUT, we thus need to take into account this negative bias.

To study the effect of negative bias on the confidence interval, we address first the relation between the actual mutation score and the mutation score computed with the reduced test suite when the entire set of mutants for the SUT is executed. A mutant killed by the entire test suite has a probability P_{KErr} of not being killed by the reduced test suite. The probability P_{KErr} can be estimated as the proportion of mutants (erroneously) not killed by the reduced test suite

$$P_{KErr} = \frac{|E_R|}{|M|} \tag{2.2}$$

with E_R being the subset of mutants that are killed by the entire test suite but not by the reduced test suite, and M being the full set of mutants for the SUT.

The mutation score for the reduced test suite (MS_R) can be computed as

$$MS_R = \frac{|K| - |E_R|}{|M|} = \frac{|K|}{|M|} - \frac{|E_R|}{|M|} = MS - \frac{|E_R|}{|M|} = MS - P_{KErr} \tag{2.3}$$

where K is the set of mutants killed by the whole test suite, M is the set of all the mutants of the SUT, and MS is the actual mutation score. Consequently, the actual mutation score can be computed as

$$MS = MS_R + P_{Err_R} \tag{2.4}$$

We now discuss the effect of a reduced test suite on the confidence interval for a mutation score estimated with mutants sampling. When mutants are sampled and tested with the entire test suite, the actual mutation score is expected to lie in the confidence interval $[L_S; U_S]$. In the presence of a reduced test suite, we can still rely on the Clopper-Pearson method to compute the confidence interval $CI_R = [L_R; U_R]$. However, we have to take into account the probability of an error in the computation of the mutation score MS_R; MS_R can be lower than MS and, based on Equation 2.4, we expect the actual mutation score to lie in an interval that is shifted with respect to the interval for MS_R:

$$CI = [L_R + P_{KErr}; U_R + P_{KErr}] \tag{2.5}$$

We can only estimate P_{KErr} since computing it would require the execution of all the mutants with the complete test suite, thus undermining our objective of reducing test executions. To do so, we can randomly select a subset M_R of mutants, on which to execute the entire test suite and identify the mutants killed by the reduced test suite. The size of the set M_R should be lower than the number of mutants we expect FSCI sampling to return, otherwise sampling would not provide any cost reduction benefit. Since, for every mutant in M_R, we can determine if it is erroneously reported as not killed by the reduced test suite R, we can estimate the probability P_{KErr} as the percentage of such mutants. As for the case of the mutation score, we assume that the binomial distribution provides a conservative estimate of the variance for P_{KErr}.

We can estimate the confidence interval for P_{KErr} using one of the methods for binomial distributions. We rely on the Wilson score method because it is known to perform well with small samples [**?**]. The value of P_{KErr} will thus lie within $CI_E = [L_E; U_E]$, with L_E and U_E indicating the lower and upper bounds of the interval.

Based on Equation 2.5, the confidence interval to be used with FSCI sampling in the presence of a reduced test suite should thus be

$$CI = [L_R + L_E; U_R + U_E] \tag{2.6}$$

The estimated mutation score is the value lying in the middle of the interval.

Since the width of the confidence interval CI (hereafter, $|CI|$) results from the sum of $|CI_R|$ and $|CI_E|$, mutation sampling with a reduced test suite may lead to the execution of a larger set of mutants.

Based on Equations 2.1 and 2.6, $|CI_R| \leq T_{CI} - |CI_E|$. Consequently, when $|CI_E| > T_{CI}$, the reduced test suite cannot lead to sufficiently accurate results. Also, a large $|CI_E|$ may prevent the identification of accurate results with a feasible number of mutants. For example, Clopper-pearson may require up to 1568 samples for a confidence interval below 0.05 [69]. We shall thus identify a threshold (T_{CE}) for the confidence interval $|CI_E|$ that enables accurate estimates with a small sample size (e.g., in the worst case, with less than 1000 samples, the sample size for related work). For this reason, starting from a minimal number of samples to estimate P_{KErr} (150 in our experiments), *MASS* keeps estimating P_{KErr} until it yields $|CI_E| \leq T_{CE}$. In our experiments we set $T_{CE} = 0.035$. To select T_{CE}, we have identified a reasonable minimal mutation score to be expected in space software (i.e., 65%) and identified, based on confidence interval estimation methods with finite population correction factor [**?**], the minimal value for $|CI_E|$ that requires a number of samples below 850 (i.e., $1000 - 150$).

When it is not possible to estimate $|CI_E| \leq T_{CE}$ or when the number of samples required to estimate $|CI_E| \leq T_{CE}$ is sufficient to accurately estimate the mutation score, the test suite can be prioritized but not reduced and the confidence interval is computed using the traditional Clopper-Pearson method, i.e., $[L_S; U_S]$.

2.1.7 Step 6: Execute prioritized subset of test cases

In Step 6, we execute a prioritized subset of test cases. We select only the test cases that satisfy the reachability condition (i.e., cover the mutated statement) and execute them in sequence. Similarly to the approach of Zhang et al. [13], we define the order of execution of test cases based on their estimated likelihood of killing a mutant. However, in our work, this likelihood is estimated differently since, as discussed above, the measurements they rely on are not applicable in the context of system-level testing and complex cyber-physical systems (see Section 1.1.4). In contrast, to minimize the impact of measurements on real-time constraints, we only collect code coverage information for a small part of the system.

We execute only covered statements assuming that the test suite is optimal with respect to code coverage. More precisely, we addume that if a statement is not covered there is a good reason for it (e.g., it depends on hardware). If a statement is not covered by the test suite, there is no chance that a mutant generated in the non-covered statement can be possibly detected by any test case. If the test suite does not reach the required coverage there is no reason to perform mutation testing, because is already known that the test suite is not good.

To reduce the number of test cases to be executed with a mutant, we should first execute the ones that more likely satisfy the necessity condition. This might be achieved by executing a test case that exercises the mutated statement with variable values not observed before. Unfortunately, in our context, the size of the

SUT and its real-time constraints prevent us from recording all the variable values processed during testing.

Therefore, we rely on code coverage to determine if two test case executions exercise the mutated statement with diverse variable values. Such coverage is collected by efficient procedures provided by compilers, thus having lower impact on execution performance than other types of dynamic analysis solutions (e.g., tracing variable values). Since, because of control- and data-flow dependencies, a different set of input values may lead to differences in code coverage, the latter helps determine if two or more test cases likely exercise a mutated statement with different variable values. To increase the likelihood that the observed differences in code coverage are due to the use of different variable values to exercise the mutated statement, we restrict the scope of code coverage analysis to the functions belonging to the component (i.e., the source file) that contains the mutated statement. Indeed, such functions typically present several control- and data-flow dependencies, thus augmenting the likelihood that a coverage difference is due to the execution of the mutated statement with a diverse set of values. Also, collecting code coverage for a small part of the system further reduces the impact of our analysis on system performance.

Based on related work, we have identified two possible strategies to characterize test case executions based on code coverage:

S1 Compare the sets of source code statements that have been covered by test cases [14].

S2 Compare the number of times each statement has been covered by test cases [16].

To determine how dissimilar two test cases are and, consequently, how likely they exercise the mutated statement with different values, we rely on widely adopted distance metrics. In the case of S1, we rely on the Jaccard and Ochiai index, which are two similarity indices for binary data and have successfully been used to compare program executions based on code coverage [**?**, **?**, **?**]. The Jaccard index is also known as ***intersection over union*** it measures similarity between sets, and is defined as the size of the intersection divided by the size of the union. The Jaccard distance measures dissimilarity between sample sets and results from subtracting the Jaccard coefficient from 1. The Ochiai index calculates cosine similarity with binary data, it is used in molecular biology and software engineering. Given two test cases T_A and T_B, the Jaccard (D_J) and Ochiai (D_O) distances are computed as follows:

$D_J(T_a, T_b) = 1 - \frac{|C_a \cap C_b|}{|C_a \cup C_b|}$ $D_O(T_a, T_b) = 1 - \frac{|C_a \cap C_b|}{\sqrt{|C_a| * |C_b|}}$, where C_a and C_b are the set of covered statements exercised by T_a and T_b, respectively.

In the case of S2, we compute the distance between two test cases by relying on the euclidean distance (D_E) and the cosine similarity distance (D_C), two popular distance metrics used in machine learning. Euclidean distance is the straight-line distance between two points in Euclidean space; precisely, the Euclidean distance between two points is the length of the line segment connecting them. In our context, the two vectors consist of the number of times the program statements had been exercised by a test. Cosine similarity measures similarity between two vectors of an inner product space. It results from the inner product of the same vectors normalized to both have length 1, which matches the the cosine of the angle between them. It is widely adopted to measure cohesion within clusters in data mining. Given two vectors V_A and V_B, whose elements capture the number of times a statement has been covered by test cases T_A and T_B, the distances D_E and D_C can be computed as follows:

$D_E = \sqrt{\sum_{i=1}^{n}(A_i - B_i)^2}$

$D_C = 1 - \frac{\sum_{i=1}^{n} A_i * B_i}{\sqrt{\sum_{i=1}^{n} {A_i}^2} * \sqrt{\sum_{i=1}^{n} {B_i}^2}}$, where A_i and B_i refer to the number of times the i-th statement had been covered by T_A and T_B, respectively.

Figure 2.2 shows the pseudocode of our algorithm for selecting and prioritizing test cases. It generates as output a prioritized test suite (***PTS***). Based on the findings of Zhang et al. [13], we first select the test case that exercises the mutated statement the highest number of times (Line 3) and add it to the prioritized test suite

(Line 4). Then, in the next iterations, the test case selected is the one with the largest distance from the closest test case already selected (Lines 10 to 13). When two or more test cases have the same distance, we select randomly among the test cases that exercise the mutated statement the most.

The algorithm iterates as long as it identifies a test case showing a difference in code coverage from the already selected test cases (Line 14).

Test cases are then executed in the selected order. During execution, we collect code coverage information and identify killed and live mutants.

```
Require: TS, the test suite of the software under test
Require: Cov, coverage information, for each test case
Require: ms, the mutated statement
Ensure: PTS, a list of test cases to be executed, sorted by priority
 1: $TS_m$ ← subset of TS that cover the mutated statement ms, based on Cov
 2: PTS ← newlist //this list is initially empty
 3: PTS ← based on Cov select from $TS_m$ the test case $t$ that exercises ms more times
 4: PTS ← PTS ∪ $t$ //include first the test case selected above
 5: repeat
 6:     for each $n$ in the set ($TS_m$ - PTS) , which is the set of test cases not already added to PTS
 7:         for each $t$ in PTS
 8:             compute the distance between $t$ and $n$
 9:         identify $t_n$ i.e., the test case $t$ with the minimal $d$
10:     among all the $t_n$ identified, select the one with the highest distance $d$
11:     if $d > 0$ //there is at least a test case with a different coverage
12:         //note: $n$ is the test case in the set ($TS_m$ - PTS) closer to $t_n$
13:         PTS ← PTS ∪ $n$
14: until $d > 0$
```

Figure 2.2: Algorithm for prioritizing test cases

2.1.8 Step 7: Discard Mutants

In this step, we identify likely nonequivalent and likely nonduplicate mutants by relying on code coverage information collected in the previous step.

Similarly to related work [16], we identify nonequivalent and nonduplicate mutants based on a threshold.

In our case, consistently with previous steps of *MASS*, we compute normalized distances based on the distance metrics D_J, D_O, D_E, and D_C. A mutant is considered nonequivalent when the distance from the original program is above the threshold T_E, for at least one test case. Similarly, a mutant is considered nonduplicate when the distance from every other mutant is above the threshold T_D, for at least one test case. For the identification of nonequivalent mutants, we consider live mutants only. To identify nonduplicate mutants, we consider both live and killed mutants; however, to avoid combinatorial explosion, we compare only mutants belonging to the same source file (indeed, mutants belonging to different files are unlikely to be redundant). Killed mutants that lead to the failure of different test cases are not duplicate, regardless of their distance.

Thresholds T_E and T_D should enable the identification of mutants that are guaranteed to be nonequivalent and nonduplicate. In particular, we are interested in the set of *live, nonequivalent, nonduplicate mutants* (hereafter, *LNEND*) and the set of *killed, nonduplicate mutants* (hereafter, *KND*). With such guarantees, the mutation score can be adopted as an adequacy criterion in safety certification processes. For example, certification agencies may require safety-critical software to reach a mutation score of 100%, which is feasible in the presence of nonequivalent mutants.

Figure 2.3 shows the algorithm for detecting nonequivalent and nonduplicate mutants. It first identify among the list of killed mutants all the non-duplicate ones (Line 1). Then it identifies the non-equivalent mutants among the list of live mutants (Line 2). Finally, it further filters the list of non-equivalent mutants to keep only the ones that appear to be nonduplicate (Line 3).

```
Require: D, the distance function to use to identify equivalent/duplicate mutants
Require: KM, list of killed mutants
Require: LM, list of live mutants
Require: $Cov_O$, coverage information for all the test cases, for the original program
Require: $Cov_M$, coverage information for all the executed test cases, for every mutant
Require: TS, list of test cases
Require: TR, test results, for all the executions
Ensure: KND, a list of killed, non-duplicate mutants
Ensure: LNEND, a list of live, non-equivalent, non-duplicate mutants
KND ← identifyNonDuplicateMutants(KM, TS, TR, $Cov_M$)
LNE ← identifyNonEquivalentMutants(LND, TS, $Cov_M$, $Cov_O$)
LNEND ← identifyDuplicateMutants(LNE, TS, $Cov_M$)
procedure identifyNonDuplicateMutants(M, TS, TR, $Cov_M$)//$M$ is a list of mutants, TS, TR, and $Cov_M$ are defined above
    ND ← emptyset
    k1 ← extract and remove first element of M
    ND ← ND ∪ k1
    while M not empty do
        k2 ← extract and remove first element of M
        for mutant $k1$ in ND do
            duplicate = TRUE
            for test case $t$ in TS do
                if $t$ has different result in $k1$ and $k2$ then
                    duplicate = FALSE
                    break
                else
                    $cov_{k1t}$ ← extract coverage information for test case $t$ executed with mutant $k1$
                    $cov_{k2t}$ ← extract coverage information for test case $t$ executed with mutant $k2$
                    if $D(cov_{k1t}, cov_{k2t}) > T_R$ then
                        duplicate = FALSE
                        break
                    end if
                end if
            end for
            if duplicate == FALSE then
                break //No need to compare with all the mutants if we know that it is not duplicate
            end if
        end for
        if duplicate == FALSE then
            ND ← ND ∪ k2
        end if
    end while
    return ND
end procedure
procedure identifyNonEquivalentMutants(M, TS, $Cov_M$, $Cov_O$)//$M$ is a list of mutants, TS, and $Cov_M$, and $Cov_O$ are defined above
    NE ← emptyset
    while M not empty do
        m ← extract and remove first element of M
        for test case $t$ in TS do
            $cov_m$ ← extract coverage information for test case $t$ executed with mutant $m$
            $cov_o$ ← extract coverage information for test case $t$ executed with original program
            if $D(cov_m, cov_o) > T_E$ then
                equivalent = FALSE
                break
            end if
        end for
        if equivalent == FALSE then
            NE ← NE ∪ m
        end if
    end while
    return NE
end procedure
```

Figure 2.3: Algorithm for identifying non-equivalent and non-duplicate mutants

2.1.9 Step 8: Compute Mutation Score

The ***mutation score*** (MS) is computed as the percentage of killed nonduplicate mutants over the number of nonequivalent, nonduplicate mutants identified in Step 7):

$$MS = \frac{|KND|}{|LNEND| + |KND|} \tag{2.7}$$

```
flag T_INT_IsConstraintValid(const T_INT* pVal, int* pErrCode)
{
    flag ret = TRUE;
    (void)pVal;

    ret = ((*(pVal)) <= 50UL);

    klee_semu_GenMu_Mutant_ID_Selector_Func(1,2);
    *pErrCode = ret ? (klee_semu_GenMu_Mutant_ID_Selector==2?
    ((-1)):
    (klee_semu_GenMu_Mutant_ID_Selector==1?
    (1):
    (0))) :  ERR_T_INT;
    klee_semu_GenMu_Post_Mutation_Point_Func(0,0);
    klee_semu_GenMu_Post_Mutation_Point_Func(1,2);

    return ret;
}
```

Listing 2.1: Meta-Mutant of function T_INT_IsConstraintValid.

```
flag T_INT_IsConstraintValid(const T_INT* pVal, int* pErrCode)
{
    flag ret = TRUE;
    (void)pVal;

    ret = ((*(pVal)) <= 50UL);
    *pErrCode = ret ? 1 :  ERR_T_INT;

    return ret;
}
```

Listing 2.2: Mutant 1 of function T_INT_IsConstraintValid.

2.2 Code-driven Mutation Testing: SEMuS

2.2.1 Overview

To achieve ***test suite augmentation*** (i.e., automatically generate test cases that kill mutants), in FAQAS, we have implemented an extension of SEMu that we call SEMu for Space Software (***SEMuS***).

In the FAQAS context, we cannot use SEMu as it is, since it requires mutants to be compiled with the ***LLVM compiler*** in LLVM-IR format. As explained previously (1) our case studies relies on compiler pipelines (e.g., RTEMS) that include architecture-specific optimizations that are not supported by LLVM, (2) there is no guarantee that the compiled objects produced by LLVM are equivalent to those produced by the original compiler, and (3) LLVM optimizations are often applied at runtime, which is infeasible when the software under test needs to be executed within a dedicated simulator (e.g., a SVF facility). The practical consequence is that the test suites of our case study systems, and, more in general, integration and system test suites of space software, cannot be used to test LLVM versions of the SUT. Therefore, they cannot be used to identify killed and live mutants, as a consequence, it is not possible to determine which additional test cases need to be generated. The solution to this problem consists of first executing ***MASS*** to identify killed and live mutants, and then compile only the live mutants into LLVM-IR format. To minimize the limitations above, instead of compiling the whole software, we compile only the mutated function and its dependencies, which enables the generation of unit test cases, which shall be sufficient to ensure the quality of the system under test in this context (e.g., unit test cases are normally used to ensure high code coverage). To enable the compilation of such mutants, we have extended MASS to generate both a meta-mutant processed by SEMu and mutants that, after mutation analysis, can be traced back to such meta-mutant.

TODO: Update the structure to reflect the presentation during the review meeting

Listing 2.1 introduces an example of a ***meta-mutant***, while Listings 2.2 and 2.3 introduce an example of single mutants generated by MASS for the same statement. The meta-mutant of Listing 2.1 shows how KLEE-SEMu introduces two mutations within the same source code; to this end, KLEE-SEMu use a set of functions for controlling the mutation at runtime within the meta-mutant, these functions are:

```
flag T_INT_IsConstraintValid(const T_INT* pVal, int* pErrCode)
{
    flag ret = TRUE;
    (void)pVal;

    ret = ((*(pVal)) <= 50UL);
    *pErrCode = ret ? (-1) :  ERR_T_INT;

    return ret;
}
```

Listing 2.3: Mutant 2 of function T_INT_IsConstraintValid.

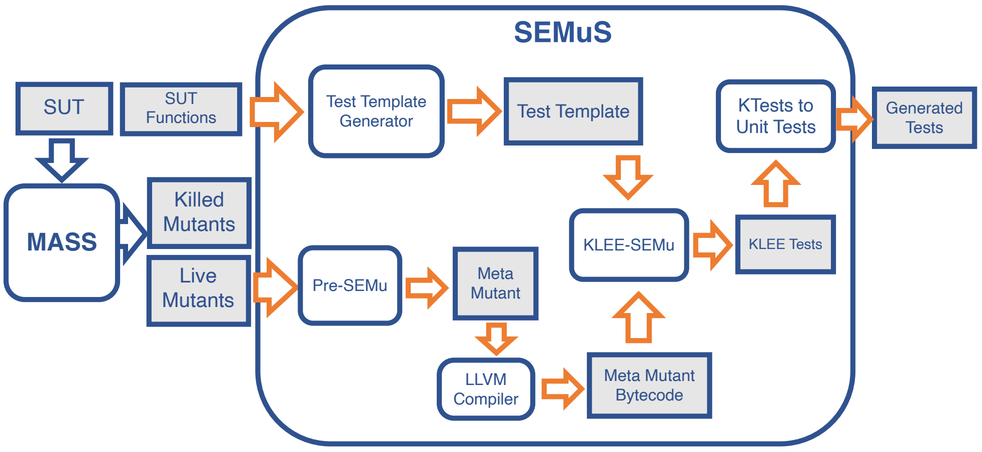

Figure 2.4: FAQAS-SEMuS Architecture and Workflow

- `klee_semu_GenMu_Mutant_ID_Selector_Func`: function that takes two mutant IDs as arguments, representing a range of mutant IDs.
- `klee_semu_GenMu_Mutant_ID_Selector`: global variable that contains the ID of the mutant to be activated at runtime.
- `klee_semu_GenMu_Post_Mutation_Poin_Func`: This enable SEMu to do conservative pruning and remove the mutant states that are non-infected.

As shown in Listing 2.1, mutant 2 from Listing 2.3 is represented on line 10, and mutant 1 from Listing 2.2 is represented in line 12.

Figure 2.4 shows the architecture of SEMuS and how it interacts with MASS. SEMuS consists of five components, which are *Test Template Generator*, *Pre-SEMu*, *KLEE-SEMu*, ***KTest to Unit Test***, and ***LLVM***. They enable the adoption of SEMu in the space context. In particular, SEMuS (1) automates the generation of a test template including symbolic variables that guides the generation of test inputs, and (2) compiles the test template and the mutant into the format required by SEMu (i.e., LLVM-IR).

The ***Test Template Generator*** (TTG) component automates the generation of templates for the symbolic execution search. The component receives as inputs the SUT source code and the list of SUT functions. Listing 2.4 shows an example of a test template generated by the TTG. The TTG generates a template for every SUT function; as shown in Listing 2.4 the component parses the function arguments and declares them symbolic through use of the KLEE function `klee_make_symbolic`. Then, it adds a call to the function under analysis with symbolic values, and it saves the output into the variable `result_faqas_semu`. Finally, the TTG adds a return call with the `result_faqas_semu` variable value.

The ***Pre-SEMu*** component generates ***mutant schemata***; specifically, the component includes and compiles

```c
int main(int argc, char** argv) {
    // Declare variable to hold function returned value
    _Bool result_faqas_semu;
    // Declare arguments and make input ones symbolic
    unsigned long pVal;
    int pErrCode;
    klee_make_symbolic(&pVal, sizeof(pVal), "pVal"); // Call function under test
    result_faqas_semu = T_INT_IsConstraintValid(&pVal, &pErrCode); // Make some output
    printf("FAQAS-SEMU-TEST_OUTPUT: %d\n", pErrCode);
    printf("FAQAS-SEMU-TEST_OUTPUT: %d\n", result_faqas_semu);
    return (int)result_faqas_semu;
}
```

Listing 2.4: SEMuS test template.

```
ktest file : 'test000001.ktest'
args       : ['/MakeSym-TestGen-Input/direct/T_INT_IsConstraintValid/test.MetaMu.bc']
num objects: 2
object    0: name: b'model_version'
object    0: size: 4
object    0: data: b'\x01\x00\x00\x00'
object    1: name: b'pVal'
object    1: size: 8
object    1: data: b'\x00\x00\x00\x00\x00\x00\x00\x00'
```

Listing 2.5: Klee-test output

all the live mutants (i.e., MASS output) into a single bytecode file named the *Meta Mutant*. At runtime, SEMu will select which mutant to consider for test generation based on a parameter. The compilation of the Meta Mutant into LLVM bitcode is supported by the *LLVM* compiler infrastructure.

KLEE-SEMu is the underlying test generation component, previously described in Section **??**. This component receives as inputs the *LLVM bitcode Meta Mutant* and the *Test Template* for the function under test, and proceeds to apply dynamic symbolic execution to generate test inputs to kill the mutants. The output of this component are the *KLEE tests*.

A ***KLEE test*** is a binary file that contains information about the execution of KLEE such as the entry point of the analysis, and the generated test inputs.

An example of a KLEE test is presented in Listing 2.5. The field *args* report the entry point of the analysis; in this case, the test generation was performed for live mutants present in the function `T_INT_IsConstraintValid`, which SEMuS stores in a dedicated folder. The fields named *object* provide information about the outputs generated by KLEE (e.g., the generated test inputs). For each object, the KLEE test provides a *name* (usually the name of the symbolic variable), its *size*, and the actual *value* generated by KLEE through constraint solving (usually this is reported in binary form). Objects are numbered. Object number *0* reports information about the data structure used by KLEE, that is, the version of the structure. The other objects report information about the generated test inputs. Our example shows that one value of size 8 was generated for the variable `pVal`, the data field shows the binary representation of the `pVal` variable, in this case `pVal=0`.

The component ***KTest to Unit Test*** (KTU) converts a KLEE test into a readable, executable C test case. Similar to the TTG component, the KTU uses a template to hold the specifics variables of each function under test. Listing 2.6 shows an example of a test case generated for a mutant present in the function `T_INT_IsConstraintValid`. For instance, line 20 shows that the variable `pVal` is initially filled with 0 (to clean it), then in line 21, it is filled with the value stored in the variable `pVal_faqas_semu_test_data`, which holds the binary output produced by KLEE. In line 25, the function under test is invoked with the concrete value of `pVal`. Finally, the KTU print the function return value and the value of every variable passed by reference. At the current stage, KTU does not generate assertions. Assertions need to be implemented by engineers in place of the generated *printfs* because only the engineers can know, based on specifications, what are the values to be expected at the end of the test case execution.

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

#include "asn1crt.c"
#include "asn1crt_encoding.c"
#include "asn1crt_encoding_uper.c"


int main(int argc, char** argv)
{
    (void)argc;
    (void)argv;

    // Declare variable to hold function returned value
    _Bool result_faqas_semu;

    // Declare arguments and make input ones symbolic
    unsigned long pVal;
    int pErrCode;
    memset(&pVal, 0, sizeof(pVal));
    const unsigned char pVal_faqas_semu_test_data[] = {0x00, 0x00, 0x00, 0x00, 0x00, 0x00, 0x00, 0x00};
    memcpy(&pVal, pVal_faqas_semu_test_data, sizeof(pVal)); // Unsigned val is 0

    // Call function under test
    result_faqas_semu = T_INT_IsConstraintValid(&pVal, &pErrCode);

    // Make some output
    printf("FAQAS-SEMU-TEST_OUTPUT: pErrCode = %d\n", pErrCode);
    printf("FAQAS-SEMU-TEST_OUTPUT: result_faqas_semu = %d\n", result_faqas_semu);
    return (int)result_faqas_semu;
}
```

Listing 2.6: Generated test case

Index

Bibliography

[1] O. E. Cornejo Olivares, F. Pastore, and L. Briand, "Mutation analysis for cyber-physical systems: Scalable solutions and results in the space domain," *IEEE Transactions on Software Engineering*, pp. 1–1, 2021.

[2] Cloudbees, "CHECK c test framework," 2020. [Online]. Available: https://libcheck.github.io/check/

[3] European Space Agency, "ExoMars 2016 - Schiaparelli Anomaly Inquiry," *DG-I/2017/546/TTN*, 2017. [Online]. Available: http://exploration.esa.int/mars/59176-exomars-2016-schiaparelli-anomaly-inquiry/

[4] European Cooperation for Space Standardization., "ECSS-Q-ST-80C Rev.1 ? Software product assurance." 2017. [Online]. Available: http://ecss.nl/standard/ecss-q-st-80c-rev-1-software-product-assurance-15-february-2017/

[5] ——, "ECSS-E-ST-40C ? Software general requirements." 2009. [Online]. Available: http://ecss.nl/standard/ecss-e-st-40c-software-general-requirements/

[6] European Space Agency, "ESA ISVV Guide, issue 2.0, 29/12/2008." 2008. [Online]. Available: ftp://ftp.estec.esa.nl/pub/wm/anonymous/wme/ecss/ESAISVVGuideIssue2.029dec2008.pdf

[7] R. A. DeMillo, R. J. Lipton, and F. G. Sayward, "Hints on test data selection: Help for the practicing programmer," *Computer*, vol. 11, no. 4, pp. 34–41, April 1978.

[8] M. Papadakis, C. Henard, M. Harman, Y. Jia, and Y. Le Traon, "Threats to the validity of mutation-based test assessment," in *Proceedings of the 25th International Symposium on Software Testing and Analysis*. ACM, 2016, pp. 354–365.

[9] L. Madeyski, W. Orzeszyna, R. Torkar, and M. Jozala, "Overcoming the equivalent mutant problem: A systematic literature review and a comparative experiment of second order mutation," *IEEE Transactions on Software Engineering*, vol. 40, no. 1, pp. 23–42, 2013.

[10] D. Shin, S. Yoo, and D. Bae, "A theoretical and empirical study of diversity-aware mutation adequacy criterion," *IEEE Transactions on Software Engineering*, vol. 44, no. 10, pp. 914–931, Oct 2018.

[11] L. Zhang, M. Gligoric, D. Marinov, and S. Khurshid, "Operator-based and random mutant selection: Better together," in *Proceedings of the 28th IEEE/ACM International Conference on Automated Software Engineering*. IEEE Press, 2013, pp. 92–102.

[12] R. Gopinath, A. Alipour, I. Ahmed, C. Jensen, and A. Groce, "How hard does mutation analysis have to be, anyway?" in *2015 IEEE 26th International Symposium on Software Reliability Engineering (ISSRE)*. IEEE, 2015, pp. 216–227.

[13] L. Zhang, D. Marinov, and S. Khurshid, "Faster mutation testing inspired by test prioritization and reduction," in *Proceedings of the 2013 International Symposium on Software Testing and Analysis*, 2013, pp. 235–245.

[14] B. J. Grün, D. Schuler, and A. Zeller, "The impact of equivalent mutants," in *2009 International Conference on Software Testing, Verification, and Validation Workshops*. IEEE, 2009, pp. 192–199.

[15] D. Schuler and A. Zeller, "(un-) covering equivalent mutants," in *2010 Third International Conference on Software Testing, Verification and Validation*. IEEE, 2010, pp. 45–54.

[16] ——, "Covering and uncovering equivalent mutants," *Software Testing, Verification and Reliability*, vol. 23, no. 5, pp. 353–374, 2013.

[17] D. Schuler, V. Dallmeier, and A. Zeller, "Efficient mutation testing by checking invariant violations," in *Proceedings of the eighteenth international symposium on Software testing and analysis*. ACM, 2009, pp. 69–80.

[18] O. Givehchi, K. Landsdorf, P. Simoens, and A. W. Colombo, "Interoperability for industrial cyber-physical systems: An approach for legacy systems," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 6, pp. 3370–3378, 2017.

[19] V. Jirkovsky, M. Obitko, and V. Marik, "Understanding data heterogeneity in the context of cyber-physical systems integration," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 2, pp. 660–667, 2017.

[20] NASA, "Mars Climate Orbiter, Spacecraft lost," https://solarsystem.nasa.gov/missions/mars-climate-orbiter/in-depth/, 1998.

[21] M. E. Delamaro, J. Offutt, and P. Ammann, "Designing deletion mutation operators," in *2014 IEEE Seventh International Conference on Software Testing, Verification and Validation*. IEEE, 2014, pp. 11–20.

[22] M. E. Delamaro, L. Deng, V. H. S. Durelli, N. Li, and J. Offutt, "Experimental evaluation of sdl and one-op mutation for c," in *2014 IEEE Seventh International Conference on Software Testing, Verification and Validation*. IEEE, 2014, pp. 203–212.

[23] Y. Jia and M. Harman, "An analysis and survey of the development of mutation testing," *IEEE transactions on software engineering*, vol. 37, no. 5, pp. 649–678, 2010.

[24] M. Papadakis, M. Kintis, J. Zhang, Y. Jia, Y. Le Traon, and M. Harman, "Mutation testing advances: an analysis and survey," in *Advances in Computers*. Elsevier, 2019, vol. 112, pp. 275–378.

[25] A. J. Offutt and J. Pan, "Automatically detecting equivalent mutants and infeasible paths," *Software testing, verification and reliability*, vol. 7, no. 3, pp. 165–192, 1997.

[26] M. Papadakis, D. Shin, S. Yoo, and D.-H. Bae, "Are mutation scores correlated with real fault detection? a large scale empirical study on the relationship between mutants and real faults," in *2018 IEEE/ACM 40th International Conference on Software Engineering (ICSE)*. IEEE, 2018, pp. 537–548.

[27] T. T. Chekam, M. Papadakis, Y. Le Traon, and M. Harman, "An empirical study on mutation, statement and branch coverage fault revelation that avoids the unreliable clean program assumption," in *2017 IEEE/ACM 39th International Conference on Software Engineering (ICSE)*, 2017, pp. 597–608.

[28] P. Ammann and J. Offutt, *Introduction to software testing*. Cambridge University Press, 2016.

[29] J. A. Offut and S. D. Lee, "An Empirical Evaluation of Weak Mutation," *IEEE Transactions on Software Engineering*, vol. 20, no. 5, pp. 337–344, 1994.

[30] M. Woodward and K. Halewood, "From weak to strong, dead or alive? an analysis of some mutation testing issues," in *[1988] Proceedings. Second Workshop on Software Testing, Verification, and Analysis*, 1988, pp. 152–158.

[31] P. R. Mateo, M. P. Usaola, and J. L. F. Aleman, "Validating second-order mutation at system level," *IEEE Transactions on Software Engineering*, vol. 39, no. 4, pp. 570–587, 2012.

[32] A. J. Offutt, A. Lee, G. Rothermel, R. H. Untch, and C. Zapf, "An experimental determination of sufficient mutant operators," *ACM Transactions on Software Engineering and Methodology (TOSEM)*, vol. 5, no. 2, pp. 99–118, 1996.

[33] G. Rothermel, R. H. Untch, and C. Zapf, "An experimental determination of sufficient mutant operators a. jefferson offutt ammei lee george mason university," *ACM Transactions on software engineering methodology*, vol. 5, no. 2, pp. 99–118, 1996.

[34] J. H. Andrews, L. C. Briand, and Y. Labiche, "Is mutation an appropriate tool for testing experiments?" in *Proceedings of the 27th international conference on Software engineering*. ACM, 2005, pp. 402–411.

[35] M. Kintis, M. Papadakis, Y. Jia, N. Malevris, Y. Le Traon, and M. Harman, "Detecting trivial mutant equivalences via compiler optimisations," *IEEE Transactions on Software Engineering*, vol. 44, no. 4, pp. 308–333, 2017.

[36] A. Siami Namin, J. H. Andrews, and D. J. Murdoch, "Sufficient mutation operators for measuring test effectiveness," in *Proceedings of the 30th international conference on Software engineering*. ACM, 2008, pp. 351–360.

[37] R. Just, D. Jalali, L. Inozemtseva, M. D. Ernst, R. Holmes, and G. Fraser, "Are mutants a valid substitute for real faults in software testing?" in *Proceedings of the 22nd ACM SIGSOFT International Symposium on Foundations of Software Engineering*, ser. FSE 2014. New York, NY, USA: Association for Computing Machinery, 2014, p. 654–665.

[38] M. Kintis, M. Papadakis, A. Papadopoulos, E. Valvis, N. Malevris, and Y. Le Traon, "How effective are mutation testing tools? An empirical analysis of Java mutation testing tools with manual analysis and real faults," *Empirical Software Engineering*, vol. 23, no. 4, pp. 2426–2463, aug 2018. [Online]. Available: https://doi.org/10.1007/s10664-017-9582-5

[39] L. Deng, J. Offutt, and N. Li, "Empirical evaluation of the statement deletion mutation operator," in *2013 IEEE Sixth International Conference on Software Testing, Verification and Validation*, 2013, pp. 84–93.

[40] Y. Jia and M. Harman, "Higher order mutation testing," *Information and Software Technology*, vol. 51, no. 10, pp. 1379–1393, 2009.

[41] M. Kintis, M. Papadakis, and N. Malevris, "Evaluating mutation testing alternatives: A collateral experiment," in *2010 Asia Pacific Software Engineering Conference*. IEEE, 2010, pp. 300–309.

[42] A. J. Offutt, "Investigations of the software testing coupling effect," *ACM Transactions on Software Engineering and Methodology (TOSEM)*, vol. 1, no. 1, pp. 5–20, 1992.

[43] M. Papadakis and N. Malevris, "An empirical evaluation of the first and second order mutation testing strategies," in *2010 Third International Conference on Software Testing, Verification, and Validation Workshops*. IEEE, 2010, pp. 90–99.

[44] M. Papadakis, N. Malevris, and M. Kintis, "Mutation testing strategies-a collateral approach." in *ICSOFT (2)*, 2010, pp. 325–328.

[45] R. H. Untch, A. J. Offutt, and M. J. Harrold, "Mutation analysis using mutant schemata," in *ACM SIGSOFT Software Engineering Notes*, vol. 18, no. 3. ACM, 1993, pp. 139–148.

[46] M. Papadakis and N. Malevris, "Automatic mutation test case generation via dynamic symbolic execution," in *2010 IEEE 21st International Symposium on Software Reliability Engineering*. IEEE, 2010, pp. 121–130.

[47] M. Becker, D. Baldin, C. Kuznik, M. M. Joy, T. Xie, and W. Mueller, "Xemu: an efficient qemu based binary mutation testing framework for embedded software," in *Proceedings of the tenth ACM international conference on Embedded software*, 2012, pp. 33–42.

[48] Y. Crouzet, H. Waeselynck, B. Lussier, and D. Powell, "The sesame experience: from assembly languages to declarative models," in *Second Workshop on Mutation Analysis (Mutation 2006-ISSRE Workshops 2006)*. IEEE, 2006, pp. 7–7.

[49] Y.-S. Ma, J. Offutt, and Y.-R. Kwon, "Mujava: a mutation system for java," in *Proceedings of the 28th international conference on Software engineering*. ACM, 2006, pp. 827–830.

[50] A. Derezinska and K. Kowalski, "Object-oriented mutation applied in common intermediate language programs originated from c," in *2011 IEEE Fourth International Conference on Software Testing, Verification and Validation Workshops*. IEEE, 2011, pp. 342–350.

[51] F. Hariri, A. Shi, H. Converse, S. Khurshid, and D. Marinov, "Evaluating the effects of compiler optimizations on mutation testing at the compiler ir level," in *2016 IEEE 27th International Symposium on Software Reliability Engineering (ISSRE)*. IEEE, 2016, pp. 105–115.

[52] F. Hariri, A. Shi, V. Fernando, S. Mahmood, and D. Marinov, "Comparing mutation testing at the levels of source code and compiler intermediate representation," in *2019 12th IEEE Conference on Software Testing, Validation and Verification (ICST)*. IEEE, 2019, pp. 114–124.

[53] A. Denisov and S. Pankevich, "Mull it over: mutation testing based on llvm," in *2018 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*. IEEE, 2018, pp. 25–31.

[54] Cobham Gaisler, "RTEMS Cross Compilation System," https://www.gaisler.com/index.php/products/operating-systems/rtems, 2020.

[55] M. E. Delamaro, J. C. Maldonado, and A. Mathur, "Proteum-a tool for the assessment of test adequacy for c programs user's guide," in *PCS*, vol. 96, 1996, pp. 79–95.

[56] K. N. King and A. J. Offutt, "A fortran language system for mutation-based software testing," *Software: Practice and Experience*, vol. 21, no. 7, pp. 685–718, 1991.

[57] S. Tokumoto, H. Yoshida, K. Sakamoto, and S. Honiden, "Muvm: Higher order mutation analysis virtual machine for c," in *2016 IEEE International Conference on Software Testing, Verification and Validation (ICST)*. IEEE, 2016, pp. 320–329.

[58] L. Zhang, S.-S. Hou, J.-J. Hu, T. Xie, and H. Mei, "Is operator-based mutant selection superior to random mutant selection?" in *Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering-Volume 1*. ACM, 2010, pp. 435–444.

[59] P. Tchebichef, "Des valeurs moyennes," *Journal de Mathématiques Pures et Appliquées*, vol. 2, no. 12, p. 177–184, 1867.

[60] R. R. Bahadur, *A Representation of the Joint Distribution of Responses to N Dichotomous Items*. Stanford University Press, 1961, pp. 158–168.

[61] L. L. Kupper and J. K. Haseman, "The Use of a Correlated Binomial Model for the Analysis of Certain Toxicological Experiments," *Biometrics*, vol. 34, no. 1, p. 69, 1978.

[62] T.-H. Ng, "A new class of modified binomial distributions with applications to certain toxicological experiments," *Communications in Statistics - Theory and Methods*, vol. 18, no. 9, pp. 3477–3492, 1989. [Online]. Available: https://doi.org/10.1080/03610928908830104

[63] P. A. G. Van Der Geest, "The binomial distribution with dependent bernoulli trials," *Journal of Statistical Computation and Simulation*, vol. 75, no. 2, pp. 141–154, 2005.

[64] N. F. Zhang, "The Use of Correlated Binomial Distribution in Estimating Error Rates for Firearm Evidence Identification," vol. 124, no. 124026, pp. 1–16, 2019.

[65] R. V. Krejcie and D. W. Morgan, "Determining sample size for research activities," *Educational and Psychological Measurement*, vol. 30, no. 3, pp. 607–610, 1970. [Online]. Available: https://doi.org/10.1177/001316447003000308

[66] W. G. Cochran, *Sampling Techniques*. New York, USA: John Wiley & Sons, 1977.

[67] J. E. Bartlett, J. W. Kotrlik, and C. C. Higgins, "Organizational research: Determining appropriate sample size in survey research," *Information Technology, Learning, and Performance Journal*, vol. 19, no. 1, pp. 43–50, 2001.

[68] K. Krishnamoorthy and J. Peng, "Some properties of the exact and score methods for binomial proportion and sample size calculation," *Communications in Statistics - Simulation and Computation*, vol. 36, no. 6, pp. 1171–1186, 2007. [Online]. Available: https://doi.org/10.1080/03610910701569218

[69] "Sample size for estimating a binomial proportion: comparison of different methods," *Journal of Applied Statistics*, vol. 39, no. 11, pp. 2453–2473, nov 2012. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/02664763.2012.713919

[70] A. Wald, "Sequential tests of statistical hypotheses," *Ann. Math. Statist.*, vol. 16, no. 2, pp. 117–186, 06 1945. [Online]. Available: https://doi.org/10.1214/aoms/1177731118

[71] E. S. William Hsu, Mehmet Sahinoglu, "An experimental approach to statistical mutation-based testing," Software Engineering Research Center, West Lafayette, IN 47907, Tech. Rep. SERC-TR-63-P, April 1990.

[72] J. Frey, "Fixed-width sequential confidence intervals for a proportion," *The American Statistician*, vol. 64, no. 3, pp. 242–249, 2010. [Online]. Available: https://doi.org/10.1198/tast.2010.09140

[73] Z. Chen and X. Chen, "Exact Group Sequential Methods for Estimating a Binomial Proportion," *Journal of Probability and Statistics*, vol. 2013, p. 603297, 2013. [Online]. Available: https://doi.org/10.1155/2013/603297

[74] T. Yaacoub, G. V. Moustakides, and Y. Mei, "Optimal stopping for interval estimation in bernoulli trials," *IEEE Transactions on Information Theory*, vol. 65, no. 5, pp. 3022–3033, 2019.

[75] S. E. Vollset, "Confidence intervals for a binomial proportion," *Statistics in Medicine*, vol. 12, no. 9, pp. 809–824, 1993. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4780120902

[76] C. J. Clopper and E. S. Pearson, "The use of confidence or fiducial limits illustrated in the case of the binomial," *Biometrika*, vol. 26, no. 4, pp. 404–413, 1934. [Online]. Available: http://www.jstor.org/stable/2331986

[77] R. Just, G. M. Kapfhammer, and F. Schweiggert, "Using non-redundant mutation operators and test suite prioritization to achieve efficient and scalable mutation analysis," in *2012 IEEE 23rd International Symposium on Software Reliability Engineering*. IEEE, 2012, pp. 11–20.

[78] M. Papadakis and N. Malevris, "Automatically performing weak mutation with the aid of symbolic execution, concolic testing and search-based testing," *Software Quality Journal*, vol. 19, no. 4, p. 691, 2011.

[79] T. A. Budd and D. Angluin, "Two notions of correctness and their relation to testing," *Acta Informatica*, vol. 18, no. 1, pp. 31–45, 1982. [Online]. Available: https://doi.org/10.1007/BF00625279

[80] M. Papadakis, Y. Jia, M. Harman, and Y. Le Traon, "Trivial compiler equivalence: A large scale empirical study of a simple, fast and effective equivalent mutant detection technique," in *Proceedings of the 37th International Conference on Software Engineering-Volume 1*. IEEE Press, 2015, pp. 936–946.

[81] D. Holling, S. Banescu, M. Probst, A. Petrovska, and A. Pretschner, "Nequivack: Assessing mutation score confidence," in *2016 IEEE Ninth International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*. IEEE, 2016, pp. 152–161.

[82] T. T. Chekam, M. Papadakis, M. Cordy, and Y. L. Traon, "Killing stubborn mutants with symbolic execution," *ACM Trans. Softw. Eng. Methodol.*, vol. 30, no. 2, Jan. 2021. [Online]. Available: https://doi.org/10.1145/3425497

[83] H. Riener, R. Bloem, and G. Fey, "Test case generation from mutants using model checking techniques," in *2011 IEEE Fourth International Conference on Software Testing, Verification and Validation Workshops*. IEEE, 2011, pp. 388–397.

[84] V. Herdt, D. Große, H. M. Le, and R. Drechsler, "Early concolic testing of embedded binaries with virtual prototypes: A risc-v case study*," in *2019 56th ACM/IEEE Design Automation Conference (DAC)*, 2019, pp. 1–6.

[85] OASIS, "OASIS Web Services Business Process Execution Language (WSBPEL) TC," https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=wsbpel, 2007.

[86] ECMA, "ECMA-404 The JSON Data Interchange Standard." https://www.json.org, 2021.

[87] R. Natella, D. Cotroneo, and H. S. Madeira, "Assessing dependability with software fault injection: A survey," *ACM Computing Surveys (CSUR)*, vol. 48, no. 3, p. 44, 2016.

[88] M. Harman, Y. Jia, and W. B. Langdon, "A manifesto for higher order mutation testing," in *2010 Third International Conference on Software Testing, Verification, and Validation Workshops*. IEEE, 2010, pp. 80–89.

[89] M. E. Delamaro, J. Maidonado, and A. P. Mathur, "Interface mutation: An approach for integration testing," *IEEE transactions on software engineering*, vol. 27, no. 3, pp. 228–247, 2001.

[90] M. Grechanik and G. Devanla, "Mutation integration testing," in *2016 IEEE International Conference on Software Quality, Reliability and Security (QRS)*, 2016, pp. 353–364.

[91] Y. Jiang, S.-S. Hou, J.-H. Shan, L. Zhang, and B. Xie, "Contract-based mutation for testing components," in *21st IEEE International Conference on Software Maintenance (ICSM'05)*, 2005, pp. 483–492.

[92] P. R. Mateo, M. P. Usaola, and J. Offutt, "Mutation at system and functional levels," in *2010 Third International Conference on Software Testing, Verification, and Validation Workshops*. IEEE, 2010, pp. 110–119.

[93] T. K. Tsai, M.-C. Hsueh, H. Zhao, Z. Kalbarczyk, and R. K. Iyer, "Stress-based and path-based fault injection," *IEEE Transactions on Computers*, vol. 48, no. 11, pp. 1183–1201, 1999.

[94] J. H. Barton, E. W. Czeck, Z. Z. Segall, and D. P. Siewiorek, "Fault injection experiments using fiat," *IEEE Transactions on Computers*, vol. 39, no. 4, pp. 575–582, 1990.

[95] S. Han, K. G. Shin, and H. A. Rosenberg, "Doctor: An integrated software fault injection environment for distributed real-time systems," in *Proceedings of 1995 IEEE International Computer Performance and Dependability Symposium*. IEEE, 1995, pp. 204–213.

[96] S. Dawson, F. Jahanian, T. Mitton, and T.-L. Tung, "Testing of fault-tolerant and real-time distributed systems via protocol fault injection," in *Proceedings of Annual Symposium on Fault Tolerant Computing*. IEEE, 1996, pp. 404–414.

[97] D. Di Nardo, F. Pastore, A. Arcuri, and L. Briand, "Evolutionary robustness testing of data processing systems using models and data mutation (t)," in *2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 2015, pp. 126–137.

[98] D. Di Nardo, F. Pastore, and L. Briand, "Generating complex and faulty test data through model-based mutation analysis," in *2015 IEEE 8th International Conference on Software Testing, Verification and Validation (ICST)*. IEEE, 2015, pp. 1–10.

[99] R. Matinnejad, S. Nejati, L. C. Briand, and T. Bruckmann, "Test generation and test prioritization for simulink models with dynamic behavior," *IEEE Transactions on Software Engineering*, vol. 45, no. 9, pp. 919–944, Sep. 2019.

[100] Peach Tech, "Peach Fuzzer." [Online]. Available: https://www.peach.tech

[101] A. K. Ghosh, M. Schmid, and V. Shah, "Testing the robustness of windows nt software," in *Proceedings Ninth International Symposium on Software Reliability Engineering (Cat. No. 98TB100257)*. IEEE, 1998, pp. 231–235.

[102] P. Godefroid, A. Kiezun, and M. Y. Levin, "Grammar-based whitebox fuzzing," in *Proceedings of the 29th ACM SIGPLAN Conference on Programming Language Design and Implementation*, ser. PLDI ?08. New York, NY, USA: Association for Computing Machinery, 2008, p. 206?215.

[103] P. Godefroid, M. Y. Levin, and D. Molnar, "Sage: whitebox fuzzing for security testing," *Communications of the ACM*, vol. 55, no. 3, pp. 40–44, 2012.

[104] E. Bounimova, P. Godefroid, and D. Molnar, "Billions and billions of constraints: Whitebox fuzz testing in production," in *Proceedings of the 2013 International Conference on Software Engineering*. IEEE Press, 2013, pp. 122–131.

[105] V.-T. Pham, M. Böhme, and A. Roychoudhury, "Model-based whitebox fuzzing for program binaries," in *2016 31st IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 2016, pp. 543–553.

[106] C. Cadar, D. Dunbar, D. R. Engler *et al.*, "Klee: Unassisted and automatic generation of high-coverage tests for complex systems programs." in *OSDI*, vol. 8, 2008, pp. 209–224.

[107] A. Biere, A. Cimatti, E. M. Clarke, and Y. Zhu, "Symbolic model checking without BDDs," in *proceedings of the International Conference Tools and Algorithms for Construction and Analysis of Systems*, ser. LNCS, vol. 1579. Springer, 1999, pp. 193–207.

[108] O. Sery, G. Fedyukovich, and N. Sharygina, "FunFrog: Bounded model checking with interpolation-based function summarization," in *proceedings of the International Symposium on Automated Technology for Verification and Analysis*, ser. LNCS, vol. 7561. Springer, 2012, pp. 203–207.

[109] J. P. M. Silva and K. A. Sakallah, "Grasp: A search algorithm for propositional satisfiability," *IEEE Transactions on Computers*, vol. 48, no. 5, pp. 506–521, 1999.

[110] N. Eén and N. Sörensson, "An extensible SAT-solver," in *proceedings of the International Conference on Theory and Applications of Satisfiability Testing*, ser. LNCS, vol. 2919. Springer, 2004, pp. 502–518.

[111] L. Bordeaux, Y. Hamadi, and L. Zhang, "Propositional satisfiability and constraint programming: A comparative survey," *ACM Comput. Surv.*, vol. 38, no. 4, p. 12?es, Dec. 2006. [Online]. Available: https://doi.org/10.1145/1177352.1177354

[112] T. T. Chekam, M. Papadakis, M. Cordy, and Y. L. Traon, "Killing stubborn mutants with symbolic execution," *ACM Transactions on Software Engineering and Methodology (TOSEM)*, vol. 30, no. 2, pp. 1–23, 2021.

[113] S. Person, M. B. Dwyer, S. Elbaum, and C. S. Pasareanu, "Differential symbolic execution," in *Proceedings of the 16th ACM SIGSOFT International Symposium on Foundations of software engineering*, 2008, pp. 226–237.

[114] M. Harman, Y. Jia, and W. B. Langdon, "Strong higher order mutation-based test data generation," in *Proceedings of the 19th ACM SIGSOFT symposium and the 13th European conference on Foundations of software engineering*. ACM, 2011, pp. 212–222.

[115] K. Ayari, S. Bouktif, and G. Antoniol, “Automatic mutation test input data generation via ant colony,” in *Proceedings of the 9th annual conference on Genetic and evolutionary computation*, 2007, pp. 1074–1081.

[116] G. Fraser and A. Arcuri, “Achieving scalable mutation-based generation of whole test suites,” *Empirical Software Engineering*, vol. 20, no. 3, pp. 783–812, 2015.

[117] E. T. Barr, M. Harman, P. McMinn, M. Shahbaz, and S. Yoo, “The oracle problem in software testing: A survey,” *IEEE Transactions on Software Engineering*, vol. 41, no. 5, pp. 507–525, 2015.

[118] R. A. Oliveira, U. Kanewala, and P. A. Nardi, “Chapter three - automated test oracles: State of the art, taxonomies, and trends,” ser. Advances in Computers, A. Memon, Ed. Elsevier, 2014, vol. 95, pp. 113 – 199. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780128001608000036

[119] G. Fraser and A. Zeller, “Mutation-driven generation of unit tests and oracles,” *IEEE Transactions on Software Engineering*, vol. 38, no. 2, pp. 278–292, 2011.

[120] C. Pacheco, S. K. Lahiri, M. D. Ernst, and T. Ball, “Feedback-directed random test generation,” in *ICSE 2007, Proceedings of the 29th International Conference on Software Engineering*, Minneapolis, MN, USA, May 2007, pp. 75–84.

[121] T. Xie, “Augmenting automatically generated unit-test suites with regression oracle checking,” in *ECOOP 2006 –Object-Oriented Programming*, D. Thomas, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 380–403.

[122] K. Taneja and T. Xie, “Diffgen: Automated regression unit-test generation,” in *Proceedings of the 2008 23rd IEEE/ACM International Conference on Automated Software Engineering*, ser. ASE ?08. USA: IEEE Computer Society, 2008, p. 407?410. [Online]. Available: https://doi.org/10.1109/ASE.2008.60

[123] M. Staats, G. Gay, and M. P. E. Heimdahl, “Automated oracle creation support, or: How i learned to stop worrying about fault propagation and love mutation testing,” in *2012 34th International Conference on Software Engineering (ICSE)*, June 2012, pp. 870–880.

[124] F. Pastore and L. Mariani, “Zoomin: Discovering failures by detecting wrong assertions,” in *2015 IEEE/ACM 37th IEEE International Conference on Software Engineering*, vol. 1, May 2015, pp. 66–76.