SEARCH-BASED STRESS TEST: AN APPROACH APPLING EVOLUTIONARY ALGORITHMS AND TRAJECTORY METHODS

A THESIS
SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE
OF UNIVERSIDADE DE FORTALEZA
(UNIFOR)
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF SCIENCE

Francisco Nauber Bernardo Gois
July 2017

I certify that I have read this thesis and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Science.

D.Sc. Pedro Porfírio Muniz de Farias (UNIFOR) Principal Adviser

I certify that I have read this thesis and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Science.

D.Sc. André Luís Vasconcelos Coelho (UNIFOR) Co-Adviser

I certify that I have read this thesis and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Science.

D. Sc. Pedro de Alcantara dos Santos Neto (UFPI)

I certify that I have read this thesis and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Science.

D. Sc. João Paulo Pordeus Gomes (UFC)

I certify that I have read this thesis and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Science.

Ph.D. Americo Tadeu Falcone Sampaio (UNIFOR)

I certify that I have read this thesis and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Science.

Docteur Napoleão Vieira Nepomuceno (UNIFOR)

Learn from yesterday, live for today, hope for tomorrow. The important thing is not to stop questioning.

Preface

The doctoral thesis at hand encompasses three research papers. First research paper discuss about use a hybrid single metaheuristics in stress tests.

Acknowledgments

I would like to thank...

Abstract

Some software systems must respond to thousands or millions of concurrent requests. These systems must be properly tested to ensure that they can function correctly under the expected load. A common use of stress testing is to find test scenarios that produce execution times that violate the timing constraints specified. In this context, search-based testing is seen as a promising approach for verifying timing constraints. In this thesis, We proposed hybrid metaheuristic approach that uses genetic algorithms, simulated annealing, and tabu search algorithms in a collaborative model using Q-Learning to improve stress search-based testing and automation. A tool named IAdapter, a JMeter plugin used for performing search-based stress tests, was developed. Four experiments were conducted to validate the proposed approach.

Related Publications

The following publications are related to this thesis:

- **N. Gois, P. Porfirio and A. Coelho**. A multi-objective metaheuristic approach to search-based stress testing. In Proceedings of the 2017
- **N. Gois, P. Porfirio, A. Coelho, and T. Barbosa**. Improving Stress Search Based Testing using a Hybrid Metaheuristic Approach. In Proceedings of the 2016 Latin American Computing Conference (CLEI), pages 718–728, 2016 [51].

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This chapter briefly introduces this work. It presents the motivations, the objectives, contributions and the structure of this work.

Performance problems such as high response times in software applications have a significant effect on the customers' satisfaction. The explosive growth of the Internet has contributed to the increased need of applications that perform at an appropriate speed. Performance problems are often detected late in the application life cycle, and the later they are discovered, the greater the cost is to fix them. The use of stress testing is an increasingly common practice owing to the increasing number of users. In this scenario, the inadequate treatment of a workload, generated by concurrent or simultaneous access due to several users, can result in highly critical failures and negatively affect the customers' perception of the company [66] [82] [122].

Software testing is an expensive and difficult activity. The exponential growth in the complexity of software makes the cost of testing to only continue to rise. Test case generation can be seen as a search problem. The test adequacy criterion is transformed into a fitness function and a set of solutions in the search space are evaluated with respect to the fitness function using a metaheuristic search technique. Search-based software testing is the application of metaheuristic search techniques to generate software tests cases or perform test executions [2] [47].

Search-based testing (SBST) is seen as a promising approach to verify timing constraints [2]. A common objective of a load search-based test is to find scenarios that produce execution times that violate the specified timing constraints [105]. Experiments involving search-based tests are inherently complex and typically time-consuming to set up and execute. Such experiments are also extremely difficult to repeat. People who might want to duplicate published results, for example, must devote substantial resources to setting up and the environmental conditions are likely to be substantially different.

Usually, search-based test methods are based on single objective optimization. Multi-objective evolutionary algorithms (MOEAs) are widely used for solving multi-objective problems (MOPs) because they produce a complete set of solutions in a single run. The NSGA-II is a genetic algorithm (GA) based on obtaining a

new offspring population from the original one by applying the typical genetic operators (selection, crossover, and mutation); then, the individuals in the two populations are sorted according to their rank, and the best solutions are chosen to create a new population.

This thesis addresses the use of hybrid and multi-objective metaheuristics in conjunction with reinforcement learning techniques in search-based tests. Reinforcement learning (RL) refers to both a learning problem and a subfield of machine learning. As a learning problem, it refers to learning to control a system so as to maximize some numerical value which represents a long-term objective. The basic idea of Reinforcement learning is simply to capture the most important aspects of the real problem, facing a learning agent interacting with its environment to achieve a goal [106]. Reinforcement learning is learning what to do—how to map situations to actions—so as to maximize a numerical reward signal. The learner needs to discover which actions yield the most reward by trying them [106].

A tool named IAdapter (www.iadapter.org, github.com/naubergois/newiadapter), a JMeter plugin for performing search-based load tests, was extended [51]. Several experiments was conducted to validate the proposed approach. The experiment uses the NSGA-II algorithm with one objective: discover application scenarios where there is a high response time for a small number of users. The relevance of finding scenarios with high response times is to enable corrective actions before the application under test is released in a production environment.

1.1 Motivation

There is strong empirical evidence, that deficient testing of both functional and nonfunctional properties is one of the major sources of software and system errors. In 2002, NIST report found that more than one-third of these costs of software failure could be eliminated by an improved testing infrastructure. Automation of testing is a crucial concern. Through automation, large-scale thorough testing can become practical and scalable. However, the automated generation of test cases presents challenges. The general problem involves finding a (partial) solution to the path sensitization problem. That is, the problem of finding an input to drive the software down a chosen path [58] [30].

Software performance is a pervasive quality, because it is affected by every aspect of the design, code, and execution environment. Performance failures occur when a software product is unable to meet its overall objectives due to inadequate performance. Such failures negatively impact the projects by increasing costs, decreasing revenue or both [114]. Stress testing of enterprise applications is manual,laborious, costly, and not particulary effective. When running many different test cases and observing application's behavior, testers intuitively sense that there are certain properties of test cases that are likely to reveal performance bugs [53]. Manual analysis of load testing is inefficient and error prone due to limited knowledge of test analyst about system under test [12].

Software testing is a expensive and difficult activity. The exponential growth in the complexity of software makes the cost of testing has only continued to rise. Test case generation can be seen as a search problem.

The test adequacy criterion is transformed into a fitness function and a set of solutions in the search space are evaluated with respect to the fitness function using a metaheuristic search technique. Search-based software testing is the application of metaheuristic search techniques to generate software tests cases or perform test execution [2] [47].

Experimentation is important to realistically and accurately test and evaluate search-based stress tests. Experimentation on algorithms is usually made by simulation. Experiments involving search based tests are inherently complex and typically time-consuming to set up and execute. Such experiments are also extremely difficult to repeat. People who might want to duplicate published results, for example, must devote substantial resources to setting up and the environmental conditions are likely to be substantially different.

Below we briefly describe the related research, thesis focus and practices on stress testing.

1.1.1 Thesis Focus and state of Research on the Search-Based Stress Testing

Bianchi et al. presents a classification schema and a general framework for the approaches to test concurrent systems. The schema shows that the test inputs can be test cases or a system model and the selection of the tests can be performed by space exploration or based on properties [20]. The focus of this thesis is use the Test Cases as input performed by a space exploration algorithm. The use of space exploration algorithms on tests is named Search-based tests. Fig 1.1 presents the summary results of the survey presented in the chapter 2. The two largest research areas on stress tests are model-based tests and search-based tests.

In the academic context, a number of studies proving the efficacy of metaheuristics to automate test execution can be found in literature. Figure 1.2 shows the growth in papers published on SBST and SBSE. The data is taken from the SBSE repository (`http://crestweb.cs.ucl.ac.uk/resources/sbse_repository/`). The aim of the SBSE repository is to contain every SBSE paper. Although no repository can guarantee 100% precision and recall, the SBSE repository has proved sufficiently usable that it has formed the basis of several other detailed analyses of the literature [57].

SBST has made many achievements, and demonstrated its wide applicability and increasing uptake. Nevertheless, there are pressing open problems and challenges that need more attention like to extend SBST to test non-functional properties, a topic that remains relatively under-explored, compared to structural testing. The Fig. 1.3 shows the non-funtional SBST by year [5] [57].

There are many kinds of non-functional search based tests [2]:

- Execution time: The application of evolutionary algorithms to find the best and worst case execution times (BCET, WCET).
- Quality of service: uses metaheuristic search techniques to search violations of service level agreements (SLAs).
- Security: apply a variety of metaheuristic search techniques to detect security vulnerabilities like detecting buffer overflows.

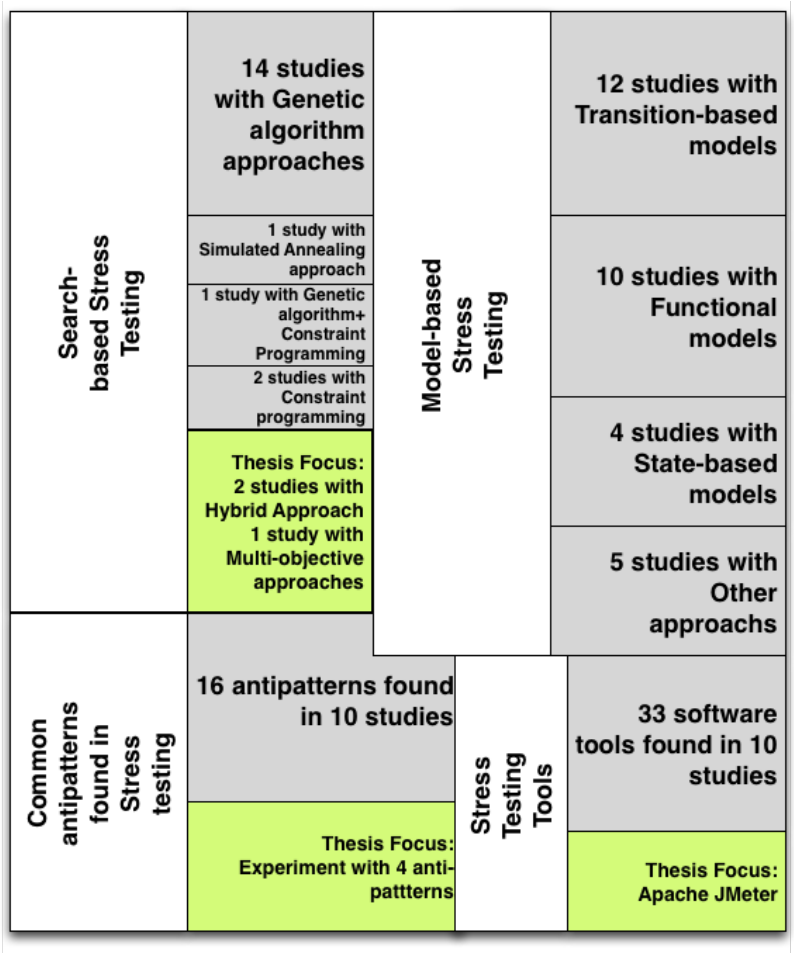


Figure 1.1: Summary of state of art

- Usability: concerned with construction of covering array which is a combinatorial object.
- Safety: Safety testing is an important component of the testing strategy of safety critical systems where the systems are required to meet safety constraints.

There is a great difficulty in comparing some approaches present in the state of the art, due to the lack of availability of the tools used. The present research intends to compare the proposed approach with other methods based on the use of single or multiobjective metaheuristics.A variety of metaheuristic search techniques are found to be applicable for non-functional testing including simulated annealing, tabu search, genetic algorithms, ant colony methods, grammatical evolution, genetic programming and swarm intelligence methods. However, most research studies are limited to making prototypes [2]. Fig. 1.4 shows a comparison between the range of metaheuristcs and the type of non-functional search based test.The Data comes from Afzal et al. [2]. Afzal's work adds to some of the latest research in this area ([44] [46] [33] [34] [4] [51]). The

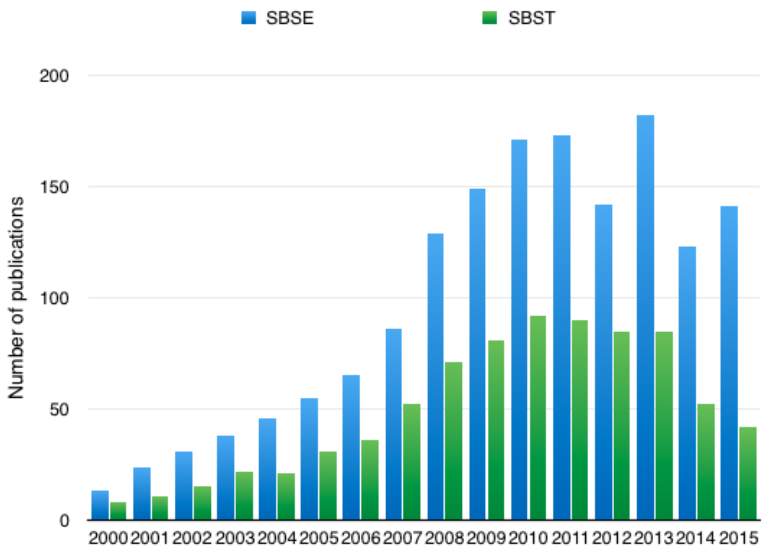


Figure 1.2: Number of publications in SBSE and SBST by Year. Data comes from the Harman et al., Afzal et al. and the SBSE repository [2] [57]

thesis focus is use hybrid and multiobjective metaheuristics in Quality of Service and Execution Time tests (Search-based stress testing).

1.1.2 State of Industrial Practices on Stress Tests

The stress testing process in the industry still follows a non-automated and ad-hoc model where the designer or tester is responsible for running the tests analyzing the results and deciding which new tests should be performed [72].

Typically, performance testing is accomplished using test scripts, which are programs that test designers write to automate testing. These test scripts performs actions or mimicking user actions on GUI objects of the system to feed input data. Current approaches to load testing suffer from limitations. Their cost-effectiveness is highly dependent on the particular test scenarios that are used yet there is no support for choosing those scenarios. A poor choice of scenarios could lead to underestimating system response time thereby missing an opportunity to detect a performance [53].

1.2 Research Hypothesis

Our underlying research hypothesis is as follows:

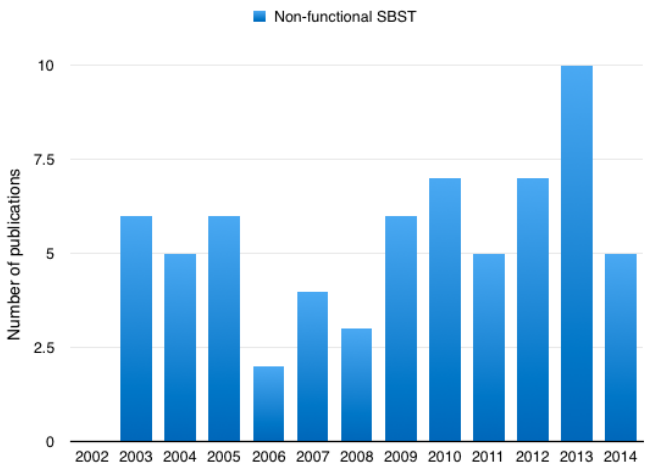


Figure 1.3: Number of publications in non-functional SBST by Year. Data comes from the Harman et al., Afzal et al. and the SBSE repository [2] [57]

The use of metaheuristics and hybrid metaheuristics in combination with Q-learning improving the choice of new test cases for each interaction, finding best or worst case scenarios.

The purpose of this thesis is to show the validity of this hypothesis through the development of a testbed tool, algorithms that use hybrid metaheuristics and the Q-learning technique and application of validation experiments. This thesis will be useful for load test practitioners and software engineering researchers interested in large-scale testing software systems.

1.3 Contributions

The main contributions of this thesis are follows:

- Hybrid algorithm approach using Tabu Search, Simulated Annealing and Genetic Algorithms for search-based stress testing (Chapter 4) [51].
- Hybrid algorithm with Q-Learning approach (Chapter 5).
- Comprehensive investigation on the use of multi-objective metaheuristics on search-based stress testing (Chapter 6).

The secondary contributions of this thesis are follows:

- The IAdapter tool

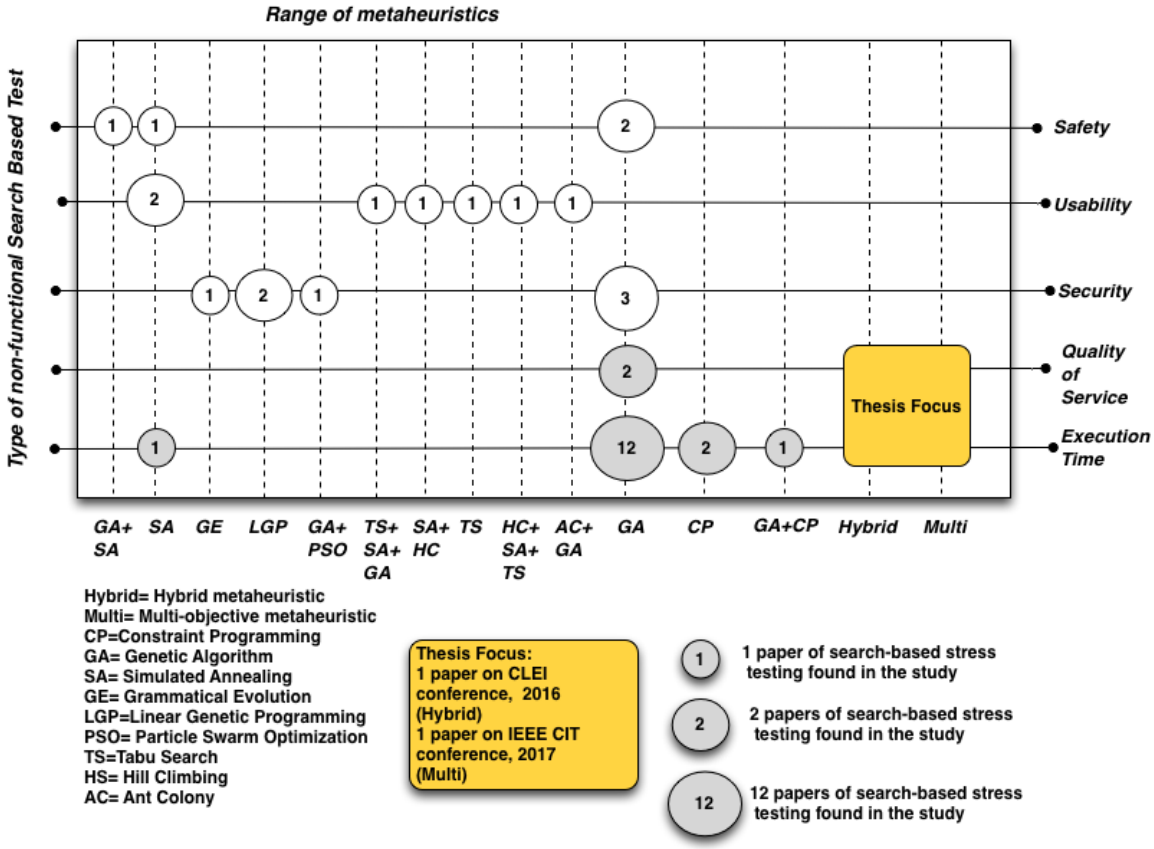


Figure 1.4: Range of metaheuristics by Type of non-functional Search Based Test[2].

1.4 Thesis Outline

This thesis is organized as follows.

Chapter 3 reviews the main existing approaches in the research area referring to the search-based stress testing. In particular, two categories of approaches are out-lined: (i) Search-Based Stress Testing on Safety-critical systems; (ii) Search-Based Stress Testing on Industrial systems. In the context of the search-based process two metrics can be applied: processor cycles or response time.

Chapter 2

A Survey on Stress Testing Software Systems

This chapter surveys the state of the art literature in stress testing research. The thesis extends the survey presented by Jiang et al. [66] and Afzal et al. [2] to the Stress Testing context . This survey will be useful for stress testing practitioners and software engineering researchers with interests in testing and analyzing software systems. The paper use the systematic review method proposed by Kitchenham [69].

The aim of a systematic review is to find as many primary studies relating to the research question as possible using an unbiased search strategy. The rigour of the search process is one factor that distinguishes systematic reviews from traditional reviews [69]. The systematic review is based on a comprehensive set of 97 articles obtained after a multi-stage selection process and have been published in the time span 1994–2016.

2.0.1 Planning a Systematic Review

A systematic review of the literature details a protocol describing the process and the methods to be applied. The most important activity during the planning phase is the formulation of research questions. To Kitchenham, before undertaking a systematic review researchers must ensure that it is necessary and the protocol should be able to answer some questions [76]:

- What are the objectives of this review?
- What sources were searched to identify primary studies? Were there any restrictions?
- What were the criteria for inclusion / exclusion and how they are applied?
- What criteria were used to evaluate the quality of the primary studies?
- How were the quality criteria applied?

- How was the data extracted from primary studies?
- What were the differences between studies investigated?
- Because the data were combined?

2.0.2 Research Questions

In order to examine the evidence of stress testing properties, we proposed the following two research questions:

- How modeling a stress test?
- What are the main problems found by stress tests?

2.0.3 Generation of search strategy

The population in this study is the domain of software testing. Intervention includes application of stress test techniques to test different types of non-functional properties. The primary studies used in this review were obtained from searching databases of peer-reviewed software engineering research that met the following criteria:

- Contains peer-reviewed software engineering journals articles, conference proceedings, and book chapters.
- Contains multiple journals and conference proceedings, which include volumes that range from 1996 to 2017.
- Used in other software engineering systematic reviews.

The resulting list of databases was:

- ACM Digital Library
- Google Scholar
- IEEE Electronic Library
- Inspec
- Scirus (Elsevier)
- SpringerLink

The search strategy was based on the following steps:

- Identification of alternate words and synonyms for terms used in the research questions. This is done to minimize the effect of differences in terminologies.
- Identify common stress testing properties for searching.
- Use of Boolean OR to join alternate words and synonyms.
- Use of Boolean AND to join major terms

We used the following search terms:

- Load Testing: load test, Load Testing
- Stress Testing: stress test, stress testing
- Performance Testing: performance tests
- Test tools: jmeter, load runner, performance tester

2.0.4 Study selection criteria and procedures for including and excluding primary studies

The idealized selection process was done in two parts: an initial document selection of the results that could reasonably satisfy the selection criteria based on a title and the articles abstract reading, followed by a final selection of the initially selected papers based on the introduction and conclusion reading of the papers. The following exclusion criteria is applicable in this review, i.e. exclude studies that:

- Do not relate to stress testing.
- Do not relate to load testing tool.
- Do not relate to load/stress testing model.

From 366 initial papers, 97 papers was selected.

2.0.5 Data Synthesis

Data synthesis involves collating and summarising the results of the included primary studies. Synthesis can be descriptive (non-quantitative). The studies was categorized by:

- Type of stress test properties;
- Type of research paper (Thesis, Journal Article, Conference Paper, Book Section or Book)
- Methodology used by the test (Model based Test, FOREPOST, Search-based Tests)

2.1 Load, Performance and Stress Testing

Load, performance, and stress testing are typically done to locate bottlenecks in a system, to support a performance-tuning effort, and to collect other performance-related indicators to help stakeholders get informed about the quality of the application being tested [97] [27].

Typically, the most common kind of performance testing for Internet applications is load testing. Application load can be assessed in a variety of ways [87]:

- Concurrency. Concurrency testing seeks to validate the performance of an application with a given number of concurrent interactive users [87].
- Stress. Stress testing seeks to validate the performance of an application when certain aspects of the application are stretched to their maximum limits. This can include maximum number of users, and can also include maximizing table values and data values [87].
- Throughput. Throughput testing seeks to validate the number of transactions to be processed by an application during a given period of time. For example, one type of throughput test might be to attempt to process 100,000 transactions in one hour [87].

The performance testing aims at verifying a specified system performance. This kind of test is executed by simulating hundreds of simultaneous users or more over a defined time interval [35]. The purpose of this assessment is to demonstrate that the system reaches its performance objectives [97]. Term often used interchangeably with “stress” and “load” testing. Ideally “performance” testing is defined in requirements documentation or QA or Test Plans [72].

In a load testing, the system is evaluated at predefined load levels [35]. The aim of this test is to determine whether the system can reach its performance targets for availability, concurrency, throughput, and response time. Load testing is the closest to real application use [82]. A typical load test can last from several hours to a few days, during which system behavior data like execution logs and various metrics are collected [2].

Stress testing investigates the behavior of the system under conditions that overload its resources. The stress testing verifies the system behavior against heavy workloads [97] [72], which are executed to evaluate a system beyond its limits, validate system response in activity peaks, and verify whether the system is able to recover from these conditions. It differs from other kinds of testing in that the system is executed on or beyond its breakpoints, forcing the application or the supporting infrastructure to fail [35] [82].

The main difference between load tests, performance tests and stress tests are:

- Performance tests demonstrate that the system reaches its performance objectives.
- Load tests necessary use a load (concurrent or simultaneous users).
- Stress tests differs from other kinds of testing in that the system is executed on or beyond its breakpoints, forcing the application or the supporting infrastructure to fail.

2.1.1 Stress Test Process

Contrary to functional testing, which has clear testing objectives, Stress testing objectives are not clear in the early development stages and are often defined later on a case-by-case basis. The Fig. 2.1 shows a common Load, Performance and Stress test process [66].

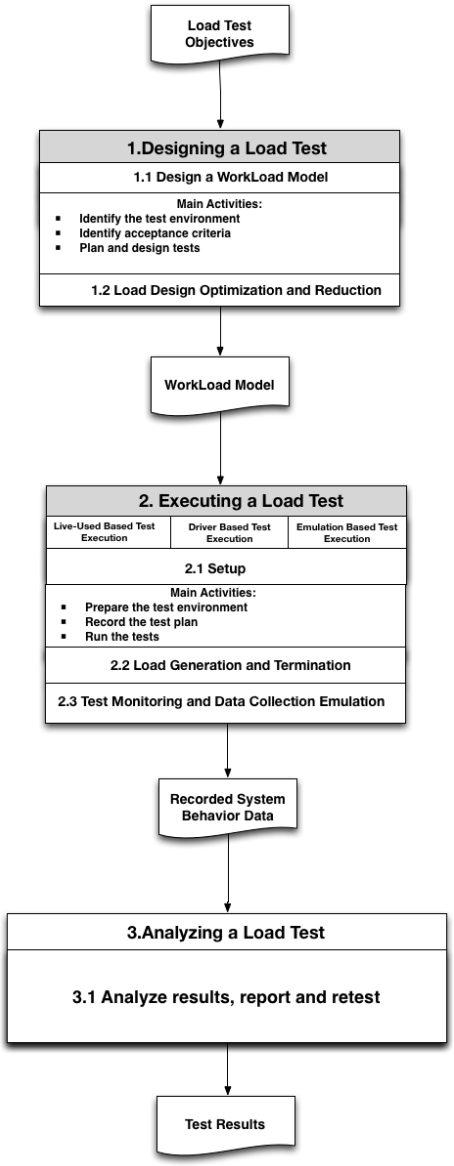


Figure 2.1: Load, Performance and Stress Test Process [66][40]

The goal of the load design phase is to devise a load, which can uncover non-functional problems. Once the load is defined, the system under test executes the load and the system behavior under load is recorded.

Load testing practitioners then analyze the system behavior to detect problems [66].

Once a proper load is designed, a load test is executed. The load test execution phase consists of the following three main aspects: (1) Setup, which includes system deployment and test execution setup; (2) Load Generation and Termination, which consists of generating the load; and (3) Test Monitoring and Data Collection, which includes recording the system behavior during execution[66].

The core activities in conducting an usual Load, Performance and Stress tests are [40]:

- Identify the test environment: identify test and production environments and knowing the hardware, software, and network configurations helps derive an effective test plan and identify testing challenges from the outset.
- Identify acceptance criteria: identify the response time, throughput, and resource utilization goals and constraints.
- Plan and design tests:identify the test scenarios.In the context of testing, a scenario is a sequence of steps in an application. It can represent a use case or a business function such as searching a product catalog, adding an item to a shopping cart, or placing an order [27]. This task includes a description of the speed, availability, data volume throughput rate, response time, and recovery time of various functions, stress, and so on. This serves as a basis for understanding the level of performance and stress testing that may be required to each test scenario [72].
- Prepare the test environment: configure the test environment, tools, and resources necessary to conduct the planned test scenarios.
- Record the test plan: record the planned test scenarios using a testing tool.
- Run the tests: Once recorded, execute the test plans under light load and verify the correctness of the test scripts and output results.
- Analyze results, report, and retest: examine the results of each successive run and identify areas of bottleneck that need addressing.

Utting et al. presents four classic test process [115]:

- Manual Testing process;
- Capture/Replay Testing process;
- Script-based Testing process;
- Keyword-Driven Automated Testing process;

In Manual Testing process the execution is done manually for each test case. The tester follows a test case and interacts directly with the application under test. The manual-designing of the tests is time-consuming and not ensure systematic coverage of the application. Capture/Replay attemps to reduce the cost of test re-execution by capturing interactions with the application during a test execution session and replay those interactions in later test executions [115].

A test script is an executable script that runs one or more test cases. The Script-based testing process tries to resolve the test execution problem by automating it. Keyword-driven testing, or action-word testing, takes this a step further by using action keywords in the test cases, in addition to data. Each action keyword corresponds to a fragment of a test script [115].

2.1.2 Stress Test Execution

The stress test execution consists of deploy the system and setup test execution ; generating the workloads according to the configurations and terminating the load when the load test is completed and recording the system behavior. There are three general approaches of load test executions [82][66]:

- Live-User Based Executions: The test examines a system's behavior when the system is simultaneously used by many users or execute a load test by employing a group of human testers.
- Driver Based Executions: The driver based execution approach automatically generate thousands or millions of concurrent requests for a long period of time using a software tool.
- Emulation Based Executions: The emulation based load test execution approach performs the load testing on special platforms and doesn't require a fully functional system and conduct load testing.

Usually, a stress test execution it is performed with Driver Based Executions approach [40] [81] [117]. There are three categories of load drivers [66]:

- Benchmark Suite: specialized load driver, designed for one type of system. For example, LoadGen is a load driver specified used to load test the Microsoft Exchange MailServer.
- Centralized Load Drivers: refer to a single load driver, which generates the load.
- Peer-to-peer Load Drivers: refer to a set of load drivers, which collectively generate the target testing load. Peer-to-peer load drivers usually have a con- troller component, which coordinates the load generation among the peer load drivers.

A stress test need to perform hundreds or thousands of concurrent requests to the application under test. Automated tools are needed to carry out serious load, stress, and performance testing. Sometimes, there is simply no practical way to provide reliable, repeatable performance tests without using some form of automation. The aim of any automated test tool is to simplify the testing process.

Stress Testing Tools

There are several tools to execution of stress testing. Stress testing tools are software products based on workload models to generate request sequences similar to real requests. They are designed and implemented as versatile software tools for performing tuning or capacity planning studies. Ussualy, the tool functions is semi-automated, whereas the execution of the tests itself is performed by a tool, the choice of scenarios to be executed as well as the decision to start new execution batteries are activities of the test designer or tester. Normally, load test tools uses test scripts. Test scripts are written in a GUI testing framework or a backend server-directed performance tool such as JMeter. These frameworks are the basis on which performance testing is mostly done in industry. Performance test scripts imitate large numbers of users to create a significant load on the application under tests. Stress testing tools typically have the following components [53] [82]:

- Scripting module: Enable recording of end-user activities in different middleware protocols;
- Test management module: Allows the creation of test scenarios;
- Load injectors: Generate the load with multiple workstations or servers;
- Analysis module: Provides the ability to analyse the data collected by each test interation.

Comparing stress test tools is a laborious and difficult task since they offer a large amount and diversity of features [38]. In next subsection we present studies that contrast stress testing tools according to a wide set of features and capabilities, focusing on their ability to realize search-based tests or have learning capacities. In the following subsection, we present details about the JMeter tool and the reasons why it was chosen as object of the present research.The stress test tools was categorized in three different groups [81]:

- Benchmarks that model the client and server paradigm in Web context.
- Software products to evaluate performance and functionality of a given Web application, such as LoadRunner, WebLOAD and JMeter.
- Testing tools and other approaches for traffic generation based on HTTP traces.

Comparative test tool studies

Illes et al. present a systematic approach for evaluation criteria for test tools. Using the TORE methodology the study identify activities which potentially could be automated or at least supported by a test tool. The study evaluate three tools: WinRunner, Rational Robot and HTTrace. WinRunner is distributed by Mercury. Rational Robot is distributed by IBM . HTTrace is not a commercial product, but developed for internal use for the company i-TV-T. The study evaluate the tools in the tasks: TestPlanning and monitoring, designing test cases, constructing test cases, executing test cases and analysing test cases. Key features of all three test tools are the construction of test cases by capturing and subsequently editing the test scripts and the execution

of the recorded test scripts. WinRunner and Rational Robot can be extended to provide test planning and monitoring as well as defect and reporting facilities. HTTrace's strength lies on testing database applications by allowing the reset of consistent database states. Additionally, all three tools can be extended to provide support for testing quality attributes of the system under test [62].

Tables 2.1,2.2 summarizes the studied workloads generators as well as the grade (full or partial) in which they fulfill the features described bellow. None of the presented tools uses heuristic or learning resources when choosing the scenarios to be tested or the workloads to be applied in the test.

Penã-Ortiz et al. present a study where stress test tools are compared using 12 features [81]:

- Distributed architecture. This refers to the ability to distribute the generation process among different nodes.
- Analytical-based architecture. This feature represents the capability to use analytical and mathematical models to define the workload.
- Business-based architecture. When defining a testing environment, the simulator architecture should implement the same features as the real environment.
- Client parameterization. This is the ability to parameterize generator nodes.
- Workload types. Some generators organize the workload in categories or types.
- Testing the Web application functionality (functional testing).
- Multiplatform refers to a software package that is implemented in multiple types of computer platforms, interoperating among them.
- Differences between LAN and WAN.
- The generator should be a friendly application.
- The load test tool has performance reports.
- The load test tool is open-source.
- Users' dynamism. The users has the hability of change they behaviour in during the test.

Benchmarks group

WebStone was designed by Silicon Graphics in 1996 to measure the performance of Web server software and hardware products. Nowadays, both executable and source actualized code for WebStone are available for free. The benchmark generates a Web server load by simulating multiple Web clients navigating a website. All the testing done by the benchmark is controlled by a Webmaster, which is a program that can be run on one of the client computers or on a different one [81] [112] .

Table 2.1: Benchmarks group

Feature/Tool	WebStone	SpecWeb	SURGE	Web Polygraph	TPC-W
Analytical-Based Architecture	Full support	Full support	Full support	Full support	Full support
Distributed Architecture	Full support	Full support	Full support	Full support	
Business-Based Architecture		Partial Suport		Partial Suport	Full support
Client Parameterization	Full support	Full support		Partial Suport	Full support
Workload Types		Full support			Full support
Functional Testing					
LAN and WAN					
Multi-platform	Full support	Full support	Full support	Full support	Full support
Ease of Use					
Performance Reports	Partial Suport	Full support		Full support	Full support
Open Source	Full support		Full support	Partial Suport	Full support
User's Dynamism					Partial Suport

Table 2.2: Software products

Feature/Tool	LoadRunner	WebLOAD	JMeter
Analytical-Based Architecture	Partial Suport	Partial Suport	Partial Suport
Distributed Architecture	Full support	Full support	Full support
Business-Based Architecture	Full support	Full support	Full support
Client Parameterization	Full support	Full support	Full support
Workload Types	Full support	Full support	Partial Suport
Functional Testing	Full support	Full support	Partial Suport
LAN and WAN			
Multi-platform	Full support	Full support	Full support
Ease of Use	Full support	Full support	Partial Suport
Performance Reports	Partial Suport	Full support	Full support
Open Source			Partial Suport
User's Dynamism	Partial Suport	Partial Suport	Partial Suport

TPC Benchmarkt (TPC-W) is a transactional Web benchmark defined by the Transaction Processing Performance Council that models a representative e-commerce evaluating the architecture performance on a generic profile.The models uses a remote browser emulator to generate requests to server under test. TPC-W adopts the CBMG model to define the workloads in spite of this model only characterizing user dynamic behavior partially. The remote browser emulators are located in the client side and generate workload towards the e-commerce Web application, which is located in the server side (e-commerce server) [81] [80].

Open STA is an open source software developed in C++, and released under the GPL licence. OpenSTA provides a script language which permits to simulate the activity of a user. This language can describe HTTP/S scenario and all the test executions is managed in a graphical interface. The composition of the test is very simple, allowing the tester choose scripts for a test and a remote computer that will execute each test.

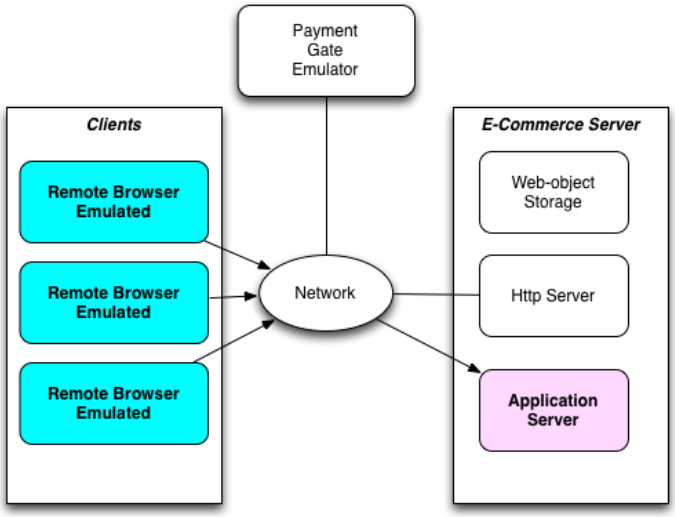


Figure 2.2: TPC-W architecture [81] [80]

Software Products

LoadRunner is one of the most popular industry-standard software products for functional and performance testing. It was originally developed by Mercury Interactive, but nowadays it is commercialized by Hewlett-Packard. LoadRunner supports the definition of user navigations, which are represented using a scripting language. The basic steps are recorded, creating a shell script. Next, this script is then taken off-line, and undergoes further manual steps such as data parameterization and correlations. Finally, the desired performance scripts are obtained after adding transactions and any other required logic (Fig. 2.3). LoadRunner scripting only permits partial reproduction of user dynamism when generating Web workload, because it cannot define either advanced interactions of users, such as parallel browsing behavior, or continuous changes in user's behaviors [81].

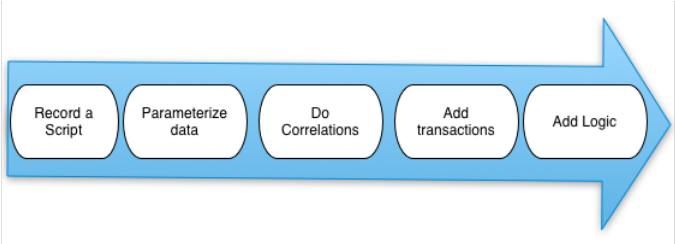


Figure 2.3: Load Runner Scripting

WebLOAD is a software tool for Web performance commercialized by RadView. It is oriented to explore the performance of critical Web applications by quantifying the utilization of the main server resources. The tool creates scenarios that try to mimic the navigations of real users. To this end, it provides facilities to

record, edit and debug test scripts, which are used to define the scenarios on workload characterization. The execution environment is a console to manage test execution, whose results are analyzed in the Analytics application. Since WebLOAD is a distributed system, it is possible to deploy several load generators to reproduce the desired load. Load generators can also be used as probing clients where a single virtual user is simulated to evaluate specific statistics of a single user. These probing clients resemble the experience of a real user using the system while it is under load [81].

Cloud testing tools

Nachiyappan and Justus show a set of other tools perform Cloud testing. Cloud testing is a form of evaluation methodology in which the applications to be tested uses cloud as a computing environment and its infrastructure to simulate real world traffic by using existing cloud computing technologies. Cloud testing are challenged by several problems such as limited budget, meeting deadlines, High cost per test, large number of test cases, little reuse of tests and geographical distributions of users. Blitz is a load-testing tool from the cloud to the cloud. Blitz have no client to install and it is unable to test applications behind firewalls or otherwise protected from the Internet. Blaze Meter is a cloud application based on JMeter scripts that allow stress and load tests on the cloud [85].

SOASTA CloudTest is a production performance testing tool for Web applications. It can simulate thousands of virtual users visiting website simultaneously, using either private or public cloud infrastructure service. The worker nodes can be distributed across public and private clouds to cooperate in a large load testing. Test results from distributed test agents are integrated for analysis [16].

Apache JMeter

Apache jmeter was the tool chosen for current research due to its open license, the use of plugins and the ease of integration with jmetal and jgap frameworks. Apache JMeter is a free open source stress testing tool. It has a large user base and offers lots of plugins to aid testing. JMeter is a desktop application designed to test and measure the performance and functional behavior of applications. The application it's purely Java-based and is highly extensible through a provided API (Application Programming Interface). JMeter works by acting as the client of a client/server application. JMeter allows multiple concurrent users to be simulated on the application [56] [40].

Apache JMeter is user friendly and a flexible open source solution for performance verification. Apache JMeter is designed in pure Java application. It is used to generate heavy loads on the servers or objects to test its strength or analyze overall performance under different load types. To briefly explain the solution how it works, as follows: An Regular Expression Extractor captures the dynamic values as mentioned above and stored in a temporary variable. The values which has been extracted and stored in temporary variables are subsequently utilized by immediate requests/re directions using HTTP samplers [68]. JMeter has components organized in a hierarchical manner. The Test Plan is the main component in a JMeter script. A typical test plan

will consist of one or more Thread Groups, logic controllers, listeners, timers, assertions, and configuration elements:

- Thread Group: Test management module responsible to simulate the users used in a test. All elements of a test plan must be under a thread group.
- Listeners: Analysis module responsible to provide access to the information gathered by JMeter about the test cases .
- Samplers: Load injectors module responsible to send requests to a server, while Logical Controllers let you customize its logic.
- Timers: allow JMeter to delay between each request.
- Assertions: test if the application under test it is returning the correct results.
- Configuration Elements: configure detais about the request protocol and test elements.

2.2 Research Question 1:How is a proper stress designed?

The design of a stress test depends intrinsically on the load model applied to the software under test. Based on the objectives, there are two general schools of thought for designing a proper load to achieve such objectives [2]:

- Designing Realistic Loads (Descriptive Workload).
- Designing Fault-Inducing Loads (Generative Workload).

In Designing Realistic Loads, the main goal of testing is to ensure that the system can function correctly once. Designing Fault-Inducing Loads aims to design loads, which are likely to cause functional or non-functional problems [2].

Stress testing projects should start with the development of a model for user workload that an application receives. This should take into consideration various performance aspects of the application and the infrastructure that a given workload will impact. A workload is a key component of such a model [82].

The term workload represents the size of the demand that will be imposed on the application under test in an execution. The metric used to measure a workload is dependent on the application domain, such as the length of the video in a transcoding application for multimedia files or the size of the input files in a file compression application [42] [82] [52].

Workload is also defined by the load distribution between the identified transactions at a given time. Workload helps researchers study the system behavior identified in several load models. A workload model can be designed to verify the predictability, repeatability, and scalability of a system [42] [82]. Workload modeling is the attempt to create a simple and generic model that can then be used to generate synthetic

workloads. The goal is typically to be able to create workloads that can be used in performance evaluation studies. Sometimes, the synthetic workload is supposed to be similar to those that occur in practice in real systems [42] [82].

There are two kinds of workload models: descriptive and generative. The main difference between the two is that descriptive models just try to mimic the phenomena observed in the workload, whereas generative models try to emulate the process that generated the workload in the first place [42].

In descriptive models, one finds different levels of abstraction on the one hand, and different levels of fidelity to the original data on the other hand. The most strictly faithful models try to mimic the data directly using the statistical distribution of the data. The most common strategy used in descriptive modeling is to create a statistical model of an observed workload. This model is applied to all the workload attributes, e.g., computation, memory usage, I/O behavior, communication, etc. [42]. Fig. 2.4 shows a simplified workflow of a descriptive model. The workflow has six phases. In the first phase, the user uses the system in the production environment. In the second phase, the tester collects the user's data, such as logs, clicks, and preferences, from the system. The third phase consists in developing a model designed to emulate the user's behavior. The fourth phase is made up of the execution of the test, emulation of the user's behavior, and log gathering.

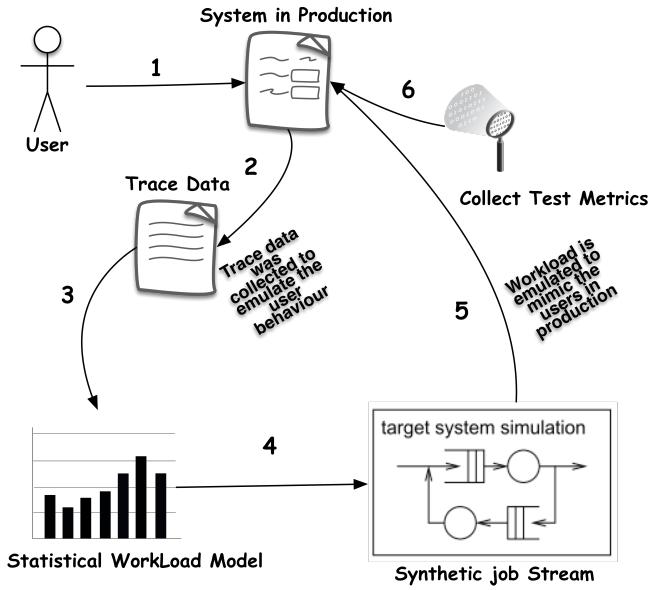


Figure 2.4: Workload modeling based on statistical data [35]

Generative models are indirect in the sense that they do not model the statistical distributions. Instead, they describe how users will behave when they generate the workload. An important benefit of the generative approach is that it facilitates manipulations of the workload. It is often desirable to be able to change the workload conditions as part of the evaluation. Descriptive models do not offer any option regarding how to

do so. With the generative models, however, we can modify the workload-generation process to fit the desired conditions [42]. The difference between the workflows of the descriptive and the generative models is that user behavior is not collected from logs, but simulated from a model that can receive feedback from the test execution (Fig. 2.5).

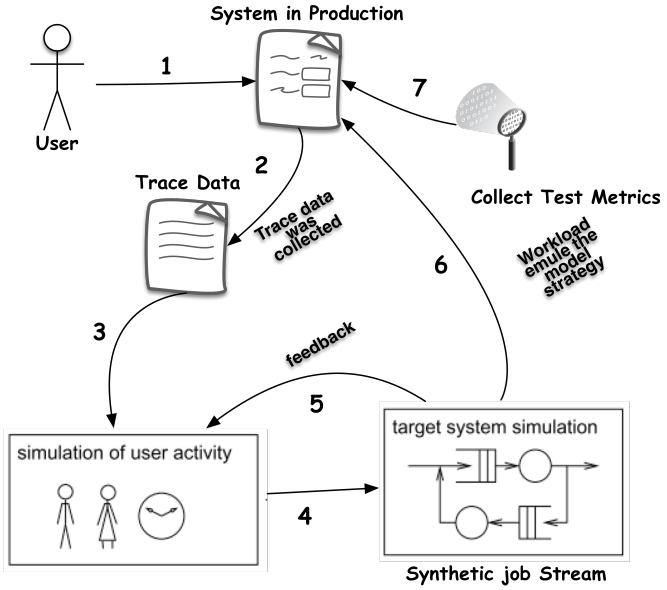


Figure 2.5: Workload modeling based on the generative model [35]

Both load models have their advantages and disadvantages. In general, loads resulting from realistic-load based design techniques (Descriptive models) can be used to detect both functional and non-functional problems. However, the test durations are usually longer and the test analysis is more difficult. Loads resulting from fault-inducing load design techniques (Generative models) take less time to uncover potential functional and non-functional problems, where the resulting loads usually only cover a small portion of the testing objectives [66]. The presented research work uses a generative model.

There are two main approaches to design generative or descriptive workloads:

- Model-based Stress testing: a usage model is proposed to simulate users' behaviors.
- Search-based Stress testing.

Search-Based Stress testing will be detail explained in the chapter 3. Six other approaches use neither a model-based test nor a search-based test.

2.2.1 Model-based Stress Testing

Model-based testing is an application of models to represent the desired behavior of a System Under Test or to represent testing strategies in a test. Some research aproaches proposes models to simulate or generate realistic loads. Model-based testing (MBT) is a variant of testing that relies on explicit behaviour models that encode the intended behaviours of a system under test. Test cases are generated from one of these models or their combination [77] [1].

The model paradgim is what paradigm and notation are used to describe the model. There are many different modelling notations that have been used for modelling the behaviour of systems for test generation purposes [77] [60].

- State-Based (or Pre/Post) Notations. These model a system as a collection of variables, which represent a snapshot of the internal state of the system, plus some operations that modify those variables. Each operation is usually defined by a precondition and a postcondition, or the postcondition may be written as explicit code that updates the state [77].

- Transition-based Notations. These focus on describing the transitions between different states of the system. Typically, they are graphical node-and-arc notations, like finite state machines (FSMs). Examples of transition-based notations used for MBT include FSMs themselves, statecharts, labelled transition systems and I/O automata [77].

- History-based Notations. These notations model a system by describing the allowable traces of its behaviour over time. Message-sequence charts and related formalisms are also included in this group. These are graphical and textual notations for specifying sequences of interactions between components [77].

- Functional Notations. These describe a system as a collection of mathematical functions. The functions may be first-order only, as in the case of algebraic specifications, or higher-order, as in notations like HOL [77]. Functional models also show the functionality of the system from the user's perspective [126]. This research also classified in the functional notation paradigm the studies that used more than one uml diagram.

- Operational Notations. These describe a system as a collection of executable processes, executing in parallel. They are particularly suited to describing distributed systems and communications protocols. Examples include process algebras such as CSP or CCS as well as Petri net notations. Slightly stretching this category, hardware description languages like VHDL or Verilog are also included in this category [77].

- Stochastic Notations. These describe a system by a probabilistic model of the events and input values and tend to be used to model environments rather than SUTs. For example, Markov chains are used to model expected usage profiles, so that the generated tests scenarios [77].

- Data-Flow Notations. These notations concentrate on the data rather than the control flow. Prominent examples are Lustre, and the block diagrams of Matlab Simulink, which are often used to model continuous systems [77].

Table 2.3 presents the papers found by the survey about model-based tests. All results was classified by model and paradigm. The most used paradigms in search-based stress testing are: Funcional-based models, Transition-based and State-based models.

Table 2.3: Summary of studies in model-based stress testing

Model	Paper	Paradigm	Year
BeliefDesire-Intention	[8]	Operational	2016
Markov-Chains	[13]	Stochastic Notation	1995
	[18]	Stochastic Notation	2007
	[15]	Stochastic Notation	1994
	[14]	Stochastic Notation	1993
MCM Model	[123]	State-based	2010
Other Aproaches	[39]	UPPAAL model checker	2013
	[10]	Model Decomposition	2015
Petri Nets	[24]	Operational	2009
Stochastic Form Model	[25]	Stochastic Notation	2007
	[80]	Stochastic Notation	2002
	[37]	Stochastic Notation	2006
State-Machine Models	[41]	State-based	2012
	[104]	State-based	2013
	[43]	Transition-based	2016
	[96]	-	2016
	[9]	Transition-based	2014
	[48]	Transition-based	2016
	[59]	Transition-based	2007
	[60]	Transition-based	2009
	[65]	State-based	2016
Symbolic Transition System	[7]	Transition-based	2013
Timed Automata	[6]	Transition-based	2013
UCML	[17]	Functional	1999
	[117]	Functional	2013
UML	[125]	Functional	2007
	[94]	Functional	2009
	[98]	Functional	2013
	[79]	Functional	2014
	[83]	Functional	2017
	[67]	Functional	2005
	[95]	Functional	2014
	[29]	Functional	2011
	[116]	Transition-based	2016
	[71]	Functional	2007

Functional Notation

All possible answers of the system, including exceptions, are defined in the functional model. The functional notation defines the authorized input values and models all the possible functional errors during execution [115]. Among the several functional models approaches, we can highlight the User Community Modeling Language (UCML). A UCML is a set of symbols that can be used to create visual system usage models and depict associated parameters [117]. The Fig. 2.6 shows a sample where all users realize a login into the application unde test. Once logged in, 40% of the users navigates on the aplication, 30% of the users realizes downloads. 20% of users realizes uploads and 10% of users performs deletions.

Garousi et al. proposes derivate Stress Test Requirements from an UML model. The input model consists of a number of UML diagrams. Some of them are standard in mainstream development methodologies and others are needed to describe the distributed architecture of the system under test (Fig. 2.9). Cai and Zeng use activity diagrams to describle variation points in use cases. Variation points describe what varies between the applications of an software product line [125]. Raulf et al. present an approach for testing of web service compositions using UML profile for Business Process Execution Language (BPEL) [94]. Schaefer et al. present the Crushinator, a framework that provides a game-independent testing tool simulating clients that perform http requests using a UML model [98]. Moscher and Fögen compares the techniques Capture and Replay (CR) and Model-Based Testing (MBT) are using a model named PLeTsPerf. PLeTsPerf describe the system under tests using use cases and activity diagrams [83].

Stochastic Notation

For many software programs, probabilistic models are a useful asset in modeling statistical behavior, such that coverage testing is possible by automating test-case selection, execution and evaluation. Ussualy, the stochastic notations are used in Markov Chains and Stochastic Formcharts.

Avritzer and Weyuker present two variants markov chain approach to realize load and stress tests and an automatic generation of load test suites approach [14] [13]. Barros et al. provide techniques for load pattern characterization via the application of Markov Chains to performance evaluation of stateful systems [18]. The work of Draheim and Weber's Formoriented analysis is a methodology for the specification of ultra-thin client based systems. Form-oriented models describe a web application as a bipartite state machine which consists of pages, actions, and transitions between them. Stochastic Formcharts are the combination of formoriented model and probability features. The Fig. 2.7 shows a sample where all users have a probability of 100% of realize a login into the application under test. Once logged in, users have a probability of 40% of navigate on the application and so on [37] [75].

One way to capture the navigational pattern within a session is through the Customer Behavior Model Graph (CBMG). Figure 2.8 depicts an example of a CBMG showing that customers may be in several different states—Home, Browse, Search, Select, Add, and Pay—and they may transition between these states as indicated by the arcs connecting them. The numbers on the arcs represent transition probabilities. A state not explicitly represented in the figure is the Exit state [80] [66] [81].

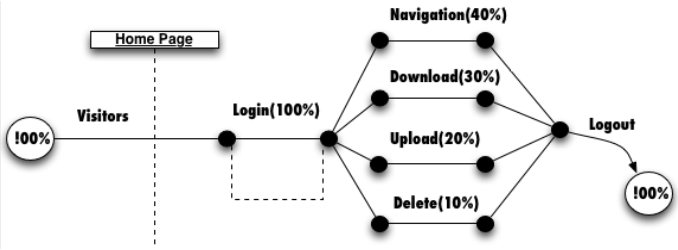


Figure 2.6: User community modeling language [117]

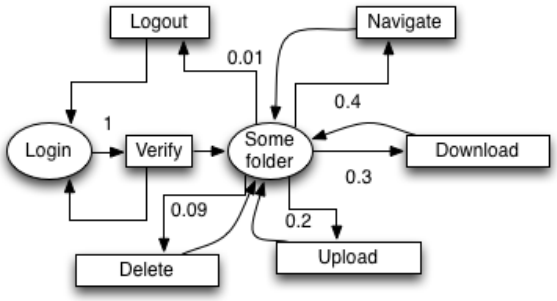


Figure 2.7: Stochastic Formcharts Example [37] [117]

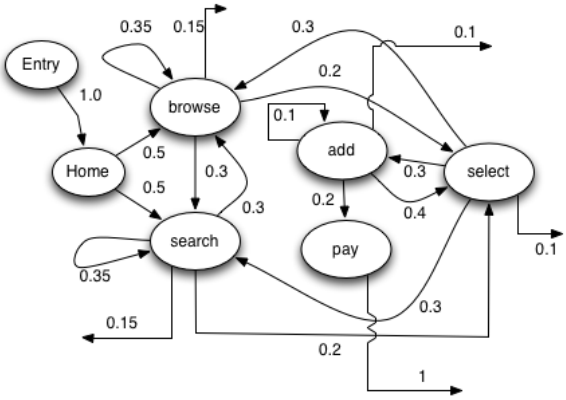


Figure 2.8: Example of a Customer Behavior Model Graph (CBMG) [80] [66] [81]

Transition-based notation

For model-based testing, the transition-based notations are the most used for developing behavioral models [115]. Broadly speaking, the transition-based notations are best for control-oriented applications. Instead of characterizing the system based on its admissible states, the system is characterized as transitions from one state to another, the properties are specified as a set of transitions functions, which map each input state to the corresponding output state. Based on the notations used, the model can be annotated with triggering events, which are conditions sufficient for the transition to take place, or guards that are necessary preconditions for the transition to be fired. The common techniques used for generating test-cases from transition-based notations are Finite State Machines (FSM), Labeled Transition Systems and UML statecharts [75].

Arantes et al. present a tool named WEB-PerformCharts that generate test cases using statecharts or FSMs [9]. Gay et al. present propose an automated steering framework that can adjust the behavior of the model to better match the behavior of the system under test to reduce the rate of false positives. The model is defined as a transition system [48]. Hessel addressed in her study two model-based problems: how to formalize a coverage criteria and how to generate a test suite to a formal timed system model [59]. Vogele et al. presents an approach that aims to automate the extraction and transformation of workload specifications for an model-based performance prediction of session-based application systems. The research also presents

transformations to the common load testing tool Apache JMeter and to the Palladio Component Model [116] [115]. The workload specification formalism (Workload Model) consists of the following components, which are detailed below and illustrated in Fig. 2.10:

- An Application Model, specifying allowed sequences of service invocations and SUT-specific details for generating valid requests.
- A set of Behavior Models,each providing a probabilistic representation of user sessions in terms of invoked services.
- A Behavior Mix, specified as probabilities for the individual Behavior Models to occur during workload gen- eration.
- A Workload Intensity that includes a function which specifies the number of concurrent users during the workload generation execution.

State-based notation

State based notations like STATECHARTS or activity diagrams describe the behavior based on abstract states and support behavior integration only using state composition [49]. In State-based or notations, the system is modeled as a collection of variables representing its state at a specific point of the execution, together with a collection of operations defined by a precondition that defines the admissible set of initial states, and a postcondition that specifies the guaranteed set of final states. Examples of such notations include the Z language, the B machine, UML's Object Constraint Language (OCL), Java Modeling Language (JML), VDM , and Spec [75].

Sridhar proposed an approach to generate test cases with MATLAB using Simulink/ Stateflow tool. After the model creation, test sequences are generated a dependency graph of that system [104]. Fang et al. developed a test case generator, from which an entire test suite can be extracted [41]. Jeong et al. propose a state transition model based to test case generation [65]. Wieczorek et al. propose an approach that uses proprietary models called Message Choreography Models (MCM) using a state-based representation [123].

2.2.2 Other Approaches

A set of other approaches dont use model-based tests or search-based tests:

- Automatic feedback, control-based, stress and load testing [19];
- Feedback-ORiEnted PerfOrmance Software Testing [73];
- PASASM : A Method for the Performance Assessment of Software Architectures [124].

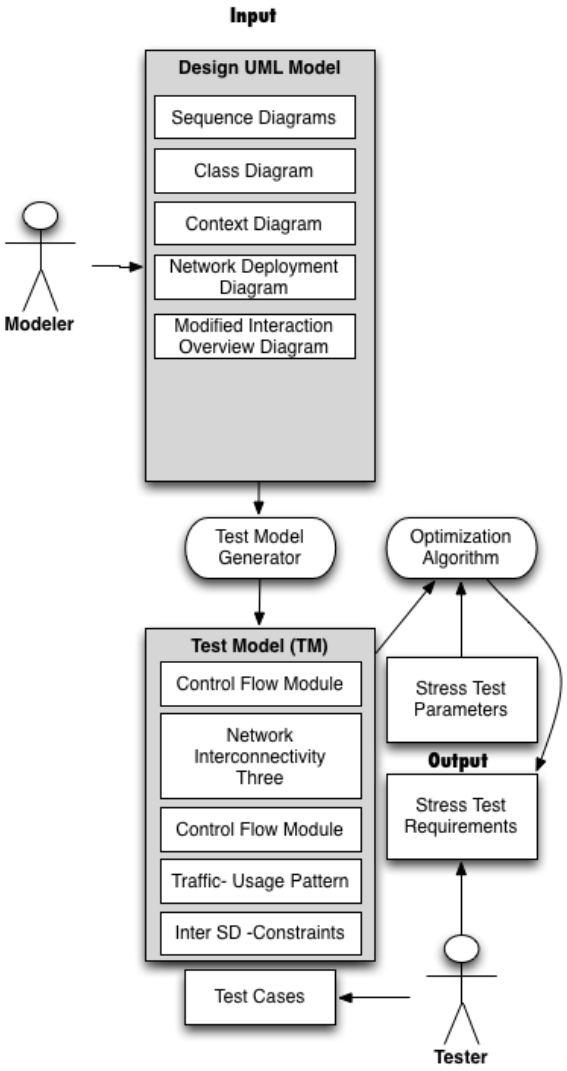


Figure 2.9: Model-based stress test methodology

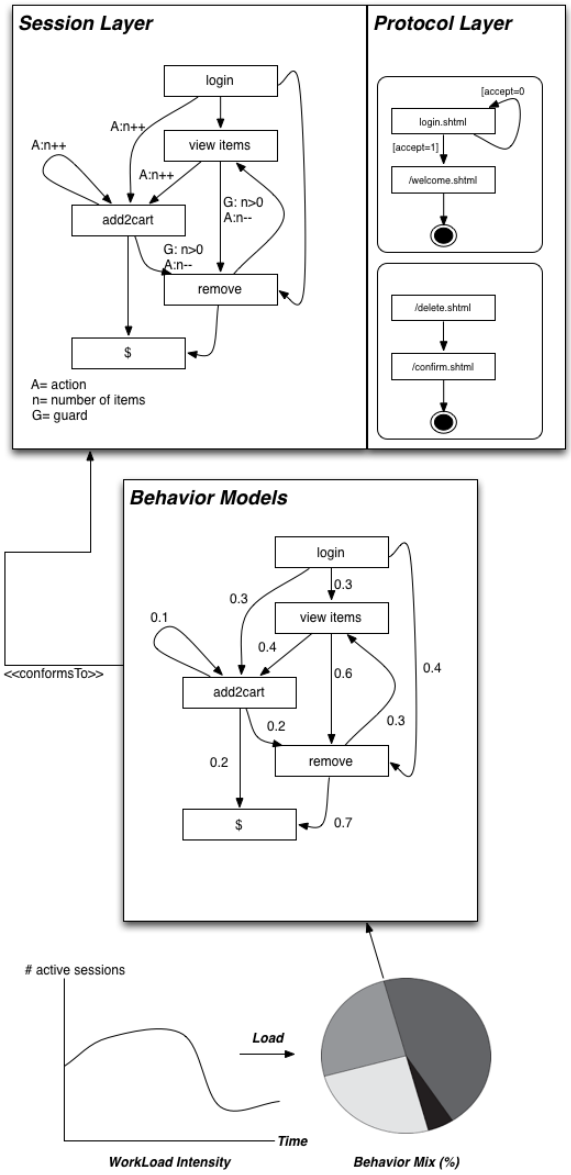


Figure 2.10: Exemplary workload model

Bayan and Cangussu present a apporach based on the application of a feedback PID (Proportional, Integral, and Derivative) controller to drive the input and make the system achieve a specified level of resource usage. For example, if the user defines the system should be tested with a memory use of 95%, starting from an initial input value, the PID controller will automatically change the input(s) until the desired level of stress has been achieved [19].

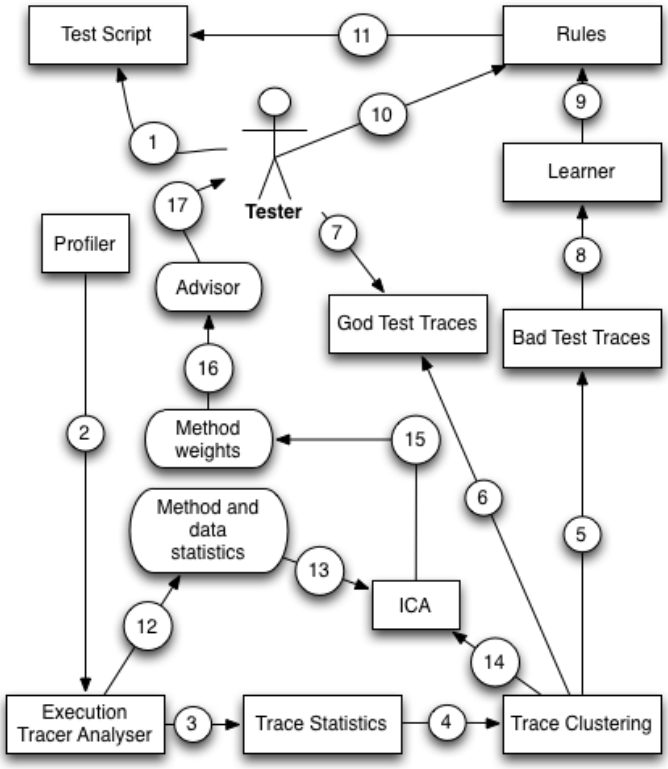


Figure 2.11: The architecture and workflow of FOREPOST

Williams and Smith describe the PASA method, a method for performance assessment of software architectures. PASA uses the principles and techniques of software performance engineering (SPE) to determine whether an architecture is capable of supporting its performance objectives [124]. Among the three approaches presented, we can highlight FOREPOST that develops a plugin of the JMeter tool to generate test cases using unsupervised learning.

Feedback-ORiEnted PerfOrmance Software Testing

Feedback-ORiEnted PerfOrmance Software Testing (FOREPOST) is an adaptive, feedback-directed learning testing system that learns rules from system execution traces and uses these learned rules to select test input data automatically to find more performance problems in applications when compared to exploratory random performance testing [53].

FOREPOST uses runtime monitoring for a short duration of testing together with machine learning techniques and automated test scripts to reduce large amounts of performance-related information collected during AUT runs to a small number of descriptive rules that provide insights into properties of test input data that lead to increased computational loads of applications.

The Fig. 2.11 presents the main workflow of FOREPOST solution. The first step, The Test Script is

written by the test engineer(1). Once the test script starts, its execution traces are collected (2) by the Profiler, and these traces are forwarded to the Execution Trace Analyzer, which produces (3) the Trace Statistics. The trace statistics is supplied (4) to Trace Clustering, which uses an ML algorithm, JRip to perform unsupervised clustering of these traces into two groups that correspond to (6) Good and (5) Bad test traces.

The user can review the results of clustering (7). These clustered traces are supplied (8) to the Learner that uses them to learn the classification model and (9) output rules. The user can review (10) these rules and mark some of them as erroneous if the user has sufficient evidence to do so. Then the rules are supplied (11) to the Test Script. Finally, the input space is partitioned into clusters that lead to good and bad test cases, to find methods that are specific to good performance test cases. This task is accomplished in parallel to computing rules, and it starts when the Trace Analyzer produces (12) the method and data statistics that is used to construct (13) two matrices (14). Once these matrices are constructed, ICA decomposes them (15) into the matrices for bad and good test cases correspondingly. Finally, the Advisor (16) determines top methods that performance testers should look at (17) to debug possible performance problems.

2.3 Research Question 2: What are the main problems found by stress tests?

Performance problems share common symptoms and many performance problems described in the literature are defined by a particular set of root causes. Fig. 2.12 shows the symptoms of known performance problems [121].

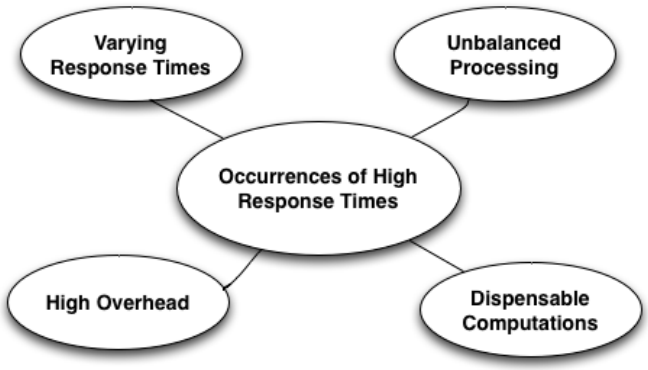


Figure 2.12: Symptoms of known performance problems [121].

There are several antipatterns that detailed features about common performance problems. Antipatterns are conceptually similar to patterns in that they document recurring solutions to common design problems. They are known as antipatterns because their use produces negative consequences.

Performance antipatterns document common performance mistakes made in software architectures or

Table 2.4: Performance antipatterns

Antipattern		**Papers**
Blob or The God Class		[122] [101] [113] [114] [28] [102]
Circuitous Treasure Hunt		[122] [113] [114] [102] [103]
Empty Semi Trucks		[122] [113] [11] [114]
Excessive Dynamic Aloccation		[113] [114] [102] [103]
More is Less		[114] [113] [102]
One-Lane Bridge		[114] [113] [102] [103]
Stifle		[122]
The Ramp		[113] [114] [102]
Tower of Babel		[113] [114]
Traffic Jam		[114] [102] [103]
Unbalanced Processing		[28] [113] [102]
	Concurrent Processing System	[114] [102]
	Piper and Filter	[114] [102]
	Extensive Process	[114] [102]
Unnecessary Processing		[102]

designs. These software Performance antipatterns have four primary uses: identifying problems, focusing on the right level of abstraction, effectively communicating their causes to others, and prescribing solutions [23]. Table 2.4 present some of the most common performance antipatterns.

Blob antipattern is known by various names, including the “god” class [8] and the “blob” [2]. Blob is an antipattern whose problem is on the excessive message traffic generated by a single class or component, a particular resource does the majority of the work in a software. The Blob antipattern occurs when a single class or component either performs all of the work of an application or holds all of the application’s data. Either manifestation results in excessive message traffic that can degrade performance [28] [101].

A project containing a “god” class usually has a single, complex controller class that is surrounded by simple classes that serve only as data containers. These classes typically contain only accessor operations (operations to get() and set() the data) and perform little or no computation of their own [101]. According to Figure 2.13 and 2.14, a hypothetical system with a BLOB problem is shown: Figure 2.13 presents a sample where the Blob class uses the features A,B,C,D,E,F and G of the hypothetical system, and Fig. 2.14 shows a static view where a complex software entity instance, i.e. Sd, is connected to other software instances, e.g. Sa, Sb and Sc, through many dependencies [114][121].

Unbalanced Processing it’s characterises for one scenario where a specific class of requests generates a pattern of execution within the system that tends to overload a particular resource. In other words the overloaded resource will be executing a certain type of job very often, thus in practice damaging other classes of jobs that will experience very long waiting times. Unbalanced Processing occurs in three different situations. The first case that cause unbalanced processing it is when processes cannot make effective use of available processors either because processors are dedicated to other tasks or because of single-threaded code. This manifestation has available processors and we need to ensure that the software is able to use them. Fig. 2.15

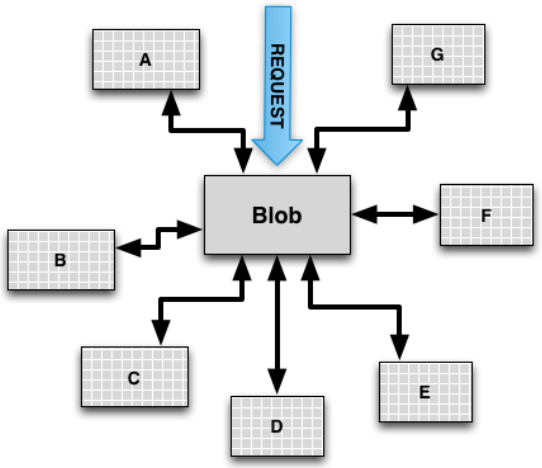


Figure 2.13: The God class[121].

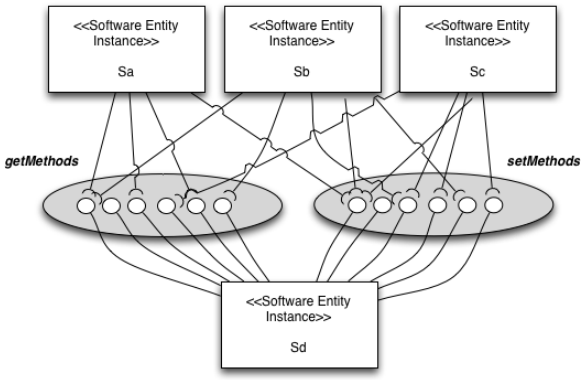


Figure 2.14: The God class[114].

shows a sample of the Unbalanced Processing. In The Fig. 2.15, four tasks are performed. The task D it is waiting for the task C conclusion that are submmited to a heavy processing situation.

The pipe and filter architectures and extensive processing antipattern represents a manifestation of the unbalanced processing antipattern. The pipe and filter architectures occurs when the throughput of the overall system is determined by the slowest filter. The Fig. 2.16 describes a software S with a Pipe and Filter Architectures problem: the operation opx is invoked in a service and the throughput of the service ($Th(S)) is lower than the required one. The extensive processing occurs when a process monopolizes a processor and prevents a set of other jobs to be executed until it finishes its computation. The Fig. 2.17 describes a software S with a Extensive Processing problem: the operations opx and opy are alternatively invoked in a service and the response time of the service ($RT(S)) is larger than the required one [114].

Circuitous Treasure Hunt antipattern occurs when software retrieves data from a first componet, uses those results in a second component, retrieves data from the second component, and so on, until the last results are obtained [103] [102]. Circuitous Treasure Hunt are typical performance antipatterns that causes unnecessarily frequent database requests. The Circuitous Treasure Hunt antipattern is a result from a bad database schema or query design. A common Circuitous Treasure Hunt design creates a data dependency between single queries. For instance, a query requires the result of a previous query as input. The longer the chain of dependencies between individual queries the more the Circuitous Treasure Hunt antipattern hurts performance [122]. The Fig. 2.18 shows a software S with a Circuitous Treasure Hunt problem: the software S generates a large number of database calls by performing several queries up to the final operation [114].

Empty Semi Trucks occurs when an excessive number of requests is required to perform a task. It may be due to inefficient use of available bandwidth, an inefficient interface, or both [11]. There are a special case of Empty Semi Trucks that occurs when many fields in a user interface must be retrieved from a remote system. Fig. shows a software S with a Empty Semi Trucks problem: the software instance Sa generates an

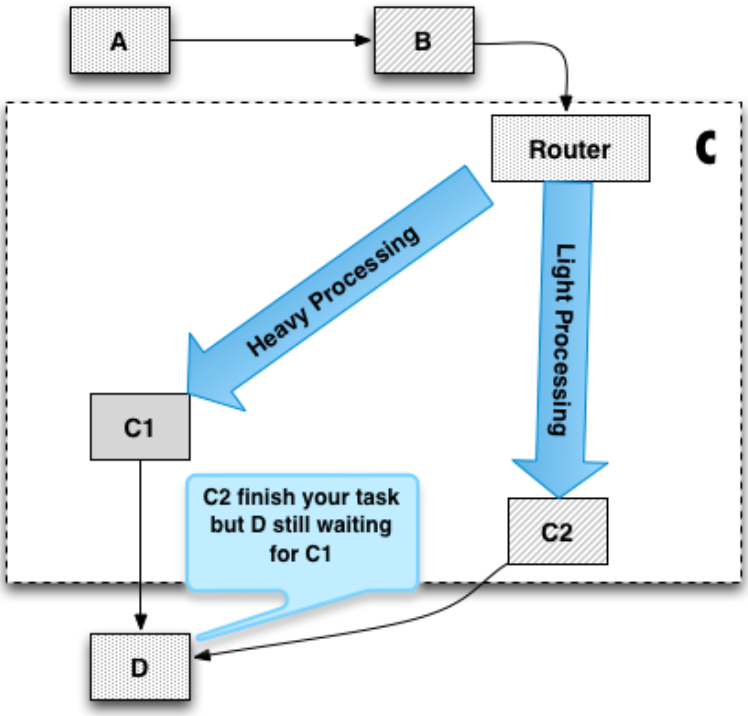


Figure 2.15: Unbalanced Processing sample [121].

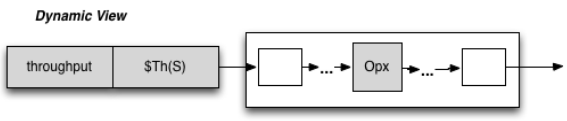


Figure 2.16: Pipe and Filter sample [114]

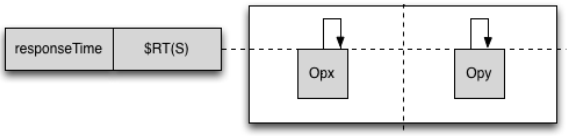


Figure 2.17: Extensive Processing sample [114].

excessive message traffic by sending a big amount of messages with low sizes, much lower than the network bandwidth, hence the network link might have a low utilization value [114].

The Tower of Babel antipattern most often occurs when information is translated into an exchange format, such as XML, by the sending process then parsed and translated into an internal format by the receiving process. When the translation and parsing is excessive, the system spends most of its time doing this and relatively little doing real work [102]. Fig. shows a system with a Tower of Babel problem: the software instances Sd performs many times the translation of format for communicating with other instances [114].

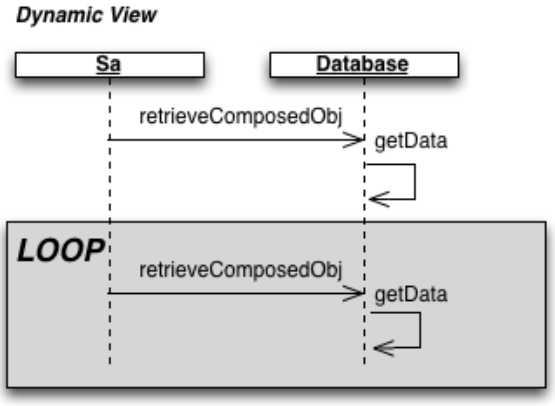


Figure 2.18: Circuitous Treasure Hunt sample [114]

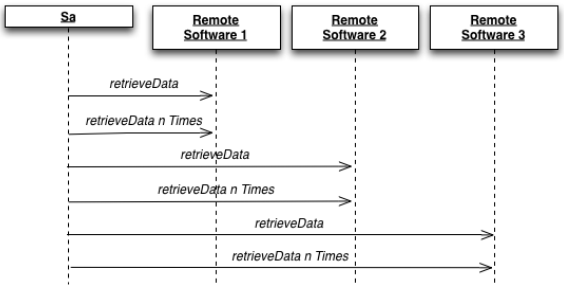


Figure 2.19: Empty Semi Trucks sample [114].

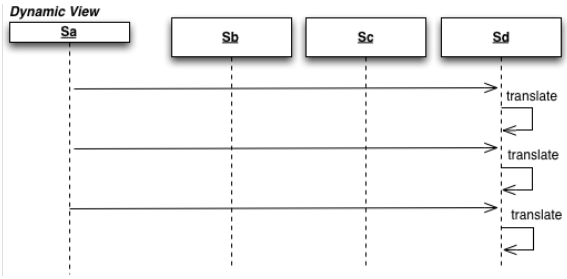


Figure 2.20: Tower of Babel sample [114]

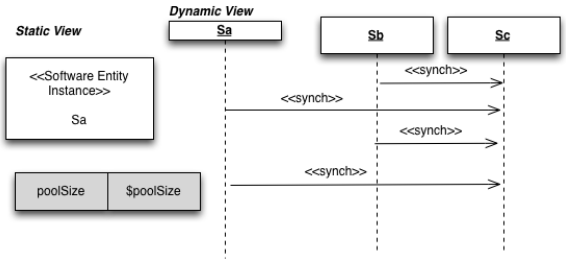


Figure 2.21: One-Lane Bridge sample [114].

One-Lane Bridge is a antipattern that occurs when one or a few processes execute concurrently using a shared resource and other processes are waiting for use the shared resource. It frequently occurs in applications that access a database. Here, a lock ensures that only one process may update the associated portion of the database at a time. This antipatterns is common when many concurrent threads or processes are waiting for the same shared resources. These can either be passive resources (like semaphores or mutexes) or active resources (like CPU or hard disk). In the first case, we have a typical One Lane Bridge whose critical resource needs to be identified. Figure 3.10 shows a system with a One-Lane Bridge problem: the software instance Sc receives an excessive number of synchronous calls in a service S and the predicted response time is higher than the required [114].

Using dynamic allocation, objects are created when they are first accessed and then destroyed when they are no longer needed. Excessive Dynamic Allocation, however, addresses frequent, unnecessary creation and destruction of objects of the same class. Dynamic allocation is expensive , an object created in memory must be allocated from the heap, and any initialization code for the object and the contained objects must be executed. When the object is no longer needed, necessary clean-up must be performed, and the reclaimed memory must be returned to the heap to avoid memory leaks [103] [102].

The Fig. 2.22 shows a Excessive Dynamic Allocation sample. This example is drawn from a call (an

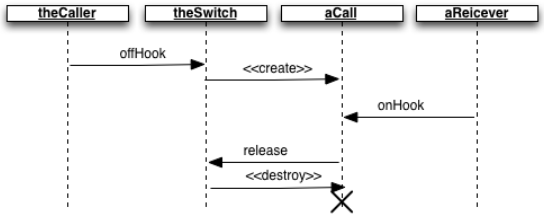


Figure 2.22: Excessive Dynamic Allocation.

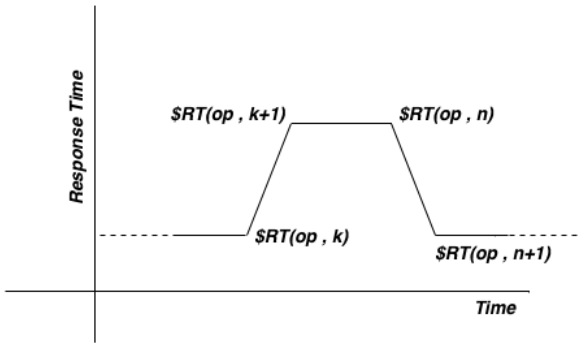


Figure 2.23: Traffic Jam Response Time [114].

offHook event), the switch creates a Call object to manage the call. When the call is completed, the Call object is destroyed. Constructing a single Call object it is not seem as excessive. A Call is a complex object that contains several other objects that must also be created. The Excessive Dynamic Allocation occurs when a switch receive hundreds of thousands of offHook events. In a case like this, the overhead for dynamically allocating call objects adds substantial delays to the time needed to complete a call.

The Traffic Jam antipattern occurs if many concurrent threads or processes are waiting for the same active resources (like CPU or hard disk). This antipatterns produces a large backlog in jobs waiting for service. The performance impact of the Traffic Jam is the transient behavior that produces wide variability in response time. Sometimes it is fine, but at other times, it is unacceptably long. Figure 2.23 describes a software with a Traffic Jam problem, the monitored response time of the operation shows a wide variability in response time which persists long [114].

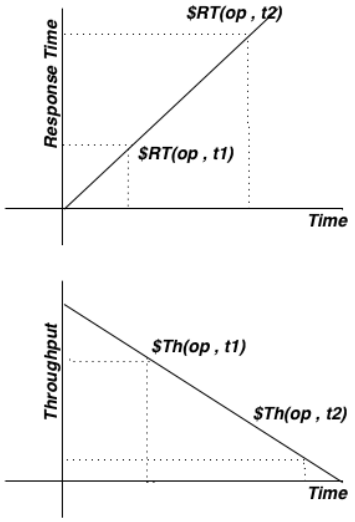


Figure 2.24: The Ramp sample [114].

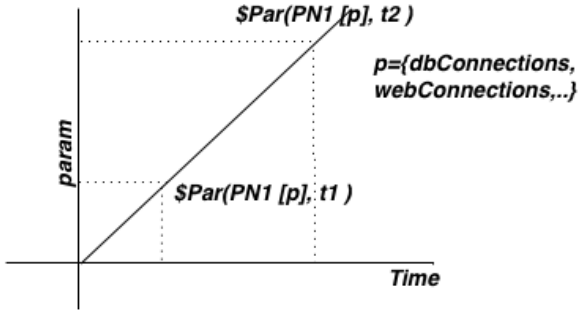


Figure 2.25: More is Less sample [114].

The Ramp it is a antipattern where the processing time increases as the system is used. The Ramp can arise in several different ways. Any situation in which the amount of processing required to satisfy a request increases over time will produce the behavior. With the Ramp antipattern, the memory consumption of the application is growing over time. The root cause is Specific Data Structures which are growing during operation or which are not properly disposed [122] [102]. Fig. 2.24 shows a system with The Ramp problem: (i) the monitored response time of the operation opx at time t1, i.e. $RT(opx, t1), is much lower than the monitored response time of the operation opx at time t2, i.e. $RT(opx, t2), with t1 < t2; (ii) the monitored throughput of the operation opx at time t1, i.e. $Th(opx, t1), is much larger than the monitored throughput of the operation opx at time t2, i.e. $Th(opx, t2), with t1 < t2.

More is less occurs when a system spends more time "thrashing" than accomplishing real work because there are too many processes relative to available resources. More is Less are presented when it is running too many programs overtime. This antipattern causes too much system paging and systems spend all their time servicing page faults rather than processing requests. In distributed systems, there are more causes.They include: creating too many database connections and allowing too many internet connection. Fig. 2.25 describes a system with a More Is Less problem: There is a processing node PN1 and the monitored runtime parameters (e.g. database connections, etc.) at time t1, i.e. $Par(PN1[p], t1), are much larger than the same parameters at time t2, i.e. $Par(PN1[p], t2), with t1 < t2.

2.4 Summary

Chapter 3

Search-Based Stress Testing

Search-based software engineering (SBSE) is the application of optimization techniques in solving software engineering problems. The applicability of optimization techniques in solving software engineering problems is suitable as these problems frequently encounter competing constraints and require near optimal solutions [2] [57].

Search Based Software Testing (SBST) is the sub-area of Search Based Software Engineering concerned with software testing. Search-based software testing is the application of metaheuristic search techniques to generate software tests. The test adequacy criterion is transformed into a fitness function and a set of solutions in the search space are evaluated with respect to the fitness function using a metaheuristic search technique [2] [5] [57].

There are many kinds of non-functional search based tests [2]:

- Execution time: The application of evolutionary algorithms to find the best and worst case execution times (BCET, WCET).
- Quality of service: uses metaheuristic search techniques to search violations of service level agreements (SLAs).
- Security: apply a variety of metaheuristic search techniques to detect security vulnerabilities like detecting buffer overflows.
- Usability: concerned with construction of covering array which is a combinatorial object.
- Safety: Safety testing is an important component of the testing strategy of safety critical systems where the systems are required to meet safety constraints.

A variety of metaheuristic search techniques are found to be applicable for non-functional testing including simulated annealing, tabu search, genetic algorithms, ant colony methods, grammatical evolution, genetic programming and swarm intelligence methods.

3.1 Metaheuristics

Metaheuristics are strategies that guide the search process to efficiently explore the search space in order to find optimal solutions. Metaheuristic algorithms are approximate and usually non-deterministic and sometimes incorporate mechanisms to avoid getting trapped in confined areas of the search space. There are different ways to classify and describe metaheuristic algorithm [21]:

- Nature-inspired vs. non-nature inspired. There are nature-inspired algorithms, like Genetic Algorithms and Ant Algorithms, and non nature-inspired ones such as Tabu Search and Iterated Local Search.
- Population-based vs. single point search (Trajectory methods). Algorithms working on single solutions are called trajectory methods, like Tabu Search, Iterated Local Search and Variable Neighborhood Search. They all share the property of describing a trajectory in the search space during the search process. Population-based metaheuristics perform search processes which describe the evolution of a set of points in the search space.
- One vs. various neighborhood structures. Most metaheuristic algorithms work on one single neighborhood structure. In other words, the fitness landscape topology does not change in the course of the algorithm. Other metaheuristics, such as Variable Neighborhood Search (VNS), use a set of neighborhood structures which gives the possibility to diversify the search by swapping between different fitness landscapes.

3.1.1 Trajectory methods

Trajectory methods are characterized by a trajectory in the search space. Two common trajectory methods are Simulated Annealing and Tabu Search.

Neighborhood

The definition of Neighborhood is a required common step for the design of any Single-Solution metaheuristic (S-metaheuristic). The neighborhood structure it is a important piece in the performance of an S-metaheuristic. If the neighborhood structure is not adequate to the problem, any S-metaheuristic will fail to solve the problem. The neighborhood function N is a mapping: $N : S \rightarrow N^2$ that assigns to each solution s of S a set of solutions $N(s) \subset$ S [108].

The neighborhood definition depends representation associated with the problem. For permutation-based representations, a usual neighborhood is based on the swap operator that consists in swapping the location of two elements s_i and s_j of the permutation [108]. The Fig. 3.1 presents a example where a set of neighbors is found by permutation.

Single-Solution Based Metaheuristics methods are characterized by a trajectory in the search space. Two common S-metaheuristics methods are Simulated Annealing and Tabu Search.

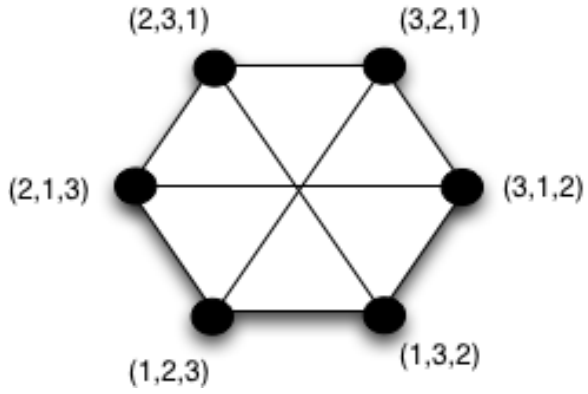


Figure 3.1: An example of neighborhood for a permutation [108].

Algorithm 1 Simulated Annealing Algorithm

```
1:  s ← GenerateInitialSolution()
2:  k ← 0
3:  Tk ← SetInitialTemperature()
4:  while termination conditions not met do
5:      s1 ← PickNeighborAtRandom(N(s))
6:      if (f(s1) < f(s)) then
7:          s ← s1
8:      else Accept s1 as new solution with probability p(s1|Tk,s)
9:      end if
10:     K ← K + 1
11:     Tk ← AdaptTemperature()
12: end while
```

Simulated Annealing

The algorithmic framework of SA is described in Alg. 1. The algorithm starts by generating an initial solution in function *GenerateInitialSolution()*. The initial temperature value is determined in function *SetInitialTemperature()* such that the probability for an uphill move is quite high at the start of the algorithm. At each iteration a solution s_1 is randomly chosen in function *PickNeighborAtRandom(N(s))*. If s_1 is better than s, then s_1 is accepted as new current solution. Else, if the move from s to s_1 is an uphill move, s_1 is accepted with a probability which is a function of a temperature parameter Tk and s [91].

Tabu Search

Tabu Search uses a tabu list to keep track of the last moves, and don't allow going back to these [50]. The algorithmic framework of Tabu Search is described in Alg. 2. The algorithm starts by generating an initial solution in function *GenerateInitialSolution()* and the tabu lists are initialized as empty lists in function *InitializeTabuLists(TL_1,...,TL_r)*. For performing a move, the algorithm first determines those solutions from the neighborhood *N(s)* of the current solution s that contain solution features currently to be found in the

Algorithm 2 Tabu Search Algorithm

$s \leftarrow GenerateInitialSolution()$
InitializeTabuLists(TL_1,...,TL_r)
while termination conditions not met **do**
    $N_a(s) \leftarrow \{s_1 \in N(s) | s_1$ does not violate a tabu condition, or it satisfies at least one aspiration condition $\}$
    $s_1 \leftarrow argmin\{f(s_2) | s_2 \in N_a(s)\}$
    UpdateTabuLists(TL_1,...,TL_r,s,s_1)
    $s \leftarrow s_1$
end while

tabu lists. They are excluded from the neighborhood, resulting in a restricted set of neighbors $N_a(s)$. At each iteration the best solution s_1 from $N_a(s)$ is chosen as the new current solution. Furthermore, in procedure *UpdateTabuLists(TL_1,...,TL_r,s,s_1)* the corresponding features of this solution are added to the tabu lists.

3.1.2 Population-based metaheuristics

Population-based metaheuristics (P-metaheuristics) could be viewed as an iterative improvement in a population of solutions. First, the population is initialized. Then, a new population of solutions is generated. Finally, this new population is integrated into the current one using some selection procedures. The search process is stopped when a stopping criterion is satisfied. Algorithms such as Genetic algorithms (GA), scatter search (SS), estimation of distribution algorithms (EDAs), particle swarm optimization (PSO), bee colony (BC), and artificial immune systems (AISs) belong to this class of metaheuristics [107].

Population-based metaheuristics are comprised of several components [61] [99] :

- a representation of the solution, refered as the chromosome;
- fitness of each chromosome, refered as objective function;
- the genetic operations of crossover and mutation which generate new offspring.

The crossover operation or recombination recombines two or more individuals to produce new individuals. Mutation or modification operators causes a self-adaptation of individuals [21]. In Search-based tests, the crossover operator creates two new test cases T1' and T2' by combining test cases from two pre-existing test cases T1 and T2 [5]. Algorithm 3 shows the basic structure of GA algorithms. In this algorithm, P denotes the population of individuals. A population of offspring is generated by the application of recombination and mutation operators and the individuals for the next population are selected from the union of the old population and the offspring population [91].

3.1.3 Hybrid Metaheuristics

A combination of one metaheuristic with components from other metaheuristics is called a hybrid metaheuristic. The concept of hybrid metaheuristics has been commonly accepted only in recent years, even if the

Algorithm 3 Genetic Algorithm

```
1: s ← GenerateInitialSolution()
2: Evaluate(P)
3: while termination conditions not met do
4:     P1 ← Recombine(P)
5:     P2 ← Mutate(P1)
6:     Evaluate(P2)
7:     P ← Select(P2, P)
8: end while
```

idea of combining different metaheuristic strategies and algorithms dates back to the 1980s. Today, we can observe a generalized common agreement on the advantage of combining components from different search techniques and the tendency of designing hybrid techniques is widespread in the fields of operations research and artificial intelligence [91].

There are two main categories of metaheuristic combinations: collaborative combinations and integrative combinations. These are presented in Fig. 3.2 [92].

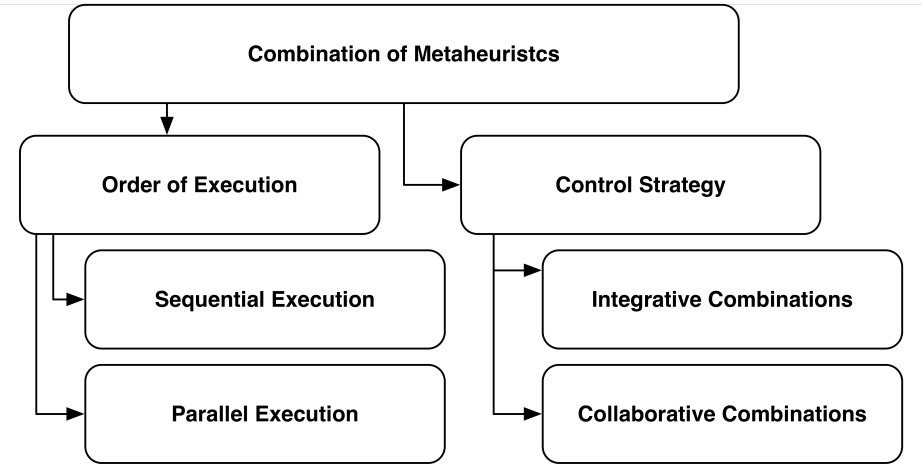


Figure 3.2: Categories of metaheuristc combinations [89]

Collaborative combinations use an approach where the algorithms exchange information, but are not part of each other. In this approach, algorithms may be executed sequentially or in parallel.

One of the most popular ways of metaheuristic hybridization consists in the use of trajectory methods inside population-based methods. Population-based methods are better in identifying promising areas in the search space from which trajectory methods can quickly reach good local optima. Therefore, metaheuristic hybrids that can effectively combine the strengths of both population-based methods and trajectory methods are often very successful [91].

3.1.4 Multi-objective heuristics

Many real optimization problems require optimizing multiple conflicting objectives with each other. There is no single optimal solution, but a set of alternative solutions. The objectives that have to be optimized are often in competition with one another and may be contradictory; we may find ourselves trying to balance the different optimization objectives of several different goals [58] [**?**]. The image of all the efficient solutions is called the Pareto front or Pareto curve or surface. The shape of the Pareto surface indicates the nature of the trade-off between the different objective functions. An example of a Pareto curve is reported in Fig. 3.3. Multi-objective optimization methods have as main purposes to minimize the distance between the non-dominated front and the Pareto optimal front and find a set of solutions that are as diverse as possible.

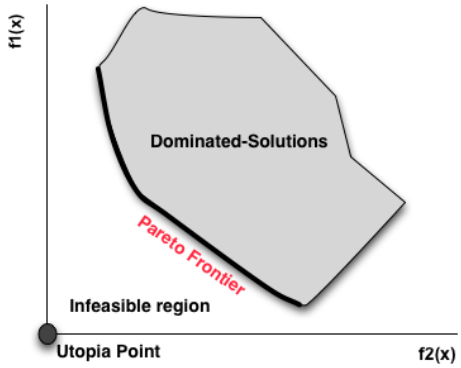


Figure 3.3: An optimized Pareto front example

What distinguishes multi-objective evolutionary algorithms from single objective metaheuristics is how they rank and select individuals in the population. If there is only one objective, individuals are naturally ranked according to this objective, and it is clear which individuals are best and should be selected as parents. In the case of multiple objectives, it is still necessary to rank the individuals, but it is no longer obvious how to do this. Most people probably agree that a good approximation to the Pareto front is characterized by:

- a small distance of the solutions to the true Pareto frontier,
- a wide range of solutions, i.e., an approximation of the extreme values, and
- a good distribution of solutions, i.e., an even spread along the Pareto frontier.

The approximation of the Pareto-optimal set involves itself two objectives: minimize the distance to the optimal front and maximize the diversity of the generated solutions. There are two fundamental issues when designing a multiobjective evolutionary algorithm: mating selection and environmental selection. The first issue is related to the question of how to guide the search towards the Pareto-optimal front. The procedure to fill the mating pool is usually randomized. The second issue is related with the question of which individuals to keep during the evolution process. In most modern EMO algorithms these two concepts are realized in the following way: Environmental selection or Mating selection [128].

In Environmental selection, an archive is maintained which contains a representation of the nondominated front among all solutions considered so far. A member of the archive is only removed if i) a solution has been found that dominates it or ii) the maximum archive size is exceeded and the portion of the front where the archive member is located is overcrowded [128].

In Mating selection, the pool of individuals is evaluated in two phases. First all individuals are compared on the basis of the Pareto dominance. Basically, the information which individuals each individual dominates, is dominated by or is indifferent to is used to define a ranking on the generation pool. Afterwards, this ranking is refined by the incorporation of density information. Various density estimation techniques are used to measure the size of the niche in which a specific individual is located [128].

NSGA-II Multi-objective heuristics

Multi-objective metaheuristics rank individuals according to the defined goals. Deb et al. proposed the non-dominated Sorting Genetic Algorithm II (NSGA-II) algorithm taking into account the need to reduce computational complexity in non-dominated classification, while introducing elitism and eliminating subjectivity in the allocation of the sharing parameter [**?**]. NSGA-II is a multi-objective algorithm, based on GAs, and implements the concept of dominance, in other words, to classify the total population in fronts according to the degree of dominance. According to NSGA-II, the individuals that are located on the first front are considered the best solutions of that generation, while in the last front are the worst. Using this concept, one can find more consistent results, located closer to the Pareto region, and that are better adapted to the type of problem.

The NSGA algorithm II applies a fitness evaluation in an initial population (Figure 3.4- ❶ and ❷). The populations are ranked using multiple tournament selections, which consist of comparing two solutions (Figure 3.4- ❸). In order to estimate the density of the solutions surrounding a particular solution in the population, the common distance between the previous solution and the posterior is calculated for each of the objectives. This distance serves as an estimate of the size of the largest cuboid that includes solution i without including any other solution of the population. A solution i beats another solution if:

- Solution i has a better rank, then $Rank_i < Rank_j$.
- Both solutions have the same rank, but i has a greater Distance than j, then $Rank_i = Rank_j$ and $Distance_i > Distance_j$.

At the end of each analysis a certain group of individuals are classified as belonging to a specific category called the front, and upon completion of the classification process, all individuals will be inserted into one of the n fronts. Front 1 is made up of all non-dominated solutions. Front 2 can be achieved by considering all non-dominated solutions excluding solutions from front 1. For the determination of front 3, solutions previously classified on front 1 and 2 are excluded, and so on until all individuals have been classified on some front.

After selection, recombination and mutation are performed as in conventional GAs (Figure 3.4- ❹). The two sets (father and son of the same dimension) are united in a single population (dimension 2) and the classification is applied in dominance fronts. In this way, elitism is guaranteed preserving the best solutions (fronts are not dominated) in the latest population (Figure 3.4- ❻).

However, not all fronts can be included in the new population. Thus, Deb et al. proposed a method called crowd distance, which combines the fronts not included in the set, to compose of the last spaces of the current population, guaranteeing the diversity of the population [**?**]. The NSGA-II algorithm creates a set of front lines, in which each front containing only non–dominating solutions. Within a front, individuals are rewarded for being 'spread out'. The algorithm also ensures that the lowest ranked individual of a front still has a greater fitness value than the highest ranked individual of the next front [**?**].

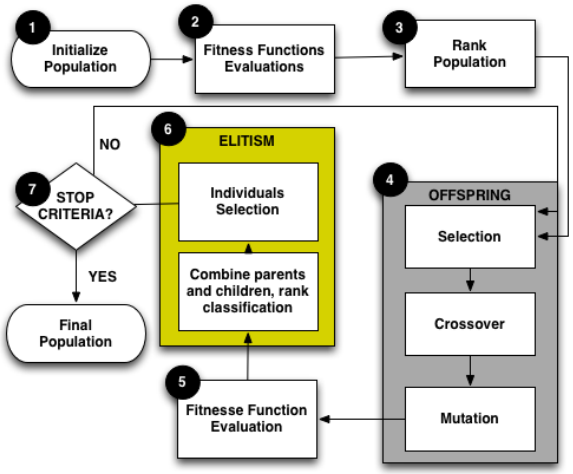


Figure 3.4: NSGA-II Algorithm

SPEA2: Improving the Strength Pareto Evolutionary Algorithm

SPEA uses a regular population and an archive. Starting with an initial population and an empty archive the following steps are performed per iteration. First, all nondominated population members are copied to the archive; any dominated individuals or duplicates are removed. If the size of the updated archive exceeds a predefined limit, further archive members are deleted by a clustering technique which preserves the characteristics of the nondominated front. Afterwards, fitness values are assigned to both archive and population members. Each individual i in the archive is assigned a strength value S(i) $\in$ [0, 1], which at the same time represents its fitness value F(i). 0 indicates a non-dominated individual,whereas a high value points out that the individual is dominated by many other ones. S(i) is the number of population members j that are dominated by or equal to i with respect to the objective values, divided by the population size plus one. The algorithmic framework of SPEA2 is described in Alg. 4. An initial population P_0 and an initial archive are created. The fitness value of all individuals are calculated in population and in the archive (external set). All

nondominated individual are copied to the new archive. Finally, the algorithm select the individual using a tournament selection [128] [109] [78].

Algorithm 4 SPEA2 Algorithm [128]

1: Read N - Population size
2: Read $\bar{N}$ - Archive size
3: Read T - Maximum number of generations
4: Generate a initial population P_0
5: Create a initial archive $\bar{P}$
6: Set T to zero
7: Calculate the fitness value of individuals in P_t
8: Calculate the fitness value of individuals in $\bar{P}_t$
9: Environmental Selection - Copy all nondominated individuals in P_t and $\bar{P}_t$ to $\bar{P}_{t+1}$
10: **if** size of $\bar{P}_{t+1}$ exceeds $\bar{N}$ **then**
11: reduce $\bar{P}_{t+1}$ by means of truncation operator
12: **else if** size of $\bar{P}_{t+1}$ less than $\bar{N}$ **then**
13: fill $\bar{P}_{t+1}$ with dominated individuals in P_t and $\bar{P}_t$
14: **end if**
15: **if** t $\gg$ T or another stopping criterion is satisfied **then**
16: set A to the set of decision vectors represented by nondominated individuals in $\bar{P}_{t+1}$
17: Stop
18: **end if**
19: Mating selection - Perform binary tournament selection with replacement on $\bar{P}_{t+1}$ in order to fill the mating pool

The main differences between SPEA2 and NSGA-II are the diversity assignment and replacement. NSGA-II uses a fast non-dominated sorting algorithm and uses Pareto optimality levels as the primary criterion to select solutions. SPEA2 derives the strength of each solution from the number of other solutions it dominates. NSGA-II uses the crowding-distance to maintain a well-spread set of solutions whereas SPEA2 applies the k-nearest neighbor approach (Figure 3.5) [109] [32].

Comparing multi-objective metaheuristics

Deb states that are two orthogonals goals for any multi-objective algorithm [31]:

- Identify solutions as close as possible to the true Pareto frontier;
- identify a diverse of sets of solutions distributed across the entire Pareto-optimal surface.

There are several metrics either closeness or diversity. Example of metrics which measure the closeness of Pareto frontier is Error ratio and Set coverage. Example of metrics which measure the diversity are the Spacing and the Spread. The Hypervolume metric measure both closeness and diversity [64].

The Hypervolume metric calculates the volume in an objective space covered by the non-dominated individuals. The hypervolume was originally proposed by Zitzler and Thiele [127]. It is especially useful when the true Pareto-optimal solution is unknown. For each solution, a hypercube is computed from a reference point and the solution as the diagonal corners of the hypercube (Figure 3.6) [64].

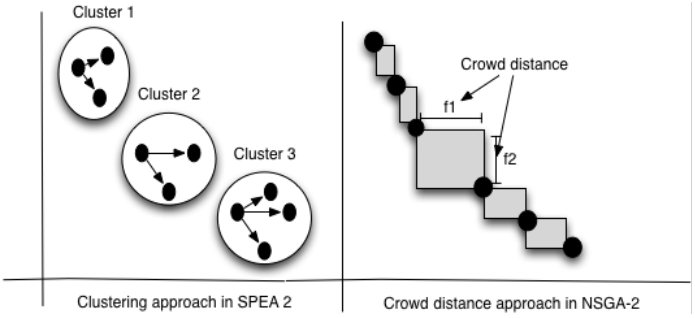


Figure 3.5: Comparison between SPEA-2 and NSGA-II [32]

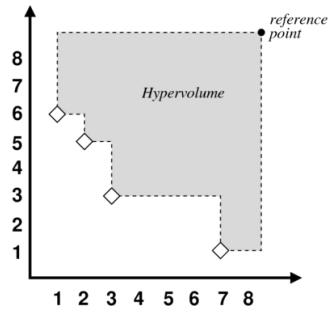


Figure 3.6: Hypervolume metric [70]

The reference point is found by constructing a vector of worst objects fitness value. The equation of the hypervolume is:

$$Hypervolume = volume(U_{i=1}^{|Q|} v_i) \tag{3.1}$$

3.1.5 Metaheuristic Noise Reduction

Software is pervasive, which raises the value of testing it [97]. Various actions outside the application under test can cause high response times such as pagination, network usage or even a software upgrade. It is necessary a noisy reduction strategy in stress test in this situations. Noisy optimization is currently receiving increasing popularity for its widespread applications in engineering optimization problems, where the objective functions are often found to be contaminated with noisy [93].

Standard Error Dynamic Resampling (SEDR), strategy has been employed for solving both noisy single and multi-objective evolutionary optimization problems. The working principle of SEDR is to add samples to a solution sequentially until the standard error of the objectives falls below a chosen threshold value [100]. It was proposed in [36] for single-objective optimization problems. In this study we apply SEDR on multi-objective problems by aggregating all objective values to a scalar value. As aggregation the median of the objective standard errors is used. The strategy is concerned with the optimal allocation of sampling budget to a trial solution based on the noise strength at its corresponding position in the search space. The contamination level of noise is captured by the standard error of the mean fitness estimate of a trial solution. The SEDR algorithm is described in Alg. 5.

SEDR noise reduction is used by this research in multi-objective scenarios experiments where the objective of the experiments besides finding the tests with the longest response time also needs to find the Pareto frontier of the application.

Algorithm 5 SEDR algorithm [100]

1: **input :** Solution s
2: Draw $b_{min} \geq 2$ initial samples of s, F(s)
3: Calculate mean of the available fitness for each of the m objectives: $\mu_i(s)$, i=1,...,m
4: Calculate standart desviation: $\sigma_i = \sqrt{\frac{1}{n-1} \sum_n^{j=1} (F_j^i - \mu_i(s))^2}$
5: Calculate the standart error: $se_i(s) = \frac{\sigma_i}{\sqrt{n}}$
6: Calculate an aggregation og the standart errors $\overline{se}(s)$
7: Stop if $\overline{se}(s) >$ threshold or $b_s \geq b_{max}$ otherwise go to step 2

3.2 Search-based Stress testing

The search for the longest execution time is regarded as a discontinuous, nonlinear, optimization problem, with the input domain of the system under test as a search space [105]. The application of SBST algorithms to stress tests involves finding the best- and worst-case execution times (B/WCET) to determine whether timing constraints are fulfilled [2].

There are two measurement units normally associated with the fitness function in a stress test: processor cycles and execution time. The processor cycle approach describes a fitness function in terms of processor cycles. The execution time approach involves executing the application under test and measuring the execution time [2] [111]. Processor cycles measurement is deterministic in the sense that it is independent of system load and results in the same execution times for the same set of input parameters. However, such a measurement is dependent on the compiler and optimizer used, therefore, the processor cycles differ from each platform. Execution time measurement is a non-deterministic approach, in which there is no guarantee of obtaining the same test inputs [2]. However, stress testing where testers have no access to the production environment should be measured by the execution time measurement [82] [2].

Table 3.1 shows a comparison between the research studies on load, performance, and stress tests presented by Afzal et al. [2]. Afzal's work was added with some of the latest research in this area ([44] [46] [33] [34] [4] [51]). The columns represent the type of tool used (prototype or functional tool), and the rows represent the metaheuristic approach used by each research study (genetic algorithm, Tabu search, simulated annealing, or a customized algorithm). The table also sorts the research studies by the type of fitness function used (execution time or processor cycles).

The studies can be grouped into two main groups: Search-Based Stress Tesing on Safety-critical systems or Search-Based Stress Testing on non Safety-critical systems.

3.2.1 Search-Based Stress Testing on Safety-critical systems

Domains such as avionics, automotive and aerospace feature safety-critical systems, whose failure could result in catastrophic consequences. The importance of software in such systems is permanently increasing due to the need of a higher system flexibility. For this reason, software components of these systems are usually subject to safety certification. In this context, software safety certification has to take into account

Table 3.1: Distribution of the research studies over the range of applied metaheuristics

	Prototypes		**Functional Tool**
	Execution Time	Processor Cycles	Execution Time
GA + SA + Tabu Search +Q-Learning + Multi-objective he			Our approach [51]
GA + SA + Tabu Search			Gois et al. 2016 [51]
GA	Alander et al.,1998 [3] Wegener et al., 1996 and [119][63] Sullivan et al., 1998 [105] Briand et al., 2005 [22] Canfora et al., 2005 [26]	Wegener and Grochtmann, [118] Mueller et al., 1998 [84] Puschner et al. [90] Wegener et al., 2000 [120] Gro et al., 2000 [55]	Di Penta et al., 2007 [86] Garoussi, 2006 [44] Garousi, 2008 [45] Garousi, 2010 [46]
Simulated Annealing (SA)			Tracey, 1998 [110]
Constraint Programming			Di Alesio et al., 2014 [34] Di Alesio et al., 2013 [33]
GA + Constraint Programming			Di Alesio et al., 2015 [4]
Customized Algorithm		Pohlheim, 1999 [88]	

performance requirements specifying constraints on how the system should react to its environment, and how it should execute on its hardware platform [33].

Usually, embedded computer systems have to fulfil real-time requirements. A faultless function of the systems does not depend only on their logical correctness but also on their temporal correctness. Dynamic aspects like the duration of computations, the memory actually needed during program execution, and other synchronisation of parallel processes are of major importance for the correct function of real-time systems [63] .

The concurrent nature of embedded software makes the order of external events triggering the system tasks is often unpredictable. Such increasing software complexity renders performance analysis and testing increasingly challenging. This aspect is reflected by the fact that most existing testing approaches target system functionality rather than performance [33]. Reactive real-time systems must react to external events within time constraints. Triggered tasks must execute within deadlines. Shousha develops a methodology for the derivation of test cases that aims at maximizing the chance of critical deadline misses [99].

The main goal of Search-Based Stress testing of Safety-critical systems is finding a combination of inputs that causes the system to delay task completion to the greastest extent possible. The followed approaches use

metaheuristics to discover the worst-case execution times. Wegener et al. [119] used GAs to search for input situations that produce very long or very short execution times. The fitness function used was the execution time of an individual measured in micro seconds [119]. Alander et al. [3] performed experiments in a simulator environment to measure extreme response times of protection relay software using genetic algorithms. The fitness function used was the response time of the tested software. The results showed that GA generated more input cases with longer response times [3].

Wegener and Grochtmann performed a experimentation to compare GA with random testing. The fitness function used was the execution duration measured in processor cycles. The results showed that, with a large number of input parameters, GA obtained more extreme execution times with less or equal testing effort than random testing [63] [118]. Gro et. al. [55] presented a prediction model which can be used to predict evolutionary testability. The research confirmed that there is a relationship between the complexity of a test object and the ability of a search algorithm to produce input parameters according to B/WCET [55]. Briand et al. [22] used GA to find the sequence of arrival times of events for aperiodic tasks, which will cause the greatest delays in the execution of the target task. A prototype tool named real-time test tool (RTTT) was developed to facilitate the execution of runs of a GA. Two case studies were conducted and results illustrated that RTTT was a useful tool to stress a system under test [22].

Pohlheim and Wegener used an extension of genetic algorithms with multiple sub-populations, each using a different search strategy. The duration of execution, measured in processor cycles, was taken as the fitness function. The GA found longer execution times for all the given modules in comparison with systematic testing [88]. Garousi presented a stress test methodology aimed at increasing the chances of discovering faults related to distributed traffic in distributed systems. The technique uses as input a specified UML 2.0 model of a system, augmented with timing information. The results indicate that the technique is significantly more effective at detecting distributed traffic-related faults when compared to standard test cases based on an operational profile [44]. Alesio, Nejati and Briand describe an approach based on Constraint Programming (CP) to automate the generation of test cases that reveal, or are likely to, task deadline misses. They evaluate it through a comparison with a state-of-the-art approach based on GAs. In particular, the study compares CP and GA in five case studies for efficiency, effectiveness, and scalability. The experimental results show that, on the larger and more complex case studies, CP performs significantly better than GA. The research proposes a tool-supported, efficient and effective approach based on CP to generate stress test cases that maximize the likelihood of task deadline misses [33].

Alesio describes stress test case generation as a search problem over the space of task arrival times. The research locates the worst-case scenarios maximizing deadline misses where each scenario characterizes a test case. The paper combines two strategies, GA and CP. The results show that, in comparison with GA and CP in isolation, GA+CP achieves nearly the same effectiveness as CP and the same efficiency and solution diversity as GA, thus combining the advantages of the two strategies. Alesio concludes that a combined GA+CP approach to stress testing is more likely to scale to large and complex systems [4].

3.2.2 Search-Based Stress Testing on non Safety-critical systems

Usually, the application of Search-Based Stress Testing on non safety-critical systems deals with the generation of test cases that causes Service Level Agreement (SLA) violations.

Tracey et al. [110] used simulated annealing (SA) to test four simple programs. The results of the research presented that the use of SA was more effective with a larger parameter space. The authors highlighted the need of a detailed comparison of various optimization techniques to explore the worst-case execution time (WCET) and the best-case execution times (BCET) of the system under test [110].

Di Penta et al. [86] used GA to create test data that violated quality of service (QoS) constraints, causing SLA violations. The generated test data included combinations of inputs. The approach was applied to two case studies. The first case study was an audio processing workflow, and the second case study, a service producing charts [86].

Gois et al. proposes a hybrid metaheuristic approach using genetic algorithms, simulated annealing, and tabu search algorithms to perform stress testing. A tool named IAdapter, a JMeter plugin used for performing search-based stress tests, was developed. Two experiments were performed to validate the solution. In the first experiment, the signed-rank Wilcoxon non-parametrical procedure was used for comparing the results. The significance level adopted was 0.05. The procedure showed that there was a significant improvement in the results with the Hybrid Metaheuristic approach. In the second experiment, the whole process of stress and performance tests, which took 3 days and about 1800 executions, was carried out without the need for monitoring by a test designer. The tool automatically selected the next scenarios to be run up to the limit of the previously established six generations [51].

Chapter 4

Stress Search Based Testing using Hybrid Metaheuristic Approach

This chapter presents the Hybrid approach proposed by Gois et al. [51]. The solution proposed by Gois et al. makes it possible to create a model that evolves during the test. A plugin called iadapter was implemented for the research. IAdapter is a JMeter plugin designed to perform search-based stress tests. The plugin is available at `www.github.com/naubergois/newiadapter`.

The proposed solution model uses genetic algorithms, tabu search, and simulated annealing in two different approaches. The study initially investigated the use of these three algorithms. Subsequently, the study will focus on other population-based and single point search metaheuristics. The first approach uses the three algorithms independently, and the second approach uses the three algorithms collaboratively (hybrid metaheuristic approach).

In the first approach , the algorithms do not share their best individuals among themselves. Each algorithm evolves in a separate way (Fig. 4.1). The second approach uses the algorithms in a collaborative mode (hybrid metaheuristic). In this approach, the three algorithms share their best individuals found (Fig. 4.2). The next subsections present details about the used metaheuristic algorithms (Representation, initial population and fitness function).

4.1 Representation

The solution representation provides a common representation for all workloads. Each workload is composed by a linear vector with 21 positions (Figure 4.3 -❶). The first position represents an metadata with the name of an individual. The next positions represent 10 scenarios and their numbers of users (Figure 4.3 -❷). The fixed-length genome approach was chosen in reason of the ease of implementation in the JMeter tool. Each scenario is an atomic operation: the scenario must log into the application, run the task goal, and undo any

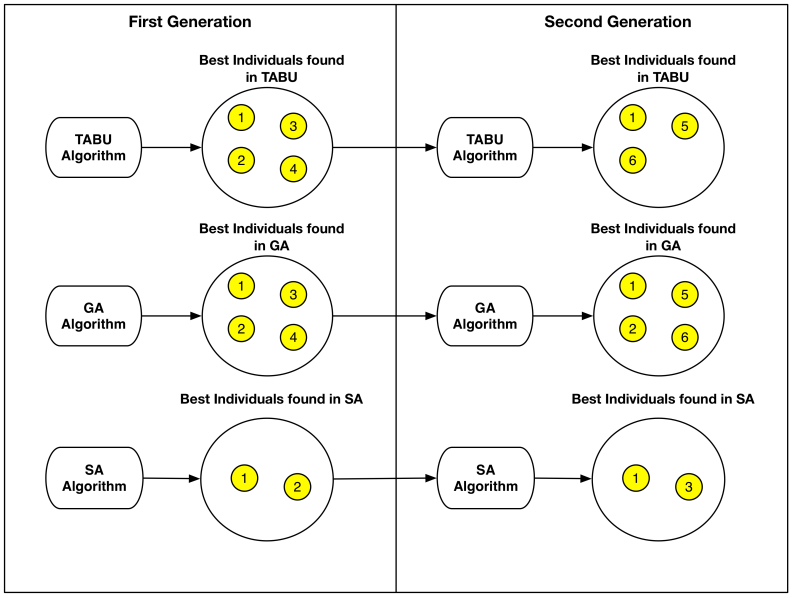


Figure 4.1: Use of the algorithms independently [51]

changes performed, returning the application to its original state.

Figure. 4.3 presents the solution representation and an example using the crossover operation. In the example, solution 1 (Figure 4.3 -❸) has the Login scenario with 2 users, the Search scenario with 4 users, Include scenario with 1 user and the Delete scenario with 2 users. After the crossover operation with solution 2 (Figure 4.3 -❹), We obtain a solution with the Login scenario with 2 users, the Search scenario with 4 users, the Update scenario with 3 users and the Include scenario with 5 users (Figure 4.3 -❺). Figure. 4.3 -❻ shows the strategy used by the proposed solution to obtain the neighbors for the Tabu search and simulated annealing algorithms. The neighbors are obtained by the modification of a single position (scenario or number of users) in the vector.

4.2 Initial population

The strategy used by the plugin to instantiate the initial population is to generate 50% of the individuals randomly, and 50% of the initial population is distributed in three ranges of values:

- Thirty percent of the maximum allowed users in the test;
- Sixty percent of the maximum allowed users in the test; and
- Ninety percent of the maximum allowed users in the test.

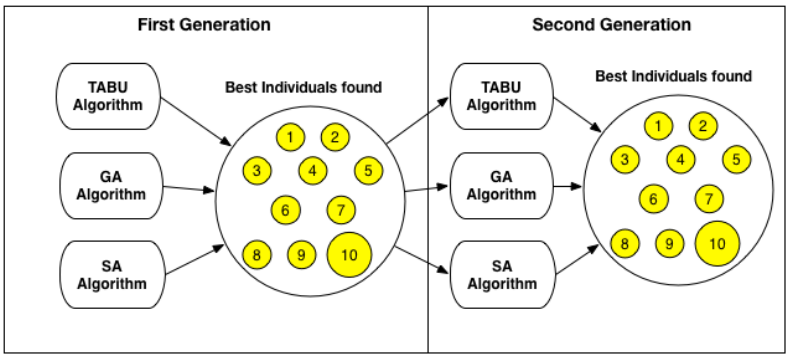


Figure 4.2: Use of the algorithms collaboratively [51]

The percentages relate to the distribution of the users in the initial test scenarios of the solution. For example, in a hypothetical test with 100 users, the solution will create initial test scenarios with 30, 60 and 90 users.

4.3 Objective (fitness) function

The proposed solution was designed to be used with independent testing teams in various situations, in which the teams have no direct access to the environment, where the application under test was installed. Therefore, the IAdapter plugin uses a measurement approach as the definition of the fitness function. The fitness function applied to the IAdapter solution is governed by the following equation:

$$\begin{aligned} fit = {} & numberOfUsersWeight * numberOfUsers \\ & -90percentileweight * 90percentiletime \\ & -80percentileweight * 80percentiletime \\ & -70percentileweight * 70percentiletime \\ & -maxResponseWeight * maxResponseTime \\ & -penalty \end{aligned} \tag{4.1}$$

The users and response time factors were chosen because they are common units of measurement in load test tools [97]. The proposed solution's fitness function uses a series of manually adjustable user-defined weights (90percentileweight, 80percentileweight, 70percentileweight, maxResponseWeight, and numberOfUsersWeight). These weights make it possible to customize the search plugin's functionality. A penalty is applied when the response time of an application under test runs longer that the service level. The penalty is calculated by the follow equation:

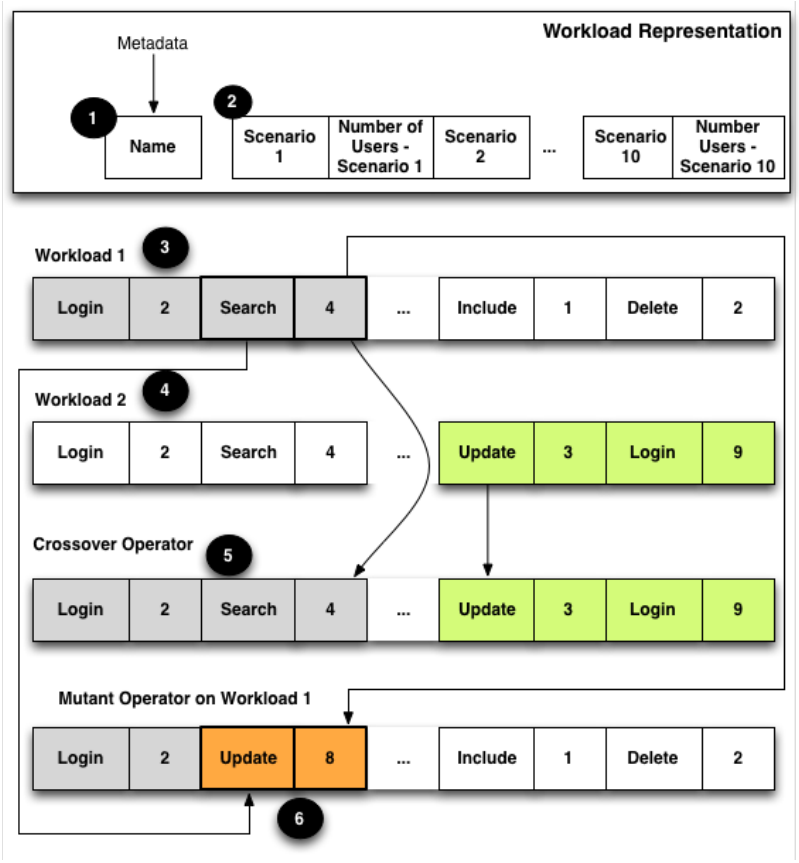


Figure 4.3: Solution representation, crossover and neighborhood operators [51]

$$\begin{aligned} penalty &= 100 * \Delta \\ \Delta &= (t_{CurrentResponseTime} - t_{MaximumResponseTimeExpected}) \end{aligned} \tag{4.2}$$

4.4 Experiments with Hybrid Algorithm

This section presents three experiments. The first one was performed on an emulated component, the second one was performed using an installed Moodle application and the third one was performed using four antipatterns. The experiments used the following fitness function:

$$\begin{aligned} fit = 0.9 * 90percentiletime \\ +0.1 * 80percentiletime \\ +0.1 * 70percentiletime+ \\ 0.1 * maxResponseTime+ \\ 0.2 * numberOfUsers - penalty \end{aligned} \tag{4.3}$$

This fitness function is the same function represented in the section VII with the manually adjustable user-defined weights filled out. This fitness function intended to find individuals with the highest percentile of 90%, followed by individuals with a higher percentile time of 80% and 70%, maximum response time, and number of users.

The first experiment ran for 27 generations, and the second experiment performed 6 generations, with 300 executions by generation (100 times for each algorithm), generating 300 new individuals. The experiments used an initial population of 100 individuals. The genetic algorithm used the top 10 individuals from each generation in the crossover operation. The Tabu list was configured with the size of 10 individuals and expired every 2 generations. The mutation operation was applied to 10% of the population on each generation.

4.4.1 First Experiment: Emulated Class Test

The first experiment aimed to perform performance, load, and stress testing on a simulated component. The purpose of using a simulated component was to be able to perform a greater number of generations in a shorter time available and eliminate variables such as the use of databases and application servers. The first experiment used a test class named SimulateConcurrentAccess. This class has a static variable named *x* and a set of methods that use the variable in a synchronized context (Listing 4.1). The experiment was executed using the JMeter Java Request Sampler Component with IAdapter.

Fig.4.4 presents the best results in 27 generations applied in the first experiment. The figure shows the results obtained with the algorithms with and without collaboration. The *x* axis represents the generation number, and the *y* axis represents the best fitness value obtained until the current generation. A higher value in the figure means that the scenario has a greater response time by the application under test. The results of the experiment showed that the use of cooperation between the three algorithms resulted in finding the individuals with better fitness values.

Table 4.1 presents the results obtained by the hybrid metaheuristic (HM) approach, genetic algorithm (GA), simulated annealing (SA), and Tabu search (TS) from 27 generations in the first experiment. The values are the maximum fitness value obtained by each algorithm.

The signed-rank Wilcoxon non-parametrical procedure was used for comparing the results with Z-value and W-value. The significant level adopted was 0.05. The Z-value obtained was -2.2736 and the p-value was 0.0232. The W-value obtained was 78. The critical value of W for N = 25 at p<= 0.05 was 89.The result was significant at p<= 0.05. The procedure showed that there was a significant improvement in the results with

Listing 4.1 SimulateConcurrentAccess class

```
1:  public class SimulateConcurrentAccess {
2:    @Test
3:    public void firstScenario() {
4:      synchronized (StaticClass.class) {
5:        for (int i = 0; i <= 1000; i++) {
6:          StaticClass.x += i;
7:        }
8:        StaticClass.x = 0;
9:      }
10:   }
11:
12:     @Test
13:   public void secondScenario() {
14:     synchronized (StaticClass.class) {
15:       for (int i = 0; i <= 2000; i++) {
16:         StaticClass.x += i;
17:       }
18:       StaticClass.x = 0;
19:     }
20:   }
```

the collaborative approach.

4.4.2 Second Experiment: Moodle Application Test

The second experiment used a Moodle application installed in a machine with 500 GB of hard disk space and 8 GB of memory. The study used six application scenarios:

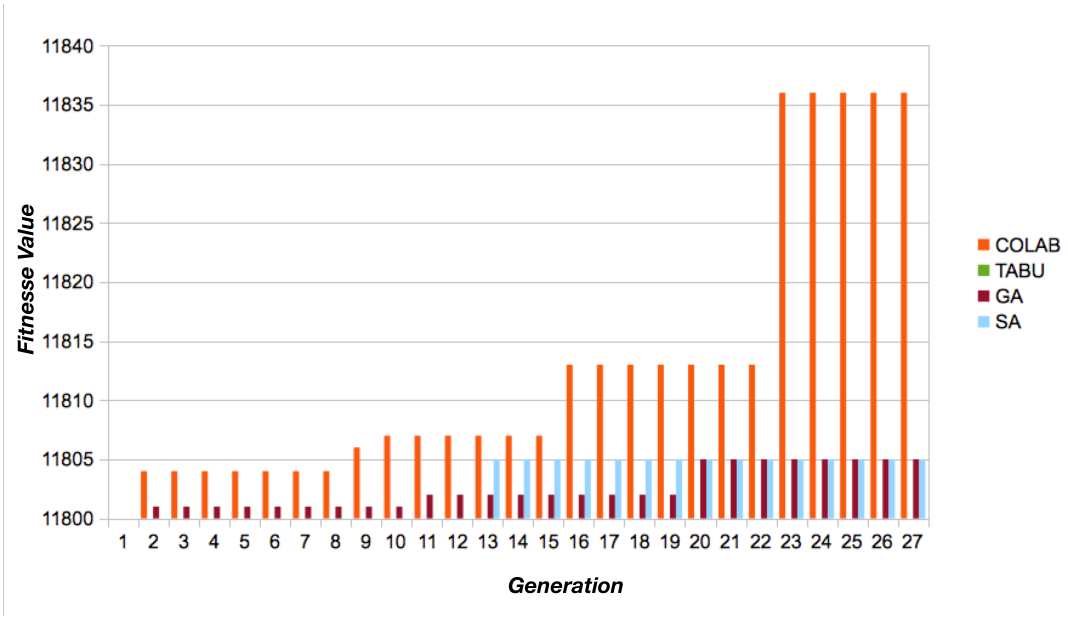
- PostDeleteMessage: This scenario posts and deletes messages in the Moodle application.
- MyHome: This scenario accesses the homepage of the user's application.
- Login: This scenario is responsible for user authentication by the application.
- Notifications: This scenario involves entering the notification page of each user.
- Start Page: This scenario shows the initial start page of the application.
- Badge: This scenario involves entering the badge page.

The maximum tolerated response time in the test was 30 seconds. Any individuals who obtained a time longer than the stipulated maximum time suffered penalties. The whole process of stress and performance tests, which took 3 days and about 1800 executions, was carried out without the need for monitoring by a test designer. The tool automatically selected the next scenarios to be run up to the limit of six generations previously established.

Table 4.2 presents the maximum fitness value obtained by the hybrid metaheuristic (HM) approach, genetic algorithm (GA), simulated annealing (SA), and Tabu search (TS) in each generation.

The small number of samples of the experiment is insufficient to give a statistical significance to the results of the Wilcoxon procedure. However, it is noted that, in four of six generations, the collaborative

Figure 4.4: Best results obtained in 27 generations



approach presented the best values. The experiment succeeded in finding 29 individuals whose maximum time expected by the application was obtained. Table 4.3 shows an example of the six individuals with the highest fitness values in the second experiment. The table shows the fitness value (Fit); the name of the scenario (Scenario); the number of users (Users); and the percentiles of 90%, 80%, and 70% (90per, 80per and 70per) in seconds.

Table 4.4 presents the percentage of genes in all test scenarios by generation with and without collaboration. Most of the genes converged to the MyHome feature, which had the highest application response time.

4.4.3 Third Experiment: AntiPatterns

In this subsection, We present the results of the experiment which we carried out to verify the best case scenarios found by the hybrid metaheuristic approach. We conducted the experiment in two phases in order to verify the effectiveness of the hybrid algorithm. Each experiment use two different antipatterns and happy scenarios. The experiment ran for 17 generations. The experiments used an initial population of 4 individuals by metaheuristic. The genetic algorithm used the top 10 individuals from each generation in the crossover operation. The Tabu list was configured with the size of 10 individuals and expired every 2 generations. The mutation operation was applied to 10% of the population on each generation. The experiments uses tabu search, genetic algorithms and the hybrid metaheuristic approach proposed by Gois et al. [51]. The objective function applied is intended to maximize the number of users and minimize the response time of

Table 4.1: Maximum value of the fitness function by algorithm

GEN	HM	TS	GA	SA
1	11238	11238	11238	11238
2	11804	11596	11801	10677
3	11787	8932	8411	10869
4	11723	9753	9611	10760
5	8164	9780	10738	4794
6	11802	9781	11086	6120
7	9985	5782	11272	11798
8	11803	11749	10084	11309
9	11806	7284	11633	10766
10	11807	9386	11717	4557
11	11802	9653	11802	11151
12	11807	10594	11793	9434
13	11802	10848	10382	11805
14	11801	11551	7219	10237
15	11807	1701	7189	9338
16	11813	6203	11758	5321
17	11805	10720	10805	11748
18	9600	6371	11698	7818
19	11733	8160	11648	11509
20	9589	9428	11805	4813
21	11800	9463	11798	10801
22	11805	11799	11804	6029
23	11836	11655	11800	3579
24	11805	11512	11803	5761
25	11804	11573	11802	9680
26	11800	11575	11403	9388
27	11805	10691	11745	9465

Table 4.2: Results obtained from the second experiment

GEN	HM	TS	GA	SA
1	32242	32242	32242	32242
2	34599	32443	26290	35635
3	35800	34896	34584	34248
4	35782	34912	32689	25753
5	35611	31833	34631	8366
6	35362	35041	33397	9706

the scenarios being tested (Best Case Scenarios). In this experiments, better fitness values means to find scenarios with more users and lower values of a response time. A penalty is applied when the response time is greater than the maximum response time expected.

Table 4.3: Example of individuals obtained in the second experiment

Id	Fit	Scenario	Users	90per	80per	70per
1	35800	MyHome	31	30	29	10
		Badges	4			
2	35795	MyHome	30	30	29	10
		Notifications	2			
		Badges	2			
3	35782	MyHome	32	30	29	10
		Badges	3			
4	35773	MyHome	22	30	29	10
		Notifications	6			
		Badges	9			
5	35771	MyHome	28	30	29	9
		Badges	6			
6	35683	MyHome	27	30	29	8
		Badges	10			

Table 4.4: Percentage of genes in each scenario by generation

Gen/ Scenarios	**Non collaboration approach**						
	Initial	1	2	3	4	5	6
Badges	20	18	16	24	15	16	17
MyHome	**15**	**59**	**55**	**48**	**53**	**50**	**52**
StartPage	15	10	12	11	20	18	19
Notifications	25	5	11	10	9	10	9
Post	8	3	1	3	1	2	1
Login	17	5	5	4	2	4	2
	Collaboration approach						
Badges	20	29	16	25	9	16	9
MyHome	**15**	**29**	**69**	**49**	**74**	**66**	**76**
StartPage	15	22	10	21	10	10	8
Nofications	25	10	1	1	2	1	3
Post	8	2	1	1	1	2	1
Login	17	8	3	3	4	5	3

4.4.4 Experiment Research Questions

The following research question is addressed: • Does the Hybrid algorithm finds scenarios without the antipatterns presented?

4.4.5 Variables

The independent variables are the test scenarios (antipatterns and happy scenarios). The dependent variable are: the number of antipatterns found in best workloads and the metaheuristic with the best fitness value.

4.4.6 Hypotheses

With regard to the antipatterns found by hybrid metaheuristic:

- H_0 (A null hypothesis) :the best workloads found in the experiments contain antipatterns
- H_1 : the best workloads found in the experiments do not contain antipatterns.

4.4.7 The Ramp and Circuitous Treasure Hunt experiment

The experiment was carried out for 8 continuous hours. All tests in the experiment were conducted without the need of a tester, automating the process of executing and designing performance test scenarios.In this experiment, Scenarios were generated with the Ramp and Circuitous Treasure antipattern as well as scenarios with Happy Scenario 1, Happy Scenario 2 and mixed scenarios. Figure 4.5 present the fitness value obtained by each metaheuristic.

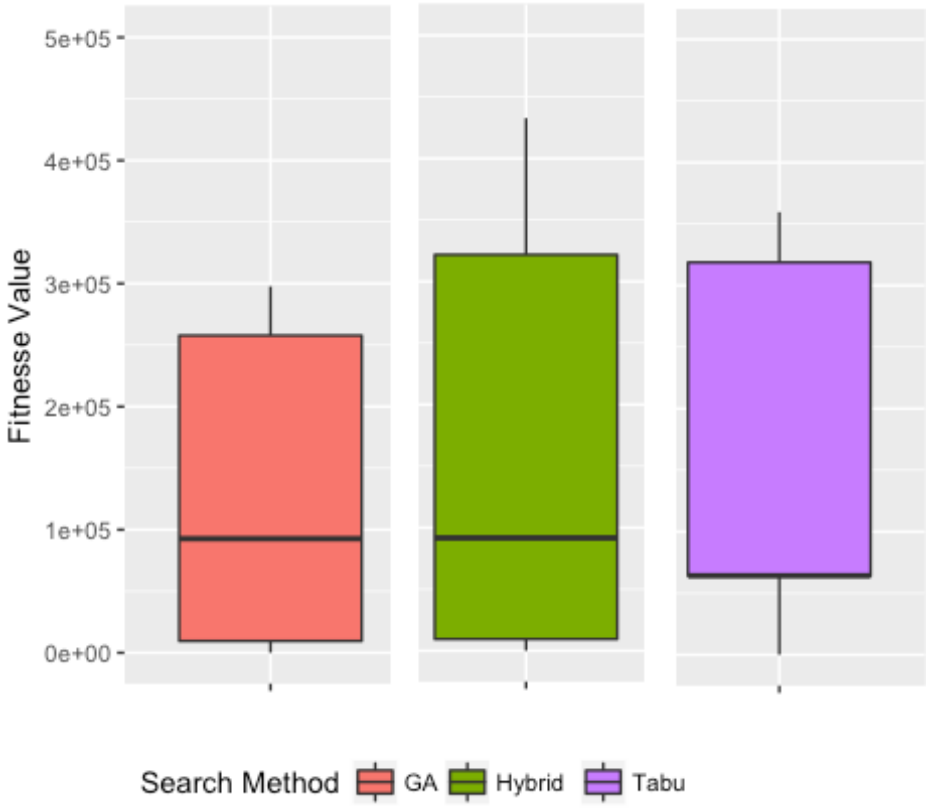


Figure 4.5: Average, median, maximum and minimal fitness value by Search Method

Table 4.5 shows 4 best individuals found in the experiment by hybrid algorithm. None of the best individuals has one of the antipatterns used in the experiment, excluding the scenarios with antipatterns.

Table 4.5: Best individuals found in the first experiment

Metaheur.	Gen.	Users	Fit	Scenarios
Hybrid	17	145	432760	Happy 1 & 2
Hybrid	17	145	432740	Happy 1 & 2
Hybrid	17	146	431760	Happy 1 & 2
Hybrid	16	143	426740	Happy 1 & 2

4.4.8 The Tower Babel and Unbalanced Processing

The experiment was carried out for 6 continuous hours. In this experiment, Scenarios were generated with Tower Babel and Unbalanced Processing antipattern as well as scenarios with Happy Scenario 1, Happy Scenario 2 and mixed scenarios.

Figure 4.6: Finesse value by generation in all tests

Table 4.6 shows the 4 best workloads found in the second experiment. Despite the fact of doing 300 conversions of the JSON format to XML. The antipattern implementation does not return a higher response time than happy paths. While happy paths returns from 10 to 15 seconds from a single user, Tower Babel antipattern has a response time of 10 to 29 seconds. None of the best individuals found implements the Unbalanced Processing antipattern.

In the second experiment, the metaheuristics converged to scenarios with a happy path and Tower Babel antipattern, excluding the scenarios with Unbalanced Processing antipattern. The hybrid metaheuristic returned individuals with higher fitness scores. The SA algorithm obtained the worst fitness values.

Table 4.6: Best individuals found by hybrid algorithm in the second experiment

Metaheur.	Gen.	Users	Fit	Scenarios
Hybrid	17	148	437780	Happy 1,2 & Tower
Hybrid	17	145	432740	Happy 1,2 & Tower
Hybrid	16	146	431800	Happy 1,2 & Tower
Hybrid	17	145	428780	Happy 1,2 & Tower

4.4.9 Threats to validity

• Construct Validity: In this work, we just evaluate the use of four antipatterns. However, several antipatterns could be applied. The testbed's common representation and the strategies used for crossover and neighborhood operators need of a better design, using an abstraction pattern to contemplate a major number of possible solutions. • Conclusion Validity: The Tower Babel antipattern was not excluded by the metaheuristics used in the experiment, requiring new studies with new approaches and experiments.

4.5 Conclusion

This chapter presented a hybrid metaheuristic approach for use in stress testing. Three experiments were performed to validate the solution. The first experiment was performed on an emulated component. The second experiment was performed using an installed Moodle application. The collaborative approach obtained better fit values in both first two experiments.

The main contributions presented in this chapter are as follows: The presentation of a hybrid metaheuristic approach for use in stress tests; the development of a JMeter plugin for search-based tests; and the automation of the stress test execution process.

In the first experiment, the signed-rank Wilcoxon non-parametrical procedure was used for comparing the results. The significant level adopted was 0.05. The procedure showed that there was a significant improvement in the results with the Hybrid Metaheuristic approach.

In the second experiment, the whole process of stress and performance tests, which took 3 days and about 1800 executions, was carried out without the need for monitoring by a test designer. The tool automatically selected the next scenarios to be run up to the limit of six generations previously established.

There is a range of future improvements in the proposed approach. Also as a typical search strategy, it is difficult to ensure that the execution times generated in the experiments represents global optimum. More experimentation is also required to determine the most appropriate and robust parameters. Lastly, there is a need for an adequate termination criterion to stop the search process.

Among the future works of the research, the use of new combinatorial optimization algorithms such as very large-scale neighborhood search is one that we can highlight

Chapter 5

Stress Search-based Testing using HybridQ approach

This chapter presents a reinforcement learning approach to optimize the choice of neighboring solutions to explore, reducing the time needed to obtain the scenarios with the longest response time in the application.The research assumes that HybridQ is more expensive than Hybrid because q-learning. The reasearch has as premise that the same application under performance tests can be submitted to more than one cycle of tests execution, reducing the cost of the exploration phase of the q-learning algorithm used.

The solution, named HybridQ, uses the GA, SA and Tabu Search algorithms in a collaborative approach. Just like most reinforcement learning problems the proposed solution works in two different phases: exploration and exploitation. The following subsections show details of the exploration and exploitation phases and the integration between metaheuristics and the Q-Learning algorithm.

5.0.1 Exploration phase

The exploration phase uses a markov model, as shown in Fig. 5.1, the proposed MDP model has three main states based on response time. A test may have a response time greater than 1.2 times the maximum response time allowed, between 0.8 and 1.2 times the maximum response time allowed or less than 0.8 times the maximum response time allowed. The values of 1.2 and 0.8 were chosen from the assumption of a tolerance margin of 20% for the application under test. This margin may be higher or lower depending on the business requirements of the application.

The algorithm maintains three different tables (Table 5.1), one for each state. The selection of which table to use depends on the response time of the application.

Algorithm 7 shows the main steps of exploration phase. The possible actions in MDP are the change of one of the test scenarios and an increase or decrease in the number of users. In line 1, the algorithm choose a random action (increase, decrease or maintain the number of users). In line 2, the algorithm choose one a

Figure 5.1: Markov Decision Process used by HybridQ

random testScenario. In lines 3 to 7, the algorithm checks if there exists a q value for the pair (action and test scenario), if not exist a q value then zero value is assigned. In line 8, the algorithm checks i the new solution increases the fitness value. A solution receives a positive reward when an action increases the fitness value and a negative reward when an action reduces the fitness value. Finally, the algorithm updates the qTable with the new q value.

Unlike the traditional approach, The update of Q values for each action also occurs in the exploitation phase. The exploration phase ends when no value of Q equals zero for a state, ie, unlike the traditional approach an agent belonging to one state may be in the exploration phase while another agent may be in the explotation phase. Table 5.1 presents hypothetical Q-values for a test. In Table 5.1, it can be observed that the agents in the Service Level state are in the exploitation phase because there is no other value of Q that equals to zero.

5.0.2 Exploitation phase

The main objective of the exploitation phase is to choose the best neighboring solution based on the Q value. The research expected that Q-Learning improve Hybrid algorithm replacing the random characteristic of the tabu search, simulated annealing and genetic algorithms operators by the direction given by the q-learning in the exploration phase. Algorithm 8 presents the main steps of exploitation phase. In first line, the algorithm

Algorithm 6 Exploration phase table selection

```
1: if responseTime < 0.8 * maxResponseTime  then
2:     return qTableBellowServiceLevel
3: end if
4: if responseTime >= 0.8 * maxResponseTime and responseTime <= 1.2* maxResponseTime then
5:     return qTableServiceLevel
6: end if
7: if responseTime > 1.2 * maxResponseTime then
8:     return qTableAboveServiceLevel
9: end if
```

Algorithm 7 HybridQ exploration phase

```
1: action ← Random.createAction()
2: testScenario ← Random.chooseTestScenario()
3: if qTable.containsKey(action + "#" + testScenario) then
4:     qValue ← qTable.get(action + "#" + testScenario);
5: else
6:     qValue ← 0
7: end if
8: if newSolution.getFitness() > oldSolution.getFitness() then
9:     qValue ← ReinforcementLearning.alpha * reward + (1 − ReinforcementLearning.alpha) * qValue
10: else
11:     qValue ← ReinforcementLearning.alpha * −reward + (1 − ReinforcementLearning.alpha) * qValue
12: end if
13: qTable.update(action + "#" + testScenario, qValue)
```

gets the original genome. In lines 2 to 11, HybridQ gets the maximum, the second maximum or the third maximum q value, depending on the random value of the random variable.The algorithm chooses one of the three largest values of q. The variation of the highest values was inserted in the algorithm to escape the local optimals. In line 12, the algorithm gets the key value in Table that have the maximum q value. In line number 13, The key is separated into two parts using the # delimiter. The first part of the key is action and the second part is the test scenario. If the action equals '*up*' value, the genome is incremented in its users. If the action equals '*down*' value, the genome is incremented in its users. Finally, the test scenario is changed and the new genome is returned.

5.0.3 Integration between metaheuristics and the Q-Learning algorithm

The Q-learning algorithm is used by Tabu Search or Simulated Annealing to obtain the neighbors and in the mutation operator of the genetic algorithm. Unlike the traditional processes of obtaining neighboring solutions such as random change and permutation, the decision to change a genome gene is made from the action that has the highest value of Q. Fig. 5.2 presents how one of the neighbors of a test is generated using Q-Learning in IAdapter. The solution uses a service called Q-Neighborhood Service to generate the neighbor from the action that has the highest value of Q.

Table 5.1: Hypothetical MDP Q-values

Above Service Level	**Scenario 1**	**Scenario 2**
Increment Users	0.2	0.0
Reduce Users	0.1	0.2
Phase	Exploration	Exploration
Service Level	**Scenario 1**	**Scenario 2**
Increment Users	0.2	0.11
Reduce Users	0.1	-0.2
Phase	Explotation	Explotation
Bellow Service Level	**Scenario 1**	**Scenario 2**
Increment Users	0.0	0.2
Reduce Users	0.1	0.0
Phase	Exploration	Exploration

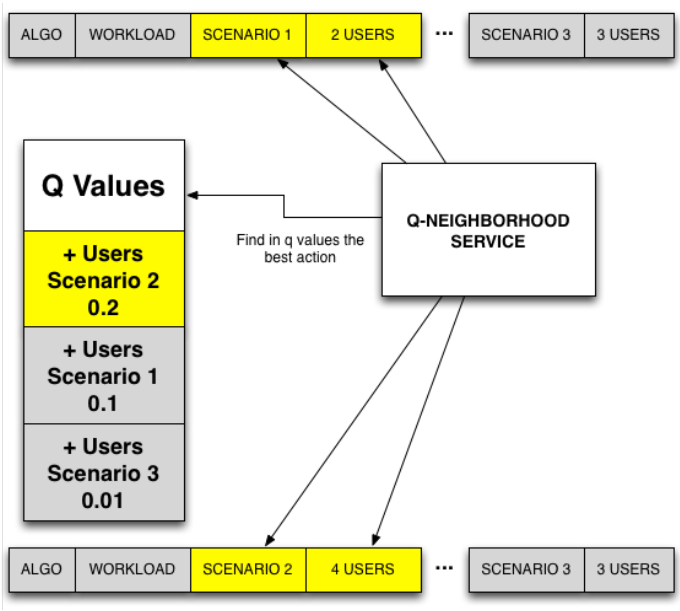


Figure 5.2: HybridQ NeighborHood Service

5.1 Experiment with HybridQ Algorithm

We conducted one experiment in order to verify the effectiveness of the HybridQ. The iterated racing procedure (irace) was applied as an automatic algorithm configuration tool for tunning metaheuristics parameters. Iterated racing is a generalization of the iterated F-race procedure to automatize the arduous task of configuring the parameters of an optimization algorithm [74]. The best parameters obtaind from irace was a population size of 5 individuals, a crossover value of 0.7551, a mutation value of 0.7947, an elitism value of 0.5356 and the maximum number of iterations of 16. The experiment ran for 16 generations in an docker

Algorithm 8 HybridQ exploitation phase

```
1: Gene[]genome ← service.getTestGenome()
2: random ← Random.nextInt(3)
3: if random==1 then
4:     qMaxValue ← qTable.getMaxValue(responseTime)
5: end if
6: if random==2 then
7:     qMaxValue ← qTable.getSecondMaxValue(responseTime)
8: end if
9: if random==3 then
10:    qMaxValue ← qTable.getThirdMaxValue(responseTime)
11: end if
12: key ← qTable.selectKey(qMaxValue)
13: String[] keySplit ← key.split('#')
14: action ← keySplit[0]
15: testScenario ← keySplit[1]
16: if action=='up' then
17:    increaseUsers(genome)
18: end if
19: if action=='down' then
20:    decreaseUsers(genome)
21: end if
22: genomePosition ← Random.nextInt(genome.length)
23: changeTestScenario(genome,testScenario,genomePosition)
```

evinronment on a server with 16 Gb of memory and 500 Gb hard disk. The experiment used an initial population of 5 individuals by metaheuristics. The genetic algorithm used the top 4 individuals from each generation in the crossover operation. The Tabu list was configured with the size of 10 individuals and expired every 2 generations. The mutation operation was applied to 79% of the population on each generation. The experiments use tabu search, genetic algorithms, simulated annealing, the hybrid metaheuristic approach proposed by Gois et al. [51] and the HybridQ approach.

The objective function applied is intended to maximize the response time of the scenarios being tested. In these experiments, better fitness values coincide with finding scenarios with higher values of response time. A penalty is applied when the response time is greater than the maximum response time expected. The experiment used the following fitness (goal) function:

$$\begin{aligned} fitness = {} & 20 * 90percentiletime \\ & 20 * 80percentiletime \\ & 20 * 70percentiletime \\ & 20 * maxResponseTime \\ & -penalty \end{aligned} \tag{5.1}$$

For the experiment an objective function with a single factor was chosen, since users and response time are conflicting factors. All tests in the experiment were conducted without the need of a tester, automating the process of executing and designing performance test scenarios.

5.1.1 Experiment Research Questions

The following research question is addressed:

- Is Q-learning technique improve the choice of neighboring solutions, improving the number of requests and the time needed to find scenarios with the longest response time in the application under test?

5.1.2 Variables

The independent variable is the algorithms used in each experiment. The dependent variables are: the optimal solution found by each algorithm, the number of requests to find optimal solution and the time of execution needed by each algorithm.

5.1.3 Hypotheses

- With regard to the optimal solution found by each algorithm:
  - $H1_0$ (null hypothesis) : The HybridQ does not find better solution than the other metaheuristic approaches.
  - $H1_1$: The HybridQ finds better solution than the one discovered by other metaheuristic approaches.
- With regard to the time consumed to find the optimal solution of each algorithm:
  - $H2_0$ (null hypothesis) : The HybridQ algorithm realizes more requests than the other algorithms in the experiments performed.
  - $H2_1$: The HybridQ algorithm does not realize more requests than the other algorithms in the executed experiments.
- With regard to the number of requests needed to find the optimal solution of each algorithm:
  - $H3_0$ (null hypothesis) : The HybridQ algorithm needs more time to find the optimal solution than the other algorithms in the experiments performed.
  - $H3_1$: The HybridQ algorithm does not need more time to find the optimal solution than the other algorithms in the experiments performed.

5.1.4 Experiment phases

The experiment was conducted in two phases. The first phase verified the number of requisitions and time required for the HybridQ exploration phase. The second phase ran the stress test using GA, Tabu Search, Simulated Annealing, Hybrid and HybridQ algorithms simultaneously.

5.1.5 OpenCart Experiment

The experiment was conducted to test the use of the HybridQ algorithm in a real implemented application. The chosen application was the OpenCart application , available at `opencart.com`. OpenCart is free open source ecommerce platform for online merchants. OpenCart works with PHP 5 and MySQL. The maximum tolerated response time in the test was 5 seconds. The whole process of stress and performance tests, which run for 2 days and with about 1.500 executions, was carried out without the need for monitoring by a test designer. The tool automatically selected the next scenarios to be run up to the limit of eleven generations previously established.

The experiments use the follow application features:

- Main page: The main page of the application.
- Search item: The application searches a product.
- Product detail: The application shows detais about one item product.
- Add to Cart: The application adds a product to shopping cart.
- View Cart:The application displays the shopping cart.
- Remove Item: The application remove item from shopping cart.

Q-Learning Training Phase

The application was submitted to 1 hour of training with the Q-learning algorithm using all test scenarios and was obtained the Table 5.2 with the values of q for response times bellow than service level. The action and state with best q-value is increment the number of users ('up') in Add to Cart feature. The learning phase required 1.431 requisitions for application under test.

Results

Figure 5.3 presents the number of requests by maximum fitness value. HybridQ algorithm obtains the maximum value of fitness: 364860 ($H1_1$ hypothesis). HybridQ obtained a solution with greater fitness value, but it needs a much greater number of requests than the other algorithms, not contemplating the hypothesis $H2_1$. GA is the algorithm that obtain the best fitness value with minor number of requests ($H1_0$ hypothesis). All

Table 5.2: Q values for response times bellow than service level

Action	Feature	Q- value	State	Feature	Q- value
up	Main Page	-0.0405763	**up**	**Add to Cart**	**0.0390237**
down	Main Page	0.00079202	down	Add to Cart	-0.00079202
same	Main Page	-0.0398	same	Add to Cart	-0.0398
up	Search Page	-0.00079202	up	View Cart	-0.0398
down	Search Page	-0.0398	down	View Cart	-0.0398
same	Search Page	-0.0398	same	View Cart	-0.0398
up	Product Detail	-0.00079202	up	Remove Item	-0.0398
down	Product Detail	-0.00079202	down	Remove Item	-0.0398
same	Product Detail	-0.0398	same	Remove Item	-0.0398

algorithms consume the same time of test (6 hours). The scenario with more fitness value has 4,8 seconds of response time and 38 users:

- 25 users on search page;
- 10 users on Add to Cart feature;
- 2 users removing items from cart;
- 1 users on Main Page.

The t-Test and Wilcoxon Rank Sum Test was applied using the R language. The test results show that HybridQ and HybriD algorithms is superior than GA, Tabu Search and Simulated Annealing with $p < 0.02$. t-Test shows that the mean of HybridQ fitness value is superior than Hybrid.

```
 1:         Welch Two Sample t-test
 2:
 3: data:  b\$MAXFIT and c\$MAXFIT
 4: t = 13.829, df = 31678, p-value < 2.2e-16
 5: alternative hypothesis: true difference in means is not equal to 0
 6: 95 percent confidence interval:
 7:  7506.846 9986.226
 8: sample estimates:
 9: mean of x mean of y
10:  322007.5  313260.9
```

5.1.6 Threats to validity

In this work, we just evaluate the use of single objective algorithm. However, several multiobjective algorithms could be applied. The experiments are performed with the configuration obtained by the irace

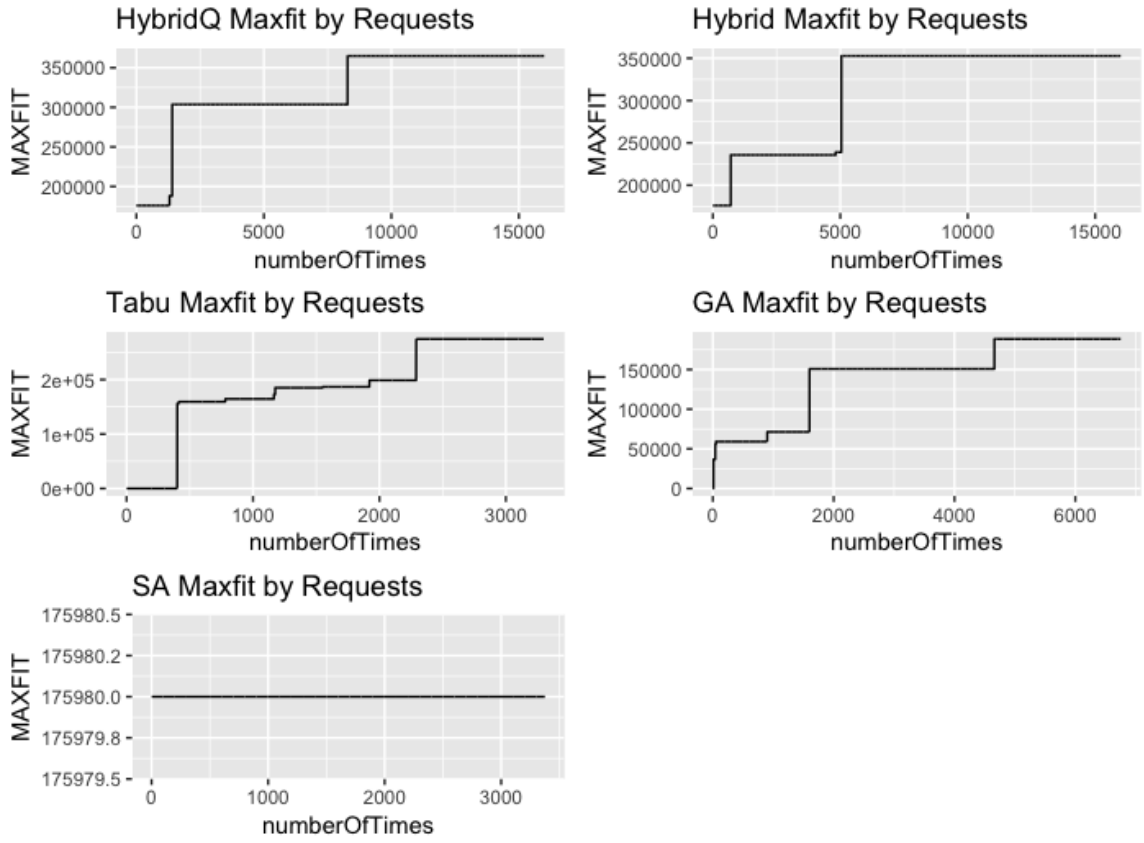


Figure 5.3: Maximum fitness value by number of requests

algorithm, however new experiments are required to verify the sensitivity of the results. It is necessary to compare the current approach with the constraint programming approaches presented in the state of art.

5.2 Conclusion

This chapter present the study that extends the article "Improving stress search based testing using a hybrid metaheuristic approach" in order to ascertain if the use of the Q-learning technique allows the meta-heuristic algorithms to improve the search for application failures One experiment was conducted to validate the proposed approach. The experiments use genetic algorithms, tabu search, simulated annealing, the Hybrid approach proposed by Gois et al. [51] and the HybridQ algorithm. The experiment ran for 17 generations. The experiment used an initial population of 5 individuals by metaheuristics. All tests in the experiment were conducted without the need of a tester, automating the execution of stress tests with the JMeter tool. HybridQ found the individuals with a greater response time. The scenario with greater fitness has 38 users with the Search Page, Add to Cart feature, Removing Item and Main Page features. GA is the algorithm that obtain

the best fitness value with minor number of requests. All algorithms consume the same time of test (6 hours). There is a range of future improvements in the proposed approach:

- Also as a typical search strategy, it is difficult to ensure that the execution times generated in the experiments represent global optimum.
- More experimentation is also required to determine the most appropriate and robust parameters. Lastly, there is a need for an adequate termination criterion to stop the search process.
- The fitness approach of the Gois et al. solution are based on two conflict measures: number of users and execution time. A multi-objective algorithm should be more adequate.
- It is necessary to compare the current approach with the constraint programming approaches presented in the state of art.

Among the future works of the research, new experiments should be performed comparing the proposed approach with the use of constraint programming.

Chapter 6

Search-based stress testing using multi-objective heuristics

This chapter present experiments to assert the benefits of multiobjective metaheuristics in search-based stress testing. Multi-objective heuristics makes it possible to create a model that performs search-based stress tests with two distinct objectives. The multiobjective algorithm implementation used an adapted implementation of the jMetal framework (`http://jmetal.sourceforge.net/`). Figure. A.5 presents the multiobjective implemented solution life cycle. Given an initial population (Figure A.5 -❶), the multiobjective algorithm implementation receives a set of workloads (Figure A.5 -❷). The multiobjective algorithm generates a new set of individuals based on crossover or/and mutant operators (Figure A.5 -❸). JMeterEngine runs each workload (Figure A.5 -❺) and the multiobjective algorithm ranks and classifies each workload based on the objective functions (Figure A.5 -❻). After all these steps the cycle begins until the maximum number of generations is reached (Figure A.5 -❼).

6.1 Experiment with multi-object NSGA-II algorithm

In this section, we present the experimental results, in which we carried out to verify the multi-objective NSGA-II implementation. The experiment was conducted to validate the use of NSGA-II multiobjective algorithm with a real implemented application. The chosen application was the JPetStore, available at `https://hub.docker.com/r/pocking/jpetstore/`. The maximum tolerated response time in the test was 500 miliseconds. The whole process of stress tests, which run for 3 days and 492 executions, was carried out without the need for monitoring by a test designer. The tool automatically selected the next scenarios to be run up to the limit of 123 previously established generations by the algorithm. The experiments use the follow application features:

- Enter in the Catalog: the application presents the catalog of pets.

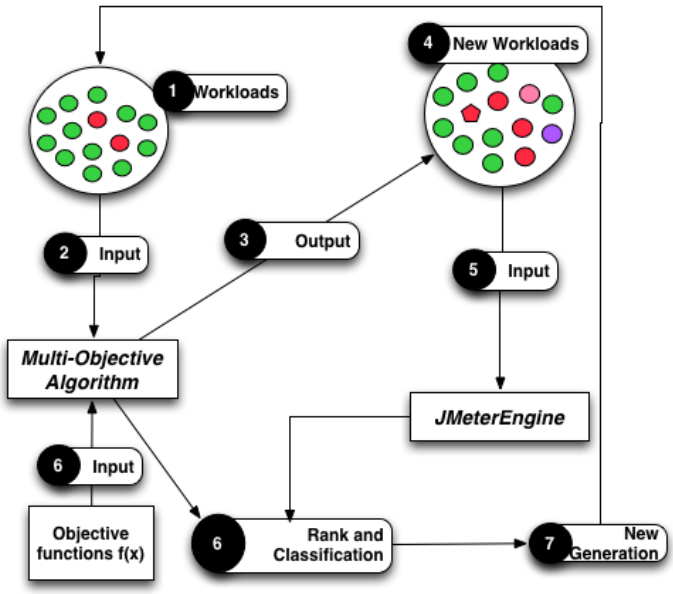


Figure 6.1: Multiobjective implemented solution life cycle

- Fish: The application shows all the fish items in stock.
- Register: a new user is registered into the system.
- Dogs: The application shows all the dog supplies in stock.
- Shopping Cart: the application displays the shopping cart.
- Add or Remove in Shopping Cart: the application adds and removes items from the shopping cart.

The experiments used an initial population of 17 individuals by metaheuristic. The genetic algorithm used the top 10 individuals from each generation in the crossover operation The mutation operation was applied to 10% of the population in each generation. The objective function applied is intended to minimize the number of users and maximize the response time of the scenarios being tested. A penalty is applied when an application under test takes a longer time to respond than the expected maximum response time. The experiment used the following objective equations:

$$objectivefunction1 = -3 * numberOfUsers - penalty \tag{6.1}$$

$$objectivefunction2 = responsetime - penalty \tag{6.2}$$

The first objective function seeks to find workloads with fewer users. The second objective function seeks to find workloads with longer response times. The penalty is calculated by the follow equation:

$$\begin{aligned} penalty &= 100 * \Delta \\ \Delta &= (t_{CurrentResponseTime} - t_{MaximumResponseTimeExpected}) \end{aligned} \tag{6.3}$$

6.1.1 Experiment Research Questions

The following research question is addressed:

- Does the NSGA-II algorithm find relevant workload scenarios according to the two test objectives?

6.1.2 Variables

The independent variable is the NSGA-II algorithm used in the test. The dependent variables are: the optimal workload scenario found by the algorithm.

6.1.3 Hypotheses

- With regard to multi-objective algorithms applied in the experiment:
  - H_0 (A null hypothesis) : The NSGA-II didn't find workloads that meet the two objective functions used in the experiment.
  - H_1 : The NSGA-II algorithm found workloads that meet the two objective functions used in the experiment.

6.1.4 Results

Fig. 6.2 and Table 6.1 present the results obtained in the experiment. The experiment found 9 optimal workloads (Pareto Frontier) that present a lower number of users with high response times. Workload number 1 with a single user accessing the dog scenario provided a response time of 245 miliseconds. Workload number 2 with a single user accessing the dog scenario, 7 users accessing the Enter Catalog feature and 3 users in register functionality, provided a response time of 400 seconds.

6.1.5 Threats to validity

In this work, we just evaluate the use of one multi-objective algorithm. However, several multi-objective algorithms could be applied. There is still a reasonable distance between the Pareto frontier and the data obtained for the second objective, and more experiments are needed to validate the results.

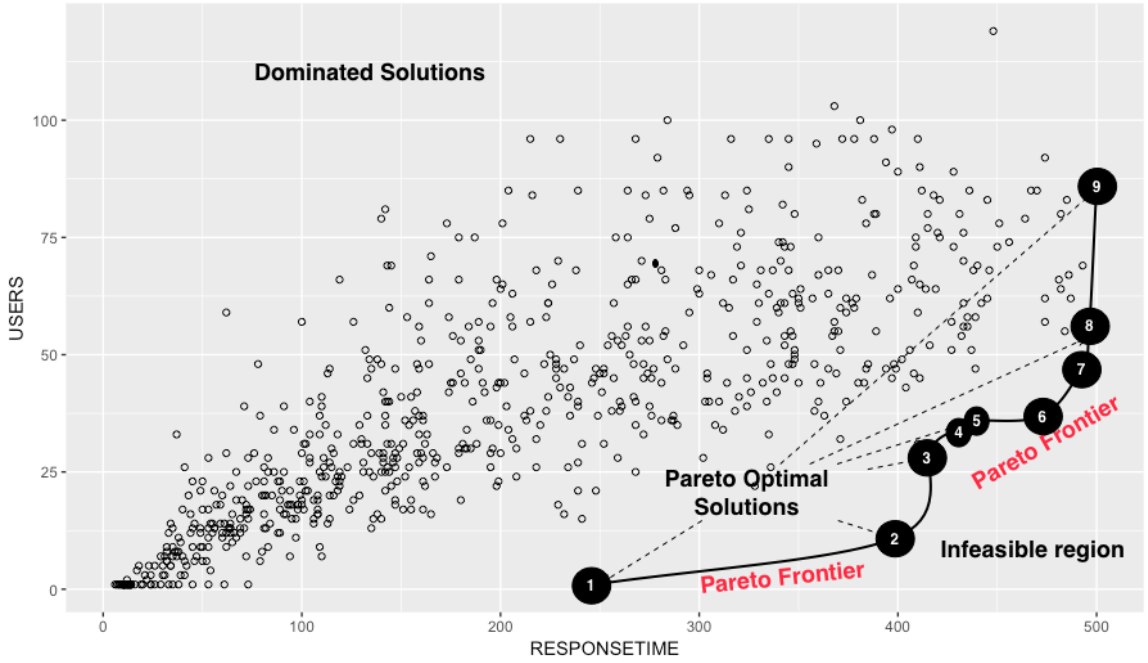


Figure 6.2: Experiment Pareto Frontier

Table 6.1: Pareto Frontier workload results

N.	OBJ. 1	OBJ. 2	Dogs Users	Enter Catalog Users	Fish Users	Register Users	Add Rem. Cart	Cart Users
❶	-3	245	1					
❷	-33	400	1	7		3		
❸	-87	416	5	15	4	5		
❹	-102	434	16	17				1
❺	-105	436	15	4	8	3	5	
❻	-111	472	7	13	7	3	7	
❼	-141	493	7	11	11	7	7	4
❽	-112	496	6		12	8	19	9
❾	-255	499		54	12		7	12

6.1.6 Experiment Conclusion

The experiment verified the use of a multi-objective algorithm in a search-based stress testing problem. A tool named IAdapter, a JMeter plugin for performing search-based load tests, was extended. One experiment was conducted to validate the proposed approach. The experiment uses the NSGA-II algorithm to discover application scenarios where there is a high response time for a small number of users. The whole process of stress tests, which ran for 3 days and 492 executions, was carried out without the need for monitoring by

a test designer. The experiment found 9 optimal workloads that present a lower number of users with high response times. The results of the experiment can help in the decision making of service levels that need to be defined for the application.

6.2 Comparative experiment with multi-object algorithms and noise reduction

In this section, we present the experimental results, in which we carried out compare four multi-objective algorithms:

- NSGA-II
- SPEA2
- PAES
- MOEAD

The experiment was conducted with the JPetStore application in two cycles with or without a noise reduction strategy. The maximum tolerated response time in the test was 10000 seconds. The whole process of stress tests, which run for 4 days and 492 executions, was carried out without the need for monitoring by a test designer. The tool automatically selected the next scenarios to be run up to the limit of 50 previously established generations by the algorithm.

The experiments uses a customized version of the SEDR algorithm for noise reduction. The algorithm runs as long as each workload has at least 3 samples or the SEDR condition is satisfied for all workloads (Figure 6.3 - ❶). The algorithm add all workloads in OUTLIERS list to the list of execution (Figure 6.3 - ❷). All workloads are performed and the SEDR condition is verified (Figure 6.3 - ❸ and ❹). If the SEDR condition is satisfied the workload is removed from OUTLIERS list (Figure 6.3 - ❺) otherwise the workload is included in the list if it is no longer present (Figure 6.3 - ❻).

6.2.1 Experiment Results

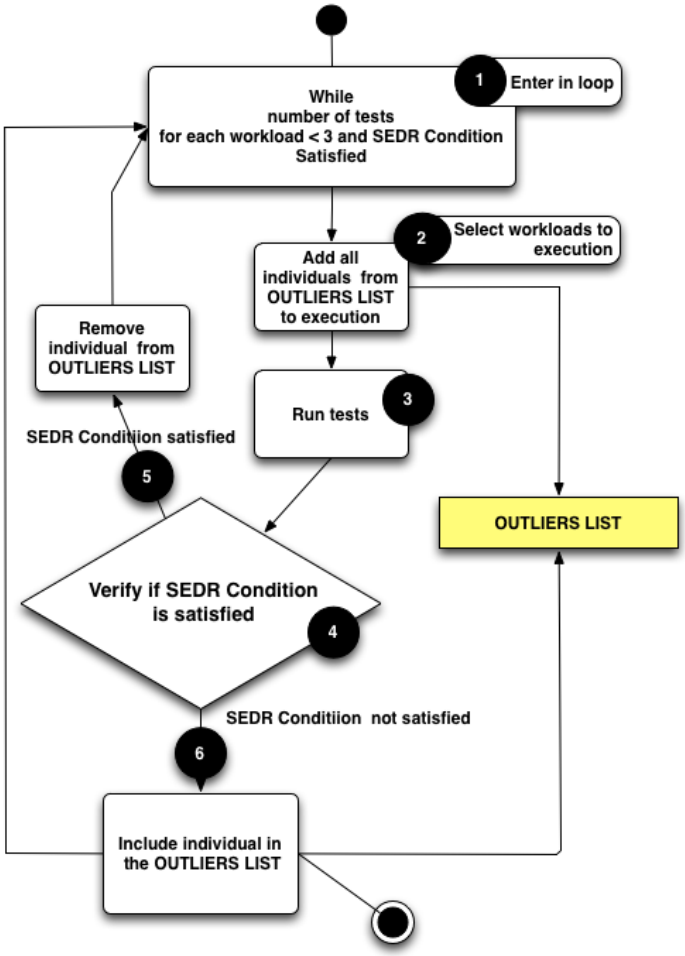


Figure 6.3: SEDR customized algorithm

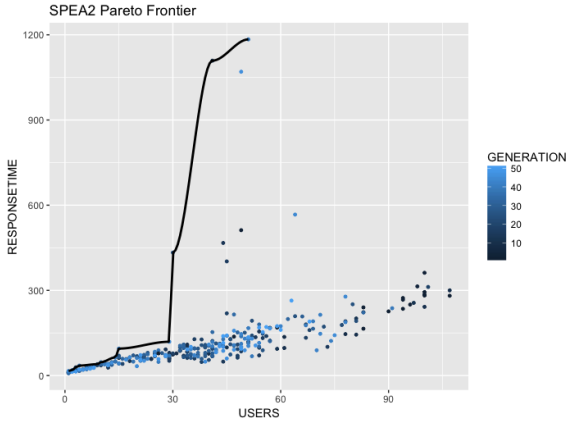


Figure 6.4: SPEA2 Pareto Frontier

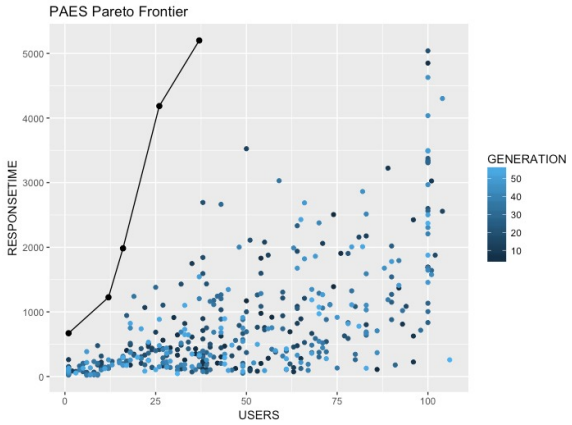


Figure 6.5: Maximum fitness value by number of requests

Chapter 7

Conclusion

In this thesis we dealt with the use of hybrid metaheuristics and Q-Learning in Stress Testing. This thesis presented an hybrid and an hybrid with Q-Learning metaheuristic approaches that combines genetic algorithms, simulated annealing, and tabu search algorithms in stress tests. A tool named IAdapter (github.com/naubergois/newiadapter), a JMeter plugin for performing search-based load tests, was developed. Six experiments were conducted to validate the proposed approach. The first experiment was performed on an emulated component. The second and third experiments are conducted in an testbed developed appplication. The fourth experiment was performed using an installed Moodle application. The fifth and sixth experiments are performned using an installed JPetStore application.

IAdapter Testbed is an open-source facility that provides software tools for search based test research. The testbed tool emulates test scenarios in a controled environment using mock objects and implementing performance antipatterns.

The main contributions of this research are as follows: The presentation of a hybrid metaheuristic using Q-learning approach for use in stress tests; the development of a Testbed tool the development of a JMeter plugin for search-based tests and the automation of the stress test execution process.

7.1 Achievements

Two experiments were performed to validate the hybrid metaheuristic,two experiments were conducted to validate the testbed tool and two experiments where conducted to validate the hybridQ metaheuristic. The experiments uses genetic, algorithms, tabu search and simulated annealing.

In the first experiment, the signed-rank Wilcoxon non-parametrical procedure was used for comparing the results. The significant level adopted was 0.05. The procedure showed that there was a significant improvement in the results with the Hybrid Metaheuristic approach.

The second and third experiments ran for 17 generations. The experiments used an initial population of 4 individuals by metaheuristic. All tests in the experiment were conducted without the need of a tester,

automating the execution of stress tests with the JMeter tool. In both experiments the hybridQ metaheuristic returned individuals with higher fitness scores. However, the Hybrid metaheuristic made twice as many requests than Tabu Search to overcome it. The SA algorithm obtained the worst fitness values. The algorithm initially used a scenario with an antipattern and found neighbors that still using the antipatterns over the 17 generations of the experiment.

In the second experiment the metaheuristics converged to scenarios with an happy path, excluding the scenarios with the use of an antipatterns.The first individual has 153 users on Happy Scenario 2, 16 users on Happy Scenario 1 and a response time of 13 seconds. None of the four best individuals has one of the antipatterns used in the experiment.

In the third experiment, the metaheuristics converged to scenarios with an happy path and Tower Babel antipattern, excluding the scenarios with Unbalanced Processing antipattern. The individual with best fitness value has 121 users on Happy Scenario 2, 171 users on Happy Scenario 1 and a response time of 11 seconds. None of the four best individuals has one of the antipatterns used in the experiment.

In the fourth experiment, the whole process of stress and performance tests, which took 3 days and about 1800 executions, was carried out without the need for monitoring by a test designer. The tool automatically selected the next scenarios to be run up to the limit of six generations previously established. The small number of samples of the experiment is insufficient to give a statistical significance to the results of the Wilcoxon procedure. However, it is noted that, in four of six generations, the collaborative approach presented the best values. The experiment succeeded in finding 29 individuals whose maximum time expected by the application was obtained.

7.2 Open Issues and future works

There is a range of future improvements in the proposed approach. Also as a typical search strategy, it is difficult to ensure that the execution times generated in the experiments represents global optimum. More experimentation is also required to determine the most appropriate and robust parameters. Lastly, there is a need for an adequate termination criterion to stop the search process.

Among the future works of the research, the use of new combinatorial optimization algorithms such as very large-scale neighborhood search is one that we can highlight.

...

...

Bibliography

[1] Model-based generation of testbeds for web services. pages 266–282, 2008.

[2] Wasif Afzal, Richard Torkar, and Robert Feldt. A systematic review of search-based testing for non-functional system properties. volume 51, pages 957–976. Elsevier B.V., 2009.

[3] Jarmo T. JT Alander, Timo Mantere, and Pekka Turunen. Genetic Algorithm Based Software Testing. In *Neural Nets and Genetic Algorithms*, 1998.

[4] Stefano D I Alesio, Lionel C Briand, Shiva Nejati, and Arnaud Gotlieb. Combining Genetic Algorithms and Constraint Programming. volume 25, 2015.

[5] Aldeida Aleti, I. Moser, and Lars Grunske. Analysing the fitness landscape of search-based software testing problems. *Automated Software Engineering*, pages 1–19, 2016.

[6] Saswat Anand, Edmund K. Burke, Tsong Yueh Chen, John Clark, Myra B. Cohen, Wolfgang Grieskamp, Mark Harman, Mary Jean Harrold, and Phil McMinn. An orchestrated survey of methodologies for automated software test case generation. volume 86, pages 1978–2001, 2013.

[7] Marco Anisetti, Claudio A. Ardagna, Ernesto Damiani, and Francesco Saonara. A test-based security certification scheme for web services. *ACM Transactions on the Web*, 7(2):1–41, 2013.

[8] Dejanira Araiza-Illan, Anthony G. Pipe, and Kerstin Eder. Model-based Test Generation for Robotic Software: Automata versus Belief-Desire-Intention Agents. pages 1–16, 2016.

[9] Alessandro Oliveira Arantes, Valdivino Alexandre De Santiago, Nandamudi Lankalapalli Vijaykumar, and Erica Ferreira De Souza. Tool support for generating model-based test cases via web. *International Journal of Web Engineering and Technology*, 9(1):62–96, 2014.

[10] Paolo Arcaini, Angelo Gargantini, and Elvinia Riccobene. Improving Model-based Test Generation by Model Decomposition. *Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering*, pages 119–130, 2015.

[11] Davide Arcelli, Vittorio Cortellessa, and Catia Trubiani. Antipattern-Based Model Refactoring for Software Performance Improvement. pages 33–42, 2012.

[12] Muhammad Arslan, Usman Qamar, Shoaib Hassan, and Sara Ayub. Automatic performance analysis of cloud based load testing of web-application & its comparison with traditional load testing. *2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, 2015-Novem(September):140–144, 2015.

[13] A. Avritzer and E.J. Weyuker. The automatic generation of load test suites and the assessment of the resulting software. *IEEE Transactions on Software Engineering*, 21(9):705–716, 1995.

[14] Alberto Avritzer and Brian Larson. Load Testing Software Using Deterministic State Testing. ISSTA '93, pages 82–88, New York, NY, USA, 1993. ACM.

[15] Alberto Avritzer and EJ Weyuker. Generating test suites for software load testing. pages 44–57, New York, New York, USA, 1994. ACM Press.

[16] Xiaoying Bai, Muyang Li, Bin Chen, Wei Tek Tsai, and Jerry Gao. Cloud testing tools. In *Proceedings - 6th IEEE International Symposium on Service-Oriented System Engineering, SOSE 2011*, pages 1–12, 2011.

[17] Scott Barber. User Community Modeling Language (UCML ™) v1 . 1 for Performance Test Workloads UCML ™ Overview. pages 1–9, 1999.

[18] Marcelo De Barros and Jing Shiau. Web services wind tunnel: On performance testing large-scale stateful web services. 2007.

[19] Mohamad Bayan and João W. Cangussu. Automatic feedback, control-based, stress and load testing. *Proceedings of the 2008 ACM symposium on Applied computing - SAC '08*, page 661, 2008.

[20] Francesco Bianchi, Alessandro Margara, and Mauro Pezze. A Survey of Recent Trends in Testing Concurrent Software Systems. *IEEE Transactions on Software Engineering*, 5589(c):1–1, 2017.

[21] C. Blum and A. Roli. Metaheuristics in combinatorial optimization: overview and conceptual comparison. *ACM Computing Surveys*, 35(3):189–213, 2003.

[22] Lionel C. Briand, Yvan Labiche, and Marwa Shousha. Stress testing real-time systems with genetic algorithms. page 1021, 2005.

[23] William H Brown, Raphael C Malveau, Hays W McCormick, and Thomas J Mowbray. *AntiPatterns: refactoring software, architectures, and projects in crisis*. John Wiley & Sons, Inc., 1998.

[24] Didier Buchs, Levi Lucio, and Ang Chen. Model checking techniques for test generation from business process models. *Reliable Software Technologies–Ada-Europe 2009*, pages 59–74, 2009.

[25] Yuhong Cai, John Grundy, and John Hosking. Synthesizing client load models for performance engineering via web crawling. *Proceedings of the twenty-second IEEE/ACM international conference on Automated software engineering - ASE '07*, page 353, 2007.

[26] Gerardo Canfora, Massimiliano Di Penta, Raffaele Esposito, and Maria Luisa Villani. An approach for QoS-aware service composition based on genetic algorithms. 2005.

[27] Microsoft Corporation. Performance Testing Guidance for Web Applications, November 2007.

[28] Vittorio Cortellessa and Laurento Frittella. A Framework for Automated Generation of Architectural Feedback from Software Performance Analysis. pages 171–185, 2007.

[29] MB da Silveira, EM Rodrigues, and AF Zorzo. Generation of Scripts for Performance Testing Based on UML Models. 2011.

[30] Leffingwell Dean and Widrig Don. Managing software requirements: A use case approach, 2003.

[31] Kalyanmoy Deb. *Multi-objective optimization using evolutionary algorithms*, volume 16. John Wiley & Sons, 2001.

[32] Kalyanmoy Deb, M. Mohan, and S. Mishra. Evaluating the epislon-domination based multiobjective evolutionary algorithm for a quick computation of Pareto-optimal solutions. *Evolutionary Computation Journal*, 13(4):501–525, 2005.

[33] S Di Alesio, S Nejati, L Briand, and A Gotlieb. Stress testing of task deadlines: A constraint programming approach. pages 158–167, 2013.

[34] Stefano Di Alesio, Shiva Nejati, Lionel Briand, and Arnaud Gotlieb. Worst-Case Scheduling of Software Tasks – A Constraint Optimization Model to Support Performance Testing. pages 813–830, 2014.

[35] Giuseppe a. Di Lucca and Anna Rita Fasolino. Testing Web-based applications: The state of the art and future trends. volume 48, pages 1172–1186, 2006.

[36] A. Di Pietro, L. While, and L. Barone. Applying evolutionary algorithms to problems with noisy, time-consuming fitness functions. *Proceedings of the 2004 Congress on Evolutionary Computation*, pages 1254–1261, 2004.

[37] D. Draheim, J. Grundy, J. Hosking, C. Lutteroth, and G. Weber. Realistic load testing of Web applications. In *Conference on Software Maintenance and Reengineering (CSMR'06)*, 2006.

[38] Elfriede Dustin, Jeff Rashka, and John Paul. *Automated Software Testing: Introduction, Management, and Performance*. 1999.

[39] Eduard Paul Enoiu, Daniel Sundmark, and Paul Pettersson. Model-based test suite generation for function block diagrams using the UPPAAL model checker. *Proceedings - IEEE 6th International Conference on Software Testing, Verification and Validation Workshops, ICSTW 2013*, pages 158–167, 2013.

[40] Bayo Erinle. *Performance Testing With JMeter 2.9*. 2013.

[41] Ling Fang, Takashi Kitamura, Thi Bich Ngoc Do, and Hitoshi Ohsaki. Formal model-based test for AUTOSAR multicore RTOS. *Proceedings - IEEE 5th International Conference on Software Testing, Verification and Validation, ICST 2012*, pages 251–259, 2012.

[42] Dror G Feitelson. *Workload Modeling for Computer Systems Performance Evaluation*. Cambridge University Press, 2013.

[43] Dharmalingam Ganesan, Mikael Lindvall, Stefan Hafsteinsson, Rance Cleaveland, Susanne L. Strege, and Walter Moleski. Experience Report: Model-Based Test Automation of a Concurrent Flight Software Bus. *Proceedings - International Symposium on Software Reliability Engineering, ISSRE*, pages 445–454, 2016.

[44] Vahid Garousi. Traffic-aware Stress Testing of Distributed Real-Time Systems based on UML Models using Genetic Algorithms. Number August, 2006.

[45] Vahid Garousi. Empirical analysis of a genetic algorithm-based stress test technique. page 1743, 2008.

[46] Vahid Garousi. A Genetic Algorithm-Based Stress Test Requirements Generator Tool and Its Empirical Evaluation. volume 36, pages 778–797, November 2010.

[47] Gregory Gay. Challenges in Using Search-Based Test Generation to Identify Real Faults in Mockito. pages 1–6.

[48] Gregory Gay, Sanjai Rayadurgam, and Mats P.E. Heimdahl. Automated Steering of Model-Based Test Oracles to Admit Real Program Behaviors. *IEEE Transactions on Software Engineering*, PP(99), 2016.

[49] H Giese, J Graf, and G Wirtz. Seamless visual object-oriented behavior modeling for distributed software systems. *Visual Languages, 1999. Proceedings. 1999 IEEE Symposium on*, pages 156–199, 1999.

[50] Fred Glover and Rafael Martí. Tabu Search. pages 1–16, 1986.

[51] N. Gois, P. Porfirio, A. Coelho, and T. Barbosa. Improving Stress Search Based Testing using a Hybrid Metaheuristic Approach. In *Proceedings of the 2016 Latin American Computing Conference (CLEI)*, pages 718–728, 2016.

[52] Marcelo Canário Gonçalves. Um Processo de Inferência de Desempenho para Apoiar o Planejamento da Capacidade de Aplicações na Nuvem Um Processo de Inferência de Desempenho para Apoiar o Planejamento da Capacidade de Aplicações na Nuvem. 2014.

[53] Mark Grechanik, Chen Fu, and Qing Xie. Automatically finding performance problems with feedback-directed learning software testing. pages 156–166. Ieee, June 2012.

[54] Amy Greenwald, Keith Hall, and R Serrano. Correlated Q-learning. Number 3, pages 84–89, 2003.

[55] Hg Gross, Bryan F Jones, and David E Eyres. Structural performance measure of evolutionary testing applied to worst-case timing of real-time systems. volume 147, pages 25–30, 2000.

[56] Emily H. Halili. *Apache JMeter: A practical beginner's guide to automated testing and performance measurement for your websites.* 2008.

[57] Mark Harman, Yue Jia, and Yuanyuan Zhang. Achievements , open problems and challenges for search based software testing. *8th IEEE International Conference on Software Testing, Verification and Validation (ICST)*, (Icst), 2015.

[58] Mark Harman and Phil McMinn. A theoretical and empirical study of search-based testing: Local, global, and hybrid search. volume 36, pages 226–247, 2010.

[59] Anders Hessel. *Model-Based Test Case Generation for Real-Time Systems*. 2007.

[60] Robert M Hierons, Kirill Bogdanov, Jonathan P Bowen, Rance Cleaveland, John Derrick, Jeremy Dick, Marian Gheorghe, Mark Harman, Kalpesh Kapoor, Paul Krause, Gerald Lüttgen, Anthony J H Simons, Sergiy Vilkomir, Martin R Woodward, and Hussein Zedan. Using formal specifications to support testing. volume 41, pages 1–76, 2009.

[61] Tzung-Pei Hong, Hong-Shung Wang, and Wei-Chou Chen. Simultaneously applying multiple mutation operators in genetic algorithms. volume 6, pages 439–455. Springer, 2000.

[62] T Illes, A Herrmann, B Paech, and J Rückert. Criteria for Software Testing Tool Evaluation – A Task Oriented View. 2:213–222, 2005.

[63] B. Jones J. Wegener, K. Grimm, M. Grochtmann, H. Sthamer. Systematic testing of real-time systems. 1996.

[64] Gerrit Janssens and José Pangilinan. Multiple criteria performance analysis of non-dominated sets obtained by multi-objective evolutionary algorithms for optimisation. *Artificial Intelligence Applications and Innovations*, pages 94–103, 2010.

[65] So-young Jeong, Cheol-jung Yoo, Hye-min Noh, and Corresponding Author. State Transition Based Test Model and Test Case Generation Technique for Embedded System: An Empirical Approach. 10(11):233–254, 2016.

[66] ZM Jiang. *Automated analysis of load testing results*. PhD thesis, 2010.

[67] Gyu Baek Kim. A method of generating massive virtual clients and model-based performance test. *Proceedings - International Conference on Quality Software*, 2005:250–254, 2005.

[68] Sandhya Kiran, Akshyansu Mohapatra, and Rajashekara Swamy. Experiences in performance testing of web applications with Unified Authentication platform using Jmeter. *2nd International Symposium on Technology Management and Emerging Technologies, ISTMET 2015 - Proceeding*, pages 74–78, 2015.

[69] Barbara Kitchenham and S Charters. Guidelines for performing Systematic Literature Reviews in Software Engineering. *Engineering*, 2:1051, 2007.

[70] Renaud Lacour, Kathrin Klamroth, and Carlos M. Fonseca. A box decomposition algorithm to compute the hypervolume indicator. *Computers and Operations Research*, pages 1–21, 2015.

[71] Chris Lenz, J Chimiak-Opoka, and Ruth Breu. Model-Driven Testing of SOA-based Software. ... *of the SEMSOA Workshop on Software ...*, 2007.

[72] William E. Lewis, David Dobbs, and Gunasekaran Veerapillai. *Software testing and continuous quality improvement*. 2005.

[73] Qi Luo, Aswathy Nair, Mark Grechanik, and Denys Poshyvanyk. FOREPOST: finding performance problems automatically with feedback-directed learning software testing. pages 1–51, 2015.

[74] Thomas Stützle Manuel López-Ibáñez, Jérémie Dubois-Lacoste, Leslie Pérez Cáceres, Mauro Birattari. The irace package: Iterated racing for automatic algorithm configuration. *Operations Research Perspectives*, 3:43–58, 2016.

[75] Raluca Marinescu, Cristina Seceleanu, Hèléne Le Guen, and Paul Pettersson. *A Research Overview of Tool-Supported Model-based Testing of Requirements-based Designs*, volume 98. 2015.

[76] Marcelo Marinho, Suzana Sampaio, Telma Lima, and Hermano de Moura. A Systematic Review of Uncertainties in Software Project Management. *International Journal of Software Engineering & Applications*, 5(6):1–21, 2014.

[77] Alexander Pretschner Mark Utting and Bruno Legeard. A taxonomy of model-based testing approaches. volume 24, pages 297–312, 2012.

[78] Rui A. Matnei Filho and Silvia R. Vergilio. A multi-objective test data generation approach for mutation testing of feature models. *Journal of Software Engineering Research and Development*, 4(1):4, 2016.

[79] Kai-Steffen Hielscher Matthias Beyer, Winfried Dulz. Performance Issues in Statistical Testing. (April 2006), 2014.

[80] Daniel A Menascé and George Mason. TPC-W : A Benchmark for E-commerce. Number June, pages 1–6, 2002.

[81] Petros Nicopolitidis Mohammad S. Obaidat and Faouzi Zarai. *Modeling and Simulation of Computer Networks and Systems Methodologies and Applications*.

[82] Ian Molyneaux. *The Art of Application Performance Testing: Help for Programmers and Quality Assurance*. "O'Reilly Media, Inc.", 1st edition, January 2009.

[83] Marco Moscher and Konrad Fögen. Facing synthetic workload generation as part of performance testing–a tools approach. *Full-scale Software Engineering/The Art of Software Testing*, page 38, 2017.

[84] F. Mueller and J. Wegener. A comparison of static analysis and evolutionary testing for the verification of timing constraints. 1998.

[85] S Nachiyappan and S Justus. Cloud Testing Tools and Its Challenges : A Comparative Study. *Procedia - Procedia Computer Science*, 50:482–489, 2015.

[86] Massimiliano Di Penta, Gerardo Canfora, and Gianpiero Esposito. Search-based testing of service level agreements. In *Proceedings of the 9th annual conference on Genetic and evolutionary computation*, pages 1090–1097, 2007.

[87] William E. Perry. *Effective methods for software testing*. 2004.

[88] Hartmut Pohlheim, Mirko Conrad, and Arne Griep. Evolutionary Safety Testing of Embedded Control Software by Automatically Generating Compact Test Data Sequences. Number 724, pages 804—-814, 2005.

[89] Jakob Puchinger and R Raidl. Combining Metaheuristics and Exact Algorithms in Combinatorial Optimization : A Survey and Classification. volume 3562, pages 41–53, 2005.

[90] P. Puschner and R. Nossal. Testing the results of static worst-case execution-time analysis. 1998.

[91] Günther R Raidl, Jakob Puchinger, and Christian Blum. Metaheuristic hybrids. In *Handbook of metaheuristics*, pages 469–496. Springer, 2010.

[92] R Raidl. A Unified View on Hybrid Metaheuristics. pages 1–12, 2006.

[93] Pratyusha Rakshit, Amit Konar, and Swagatam Das. Noisy evolutionary optimization algorithms – A comprehensive survey. *Swarm and Evolutionary Computation*, 33:18–45, 2017.

[94] Irum Rauf, Muhammad Zohaib Z Iqbal, and Zafar I Malik. Model Based Testing of Web Service Composition Using UML Profile. *2nd Workshop on Model-based Testing in Practice, MOTIP 2009*, 2009.

[95] Elder M Rodrigues, Rodrigo S Saad, Flavio M Oliveira, Leandro T Costa, Maicon Bernardino, and Avelino F Zorzo. Evaluating Capture and Replay and Model-based Performance Testing Tools: An Empirical Comparison. *Proceedings of the 8th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, pages 9:1—-9:8, 2014.

[96] Jose Lorenzo San Miguel and Shingo Takada. GUI and Usage Model-based Test Case Generation for Android Applications with Change Analysis. *Proceedings of the 1st International Workshop on Mobile Development*, pages 43–44, 2016.

[97] Corey Sandler, Tom Badgett, and TM Thomas. The Art of Software Testing. page 200. John Wiley & Sons, September 2004.

[98] Christopher Schaefer, Hyunsook Do, and Brian M. Slator. Crushinator: A framework towards game-independent testing. *2013 28th IEEE/ACM International Conference on Automated Software Engineering, ASE 2013 - Proceedings*, pages 726–729, 2013.

[99] Marwa Shousha. *Performance Stress Testing of Real-Time Systems Using Genetic Algorithms*. PhD thesis, Carleton University Ottawa, 2003.

[100] Florian Siegmund, Amos H C Ng, and Kalyanmoy Deb. A comparative study of dynamic resampling strategies for guided Evolutionary Multi-objective Optimization. *2013 IEEE Congress on Evolutionary Computation, CEC 2013*, (2013008):1826–1835, 2013.

[101] Connie U. Smith and Lloyd G. Williams. Software performance antipatterns. pages 127–136, 2000.

[102] Connie U Smith and Lloyd G Williams. More New Software Performance AntiPatterns: EvenMore Ways to Shoot Yourself in the Foot. pages 717–725, 2003.

[103] C.U. Smith and L.G. Williams. Software Performance AntiPatterns; Common Performance Problems and their Solutions. volume 2, pages 797–806, 2002.

[104] Adepu Sridhar, D. Srinivasulu, and Durga Prasad Mohapatra. Model-based test-case generation for Simulink/Stateflow using dependency graph approach. *Proceedings of the 2013 3rd IEEE International Advance Computing Conference, IACC 2013*, pages 1414–1419, 2013.

[105] Michael O Sullivan, Siegfried Vössner, Joachim Wegener, and Daimler-benz Ag. Testing Temporal Correctness of Real-Time Systems — A New Approach Using Genetic Algorithms and Cluster Analysis —. pages 1–20, 1998.

[106] Richard S. Sutton and Andrew G. Barto. Reinforcement learning. volume 3, page 322, 2012.

[107] El-Ghazali Talbi. *Metaheuristics: from design to implementation*, volume 74. John Wiley & Sons, 2009.

[108] El-Ghazali Talbi. *Metaheuristics: From Design to Implementation*, volume 53. 2013.

[109] Tommi Tervonen and United Kingdom. Evaluation of multi-objective optimization approaches for solving green supply chain design problems : Evaluation of multi-objective optimization approaches for solving green supply chain design problems. (April), 2017.

[110] N J Tracey, J a Clark, and K C Mander. Automated Programme Flaw Finding using Simulated Annealing. 1998.

[111] Nigel James Tracey. *A search-based automated test-data generation framework for safety-critical software*. PhD thesis, Citeseer, 2000.

[112] G Trent and M Sake. WebSTONE: The first generation in {HTTP} server benchmarking. 1995.

[113] Catia Trubiani. Automated generation of architectural feedback from software performance analysis results Catia Trubiani. *Language*, 2011.

[114] Catia Trubiani. PhD Thesis in Computer Science Automated generation of architectural feedback from software performance analysis results Catia Trubiani. *Language*, 2011.

[115] Mark Utting and Bruno Legeard. *Practical model-based testing: a tools approach*. Morgan Kaufmann, 2010.

[116] Christian Vogele, André van Hoorn, Eike Schulz, Wilhelm Hasselbring, and Helmut Krcmar. WESSBAS: extraction of probabilistic workload specifications for load testing and performance prediction???a model-driven approach for session-based application systems. Number October, pages 1–35. Springer Berlin Heidelberg, 2016.

[117] Xingen Wang, Bo Zhou, and Wei Li. Model-based load testing of web applications. volume 36, pages 74–86, 2013.

[118] J Wegener and M Grochtmann. Verifying timing constraints of real-time systems by means of evolutionary testing. volume 15, pages 275–298, 1998.

[119] Joachim Wegener, Harmen Sthamer, Bryan F Jones, and David E Eyres. Testing real-time systems using genetic algorithms. volume 6, pages 127–135, 1997.

[120] Harmen Wegener, Joachim and Pitschinetz, Roman and Sthamer. Automated Testing of Real-Time Tasks. 2000.

[121] Alexander Wert, Jens Happe, and Lucia Happe. Supporting swift reaction: Automatically uncovering performance problems by systematic experiments. Number May, pages 552–561, 2013.

[122] Alexander Wert, Marius Oehler, Christoph Heger, and Roozbeh Farahbod. Automatic detection of performance anti-patterns in inter-component communications. pages 3–12, 2014.

[123] S Wieczorek, A Stefanescu, and A Roth. Model-Driven Service Integration Testing - A Case Study. *Quality of Information and Communications Technology (QUATIC), 2010 Seventh International Conference on the*, pages 292–297, 2010.

[124] Lloyd G. Williams and Connie U. Smith. PASASM : A Method for the Performance Assessment of Software Architectures. *Proceedings of the third international workshop on Software and performance - WOSP '02*, (January 2002):179, 2002.

[125] Hongwei Zeng Xinying Cai. Model-based Test Generation for Software Product Line. 2007.

[126] Li Ye. Model-Based Testing Approach for Web Applications. (June), 2007.

[127] E. Zitzler and L. Thiele. Multiobjective evolutionary algorithms: a comparative case study\nand the strength Pareto approach. *IEEE Transactions on Evolutionary Computation*, 3(4):257–271, 1999.

[128] Eckart Zitzler, Marco Laumanns, and Lothar Thiele. SPEA2: Improving the Strength Pareto Evolutionary Algorithm. *Evolutionary Methods for Design Optimization and Control with Applications to Industrial Problems*, pages 95–100, 2001.

Appendix A

IAdapter

IAdapter is a JMeter plugin designed to perform search-based stress tests. Fig. A.1 presents the IAdapter Life Cycle. The main difference between IAdapter and JMeter tool is that the IAdapter provide an automated test execution where the new test scenarios are choosen by the test tool. In a test with JMeter, the tests scenarios are usually chosen by a test designer.

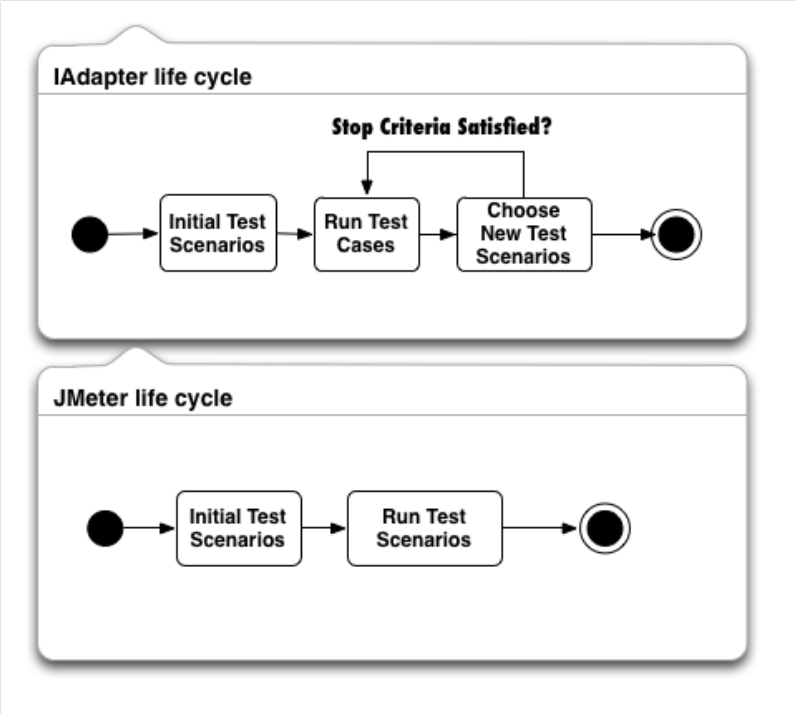


Figure A.1: IAdapter life cycle

A.1 IAdapter Visual Components

JMeter has components organized in a hierarchical manner. The IAdapter plugin provides three main components: WorkLoadThreadGroup, WorkLoadSaver, and WorkLoadController.

WorkLoadThreadGroup is a component that creates an initial population and configures the algorithms used in the IAdapter. Fig. A.2 presents the main screen of the WorkLoadThreadGroup component. The component has a name ❶, a set of configuration tabs ❷, a list of individuals by generation ❸, a button to generate an initial population ❹, and a button to export the results ❺. WorkLoadThreadGroup component uses the GeneticAlgorithm, TabuSearch and SimulateAnnealing classes. The WorkLoadSaver component is responsible for saving all data in the database. The operation of the component only requires its inclusion in the test script. WorkLoadController represents a scenario of the test. All actions necessary to test an application should be included in this component. All instances of the component require to be logged in the application under test and bring the application back to its original state.

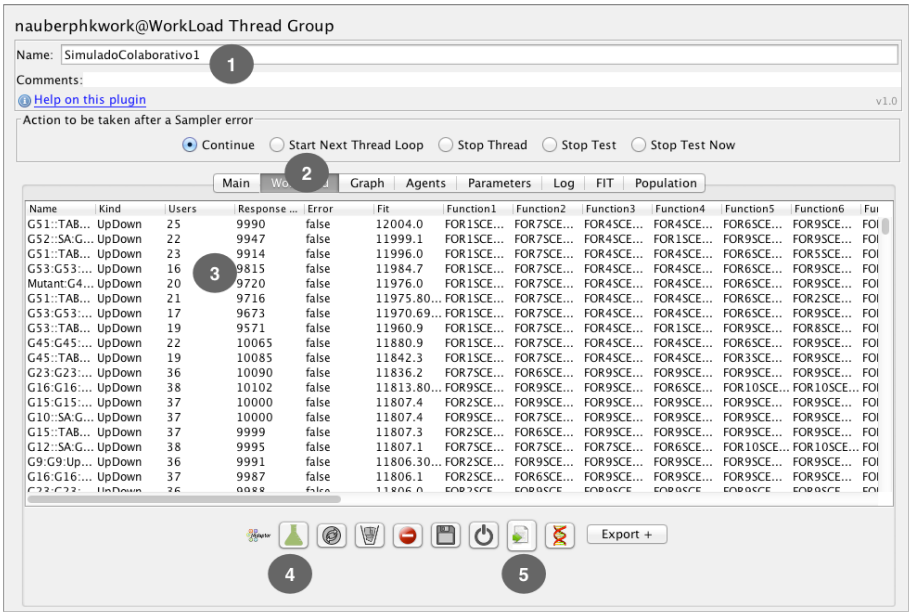


Figure A.2: WorkLoadThreadGroup component

A.2 The IAdapter architecture

In this section, We present the IAdapter main architecture. The testbed tool proposed consists of four main modules. Figure A.3 presents the main architecture of the solution proposed. The emulator module provides workloads to the Test module. The Test module uses a class loader to find all classes that extends AbstractAlgorithm in the classpath and run all workloads with each metaheuristic found. The Test Scenario library provides the scenario representation used by the metaheuristics and store the testbed results in a database.

The Operation services are responsible for finding neighbors of some workload provided as a parameter and perform crossover operations.

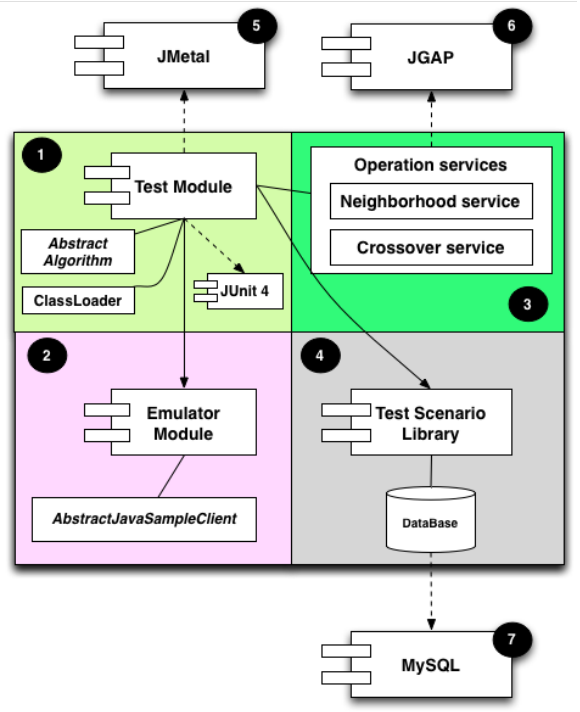


Figure A.3: IAdapter main architecture.

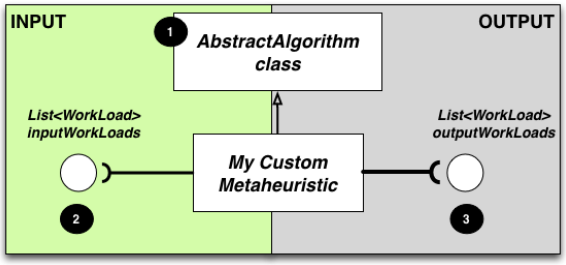


Figure A.4: Test Module class diagram.

A.2.1 Test Module

The Test Module (Figure A.3 -❶) is responsible for for the loading of all classes that extend AbstractAlgorithm in the classpath and perform the tests under the application. Figure A.4 shows the class diagram for custom and provided heuristics. All heuristic classes extends the class AbstractAlgorithm. The heuristics receives as input a list of workloads (Figure A.4 -❷) and must return a list of output workloads (the individuals selected for the next generation) (Figure A.4 -❸). Each workload represent an individual in the search space. There is an abstract class inherited from the AbstractAlgorithm named MultiObjective Algorithm, this class is used by multi-objective metaheuristics, which have a different execution flow in the system.

Figure. A.5 presents the Test Module life cycle. Given an initial population (Figure A.5 -❶), a metaheurist select a new set of workloads based on an objective function (Figure A.5 -❷). The choosen metaheurist generate a new set of individuals based on crossover or neighborhood operators (Figure A.5 -❸). JMeterEngine run each workload (Figure A.5 -❺) and the choosen metaheuristic obtain a fitness value for each workload based on some objective function (Figure A.5 -❻). Each Metaheuristic could define your own objective function. After all these steps the cycle begins until the maximum number of generations it is reached (Figure A.5 -❼).

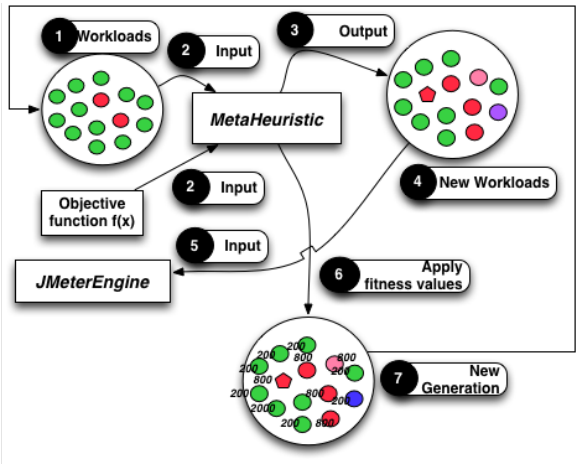


Figure A.5: Test module life cycle.

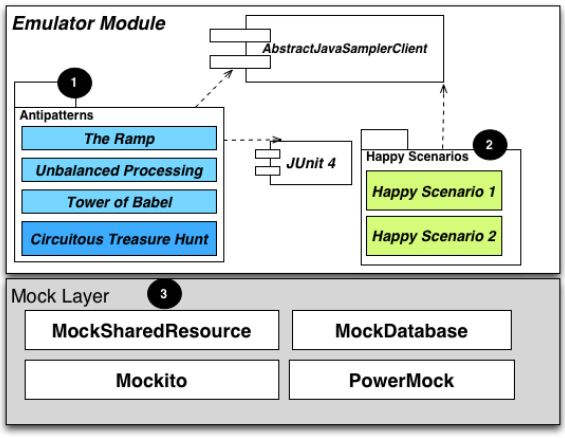


Figure A.6: Emulator module

The WorkLoadThreadGroup class is responsible for start all threads that go simulate the users in the jmeter engine. Fig A.7 presents the WorkLoadThreadGroup life cycle. First, an instance of the class waits for the user to start the test (Stopped State - Fig. A.7 - ❶). Once the test is started, the start method is called (Fig. A.7 - ❷) and the instance goes to Running state (Fig. A.7 - ❸). After running all threads, the class goes to the finished state (Fig. A.7 - ❺).

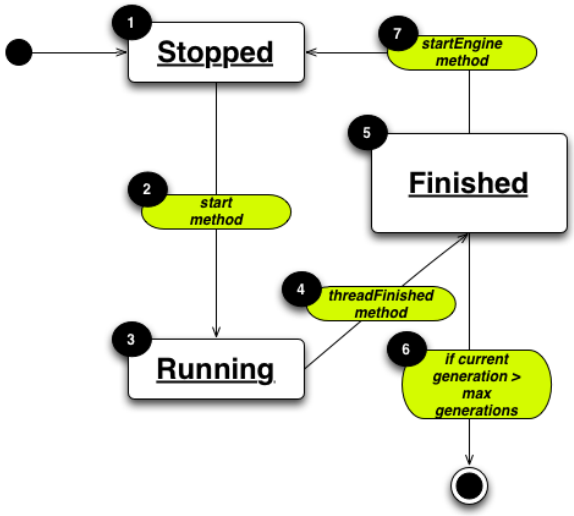


Figure A.7: WorkLoadThreadGroup class life cycle.

Figure A.8 presents the start method sequence diagram. The WorkLoad Thread Group start method is responsible for load the multi-objective weights, synchronize all threads and get the new workloads from database. Figure A.9 presents the threadFinish method, the method is triggered after the end of execution of

each Thread. The method synchronizes all threads (Fig. A.9-❸), waiting for the execution of all instances to finish, stores the results, and starts new tests until the end of all generations (Fig. A.9-❹ and ❺).

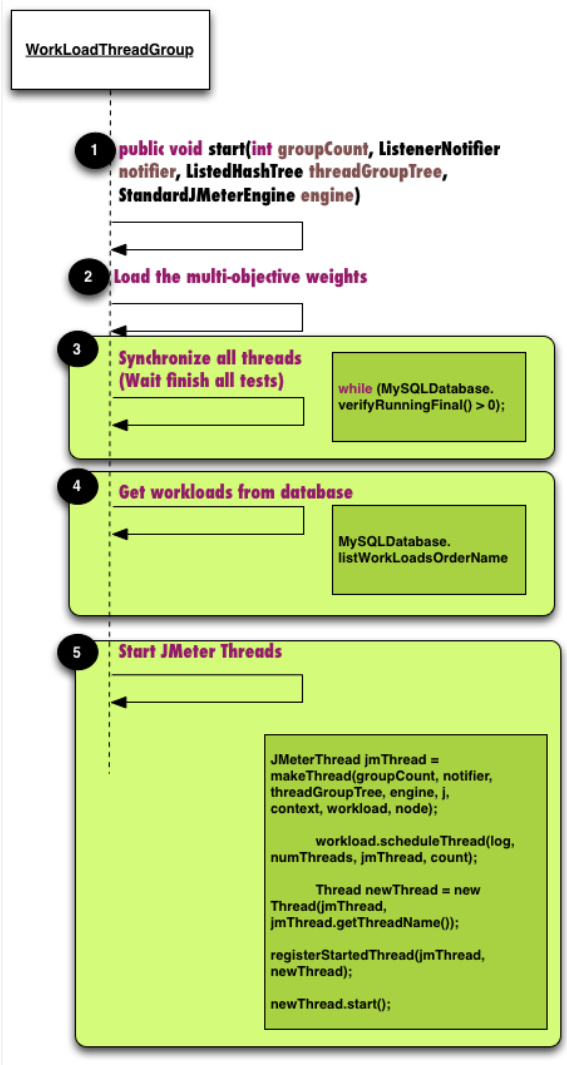


Figure A.8: WorkLoadThreadGroup start method.

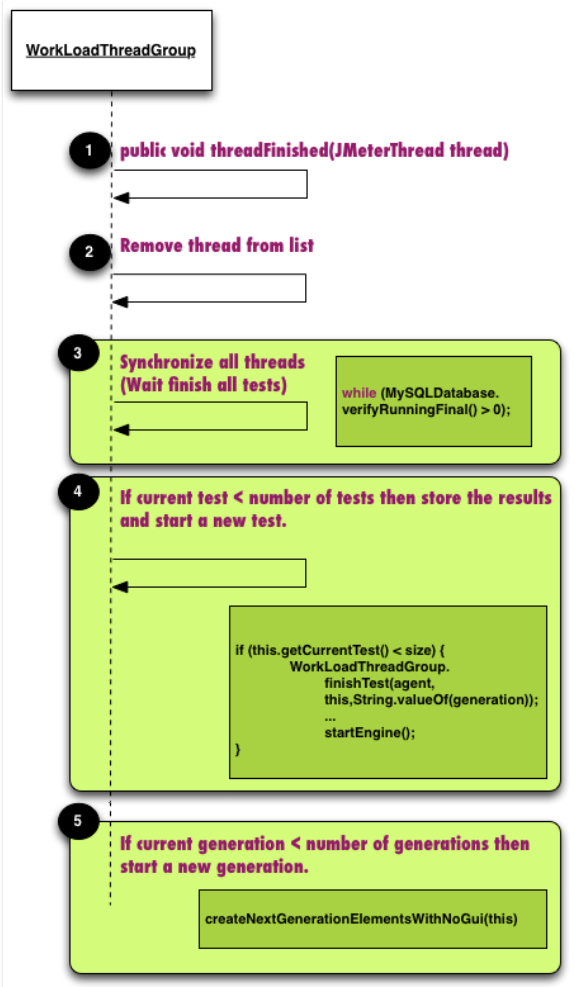


Figure A.9: WorkLoadThreadGroup threadFinished method.

A.2.2 Emulator Module

The Emulator Module is responsible for implementing and providing successful scenarios and the most common performance antipatterns (Figure A.3 -❷). All classes must extend the AbstractJavaSamplerClient class or use JUnit 4. The AbstractJavaSamplerClient class allows the creation of a JMeter Java Request. Figure A.11 presents the main features of the emulator module. The module implements 2 happy scenarios (Figure

A.11 -❷) and 4 antipatterns test scenarios (Figure A.11 -❶), in its first version. The Mock Layer provides emulated databases and components for the test scenarios. The Mock Layer use the Mockito and PowerMocks frameworks (Figure A.11 -❸).

A.2.3 Test scenario library

The representation of each individual is encapsulated in the WorkLoad class (Fig A.10). All test scenarios are referenced by string objects named *function1, function2, ..., function8, function9, and function10*.

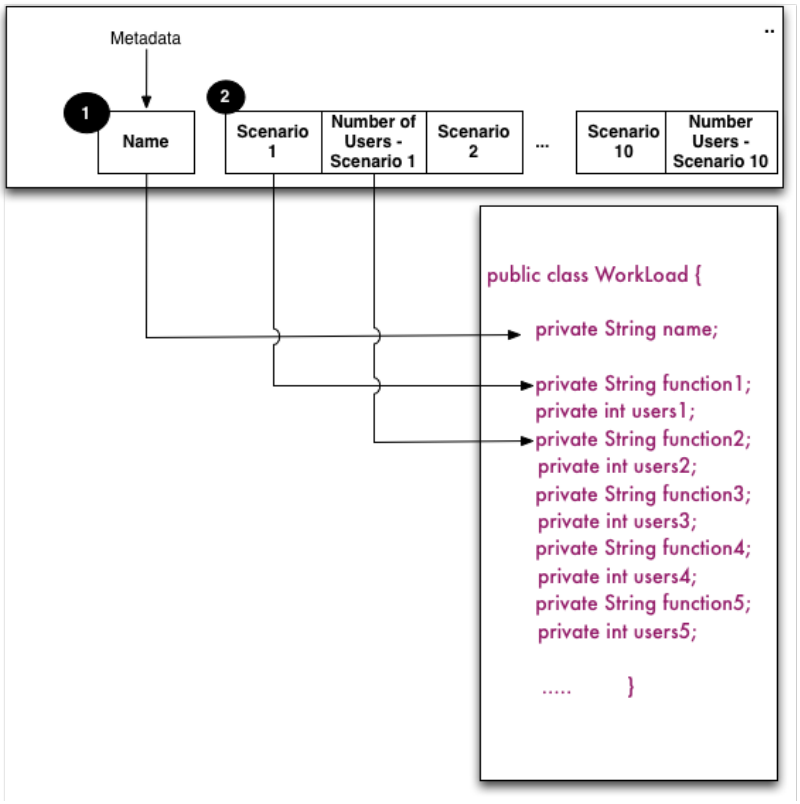


Figure A.10: WorkLoad class

A.2.4 Operation services

The services are responsible for performing some operations performed by metaheuristics.

A.2.5 External dependencies

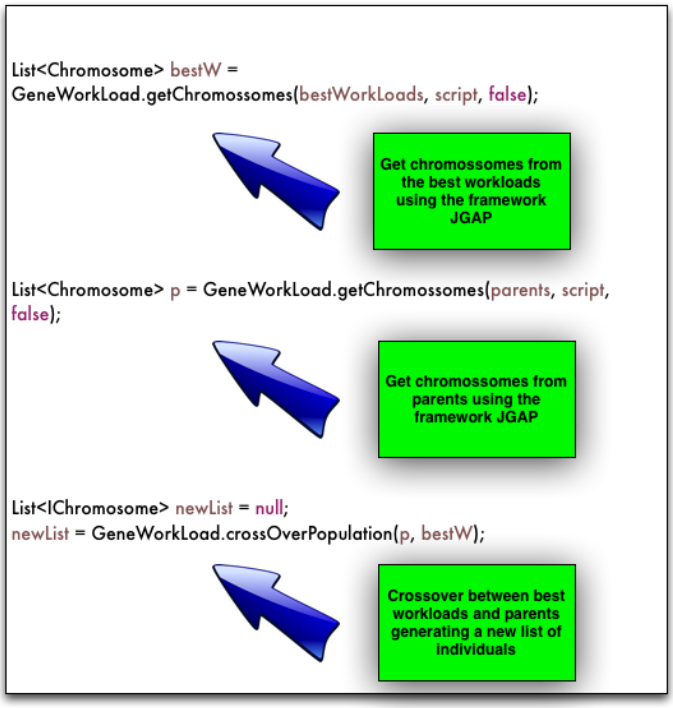


Figure A.11: WorkLoad class

Appendix B

Reinforcement Learning

Reinforcement learning (RL) refers to both a learning problem and a subfield of machine learning. As a learning problem, it refers to learning to control a system so as to maximize some numerical value which represents a long-term objective. The basic idea of Reinforcement learning is simply to capture the most important aspects of the real problem, facing a learning agent interacting with its environment to achieve a goal [106]. Reinforcement learning is learning what to do—how to map situations to actions—so as to maximize a numerical reward signal. The learner needs to discover which actions yield the most reward by trying them [106].

A typical setting where reinforcement learning operates is shown in Figure B.1: A controller receives the controlled system's state and a reward associated with the last state transition. It then calculates an action which is sent back to the system.

In Reinforcement Learning, an agent wanders in an unknown environment and tries to maximize its long term return by performing actions and receiving rewards. The challenge is to understand how a current action will affect future rewards. A good way to model this task is with Markov Decision Processes (MDP). Markov decision processes (MDPs) provide a mathematical framework for modeling decision making. In

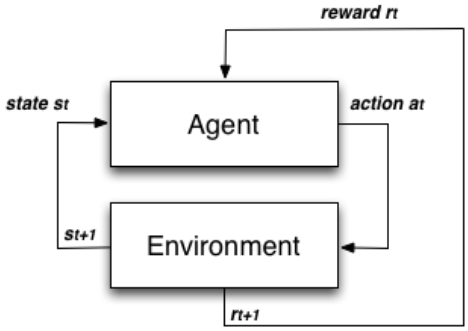


Figure B.1: Example of interation between some agent and the environment

Reinforcement Learning, all agents act in two phases: Exploration vs Explotation. In Exploration phase, the agents try to discover better action selections to improve its knowledge. In Exploitation phase, the agents try to maximize its reward, based on what is already know.

One of the challenges that arise from reinforcement learning is the trade-off between exploration and exploitation. To obtain a large reward, a reinforcement learning agent must prefer actions that it has tried in the past and found to be effective in producing reward. But to discover such actions, it has to try actions that it has not selected before. The agent has to exploit what it already knows in order to obtain a reward, but it also has to explore in order to make better action selections in the future.

Q-learning is a model-free reinforcement learning technique. Q-learning is a multiagent learning algorithm that learns equilibrium policies in Markov games, just as Q-learning learns to optimize policies in Markov decision processes [54].

Q-learning and related algorithms try to learn the optimal policy from its history of interaction with the environment. A history of an agent is a sequence of state-action-rewards.Where s_n is a state, a_n is an action and r_n is a reward:

$$< s_0, a_0, r_1, s_1, a_1, r_2, s_2, a_2, r_3, s_3, a_3, r_4, s_4.... >, \tag{B.1}$$

In Q-Learning, the system's objective is to learn a control policy $\pi = \sum_{n=0}^{\infty} \gamma^{\mathrm{n}} r_t + n$, where π is the discounted cumulative reward, γ is the discount rate (01) and r_t is the reward received after the execution of an action at time t. Figure B.2 shows the summary version of Q-Learning algorithm. The first step is to generate the initial state of the MDP. The second step is to choose the best action or a random action based on the reward, hence the actions with best rewards are chosen.

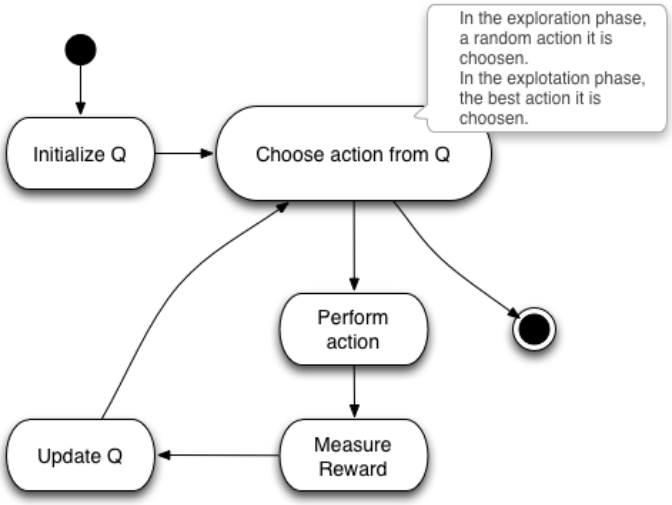


Figure B.2: Q Learning algorithm