

Improving Search-Based Stress Testing using Q-Learning and Hybrid Metaheuristic Approach

by

Francisco Nauber Bernardo Gois

Submitted to the Doutorado em Informatica Aplicada
in partial fulfillment of the requirements for the degree of

Doctor of Science in Applied Informatics

at the

Universidade de Fortaleza

February 2017

.....
D.Sc. Pedro Porfírio Muniz de Farias
Advisor - Universidade de Fortaleza

.....
D.Sc. Andre Luís Vasconcelos Coelho
Co-Advisor - Universidade de Fortaleza

.....
D.Sc. Pedro de Alcantara dos Santos Neto
External Examiner - Universidade Federal do Piauí

.....
Ph.D. Americo Tadeu Falcone Sampaio
Universidade de Fortaleza

Improving Search-Based Stress Testing using Q-Learning and Hybrid Metaheuristic Approach

by

Francisco Nauber Bernardo Gois

Submitted to the Doutorado em Informatica Aplicada
on February 16, 2017, in partial fulfillment of the
requirements for the degree of
Doctor of Science in Applied Informatics

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.

Thesis Supervisor: Pedro Porfirio Muniz de Farias
Title: Associate Professor

Acknowledgments

This is the acknowledgements section. You should replace this with your own acknowledgements.

Related Publications

The following publications are related to this thesis:

- **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 [60].

Contents

1	Introduction	17
1.1	Motivation	20
1.1.1	State of Research on the Search-Based Stress Testing	21
1.1.2	State of Industrial Practices on Stress Tests	21
1.2	Research Hypothesis	22
1.3	Thesis Contributions	22
1.3.1	Hybrid Metaheuristic Approach applied on Stress Testing	22
1.3.2	Q-Learning Metaheuristic Approach	23
1.3.3	IAdapter JMeter Plugin	23
1.3.4	IAdapter Testbed Tool	23
1.4	Thesis Organization	23
2	A Survey on Stress Testing Software Systems	24
2.1	Method	25
2.1.1	Planning a Systematic Review	26
2.1.2	Research Questions	27
2.1.3	Generation of search strategy	27
2.1.4	Study selection criteria and procedures for including and excluding primary studies	29
2.1.5	Data Synthesis	30
2.1.6	Stress Test Process	30
2.2	Research Question 1: How is a proper stress designed?	34
2.2.1	Model-based Stress Testing	37

2.2.2	Feedback-ORiEnted PerfOrmance Software Testing	41
2.3	Research Question 2: How is a stress test executed and automated?	43
2.3.1	Load Test Tools	44
2.4	Research Question 3: What are the main problems found by stress tests? . .	50
2.5	Research Question 4: How are the stress tests results analysed?	61
2.6	Conclusion	61
3	Search-Based Stress Testing	64
3.1	Introduction	64
3.2	Search-Based Testing	65
3.3	Non-functional Search-Based Testing	66
3.4	Search-Based Stress Testing	67
3.4.1	Search-Based Stress Tesing on Safety-critical systems	69
3.4.2	Search-Based Stress Testing on Industrial systems	72
4	Metaheuristics	74
4.1	Single-Solution Based Metaheuristics	75
4.1.1	Neighborhood	76
4.1.2	Simulated Annealing	77
4.1.3	Tabu Search	77
4.2	Population-based metaheuristics	78
4.2.1	Genetic Algorithms	79
4.3	Hybrid Metaheuristics	79
4.4	Multi-Objective Optimization	81
4.4.1	NSGA-II Multiobjective heuristics	82
5	Q-Learning	84
5.1	Reinforcement Learning	84
5.1.1	Markov decision processes	86
5.2	Q-Learning	87
5.3	GridWorld Example	88

5.4	Relationships Between Reinforcement Learning and Optimization	89
5.5	Software Testing with Reinforcement Learning	90
6	Improving Stress Search Based Testing using Q-Learning and Hybrid Metaheuristic Approach	92
6.1	Hybrid Approach	92
6.1.1	Representation	93
6.1.2	Initial population	94
6.1.3	Objective (fitness) function	95
6.2	Hybrid Metaheuristic with Q-Learning Approach	97
6.3	IAdapter	99
6.3.1	IAdapter Life Cycle	99
6.3.2	IAdapter Components	100
6.3.3	IAdapter Testbed Tool	101
6.3.4	Test Module	102
6.3.5	Emulator Module	103
6.3.6	Test Scenario Library	104
6.3.7	Operation services	105
7	Experiments	108
7.1	Emulated Class Test Experiment	108
7.2	Testbed Tool Experiments	111
7.2.1	The Ramp and Circuitous Treasure Hunt experiment	112
7.2.2	The Tower Babel and Unbalanced Processing experiment	118
7.3	Moodle Application Experiment	121
7.4	JPetStore Application Experiments	124
7.5	Experimental Results	127
7.5.1	Experiment Research Questions	128
7.5.2	Variables	128
7.5.3	Hypotheses	128
7.5.4	Results	129

7.5.5	Threats to validity	130
8	Conclusion	131
8.1	Achievements	132
8.2	Open Issues and future works	133
A	Tables	134
B	Figures	142

List of Figures

1-1	Possible test scenarios for a hypothetical application	18
1-2	Illustrative example showing how IAdapter should be used	19
1-3	Thesis Organization	23
2-1	Bubble chart for results of search with the keyword 'Load Testing'	26
2-2	Bubble chart for results of search with the keyword 'Load Testing'	29
2-3	Bubble chart for results of search with the keyword 'Stress Testing'	29
2-4	Word cloud for results of search with the keyword 'Load Testing'	30
2-5	Word cloud for results of search with the keyword 'Stress Testing'	30
2-6	Summary of type of research paper by year	31
2-7	Load, Performance and Stress Test Process [76][47]	33
2-8	Workload modeling based on statistical data [42]	35
2-9	Workload modeling based on the generative model [42]	36
2-10	User community modeling language [129]	39
2-11	Stochastic Formcharts Example [43] [129]	39
2-12	Example of a Customer Behavior Model Graph (CBMG) [90] [76] [91]	40
2-13	Model-based stress test methodology	40
2-14	Exemplary workload model	42
2-15	The architecture and workflow of FOREPOST	43
2-16	TPC-W architecture [91] [90]	46
2-17	Load Runner Scripting	47
2-18	Group I- Benchmarks that model the client and server paradigm in Web context. [91]	50

2-19	Group II- Software products to evaluate performance and functionality[91].	50
2-20	Group III- Testing tools and other approaches for traffic generation based on HTTP traces [91].	51
2-21	Symptoms of known performance problems [134].	52
2-22	The God class[134].	54
2-23	The God class[125].	55
2-24	Unbalanced Processing sample [134].	56
2-25	Pipe and Filter sample [125]	56
2-26	Extensive Processing sample [125].	56
2-27	Circuitous Treasure Hunt sample [125]	57
2-31	Excessive Dynamic Allocation.	57
2-32	Traffic Jam Response Time [125].	57
2-28	Empty Semi Trucks sample [125].	58
2-29	Tower of Babel sample [125]	59
2-33	The Ramp sample [125].	59
2-34	More is Less sample [125].	59
2-30	One-Lane Bridge sample [125].	60
2-35	Results summary of the systematic review.	63
3-1	Evolutionary Algorithm Search Based Test Cycle[19].	66
3-2	Range of metaheuristics by Type of non-functional Search Based Test[6].	68
3-3	Genome representation [98].	72
4-1	Classical optimization methods [121].	74
4-2	Main principles of single-based metaheuristics.	76
4-3	An example of neighborhood for a permutation [121].	76
4-4	Categories of metaheuristic combinations [102]	80
4-5	An minimization pareto front example	81
5-1	Example of interaction between some agent and the environment	85
5-2	Example of a simple MDP with three states and two actions	86

5-3	Q Learning algorithm	88
5-4	GridWorld - initial and final stage on exploration phase	89
6-1	Use of the algorithms independently	93
6-2	Use of the algorithms collaboratively	94
6-3	Solution representation and crossover example	95
6-4	Tabu search and simulated annealing neighbor strategy	96
6-5	Hybrid Metaheuristic with Q-Learning Approach	97
6-6	Markov Decision Process used by HybridQ	98
6-7	HybridQ NeighborHood Service	100
6-8	IAdapter architecture	101
6-9	IAdapter life cycle	102
6-10	WorkLoadThreadGroup component	103
6-11	Testbed main architecture.	104
6-12	Test Module class diagram.	105
6-13	Test module life cycle.	106
6-14	Emulator module	107
7-1	Best results obtained in 27 generations	110
7-2	fitness value obtained by Search Method	113
7-3	Number of requests by Search Method	113
7-4	Average, median, maximum and minimal fitness value by Search Method .	114
7-5	fitness value by generation	114
7-6	Density graph of number of users by fitness value	115
7-7	Response time by number of users of individuals with Happy Scenario 1 and Happy Scenario 2	116
7-8	Response time by number of users of individuals with the Ramp and Cir- cuitous Treasure antipatterns	116
7-9	Markov decision process of experiment with Circuitous Treasure and The Ramp antipatterns	117
7-10	Fitness value obtained by Search Method	118

7-11	Number of requests by Search Method	119
7-12	Fitness value by generation in all tests	119
7-13	Response time by generation in all tests scenarios	119
7-14	Finesse value by generation in all tests	119
7-15	Response time by number of users of individuals with Happy Scenario 1, Happy Scenario 2 and Tower Babel antipattern	120
7-16	Response time by number of users of individuals with Unbalanced Process- ing antipattern	121
7-17	Search space of individuals found in the experiment by search method . . .	125
7-18	Search space of individuals found in the experiment by fitness value . . .	125
7-19	Fitness value by generation on JPetStore First experiment	126
7-20	Fitness value by generation on JPetStore First experiment	126
7-21	Number of requests by Search Method	126
7-22	Fitness value by generation in all tests	126
7-23	Experiment Pareto Frontier	129
B-1	GridWorld initial states	143
B-2	GridWorld after four initial actions	143
B-3	GridWorld - exploration phase (Step 5 to 10)	144
B-4	GridWorld - exploration phase (Step 10 to 13)	145

List of Tables

2.1	Number of papers by type of Stress Testing	32
2.2	Performance antipatterns	53
3.1	Distribution of the research studies over the range of applied metaheuristics	69
5.1	Q-value for some states of the test [100]	91
6.1	Hypothetical MDP Q-values	99
7.1	Maximum value of the fitness function by algorithm	111
7.2	Best individuals found in the first experiment	115
7.3	Best individuals found in the second experiment	120
7.4	Results obtained from the second experiment	122
7.5	Example of individuals obtained in the second experiment	123
7.6	Percentage of genes in each scenario by generation	123
7.7	Maximum number of users of isolated test scenarios	124
7.8	Best individuals found in JPetStore first experiment	127
7.9	Pareto Frontier workloads results	130
A.1	Best individuals found in the second Testbed Tool experiments	134
A.4	My caption	134
A.2	Researches found using keyword 'Load Testing'	139
A.3	Researches found using keyword 'Stress Testing'	140

A.5 Comparison study presented by Pen~a-Ortiz on the book Modeling and Simulation of Computer Networks and Systems Methodologies [F- Fully attended feature P-Partial attended feature] [91]	141
--	-----

Chapter 1

Introduction

Many systems must support concurrent access by hundreds or thousands of users. Failure to providing scalable access to users may results in catastrophic failures and unfavorable media coverage [76].

The explosive growth of the Internet has contributed to the increased need for 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 to fix them [92].

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 [43] [76].

Stress testing determines the responsiveness, throughput, reliability, or scalability of a system under a given workload. The quality of the results of applying a given load testing to a system is closely linked to the implementation of the workload strategy. The performance of many applications depends on the load applied under different conditions. In some cases, performance degradation and failures arise only in stress conditions [54] [76].

A stress test uses a set of workloads that consist of many types of usage scenarios and a combination of different numbers of users. A load is typically based on an operational profile. Different parts of an application should be tested under various parameters and stress conditions [20]. The correct application of a stress test should cover most parts of an

application above the expected load conditions[43].

Fig. 1-1 shows an example of a system under assessment with three pages (the main page, profile page, and search page) and six possible users. From the combinations of users and application pages, various scenarios can be created, such as scenarios 1 and 2 shown in the figure. The first scenario presents a test that has passed, and the second scenario presents a test that has an HTTP 404 error.

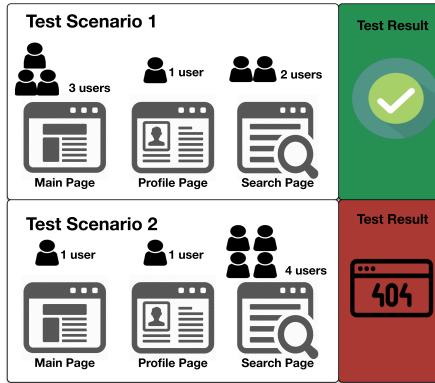


Figure 1-1: Possible test scenarios for a hypothetical application

A stress test usually lasts for several hours or even a few days and only tests a limited number of workloads. The major challenge is to find the workloads that expose a major number of errors and to discover the maximum number of users supported by an application under testing [22].

Search-based testing is seen as a promising approach to verifying timing constraints [6]. A common objective of a load search-based test is to find scenarios that produce execution times that violate the specified timing constraints [116].

This research proposes the use of a hybrid metaheuristic approach 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. Two experiments were conducted to validate the proposed approach. The first experiment was performed on an emulated component, and the second one was performed using an installed Moodle application.

Fig. 1-2 shows an example where IAdapter stress test automation finds two test scenarios. The first scenario presents a test that has an HTTP 500 error, and the second scenario

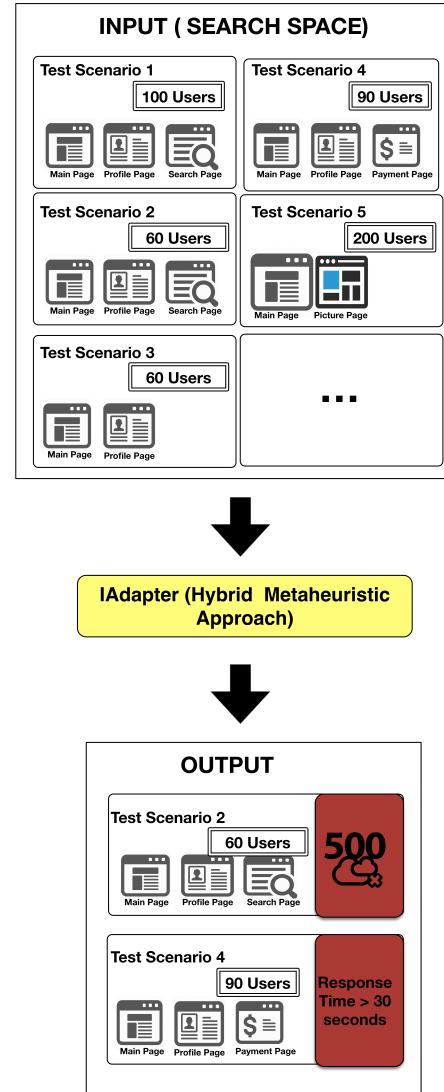


Figure 1-2: Illustrative example showing how IAdapter should be used

presents a test that has a response time higher than 30 seconds.

The research also addresses the problem of comparing the use of several metaheuristics in search based tests. When comparing a new metaheuristic to existing ones, it is advantageous to test on the problem instances already tested by previous papers. Then, results will be comparable on a by-instance basis, allowing relative gap calculations between the two heuristics. In this paper, we propose a flexible testbed tool to evaluate various diversity combining metaheuristics in search based software testing. The IAdapter Testbed is an open-source facility that provides software tools for search based test research. The

testbed tool emulates test scenarios in a controlled environment using mock objects and implementing performance antipatterns.

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 [67] [39].

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 [125].

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 [6] [55].

Performance testing of enterprise applications is manual, laborious, costly, and not particularly 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. Distilling these properties automatically into rules that describe how these properties affect performance of the application is a subgoal of our approach [62].

Experimentation is important to realistically and accurately test and evaluate search based 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 and practices on stress testing:

1.1.1 State of Research on the Search-Based Stress Testing

In the academic context, a number of studies proving the efficacy of metaheuristics to automate test execution can be found in literature. 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 [6].

Garousi and Alesio presents functional tools to Systems from safety-critical domains. Garousi presented a stress test methodology aimed at increasing chances of discovering faults related to distributed traffic in distributed systems. Alesio describe stress test case generation as a search problem over the space of task arrival times. The research search for worst case scenarios maximizing deadline misses where each scenario characterizes a test case. The paper combine two strategies, GA and Constraint Programming (CP) [52] [54] [40] [41] [12].

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 [80].

Typically, performance testing is accomplished using test scripts, which are programs that test designers write to automate testing. These test scripts performs actions or mim-

icking 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 [62].

1.2 Research Hypothesis

Our underlying research hypothesis is as follows:

The use of metaheuristics and hybrid metaheuristics in combination with Q-learning can make it possible to automate the stress test execution process, improving the choice of new test cases for each interaction and finding scenarios that maximize the number of users of the application under test and minimize response time or find scenarios with a expected response time.

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 Thesis Contributions

The major contributions of this thesis are:

1.3.1 Hybrid Metaheuristic Approach applied on Stress Testing

We present a hybrid metaheuristic approach that uses Genetic Algorithms, Simulated Annealing and Tabu Search in a collaborative mode. In the first experiment, The signed-rank Wilcoxon non-parametrical procedure showed that there was a significant improvement in the results with the Hybrid Metaheuristic approach [60].

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 hybrid metaheuristic returned individuals with higher fitness scores.

1.3.2 Q-Learning Metaheuristic Approach

1.3.3 IAdapter JMeter Plugin

1.3.4 IAdapter Testbed Tool

1.4 Thesis Organization

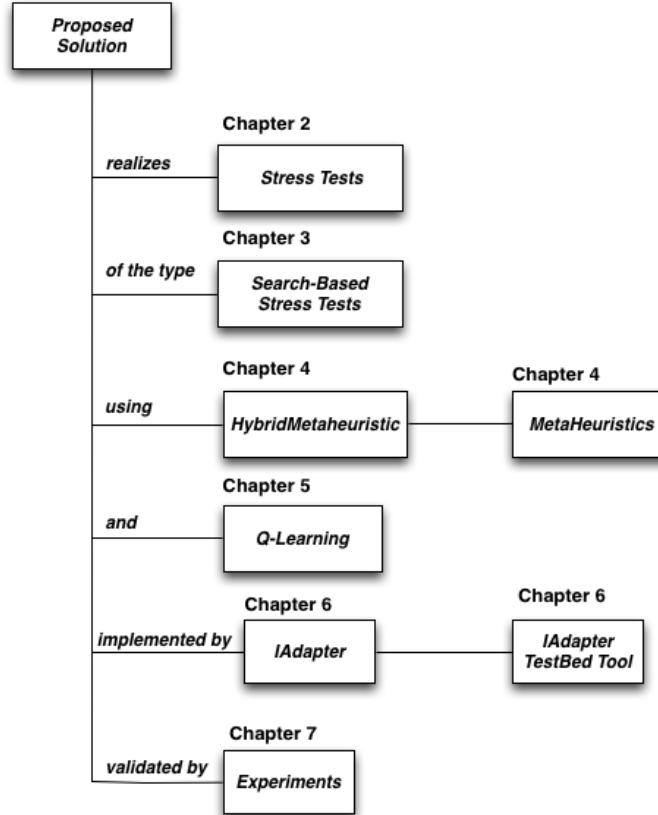


Figure 1-3: Thesis Organization

Chapter 2

A Survey on Stress Testing Software Systems

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 [107] [37].

Typically, the most common kind of performance testing for Internet applications is load testing. Application load can be assessed in a variety of ways [99]:

- Concurrency. Concurrency testing seeks to validate the performance of an application with a given number of concurrent interactive users [99].
- 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 [99].
- 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 [99].

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 [42]. The purpose of this assessment is to demonstrate that the system reaches its performance objectives [107]. Term often used interchangeably with “stress” and “load” testing. Ideally “performance” testing is defined in requirements documentation or QA or Test Plans [80].

In a load testing, the system is evaluated at predefined load levels [42]. 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 [92]. 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 [6].

Stress testing investigates the behavior of the system under conditions that overload its resources. The stress testing verifies the system behavior against heavy workloads [107] [80], 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 break-points, forcing the application or the supporting infrastructure to fail [42] [92].

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 break-points, forcing the application or the supporting infrastructure to fail.

2.1 Method

This paper surveys the state of the art literature in stress testing research. The thesis extends the survey presented by Jiang et al. [76] and Afzal et al. [6] 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 [?].

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 [?].

Figure 2-1 presents the summary result of systematic review process. 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.

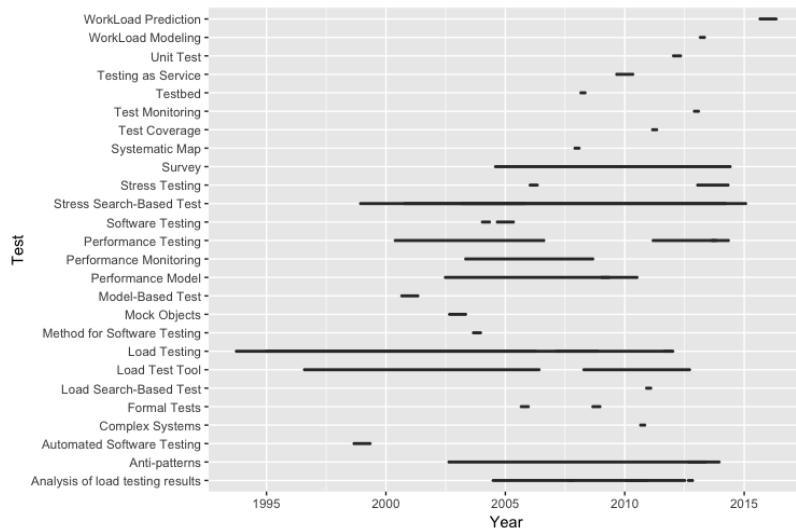


Figure 2-1: Bubble chart for results of search with the keyword 'Load Testing'

2.1.1 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 [?]:

- 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?

2.1.2 Research Questions

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

- How is a proper stress designed?
- How is a stress test executed and automated?
- What are the main problems found by stress tests?
- How are the stress tests results analysed?

2.1.3 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

All papers found are listed in <https://www.mendeley.com/community/pesquisatesteperformance>

Figures 2-2 and 2-3 show the bubble chart for 'Load Testing' and 'Stress Testing' keywords. Figures 2-4 and 2-4 present the word cloud extracted for title and abstract of the papers found with keywords 'Load testing' and 'Stress testing'.

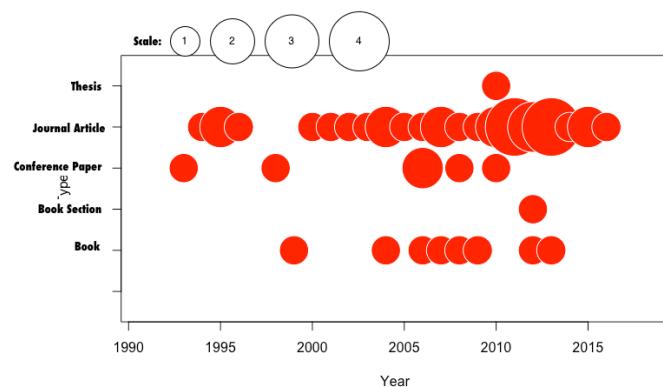


Figure 2-2: Bubble chart for results of search with the keyword 'Load Testing'

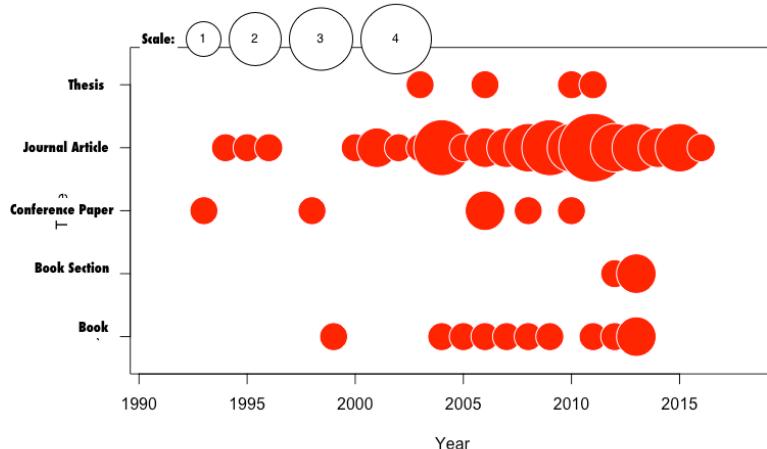


Figure 2-3: Bubble chart for results of search with the keyword 'Stress Testing'

2.1.4 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.



Figure 2-4: Word cloud for results with the keyword 'Load Testing'

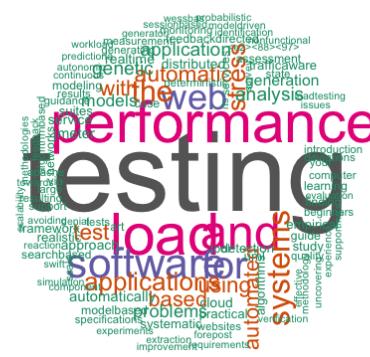


Figure 2-5: Word cloud for results of search with the keyword 'Stress Testing'

- Do not relate to load/stress testing model.

From 366 initial papers, 97 papers was selected.

2.1.5 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)

Figure 2-6 presents the type of research paper by year.

Table 2.1 show the number of papers by type of stress testing found un the systematic review.

2.1.6 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. 2-7 shows a common Load, Performance and Stress test process [76].

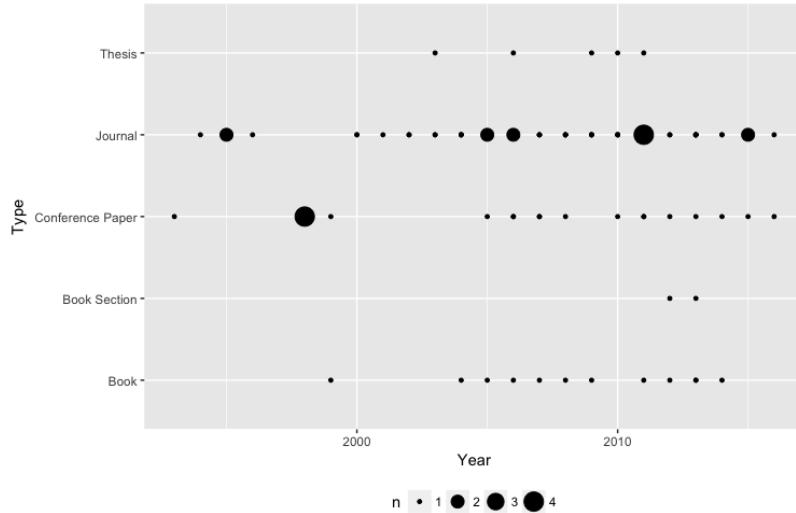


Figure 2-6: Summary of type of research paper by year

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 [76].

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[76].

The core activities in conducting an usual Load, Performance and Stress tests are [47]:

- 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

Table 2.1: Number of papers by type of Stress Testing

Analysis of load testing results	Anti-patterns	Automated Software Testing
9	5	1
Complex Systems	Formal Tests	Load Search-Based Test
1	2	1
Load Test Tool	Load Testing	Method for Software Testing
5	16	1
Mock Objects	Model-Based Test	Performance Model
1	1	5
Performance Monitoring	Performance Testing	Software Testing
2	7	2
Stress Search-Based Test	Stress Testing	Survey
21	4	5
Systematic Map	Test Coverage	Test Monitoring
1	1	1
Testbed	Testing as Service	Unit Test
1	1	1
WorkLoad Modeling	WorkLoad Prediction	FOREPOST
1	1	2

function such as searching a product catalog, adding an item to a shopping cart, or placing an order [37]. 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 [80].

- 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

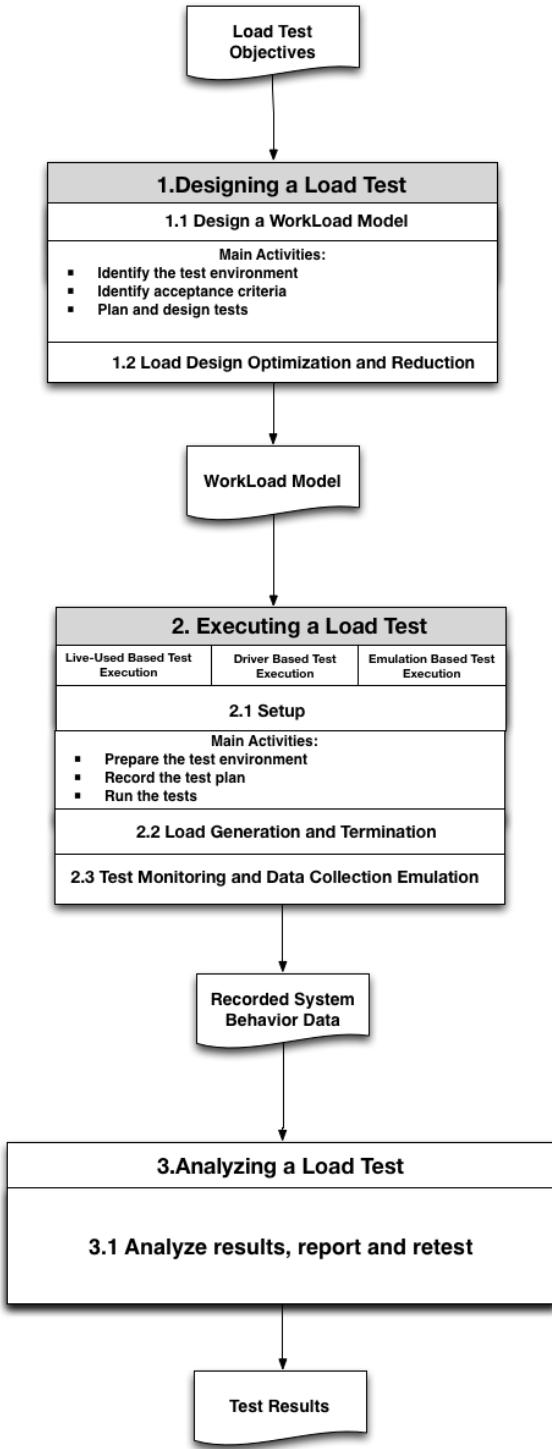


Figure 2-7: Load, Performance and Stress Test Process [76][47]

identify areas of bottleneck that need addressing.

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 [6]:

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

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 [6].

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 [92].

The term workload represents the size of the demand that will be imposed on the application under test in an execution. The metric used for 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 [48] [92] [61].

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 [48] [92].

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 [48] [92].

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 (Fig. 2-8). This model is applied to all the workload attributes, e.g., computation, memory usage, I/O behavior, communication, etc. [42]. Fig. 2-8 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.

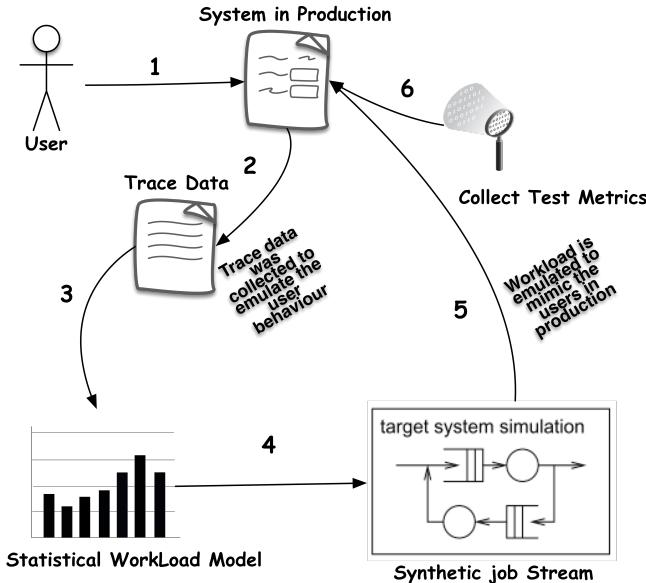


Figure 2-8: Workload modeling based on statistical data [42]

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. 2-9).

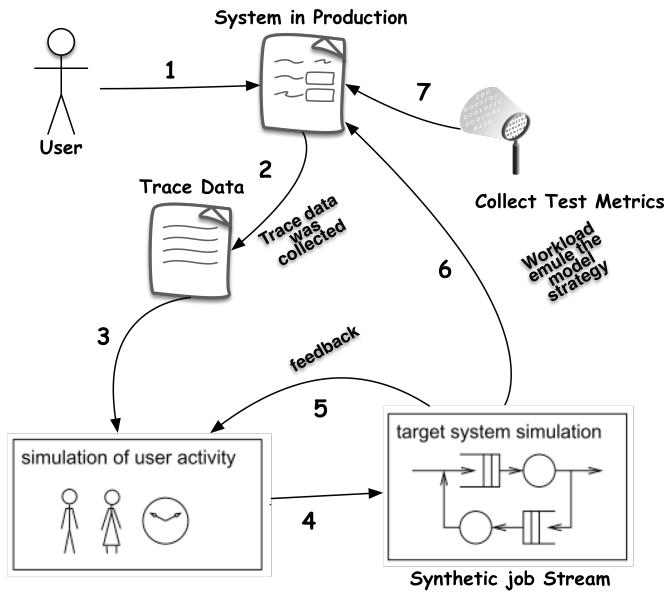


Figure 2-9: Workload modeling based on the generative model [42]

Both load model 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, the resulting loads usually only cover a small portion of the testing objectives [76]. The presented research work uses a generative model.

There are several approaches to design generative or descriptive workloads:

- Model-based Stress testing: a usage model is proposed to simulate users' behaviors.
- Feedback-Directed Learning Software Testing: is an adaptive, feedback-directed

learning testing system that learns rules from system execution [81] [135].

- Search-based Stress testing.

Search-Based Stress testing will be detail explained in the chapter 3.

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 [85] [1].

The model paradigm 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 [85] [69].

- 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 [85].
- 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 [85].
- 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 [85].

- 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 [85].
- 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 [85].
- 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 [85].
- 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 [85].

A User Community Modeling Language (UCML) is a set of symbols that can be used to create visual system usage models and depict associated parameters [129]. The Fig. 2-10 shows a sample where all users realize a login into the application unde test. Once logged in, 40% of the users navigates on the application, 30% of the users realizes downloads. 20% of users realizes uploads and 10% of users performs deletions.

Another technique to create workload models it is Stochastic Formcharts. 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. 2-11 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 [43].

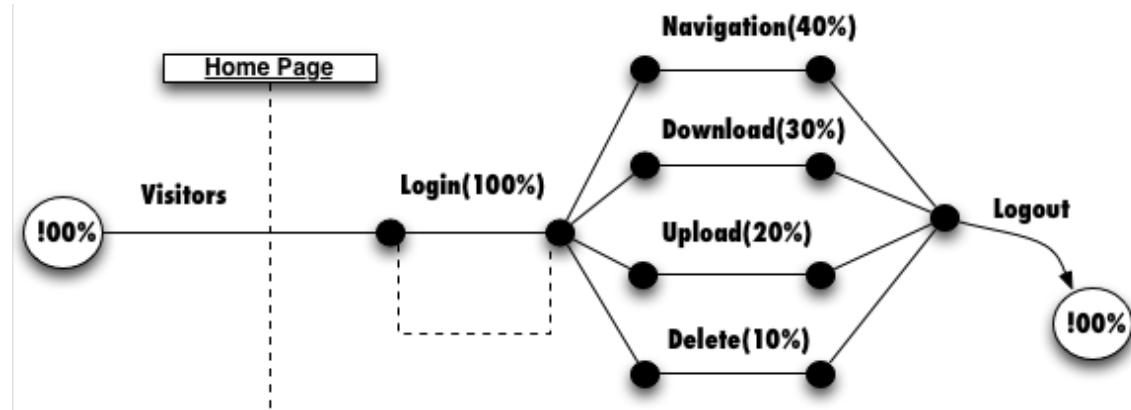


Figure 2-10: User community modeling language [129]

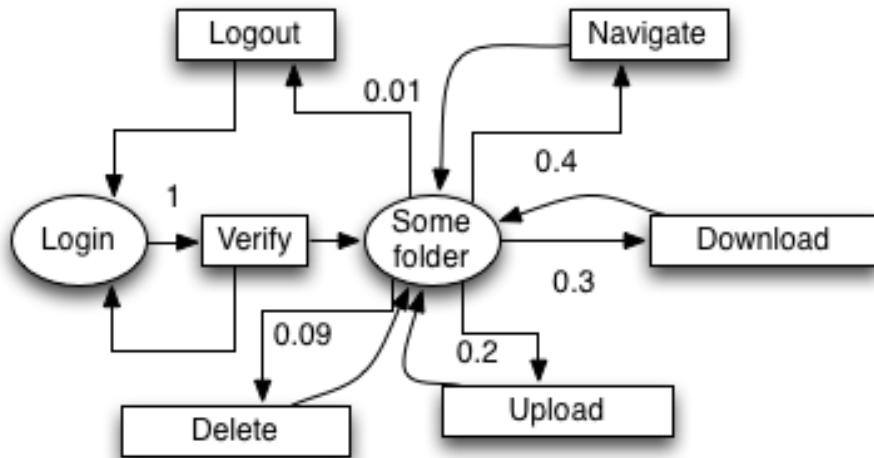


Figure 2-11: Stochastic Formcharts Example [43] [129]

One way to capture the navigational pattern within a session is through the Customer Behavior Model Graph (CBMG). Figure 2-12 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 [90] [76] [91].

Garousi et al. proposes derive 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. 2-13).

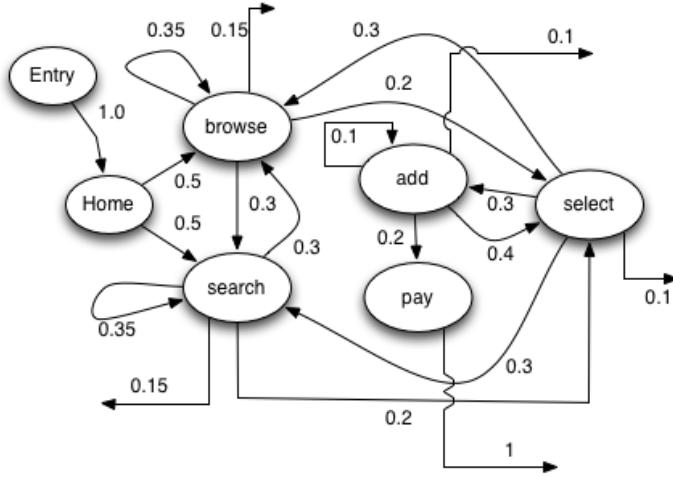


Figure 2-12: Example of a Customer Behavior Model Graph (CBMG) [90] [76] [91]

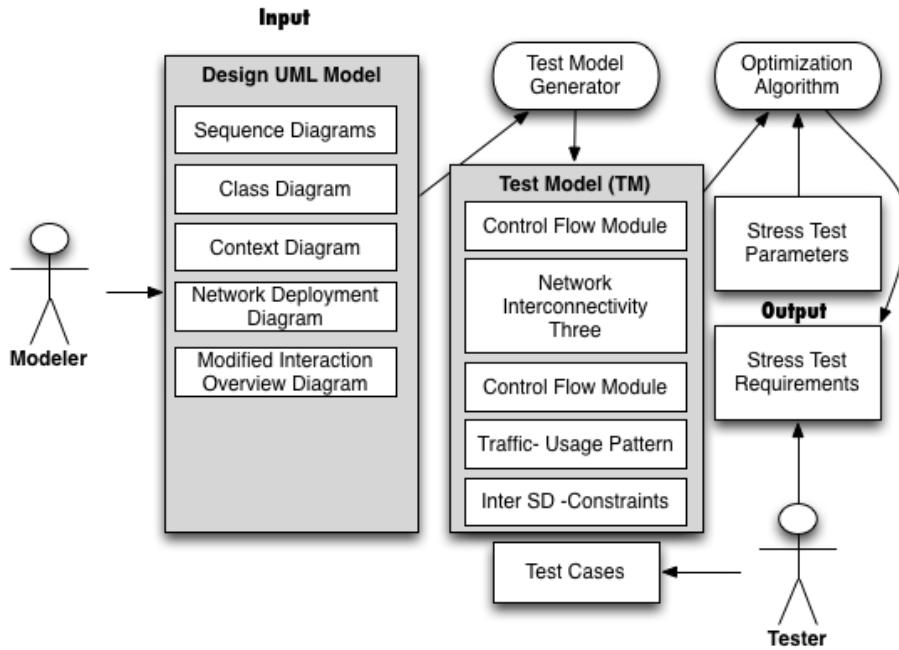


Figure 2-13: Model-based stress test methodology

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 [127]. The workload specification formalism (Workload Model) consists of the following components, which are detailed below and illustrated in Fig. 2-14:

- 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 generation.
- A Workload Intensity that includes a function which specifies the number of concurrent users during the workload generation execution.

2.2.2 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 [62].

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. 2-15 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

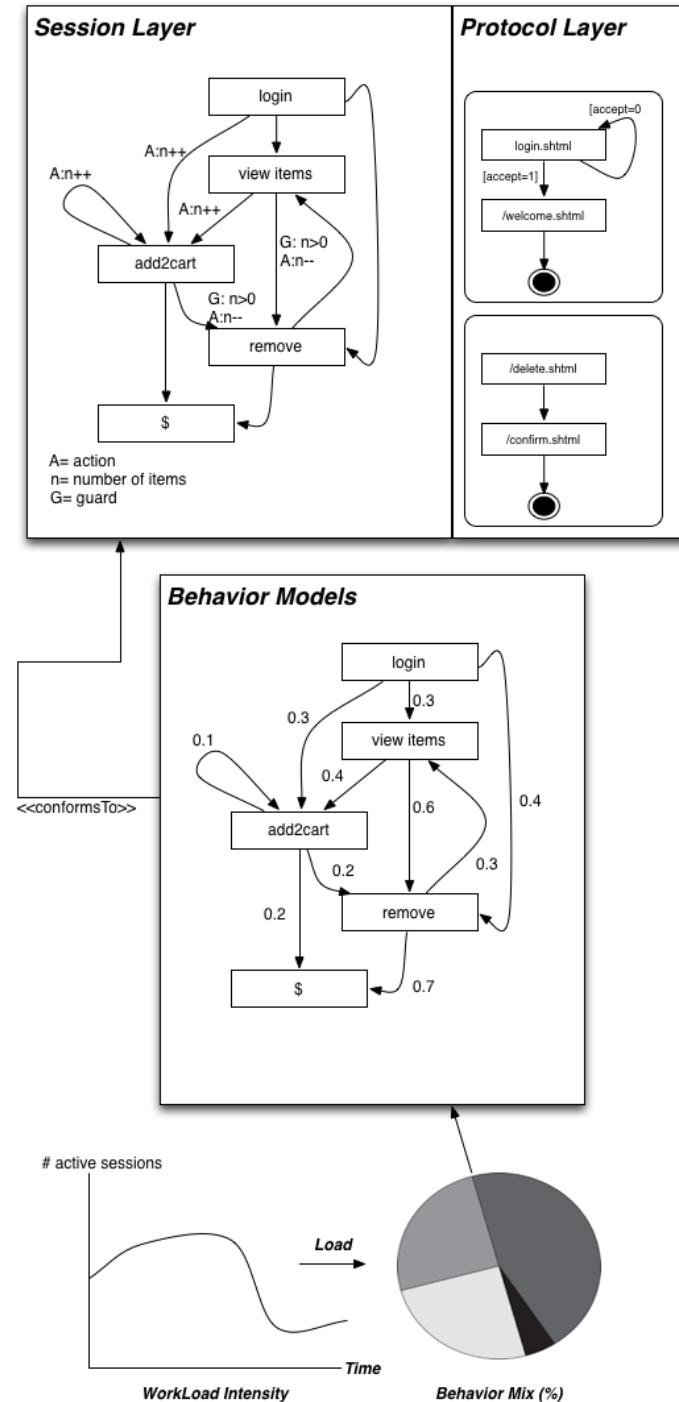


Figure 2-14: Exemplary workload model

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.

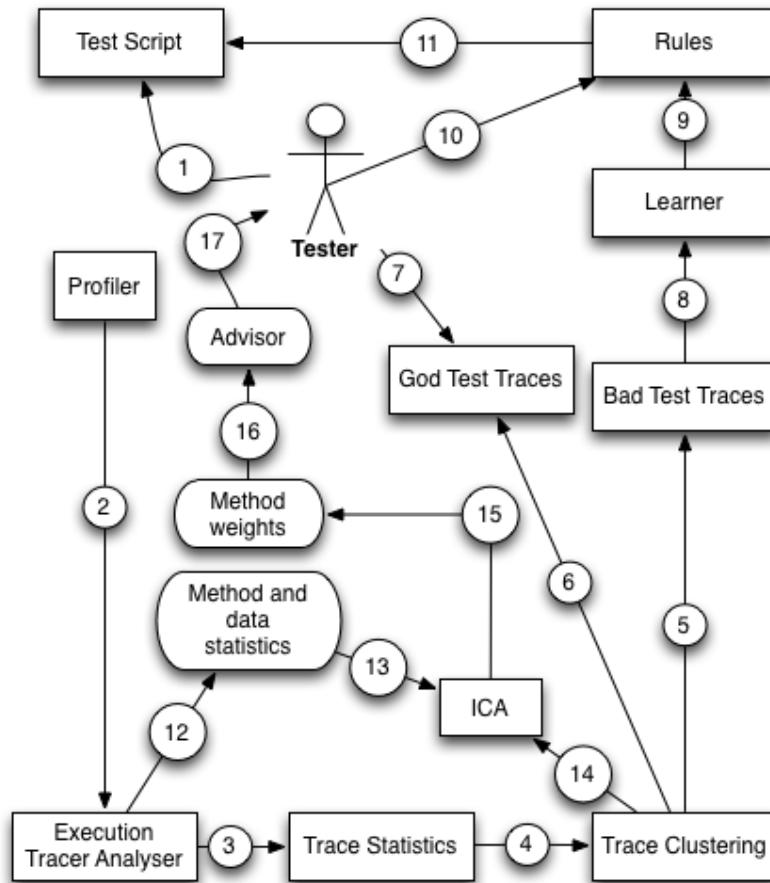


Figure 2-15: The architecture and workflow of FOREPOST

2.3 Research Question 2: How is a stress test executed and automated?

The stress test execution consists of deploy the system and steup 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 [92][76]:

- 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 [47] [91] [129]. There are three categories of load drivers [76]:

- 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 controller component, which coordinates the load generation among the peer load drivers.

2.3.1 Load Test Tools

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.

Workload generators 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. Workload generators typically have the following components [92]:

- 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 iteration.

There are several tool to execution of Stress testing. In this tools, the procedure 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 [62].

Performance testing tools, such as Rational's PerformanceStudio, allow for load testing, where the tool can be programmed to run a number of client machines simultaneously to load the client/server system and measure response time. Load testing typically involves various scenarios to analyze how the client/server system responds under various loads [45].

Comparing Web workload generators is a laborious and difficult task since they offer a large amount and diversity of features. In this section we contrast generators according to a wide set of features and capabilities, focusing on their ability to realize search-based tests or have learning capacities.

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 [91] [124].

TPC Benchmark (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 model 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) [91] [90].

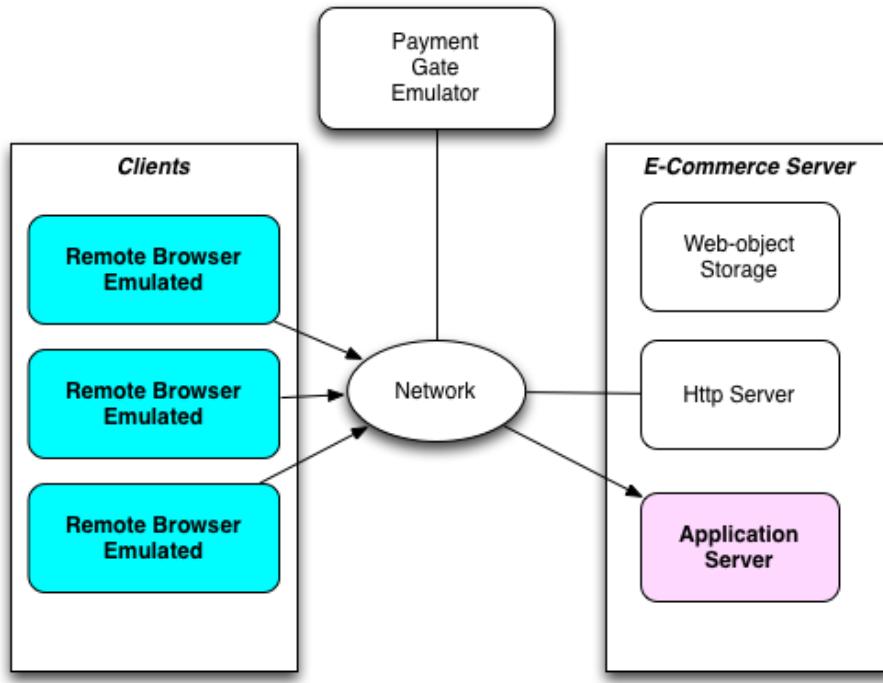


Figure 2-16: TPC-W architecture [91] [90]

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.

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. 2-17). 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 [91].

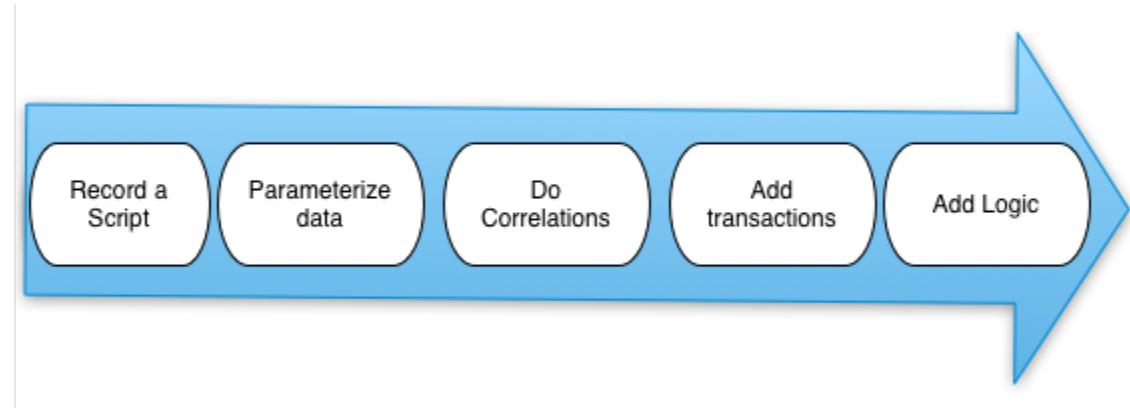


Figure 2-17: Load Runner Scripting

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 [91].

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 [65] [47].

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 details about the request protocol and test elements.

The stress test tools was categorized in three different groups [91]:

- 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.

The tools are compared using 12 features [91]:

- 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.

Figures 2-18,2-19 and 2-20 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.

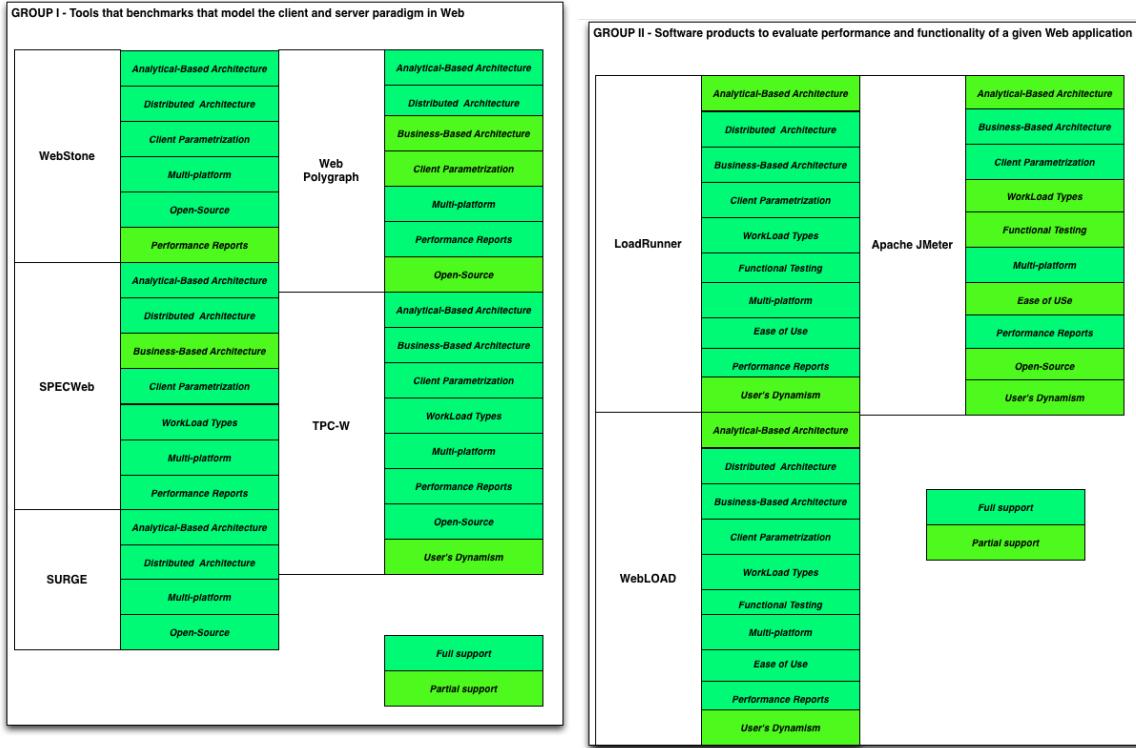


Figure 2-18: Group I- Benchmarks that model the client and server paradigm in Web context. [91]

Figure 2-19: Group II- Software products to evaluate performance and functionality[91].

2.4 Research Question 3: 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. 2-21 shows the symptoms of known performance problems [134].

There are several antipatterns that details 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 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 [34]. The

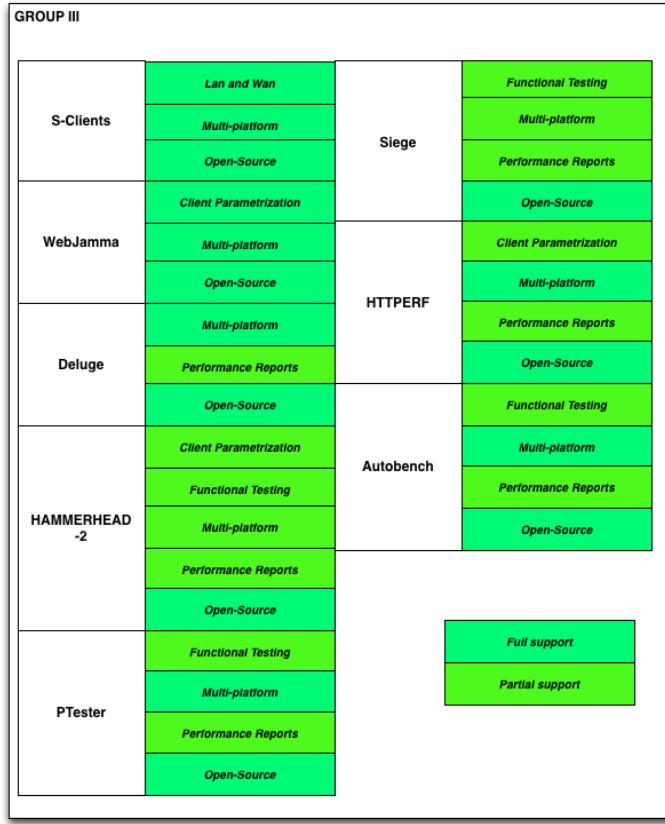


Figure 2-20: Group III- Testing tools and other approaches for traffic generation based on HTTP traces [91].

table 2.2 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 [38] [113].

A project containing a “god” class is 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 [113]. The Figures 2-22 and 2-23 describes an hypothetical system with a BLOB problem: The Fig. 2-22 presents a sample where the Blob class uses the features A,B,C,D,E,F and G of the hypothetical system; The Fig. 2-23 shows a static

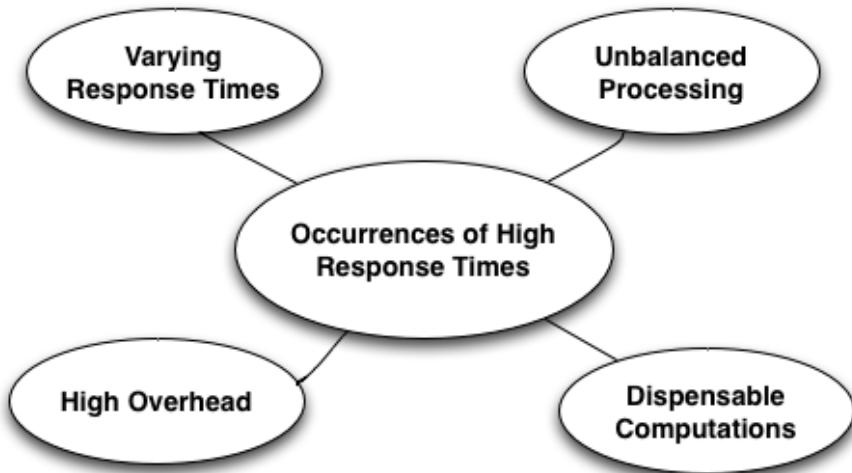


Figure 2-21: Symptoms of known performance problems [134].

view where a complex software entity instance, i.e. S_d, is connected to other software instances, e.g. S_a, S_b and S_c, through many dependencies [125][134].

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. 2-24 shows a sample of the Unbalanced Processing. In The Fig. 2-24, 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. 2-25 describes a software S with a Pipe and Filter Architectures problem: (a) Static View, there is a software entity instance, e.g. S_a, offering an operation (operation x) whose resource

Table 2.2: Performance antipatterns

antipattern	Derivations
Blob or The God Class	
Unbalanced-Processing	Concurrent processing Systems Piper and Filter Architectures Extensive Processing
Circuitous Treasure Hunt	
Empty Semi Trucks	
Tower of Babel	
One-Lane Bridge	
Excessive Dynamic Allocation	
Traffic Jam	
The Ramp	
More is Less	

demand (computation = \$compOpx, storage = \$storOpx, bandwidth = \$bandOpx) is quite high; (b) Dynamic View, 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. 2-26 describes a software S with a Extensive Processing problem: (a) Static View, there is a software entity instance, e.g. Sa, offering two operations (operation x, operation y) whose resource demand is quite unbalanced, since opx has a high demand (computation = \$compOpx, storage = \$storOpx, bandwidth = \$bandOpx), whereas opy has a low demand (computation = \$compOpy, storage = \$storOpy, bandwidth = \$bandOpy); (b) Dynamic View, 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 [125].

Circuitous Treasure Hunt antipattern occurs when software retrieves data from a first component, uses those results in a second component, retrieves data from the second component, and so on, until the last results are obtained [115] [114]. 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 Treas-

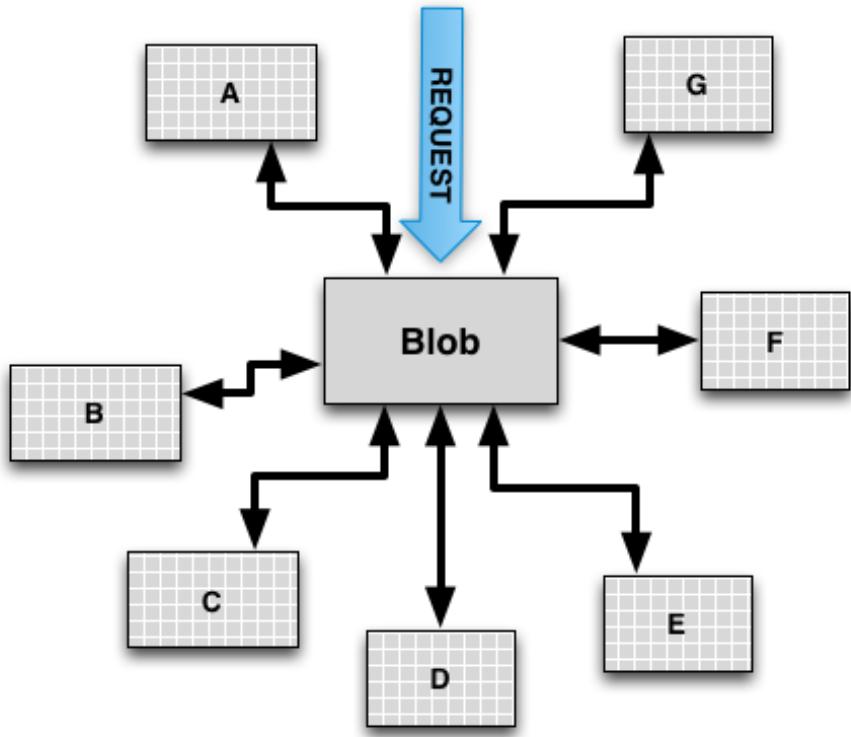


Figure 2-22: The God class[134].

sure Hunt antipattern hurts performance [135]. The Fig. 2-27 shows a software S with a Circuitous Treasure Hunt problem: (a) Static View, there is a software entity instance e.g. Sa, retrieving information from the database; (b) Dynamic View, the software S generates a large number of database calls by performing several queries up to the final operation [125].

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 [14]. 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: (a) Static View, there is a software entity instance, e.g. Sa, retrieving some information from several instances (Remote Software 1, . . . , Remote Software n); (b) Dynamic View, the software instance Sa generates an 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 [125].

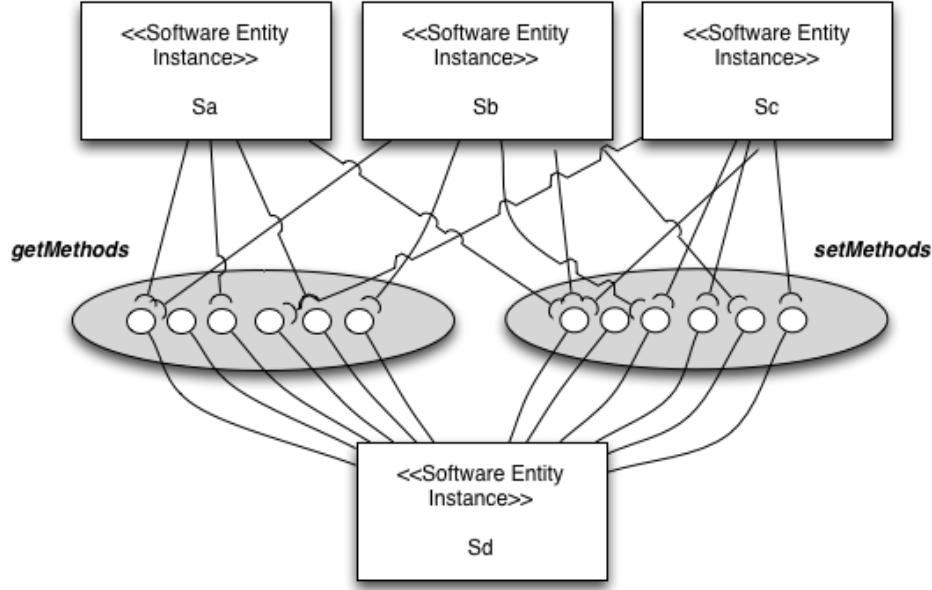


Figure 2-23: The God class[125].

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 [114]. Fig. shows a system with a Tower of Babel problem: (a) Static View, there are some software entity instances, e.g. `Sa`, `Sb`, . . . , `Sn`; (b) Dynamic View, the software instances `Sd` performs many times the translation of format for communicating with other instances [125].

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: (a) Static View, there is a software entity instance with a capac-

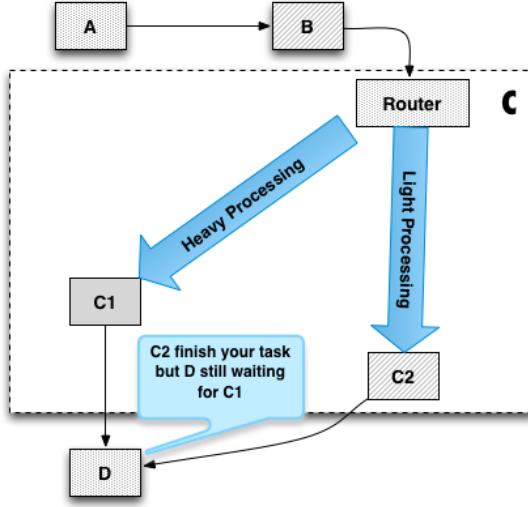


Figure 2-24: Unbalanced Processing sample [134].

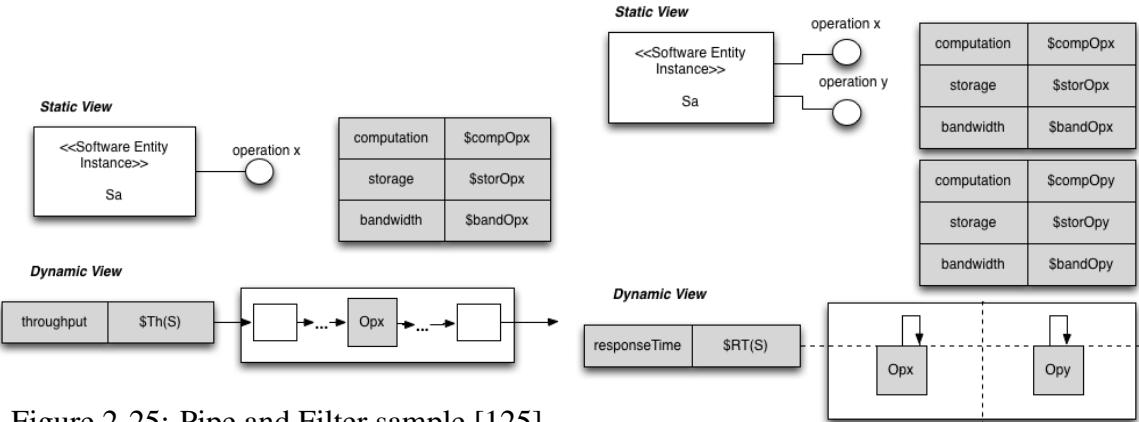


Figure 2-25: Pipe and Filter sample [125]

Figure 2-26: Extensive Processing sample [125].

ity of managing \$poolSize threads; (b) Dynamic View, 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 [125].

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 re-

Static View



Dynamic View

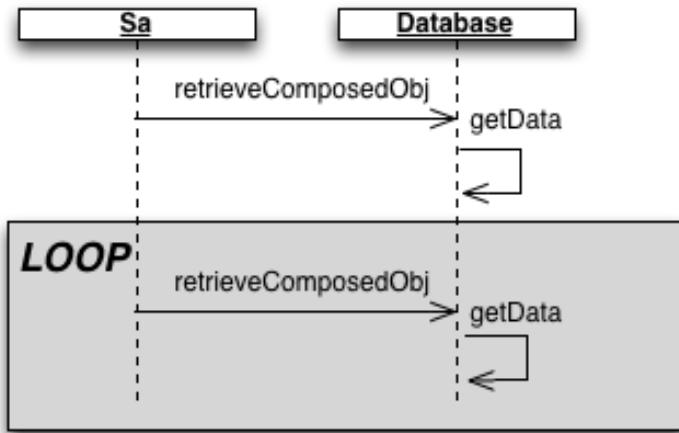


Figure 2-27: Circuitous Treasure Hunt sample [125]

claimed memory must be returned to the heap to avoid memory leaks [115] [114].

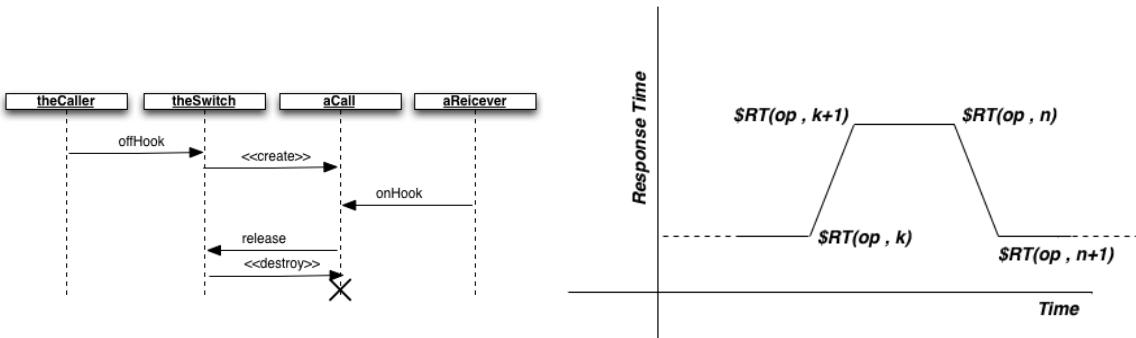


Figure 2-31: Excessive Dynamic Allocation.

Figure 2-32: Traffic Jam Response Time [125].

The Fig. 2-31 shows a Excessive Dynamic Allocation sample. This example is drawn from a call (an 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

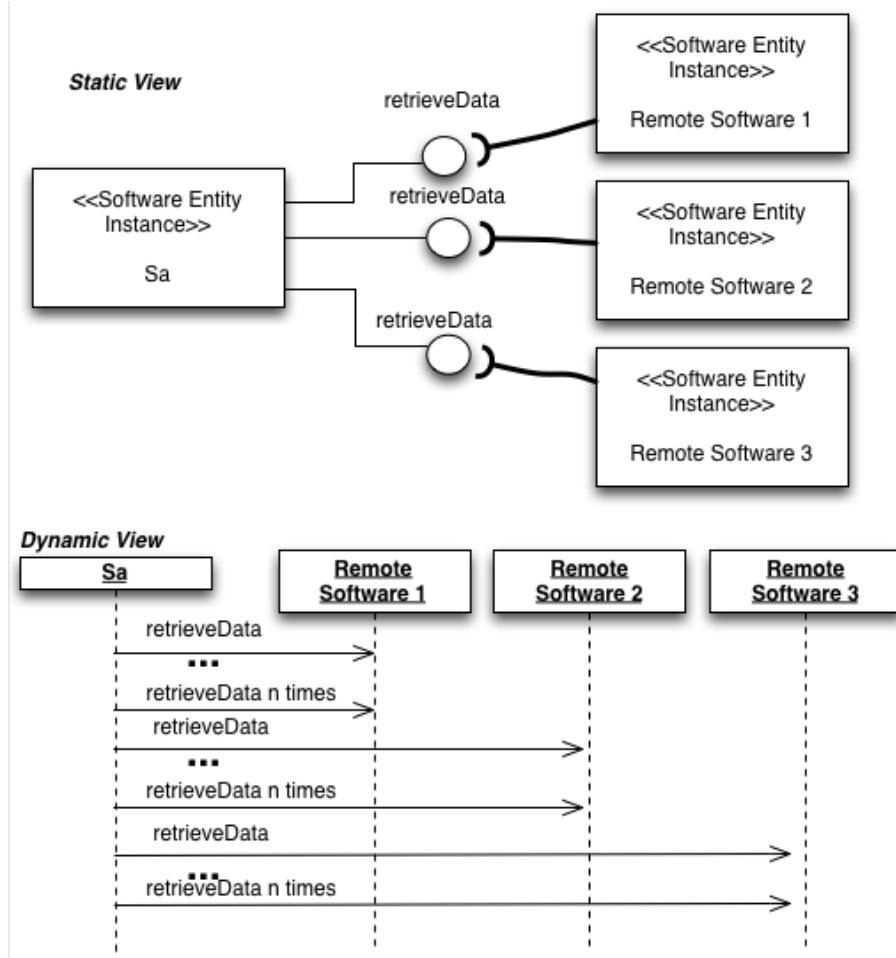


Figure 2-28: Empty Semi Trucks sample [125].

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 antipattern 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 2-32 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 [125].

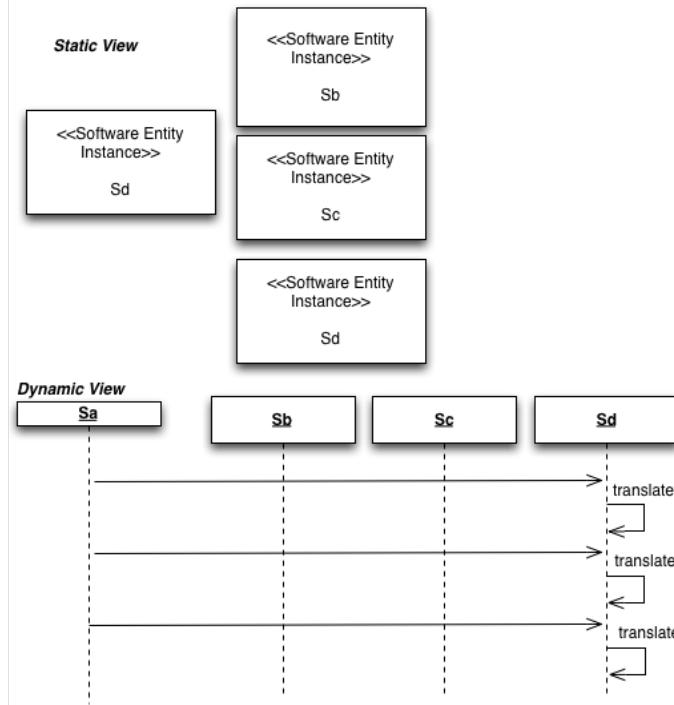


Figure 2-29: Tower of Babel sample [125]

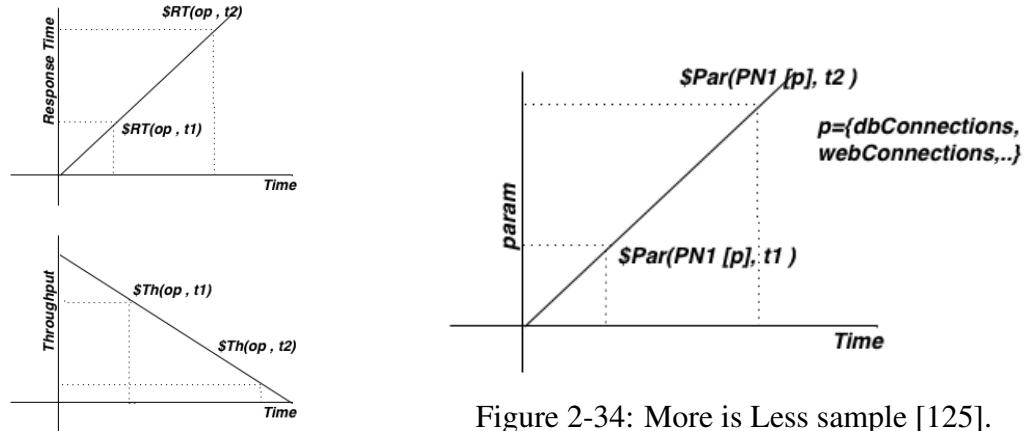


Figure 2-34: More is Less sample [125].

Figure 2-33: The Ramp sample [125].

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 [135] [114]. Fig. 2-33 shows a system with The Ramp problem: (i) the

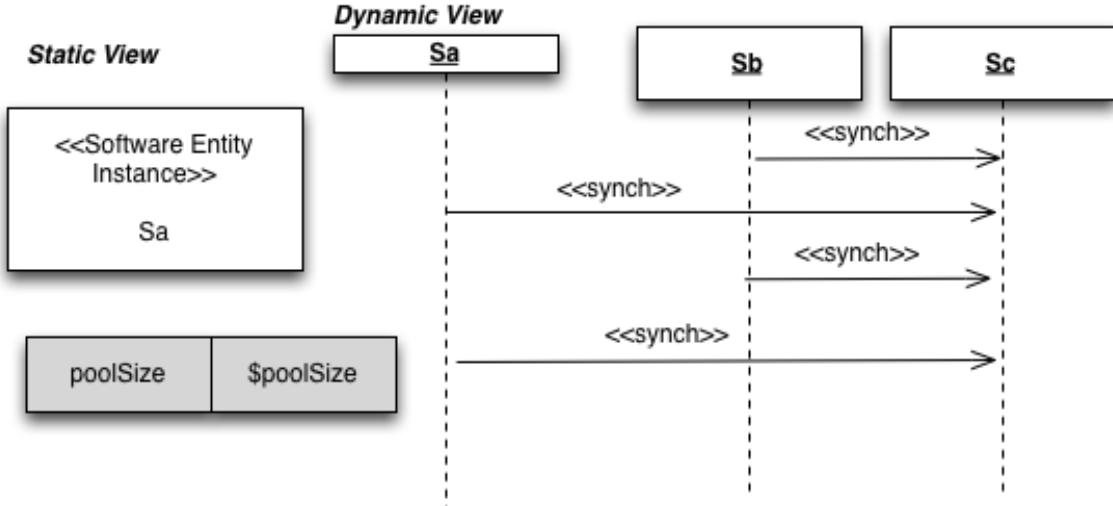


Figure 2-30: One-Lane Bridge sample [125].

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. 2-34 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 $t1 < t2$.

2.5 Research Question 4: How are the stress tests results analysed?

The system behavior recorded during the stress test execution phase needs to be analyzed to determine if there are any load-related functional or non-functional problems [76].

There can be many formats of system behavior like resource usage data or end-to-end response time, which is recorded as response time for each individual request. These types of data need to be processed before comparing against threshold values. A proper data summarization technique is needed to describe these many data instances into one number.

There are three types of data summarization techniques proposed in the literature. Jiang et al. use response time analysis as an example to describe the proposed data summarization techniques [76]:

- Maximum values;
- Average or Medium Values;
- Percentile-values.

Some researchers advocate that the 90-percentile response time is a better measurement than the average/medium response time, as the former accounts for most of the peaks, while eliminating the outliers [76].

2.6 Conclusion

This systematic review investigated the use of stress test techniques. Figure 2-35 present the results summary of the systematic review. The Test Design phase could use Realistic Load and Fault-Inducing Load. Realistic Load 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. There are several approaches to design generative or descriptive workloads.

In Model-based Stress testing, usage model is proposed to simulate users' behaviours. The model paradigm 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: State-Based (or Pre/Post) Notations; Transition-based Notations; History-based Notations; Functional Notations; Operational Notations; Stochastic Notations and Data-Flow Notations. There are many kinds of non-functional search based tests: Execution time, Quality of service, Security, Usability and Safety.

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.

Search-Based Testing is the process of automatically generating test according to a test adequacy criterion. 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. There are two measurement units normally associated with the fitness function in 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.

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 behaviour. There are three general approaches of load test executions: Live-User Based, Driver-Based and Emulation-Based. In 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. The driver based execution approach automatically generate thousands or millions of concurrent requests for a long period of time using a software tool. 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.

There are several antipatterns that details features about common performance problems. Blob is an antipattern whose problem is on the excessive message traffic generated

by a single class or component. 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.

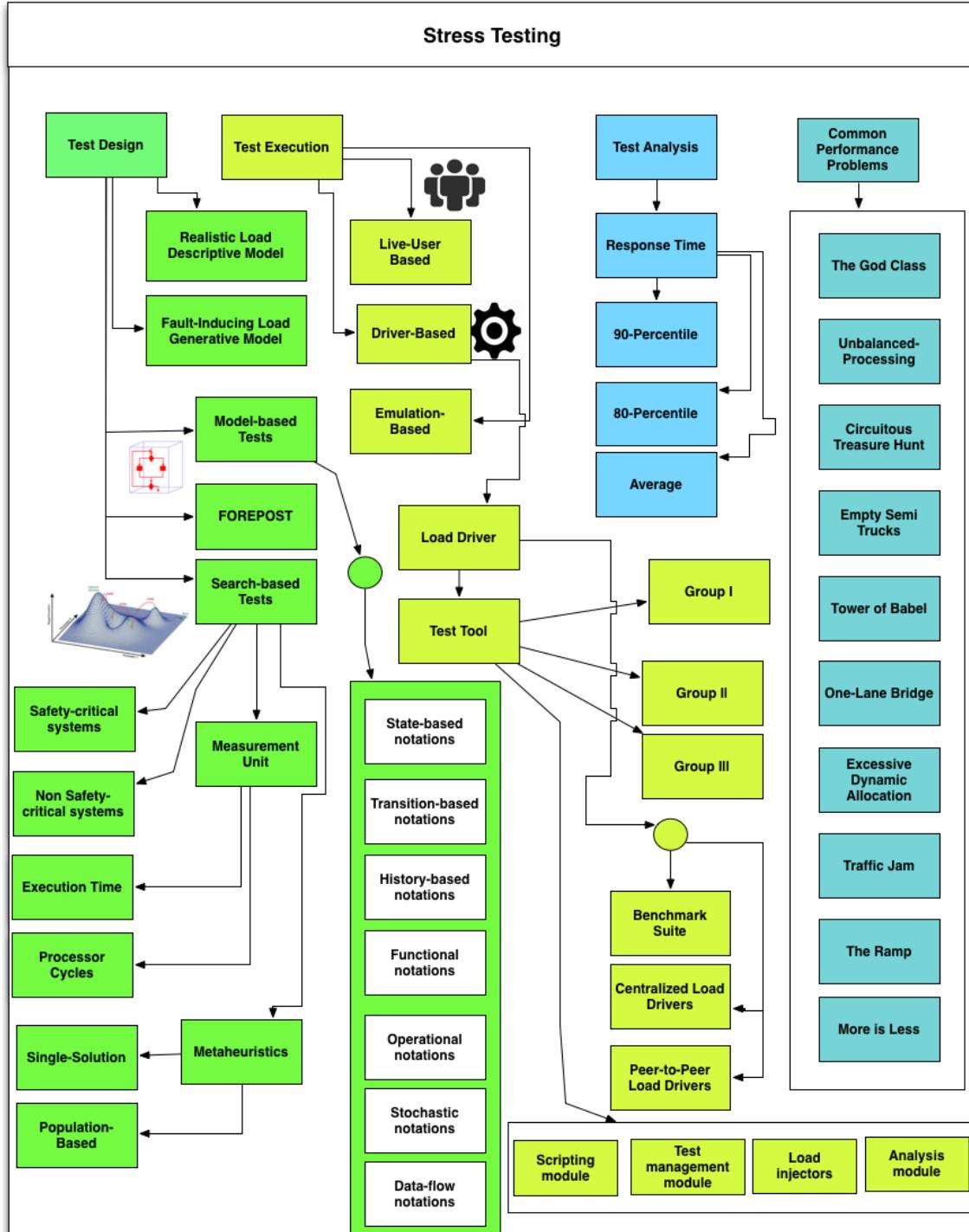


Figure 2-35: Results summary of the systematic review.

Chapter 3

Search-Based Stress Testing

The goal of this Chapter is to describe Search-Based Testing, define recurring solutions in search-based testing

3.1 Introduction

Search-based software engineering (SBSE) is the application of optimization techniques in solving software engineering problems [1,2]. The applicability of optimization techniques in solving software engineering problems is suitable as these problems frequently encounter competing constraints and require near optimal solutions [6] [66].

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. SBSE uses computational search techniques to tackle software engineering problems, typified by large complex search spaces. SBSE derives test inputs for a software system with the goal of improving various criteria. 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 [6] [13] [66].

Stress search-based testing is a application of SBST where the main goal it is to find test scenarios that produce execution times that exceed the timing constraints specified. If a temporal error is found, the test was successful [116].

3.2 Search-Based Testing

Search-Based Testing is the process of automatically generating test according to a test adequacy criterion, encoded as a fitness function, using search-based optimization algorithms, which are guided by a fitness function. The role of the fitness function is to capture a test objective that, when achieved, makes a contribution to the desired test adequacy criterion [67].

Search-Based Testing uses metaheuristic algorithms to automate the generation of test inputs that meet a test adequacy criterion. Many algorithms have been considered in the past, including Simulated Annealing, Parallel Evolutionary Algorithms [10], Evolution Strategies, Estimation of Distribution Algorithms , Scatter Search , Particle Swarm Optimization , Tabu Search and the Alternating Variable Method. An advantage of metaheuristic algorithms is that they are widely applicable to problems that are infeasible for analytic approaches. All one has to do is come up with a representation for candidate solutions and an objective function to evaluate those solution [19].

The application of metaheuristic search techniques to test case generation is a possibility which offers much benefits. Metaheuristic search techniques are high-level frameworks which utilise heuristics in order to find solutions to combinatorial problems at a reasonable computational cost. Such a problem may have been classified as NP-complete or NP-hard, or be a problem for which a polynomial time algorithm is known to exist but is not practical [88].

One of the most popular search techniques used in SBST belong to the family of Evolutionary Algorithms in what is known as Evolutionary Testing. Evolutionary Algorithms represent a class of adaptive search techniques based on natural genetics and Darwin's theory of evolution. They are characterized by an iterative procedure that works in parallel on a number of potential solutions to a problem. Figure 3-1 shows the cycle of an Evolutionary Algorithm when used in the context of Evolutionary Testing [19].

First, a population of possible solutions to a problem is created, usually at random. Starting with randomly generated individuals results in a spread of solutions ranging in fitness because they are scattered around the search-space. Next, each individual in the

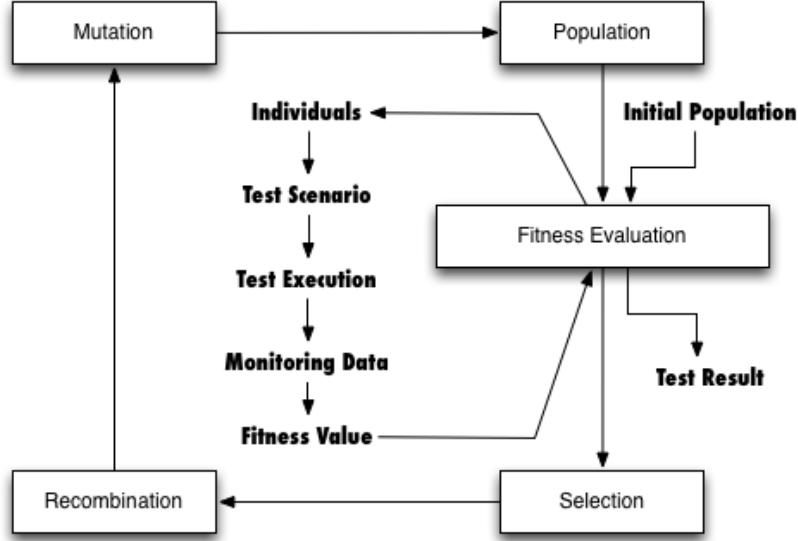


Figure 3-1: Evolutionary Algorithm Search Based Test Cycle[19].

population is evaluated by calculating its fitness via a fitness function. The principle idea of an Evolutionary Algorithm is that fit individuals survive over time and form even fitter individuals in future generations. Selected individuals are then recombined via a crossover operator. After crossover, the resulting offspring individuals may be subjected to a mutation operator. The algorithm iterates until a global optimum is reached or another stopping condition is fulfilled [19].

The fitness evaluation is the most time consuming task of SBST. However, for time consuming functional testing of complex industrial systems, minimizing the number of generated individuals may also be highly desirable. This might be done using an assumption about the "potential" of individuals in order to predict which individuals are likely to contribute to any future improvement. This prediction could be achieved by using information about similar individuals that have been executed in earlier generations.

3.3 Non-functional Search-Based Testing

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. There are many kinds of non-functional search based tests [6]:

- 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. The Fig. 3-2 shows a comparison between the range of metaheuristics and the type of non-functional search based test. The Data comes from Afzal et al. [6]. Afzal's work adds to some of the latest research in this area ([52] [54] [40] [41] [12] [60]).

3.4 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 [116]. 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 [6].

There are two measurement units normally associated with the fitness function in stress test: processor cycles and execution time. The processor cycle approach describes a fitness

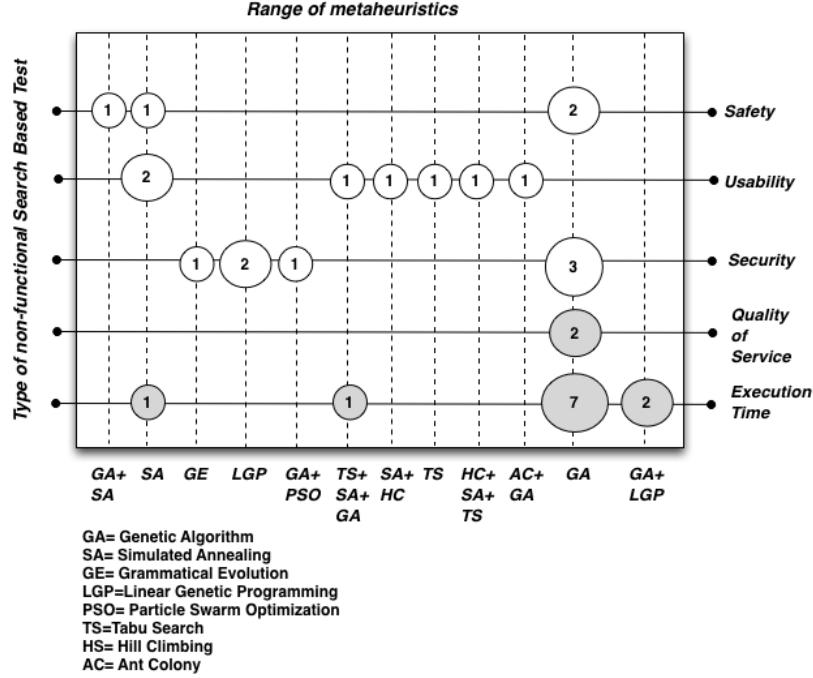


Figure 3-2: Range of metaheuristics by Type of non-functional Search Based Test[6].

function in terms of processor cycles. The execution time approach involves executing the application under test and measuring the execution time [6] [123].

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 for each platform. Execution time measurement is a non deterministic approach, there is no guarantee to get the same results for the same test inputs [6]. However, stress testing where testers have no access to the production environment should be measured by the execution time measurement [92] [6].

Table 3.1 shows a comparison between the research studies on load, performance, and stress tests presented by Afzal et al. [6]. Afzal's work adds to some of the latest research in this area ([52] [54] [40] [41] [12] [60]). 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).

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			Gois et al. 2016 [60]
GA	Alander et al., 1998 [9] Wegener et al., 1996 and 1997 [132][71] Sullivan et al., 1998 [116] Briand et al., 2005 [32] Canfora et al., 2005 [36]	Wegener and Grochtmann, 1998 [130] Mueller et al., 1998 [93] Puschner et al. [103] Wegener et al., 2000 [133] Gro et al., 2000 [64]	Di Penta, 2007 [98] Garoussi, 2006 [52] Garoussi, 2008 [53] Garoussi, 2010 [54]
Simulated Annealing (SA)			Tracey, 1998 [122]
Constraint Programming			Alesio, 2014 [41] Alesio, 2013 [40]
GA + Constraint Programming			Alesio, 2015 [12]
Customized Algorithm		Pohlheim, 1999 [101]	

The studies can be grouped into two main groups:

- Search-Based Stress Tesing on Safety-critical systems.
- Search-Based Stress Testing on Industrial systems.

3.4.1 Search-Based Stress Tesing 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 performance requirements specifying constraints on how the system should react to its environment, and how it should execute on its hardware platform [40].

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, or the synchronisation of parallel processes are of major importance for the correct function of real-time systems [71].

The concurrent nature of embedded software makes the order of external events triggering the system's tasks 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 [40].

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 [111].

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 greatest extent possible [111]. The followed approaches use metaheuristics to discover the worst-case execution times.

Wegener et al. [132] used genetic algorithms (GA) 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 microseconds [132]. Alander et al. [9] performed experiments in a simulator environment to measure response time extremes 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 [9].

Wegener and Grottmann performed a experimentation to compare GA with random testing. The fitness function used was duration of execution 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 [71] [130].

Gro et. al. [64] 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 [64].

Briand et al. [32] 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 genetic algorithm. Two case studies were conducted and results illustrated that RTTT was a useful tool to stress a system under test [32].

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[101].

Garousi presented a stress test methodology aimed at increasing 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 [52].

Alesio, Nejati and Briand describe a 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 Genetic Algorithms (GA). In particular, wthe study compares CP and GA in five case studies for efficiency, effectiveness, and scalability. The experimental results show that, on the largest 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 [40].

Alesio describe stress test case generation as a search problem over the space of task arrival times. The research search for worst case scenarios maximizing deadline misses where each scenario characterizes a test case. The paper combine two strategies, GA and Constraint Programming (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 [12].

3.4.2 Search-Based Stress Testing on Industrial 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 Agreements violations.

Tracey et al. [122] used simulated annealing (SA) to test four simple programs. The results of the research presented that the use of SA was more effective with larger parameter space. The authors highlighted the need of a detailed comparison of various optimization techniques to explore WCET and BCET of the of the system under test [122].

Di Penta et al. [98] used GA to create test data that violated 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. The second case study, a service producing charts, applied the black-box approach with fitness calculated only on the basis of how close solutions violate QoS constraint. The genome representation is presented in Fig 3-3. The representation models a wsdl request to a webservice.

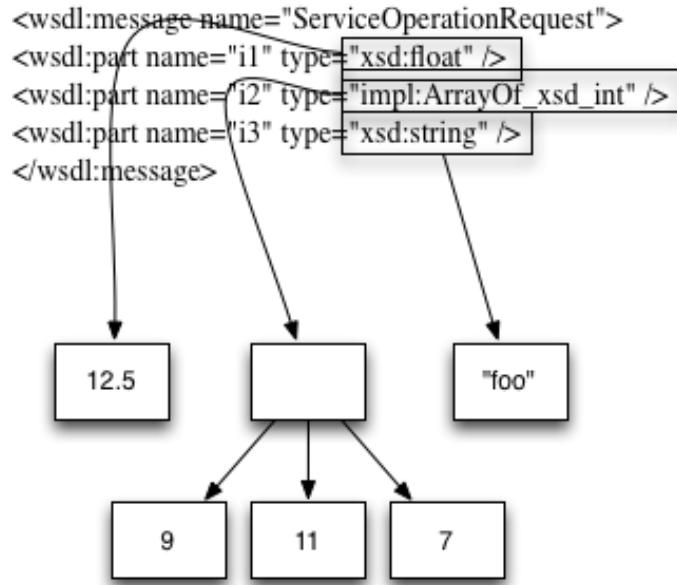


Figure 3-3: Genome representation [98].

In case of audio workflow, the GA outperformed random search. For the second case study, use of black-box approach successfully violated the response time constraint, show-

ing the violation of QoS constraints for a real service available on the Internet [98].

Gois et al. proposes an 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 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 [60].

Chapter 4

Metaheuristics

Following the complexity of the problem, it may be solved by an exact method or an approximate method. Exact methods obtain optimal solutions and guarantee their optimality. Approximate (or heuristic) methods generate high quality solutions in a reasonable time for practical use, but there is no guarantee of finding a global optimal solution [121].

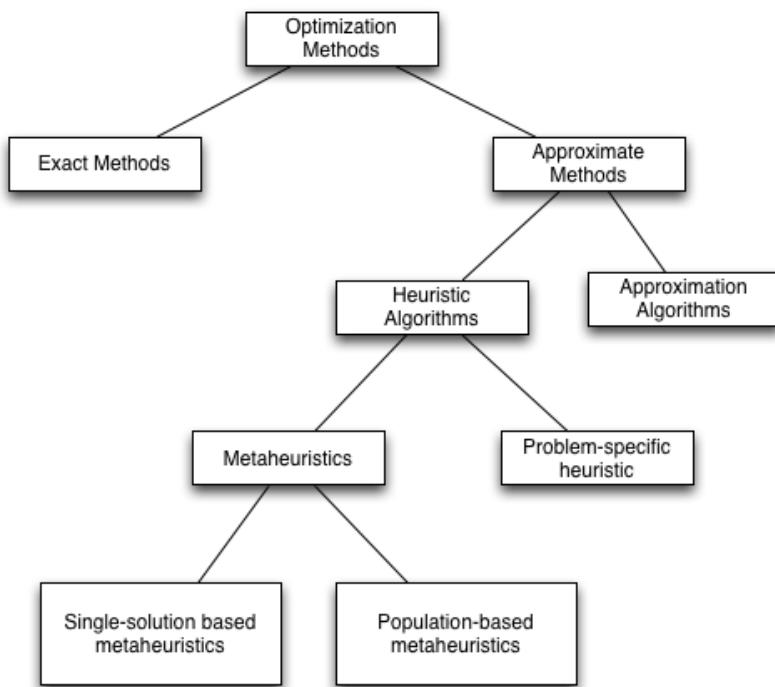


Figure 4-1: Classical optimization methods [121].

In the computer science, the term metaheuristic is accepted for general techniques

which are not specific to a particular problem. A metaheuristic is formally defined as an iterative generation process which guides a subordinate heuristic by combining intelligently different concepts for exploring and exploiting the search space [104].

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 [29]:

- 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. 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.

4.1 Single-Solution Based Metaheuristics

While solving optimization problems, single-solution based metaheuristics improve a single solution. They could be viewed as "walks" through neighborhoods or search trajectories through the search space of the problem at hand.

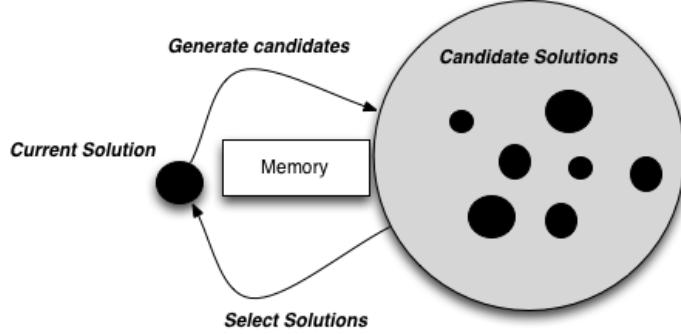


Figure 4-2: Main principles of single-based metaheuristics.

4.1.1 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 $N(s) \subset S$ [121].

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 [121]. The Fig. 4-3 presents a example where a set of neighbors is found by permutation.

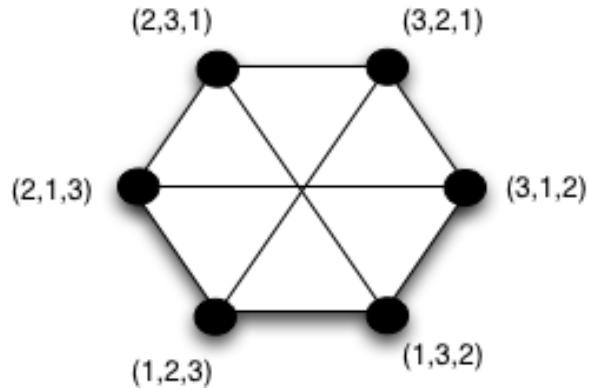


Figure 4-3: An example of neighborhood for a permutation [121].

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.

4.1.2 Simulated Annealing

Simulated Annealing (SA) is a randomized algorithm that tries to avoid being trapped in local optimum solution by assigning probabilities to deteriorating moves. The SA procedure is inspired from the annealing process of solids. SA is based on a physical process in metallurgy discipline or solid matter physics. Annealing is the process of obtaining low energy states of a solid in heat treatment [72].

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 [104].

Algorithm 1 Simulated Annealing Algorithm

```
1:  $s \leftarrow \text{GenerateInitialSolution}()$ 
2:  $k \leftarrow 0$ 
3:  $Tk \leftarrow \text{SetInitialTemperature}()$ 
4: while termination conditions not met do
5:    $s_1 \leftarrow \text{PickNeighborAtRandom}(N(s))$ 
6:   if  $(f(s_1) < f(s))$  then
7:      $s \leftarrow s_1$ 
8:   else Accept  $s_1$  as new solution with probability  $p(s_1|Tk,s)$ 
9:   end if
10:   $K \leftarrow K + 1$ 
11:   $Tk \leftarrow \text{AdaptTemperature}()$ 
12: end while
```

4.1.3 Tabu Search

Tabu Search (TS) is a metaheuristic that guides a local heuristic search procedure to explore the solution space beyond local optimal and search with short term memory to avoid cycles.

Tabu Search uses a tabu list to keep track of the last moves, and don't allow going back to these [59].

The basic idea of TS is the explicit use of search history, both to escape from local minima and to implement a strategy for exploring the search space. A basic TS algorithm uses short term memory in the form of socalled tabu lists to escape from local minima and to avoid cycles [106].

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, \dots, 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 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($TL_1, \dots, TL_r, s, s_1$)* the corresponding features of this solution are added to the tabu lists.

Algorithm 2 Tabu Search Algorithm

```

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

```

4.2 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 [119].

4.2.1 Genetic Algorithms

Genetic Algorithms could be a mean of solving complex optimization problems that are often NP Hard. GAs are based on concepts adopted from genetic and evolutionary theories. GAs are comprised of several components [70] [111] :

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

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 [104].

Algorithm 3 Genetic Algorithm

```

 $s \leftarrow GenerateInitialSolution()$ 
Evaluate(P)
3: while termination conditions not met do
     $P_1 \leftarrow Recombine(P)$ 
     $P_2 \leftarrow Mutate(P_1)$ 
    6:   Evaluate( $P_2$ )
     $P \leftarrow Select(P_2, P)$ 
end while

```

4.3 Hybrid Metaheuristics

However, in recent years it has become evident that the concentration on a sole metaheuristic is rather restrictive. A skilled combination of a metaheuristic with other optimization

techniques, a so called hybrid metaheuristic, can provide a more efficient behavior and a higher flexibility when dealing with real-world and large-scale problems [120].

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 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 [104].

There are two main categories of metaheuristic combinations: collaborative combinations and integrative combinations. These are presented in Fig. 4-4 [105].

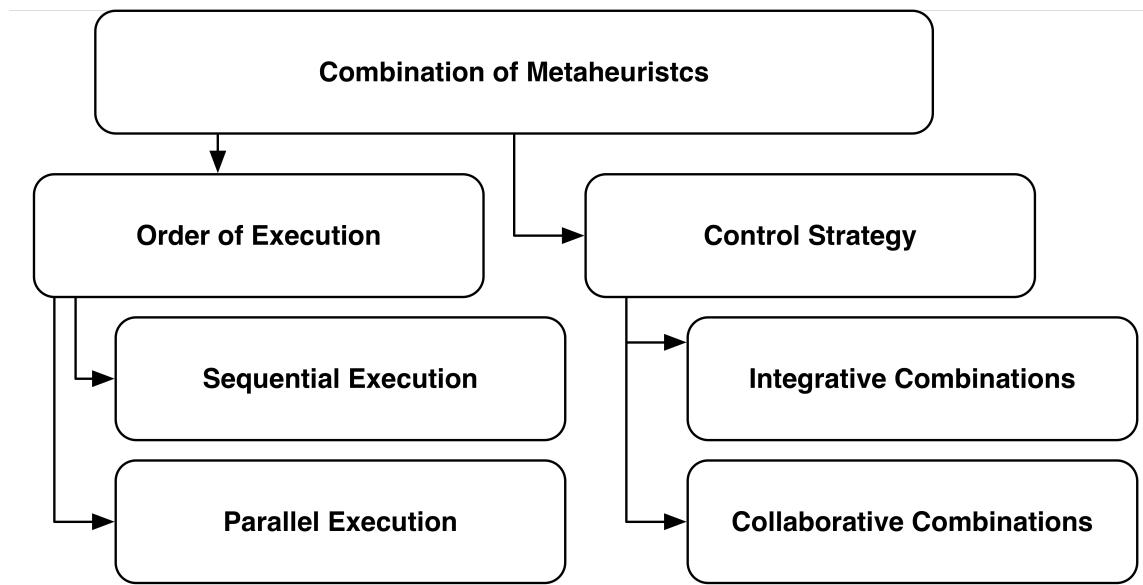


Figure 4-4: Categories of metaheuristic combinations [102]

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 [104].

The work uses a type of collaborative combination with sequential execution with two trajectory methods (Tabu Search and Simulated Annealing) and Genetic Algorithms.

4.4 Multi-Objective Optimization

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 optimised are often in competition with one another and may be contradictory; we may find ourselves trying to balance the different optimisation objectives of several different goals [67] [?].

The image of all the efficient solutions for is called 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. 4-5. Multiobjective 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.

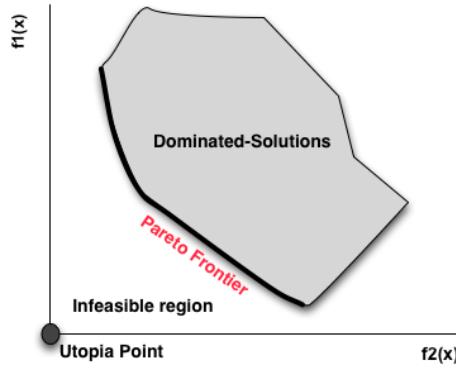


Figure 4-5: An minimization pareto front example

What distinguishes multi-objective EAs from singleobjective 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 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.

Multi-objective metaheuristics rank individuals according to how much they contribute to the above goals.

4.4.1 NSGA-II Multiobjective heuristics

Deb et al. proposed the NSGA II taking into account the need to reduce computational complexity in non-dominated classification, introduce elitism and eliminate subjectivity in the allocation of the sharing parameter [?]. NSGA II (Non-dominated Sorting Genetic Algorithm II) is a multiobjective algorithm, based on Genetic Algorithms and that implements the concept of Dominance, that is, 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, points 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 ??- **①** and **②**). The population are ranked using multiple tournament selection, which consists of comparing two solutions (Figure ??- **③**). 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 solution is calculated over 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 wins another solution if:

- Solution i has a better rank, then $Rank_i < Rank_j$.

- Both solutions have the same rank, but i has a Distance better 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 Genetic Algorithms (Figure ??- ④). 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 not dominated) in the later population (Figure ??- ⑥).

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 the last spaces of the current population, guaranteeing the diversity of the population [?]. The NSGA-II algorithm creates a set of front lines, 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 better fitness value than the highest ranked individual of the next front [?].

Chapter 5

Q-Learning

5.1 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. A typical setting where reinforcement learning operates is shown in Figure 5-1: 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.

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 [117]. Reinforcement learning is learning what to do—how to map situations to actions—so as to maximize a numerical reward signal. The learner need to discover which actions yield the most reward by trying them [117].

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), which have become the dominant approach in Reinforcement Learning. There are two types of learning problems:

- Interactive learning;

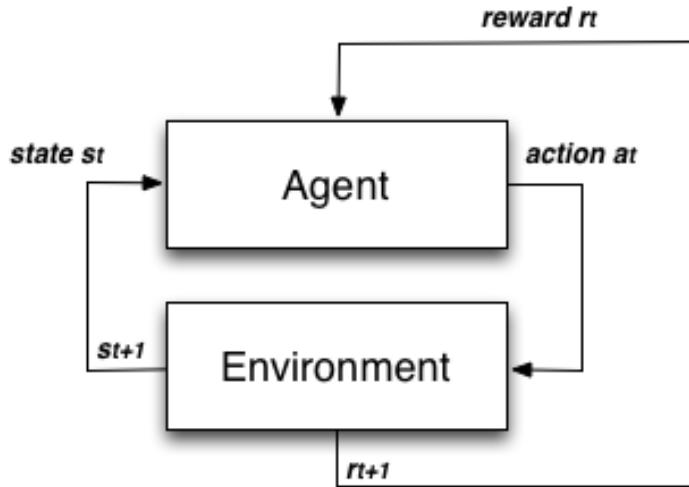


Figure 5-1: Example of interation between some agent and the environment

- Non-interactive learning.

In non-interactive learning, the natural goal is to find a good policy given a fixed number of observations. A common situation is when the sample is fixed. For example, the sample can be the result of some experimentation with some physical system that happened before learning started.

In Interactive learning, learning happens while interacting with a real system in a closed-loop fashion. A reasonable goal then is to optimize online performance, making the learning problem an instance of online learning. Online performance can be measured in different ways. A natural measure is to use the sum of rewards incurred during learning.

Interactive learning is potentially easier since the learner has the additional option to influence the distribution of the sample. However, the goal of learning is usually different in the two cases, making these problems incomparable in general.

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

One of the challenges that arise in reinforcement learning is the trade-off between exploration and exploitation. To obtain a lot of reward, a reinforcement learning agent must prefer actions that it has tried in the past and found to be effective in producing reward. But

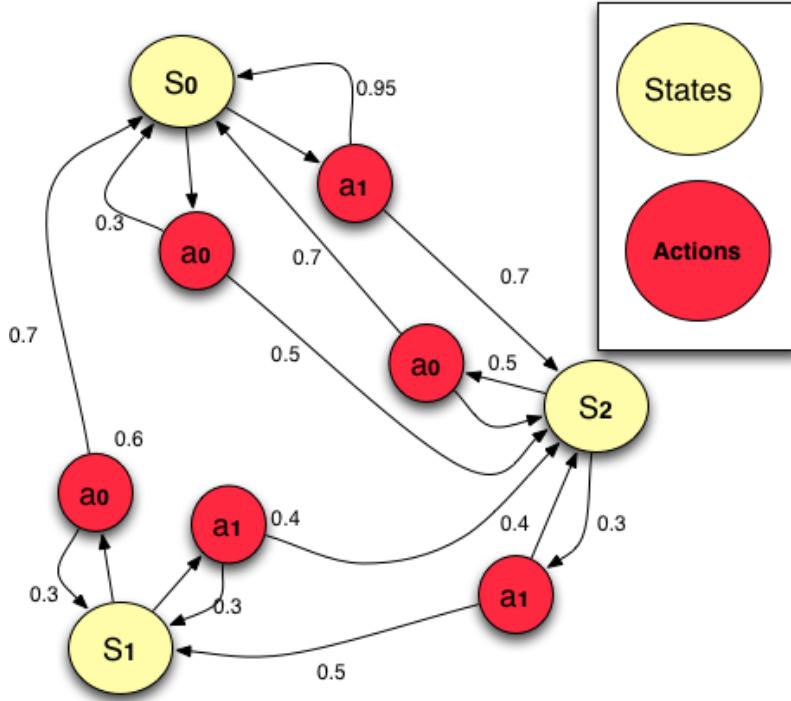


Figure 5-2: Example of a simple MDP with three states and two actions

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 reward, but it also has to explore in order to make better action selections in the future.

5.1.1 Markov decision processes

Markov decision processes (MDPs) provide a mathematical framework for modeling decision making. A countable MDP is defined as a triplet $M = (\chi, A, P_0)$ [118], where χ is a set of states, A is a set of actions. The transition probability kernel P_0 assigns to each state-action pair $(x, a) \in \chi \times A$

The six main elements of a MDP are:(1) state of the system, (2) actions, (3) transition probabilities, (4) transition rewards, (5) a policy, and (6) a performance metric [117].

The state of a system is a parameter or a set of parameters that can be used to describe a system. For example the geographical coordinates of a robot can be used to describe its state. A system whose state changes with time is called a dynamic system. Then it is not hard to see why a moving robot produces a dynamic system.

Actions are the controls allowed for an agent. Transition Probability denotes the probability of going from state i to state j under the influence of action a in one step. If an MDP has 3 states and 2 actions, there are 9 transition probabilities per action. Usually, the system receives an immediate reward ,which could be positive or negative, when it transitions from one state to another

A policy defines the learning agent's way of behaving at a given time. Roughly speaking, a policy is a mapping from perceived states of the environment to actions to be taken when in those states. It corresponds to what in psychology would be called a set of stimulus-response rules or associations. Policies mapping from states to actions.

Performance Metric: Associated with any given policy, there exists a so-called performance metric — with which the performance of the policy is judged. Our goal is to select the policy that has the best performance metric.

5.2 Q-Learning

Q-learning is a model-free reinforcement learning technique. Q-learning, it is a multiagent learning algorithm that learns equilibrium policies in Markov games, just as Q-learning learns to optimal policies in Markov decision processes [63].

Q-learning and related algorithms tries 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 it is a state, a_n it is an action and r_n is a reward:

$$< s_0, a_0, r_1, s_1, a_1, r_2, s_2, a_2, r_3, s_3, a_3, r_4, s_4, \dots >, \quad (5.1)$$

In Q-Learning, the system's objective is to learn a control policy $\pi = \sum_{n=0}^{\infty} \gamma^n r_t + n$, where π is the discounted cumulative reward, γ is the discount rate (01) and r_t is the reward received after execution an action at time t. The fig. 5-3 shows the summary version of Q-Learning algorithm. The first step it is to generate the initial state of the MDP. The second step it is to choose the best action or a random action based on the reward, the actions with best rewards are chosen.

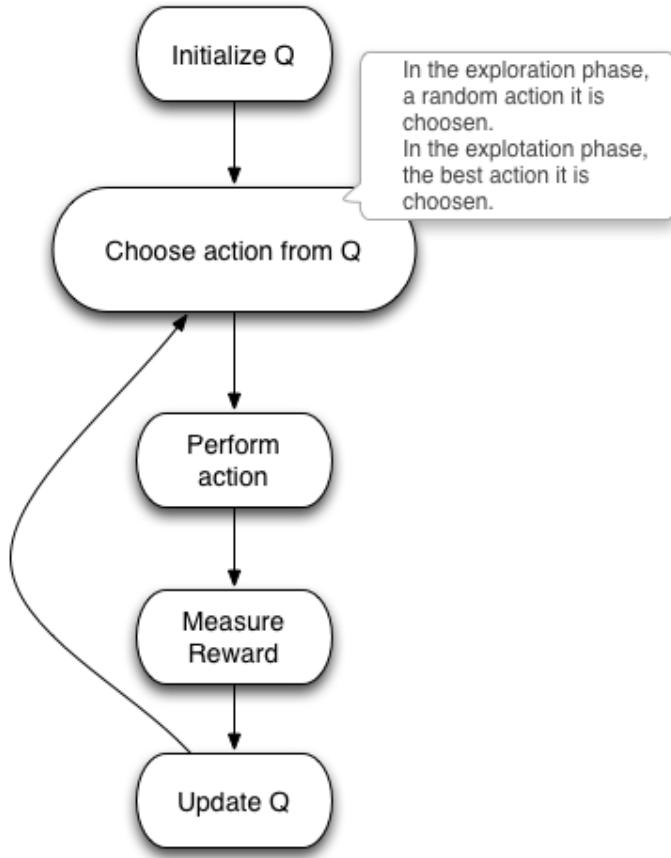


Figure 5-3: Q Learning algorithm

5.3 GridWorld Example

The GridWorld problem is an example of using reinforcement learning with Q-Learning. In this example, The agent's goal is to reach the reward state. There are one rewarding states with value +1 (Gray square at position (3,1)). The agent receives +0.02 of reward if it get closer to the reward state As the agent moves away from the reward state, he receives -0.01 of reward. The agent receive +0 points of reward if it is the same distance to the reward state.

The Fig. 5-4 shows the initial and final phase on exploration phase. The numbers in the squares shows the Q-values for four available actions: up (u), left (l), right(r) and down(d). The arrows show the optimal action based on the current value function. The initial discount rate is 0.9.

The figure B-1,B-2, B-3 and B-4 shows the states of the game in exploration phase. In

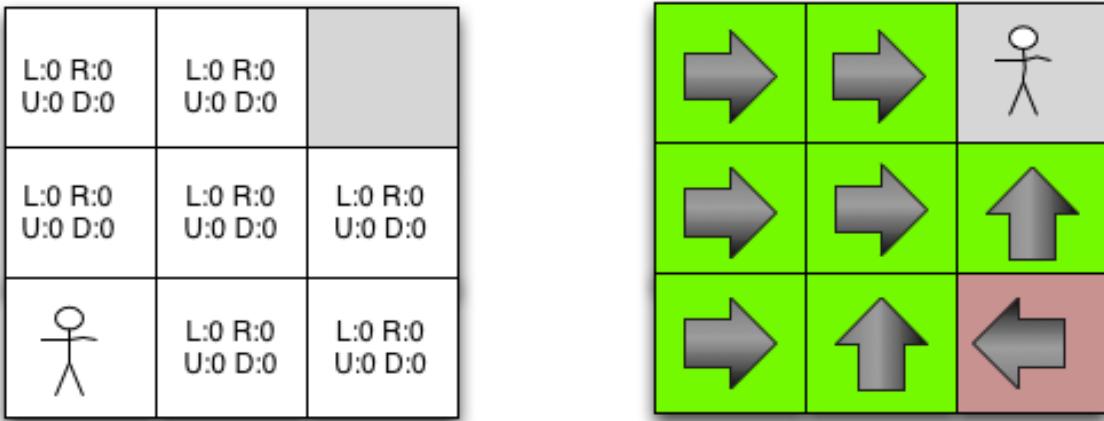


Figure 5-4: GridWorld - initial and final stage on exploration phase

the first image, all frames have the accumulated reward value equals to zero. In the next images the reward values increase or decrease as the agent moves.

5.4 Relationships Between Reinforcement Learning and Optimization

Many are the intersections between optimization and Reinforcement Learning. First, approximated versions of optimization and RL tasks contain challenging optimization tasks, the maximization operations in determining the best action when an action value function is available or the optimal choice of approximation [25]. Both optimization and Machine Learning analysts address real world problems by formulating a model, deriving the core optimization problem, and using mathematical programming to solve it [28] [31].

Gambardella et al. show an Ant metaheuristic with Q-Learning approach (Ant-Q). They experimentally investigate the functioning of Ant-Q and we show that the results obtained by Ant-Q on symmetric TSP's are competitive with those obtained by other heuristic approaches based on neural networks or local search. The research used Q-Learning to indicate how useful it is to make move in TSP for each ant [50]. Bianchi et al. present a new algorithm, called Heuristically Accelerated Q-Learning (HAQL), that allows the use of heuristics to speed up the well-known Reinforcement Learning algorithm Q-learning.

The Heuristically Accelerated Q-Learning algorithm can be defined as a way of solving the RL problem which makes explicit use of a heuristic function to influence the choice of actions during the learning process. The HAQL uses a modification Greedy rule to control exploration and exploitation phases [86].

5.5 Software Testing with Reinforcement Learning

Many software systems are reactive. The behavior of a reactive system, especially when distributed or multithreaded, can be nondeterministic. In these cases, a test suite generated offline may be infeasible.

Online testing is a formal model-based testing (MBT), where the tester uses a specification of the system's behavior to guide the testing and to detect the discrepancies between the application under test and the model. Online testing can be more appropriate than offline tests for reactive systems. The reason is that with online testing the tests may be dynamically adapted at runtime, effectively pruning the search space to include only those behaviors actually observed instead of all possible behaviors. The interaction between tester and the application under test is seen as a game where the tester chooses moves based on the observed behavior of the implementation under test [68].

Havelund et al transforms online testing problem into a special case of reinforcement learning where the frequencies of various abstract behaviors are recorded and allows to better choose controllable actions [68].

Andre et al. show a method that operationalizes the risk assessment for interoperability testing. Typically high number of possible interaction scenarios makes interoperability testing a complex task. Since it seems impossible to cover all scenarios, their relevance for being tested has to be prioritized. The method uses behavior models of the system under test and reinforcement learning techniques to obtain the relevance of single system actions for being tested [100]. The table 5.1 shows the Q-value for some states of the test.

Sato and Sugihara propose an automatic test pattern generation by applying genetic algorithms in reinforcement learning. The results of evaluation with an ATM system which assuming the existence of a bug related to a global variable are reported [109].

Table 5.1: Q-value for some states of the test [100]

State	Action	Q-value
M'1N1O1	M 1 → M 2	2.5
M'2N2O2	M 2 → M 1	1.25
M'2N2O2	M 2 → M 3	0.0
M'1N1O2	M 1 → M 2	0.5

Meinke and Niu presents an application of learning techniques to the problem of automated test case generation for numerical software. Our approach uses n-dimensional polynomial models as an algorithmically learned abstraction of the application under test. Test cases are iteratively generated by applying a satisfiability algorithm to first-order program specifications over real closed fields and iteratively refined piecewise polynomial models [89]. Kamali et al. propose a approach based on Q-learning, on top of genetic algorithms (GA) to determine the best weightings for an metaheuristic [78].

Chapter 6

Improving Stress Search Based Testing using Q-Learning and Hybrid Metaheuristic Approach

6.1 Hybrid Approach

A large number of researchers have recognized the advantages and huge potential of building hybrid metaheuristics. The main motivation for creating hybrid metaheuristics is to exploit the complementary character of different optimization strategies. In fact, choosing an adequate combination of algorithms can be the key to achieving top performance in solving many hard optimization problems [102] [30].

The proposed solution makes it possible to create a model that evolves during the test. 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 in others 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. 6-1).

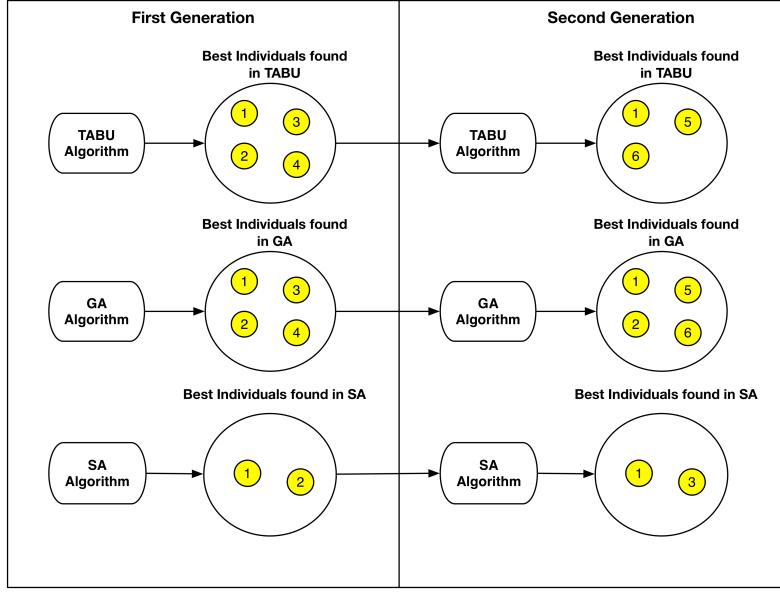


Figure 6-1: Use of the algorithms independently

The second approach uses the algorithms in a collaborative mode (hybrid metaheuristic). In this approach, the three algorithms share their best individuals found (Fig. 6-2). The next subsections present details about the used metaheuristic algorithms (Representation, initial population and fitness function).

6.1.1 Representation

The solution representation is composed by a linear vector with 23 positions. The first position represents the name of an individual. The second position represents the algorithm (genetic algorithm, simulated annealing, or Tabu search) used by the individual. The third position represents the type of test (load, stress, or performance). The next positions represent 10 scenarios and their numbers of users. Each scenario is an atomic operation: the scenario must log into the application, run the task goal, and undo any changes performed, returning the application to its original state.

Fig. 6-3 presents the solution representation and an example using the crossover operation. In the example, genotype 1 has the Login scenario with 2 users, the Form scenario with 0 users, and the Search scenario with 3 users. Genotype 2 has the Delete scenario with 10 users, the Search scenario with 0 users, and the Include scenario with 5 users. After the

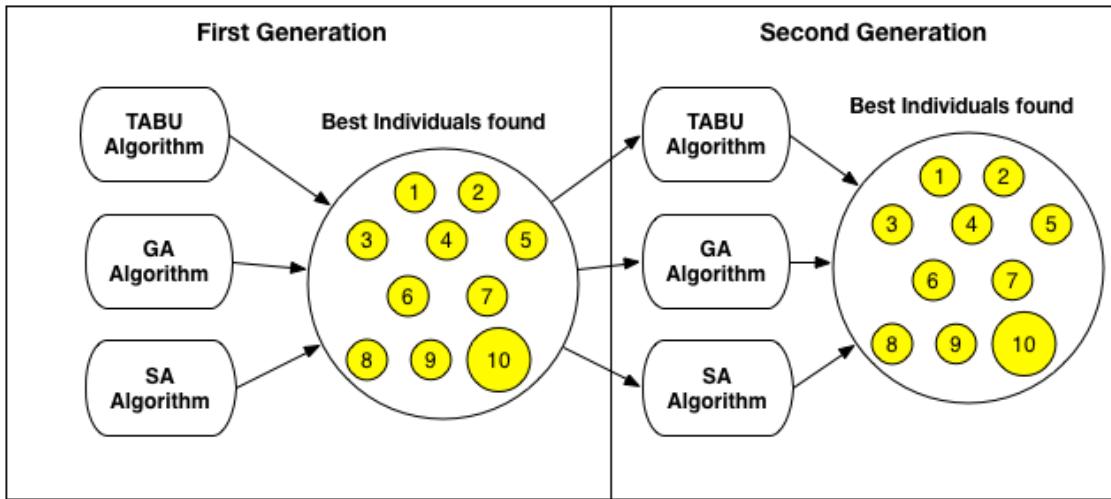


Figure 6-2: Use of the algorithms collaboratively

crossover operation, we obtain a genotype with the Login scenario with 2 users, the Search scenario with 0 users, and the Include scenario with 5 users.

Fig. 6-4 shows the strategy used by the proposed solution to obtain the representation of 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.

6.1.2 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.

The percentages relates 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.

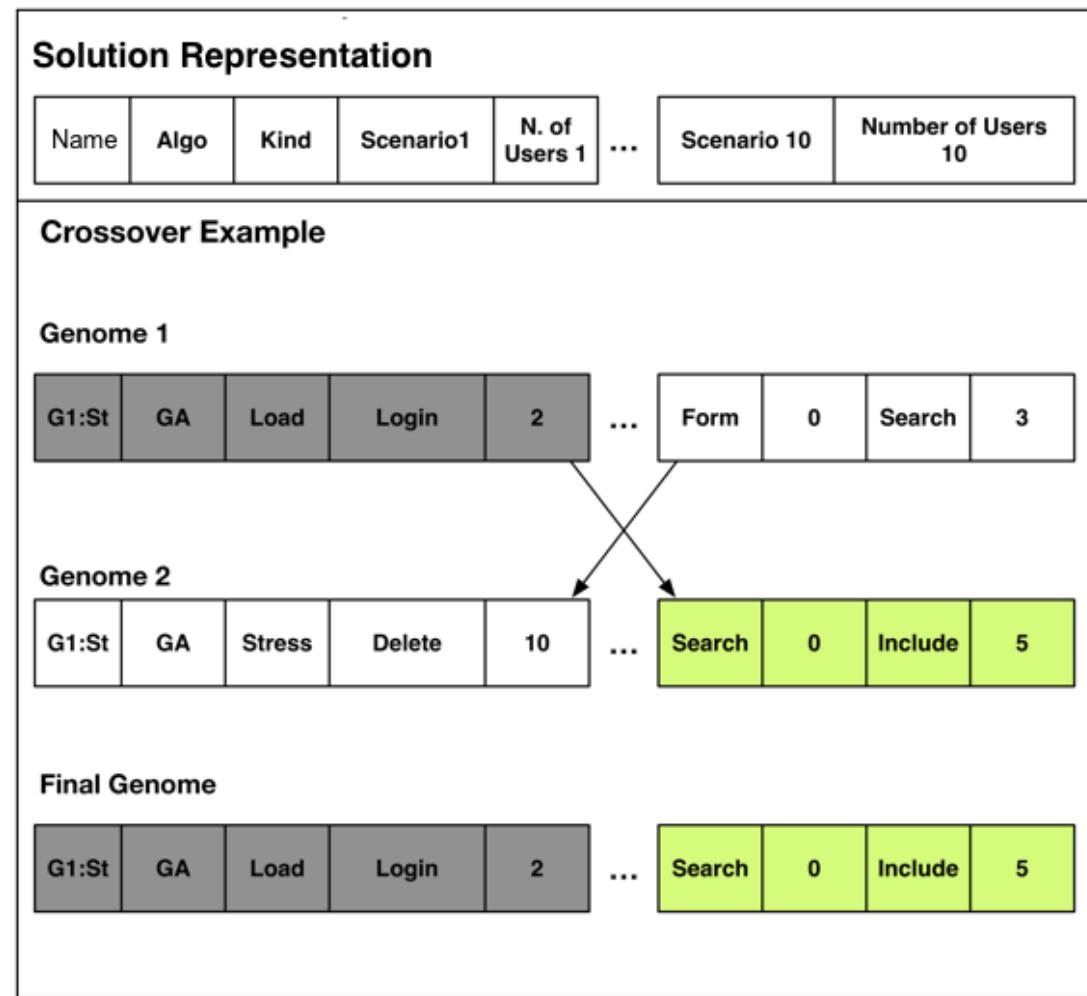


Figure 6-3: Solution representation and crossover example

6.1.3 Objective (fitness) function

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

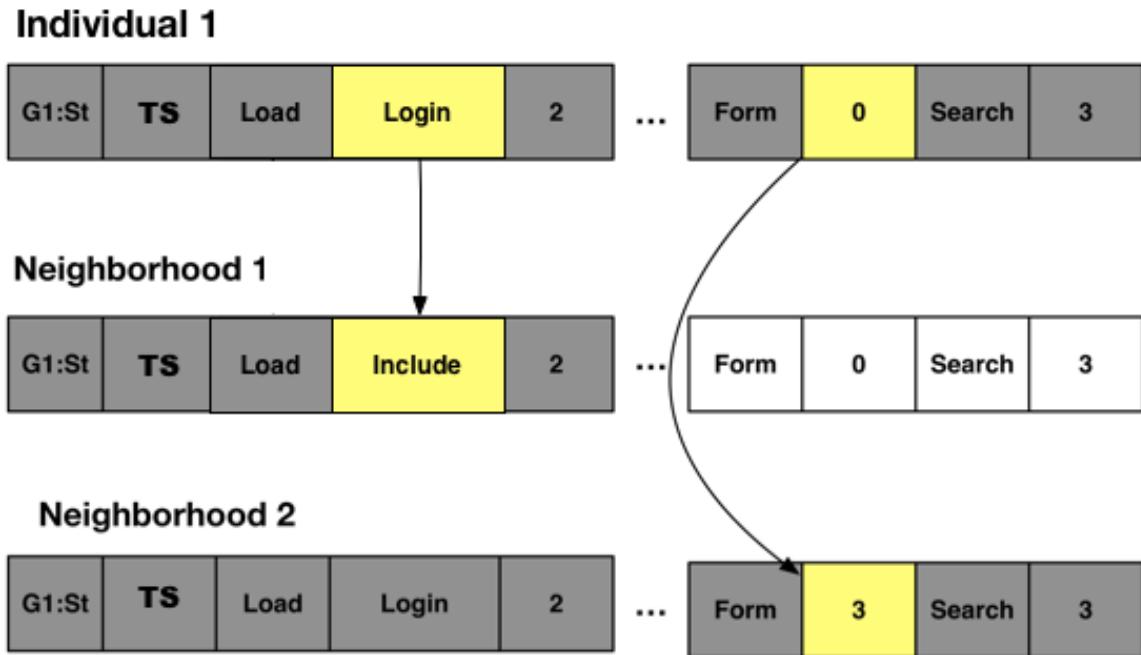


Figure 6-4: Tabu search and simulated annealing neighbor strategy

$$\begin{aligned}
 fit = & 90\text{percentileweight} * 90\text{percentiletime} \\
 & + 80\text{percentileweight} * 80\text{percentiletime} \\
 & + 70\text{percentileweight} * 70\text{percentiletime} + \\
 & maxResponseWeight * maxResponseTime + \\
 & numberOfUsersWeight * numberOfUsers - penalty
 \end{aligned} \tag{6.1}$$

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 an application under test takes a longer time to respond than the level of service. The penalty is calculated by the following equation:

$$\begin{aligned}
 penalty = & 100 * \Delta \\
 \Delta = & (t_{CurrentResponseTime} - t_{MaximumResponseTimeExpected})
 \end{aligned} \tag{6.2}$$

6.2 Hybrid Metaheuristic with Q-Learning Approach

The HybridQ algorithm uses the GA, SA and Tabu Search algorithms in a collaborative approach in conjunction with Q-Learning technique. The biggest difference between the Hybrid and HybridQ algorithms is the application of a series of modifications on individuals based on the Q-Learning algorithm before each generation.

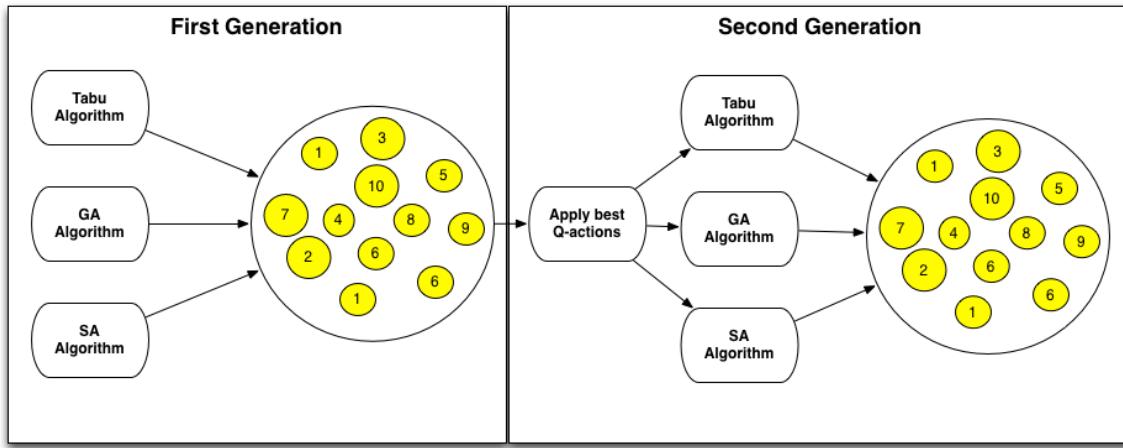


Figure 6-5: Hybrid Metaheuristic with Q-Learning Approach

Figure 6-6 shows the proposed MDP model for HybridQ. The 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. A test receives a positive reward when an action increases fitness value and a negative reward when an action reduces the fitness value. The possible actions in MDP are the change of one of the test scenarios and increase or decrease of the number of users.

Unlike the traditional approach, The update of Q values for each action also occurs in the exploitation phase. The exploit phase ends when no value of Q it is 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 exploitation phase. The table presents an hypothetical Q-values for a test. In the table 6.1, 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 equals to zero.

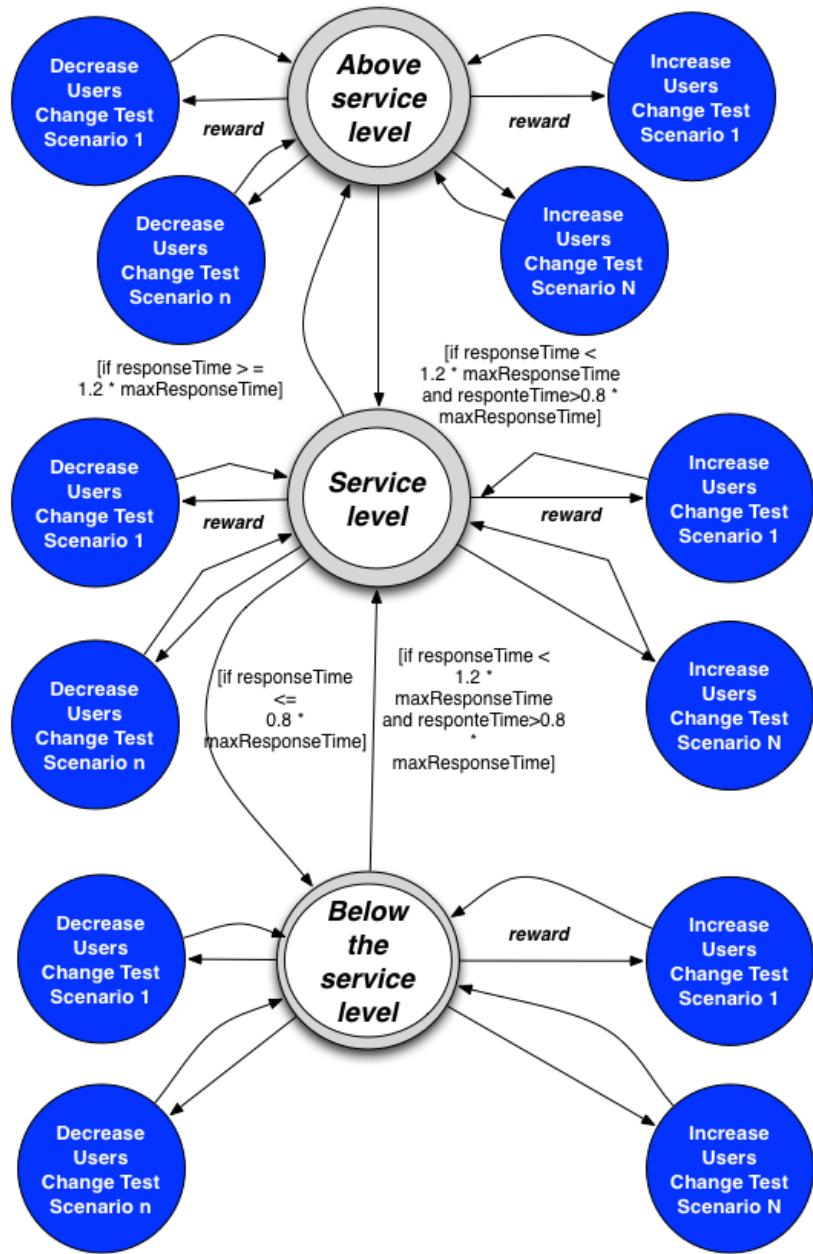


Figure 6-6: Markov Decision Process used by HybridQ

The Fig. 6-7 presents how one of the neighbors of a test is generated using Q-Learning. The solution uses a service called Q-Neighborhood Service to generate the neighbor from the action that has the highest value of Q.

Table 6.1: 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	Exploration	Exploration
Bellow Service Level	Scenario 1	Scenario 2
Increment Users	0.0	0.2
Reduce Users	0.1	0.0
Phase	Exploration	Exploration

6.3 IAdapter

IAdapter is a JMeter plugin designed to perform search-based stress tests. The plugin is available on www.iadapter.org. The IAdapter plugin implements the solution proposed in Section 5. The next subsections present details about the Apache JMeter tool, the IAdapter Life Cycle and the IAdapter Components. The IAdapter plugin provides three main components: WorkLoadThreadGroup, WorkLoadSaver, and WorkLoadController.

The Fig. 6-8 show the IAdapter architecture. All metaheuristic class implements the interface IAlgorithm. Test scenarios and test results are stored in a Mysql database. GeneticAlgorithm class uses a framework named JGAP to implement Genetic Algorithms.

The WorkLoadThreadGroup class is the Load Injection and Test Management modules, responsible to generate the initial population and uses the JMeter Engine to realize requests to server under test.

6.3.1 IAdapter Life Cycle

Fig. 6-9 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 chosen by the test tool. In a test with JMeter, the tests scenarios are usually chosen by a test designer.

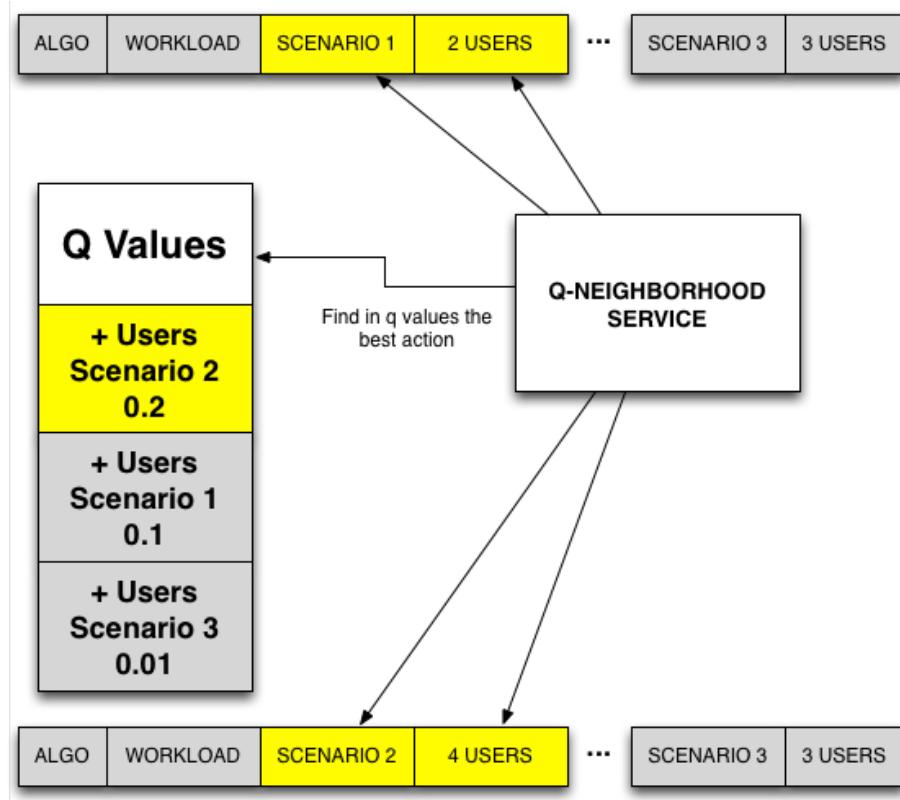


Figure 6-7: HybridQ NeighborHood Service

6.3.2 IAdapter Components

WorkLoadThreadGroup is a component that creates an initial population and configures the algorithms used in IAdapter. Fig. 6-10 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 need to login into the application under test and bring the application back to its original state.

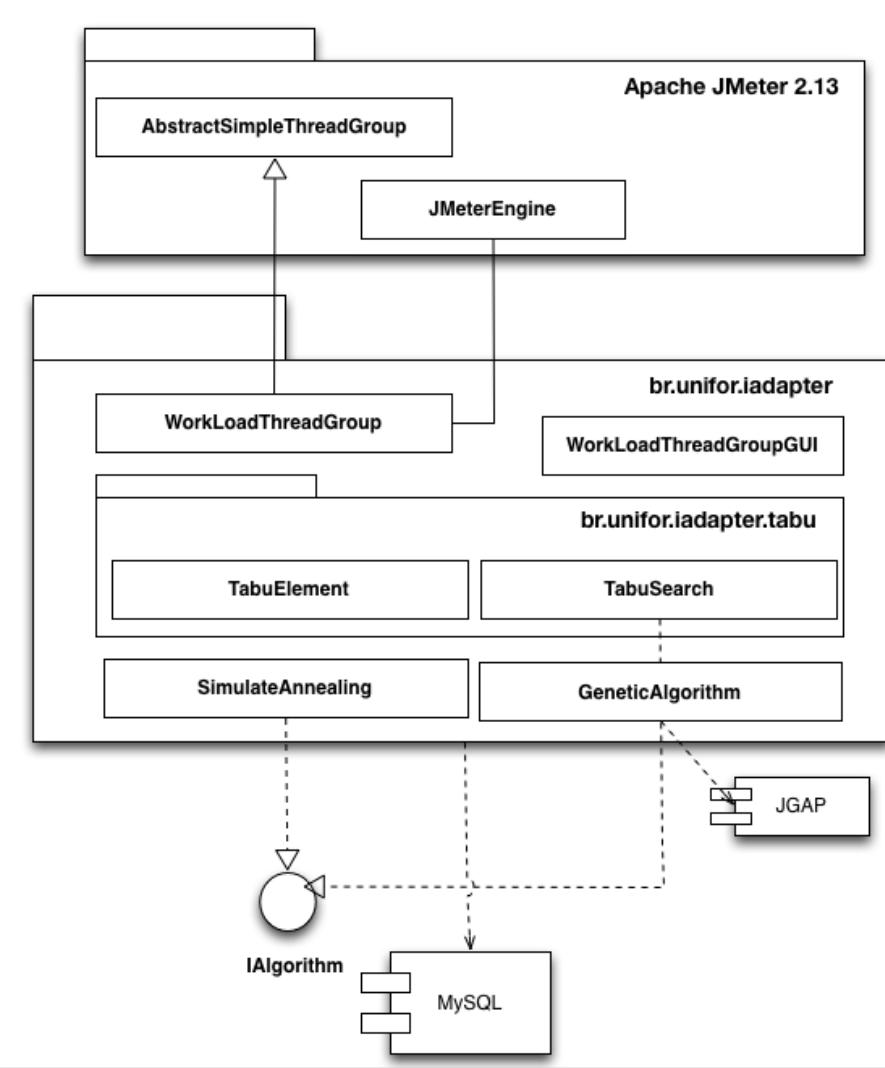


Figure 6-8: IAdapter architecture

6.3.3 IAdapter Testbed Tool

In this subsection, We devise a new testbed that has the ability to reproduce different types of web workloads. The proposed solution extends a tool named IAdapter to create a testbed tool to validate load, performance and stress search based test approaches [60]. This new testbed must accomplish three main goals. First, it must reproduce a workload by using an antipattern implementation. Second, it must be able to provide metrics with the aim of being used for research evaluation studies. Finally, it should be extensible, allowing the creation of new metaheuristic approaches.

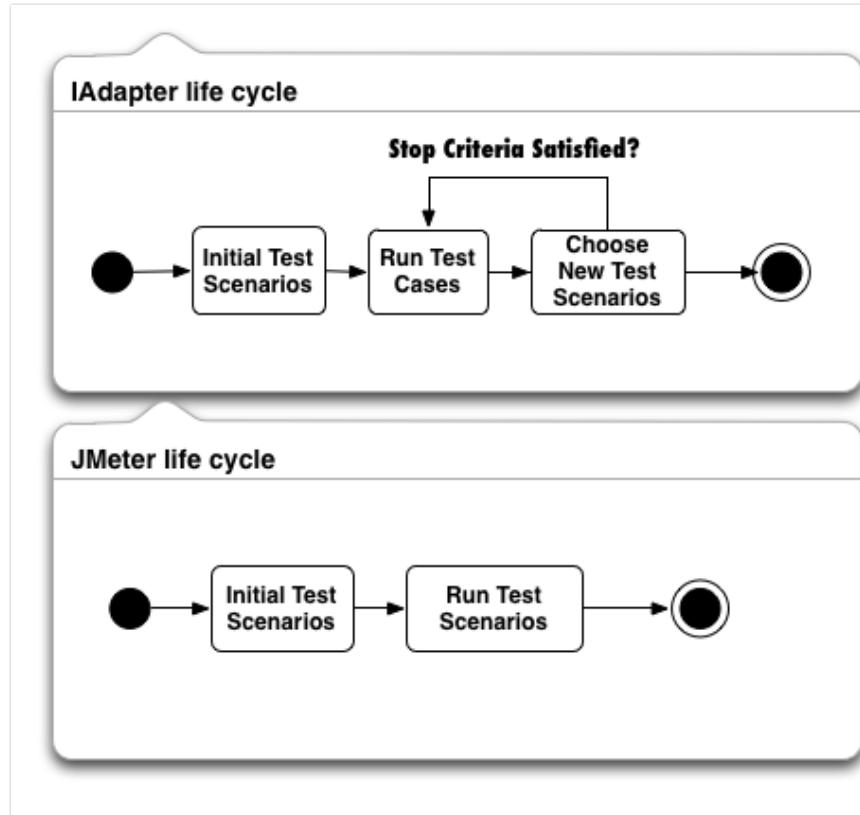


Figure 6-9: IAdapter life cycle

The testbed tool proposed consists of four main modules. Figure 6-11 presents the main architecture of the Testbed 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.

6.3.4 Test Module

The Test Module (Figure 6-11 -①) is responsible for the loading of all classes that extend AbstractAlgorithm in the classpath and perform the tests under the application. Figure 6-12 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

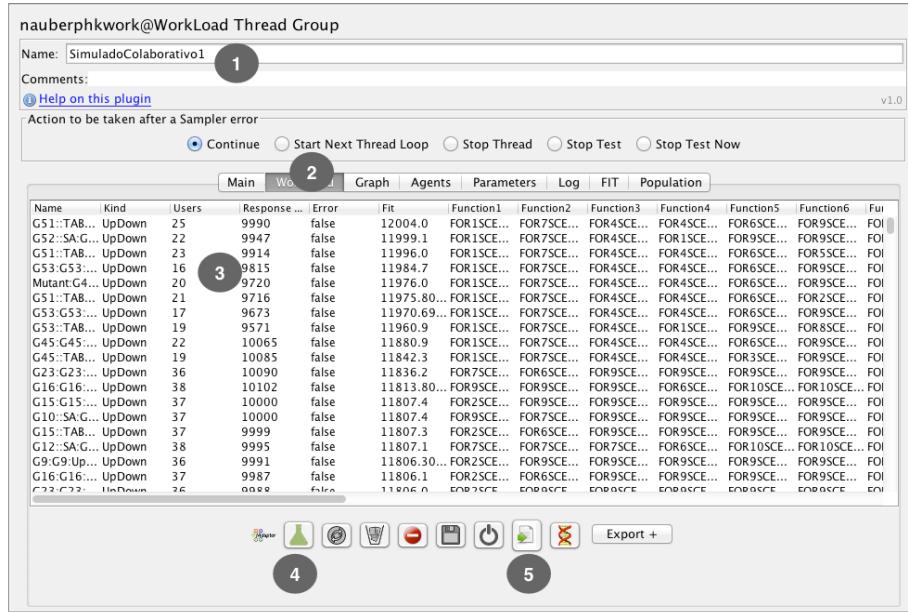


Figure 6-10: WorkLoadThreadGroup component

6-12 -②) and must return a list of output workloads (the individuals selected for the next generation) (Figure 6-12 -③). Each workload represent an individual in the search space. Figure. 6-13 presents the Test Module life cycle. Given an initial population (Figure 6-13 -①), a metaheurist select a new set of workloads based on an objective function (Figure 6-13 -②). The choosen metaheurist generate a new set of individuals based on crossover or neighborhood operators (Figure 6-13 -③). JMeterEngine run each workload (Figure 6-13 -⑤) and the choosen metaheuristic obtain a fitness value for each workload based on some objective function (Figure 6-13 -⑥). 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 6-13 -⑦).

6.3.5 Emulator Module

The Emulator Module is responsible for implementing and providing successful scenarios and the most common performance antipatterns (Figure 6-11 -②). All classes must extend the AbstractJavaSamplerClient class or use JUnit 4. The AbstractJavaSamplerClient class allows the creation of a JMeter Java Request. Figure 6-14 presents the main features of the emulator module. The module implements 2 happy scenarios (Figure 6-14 -②) and

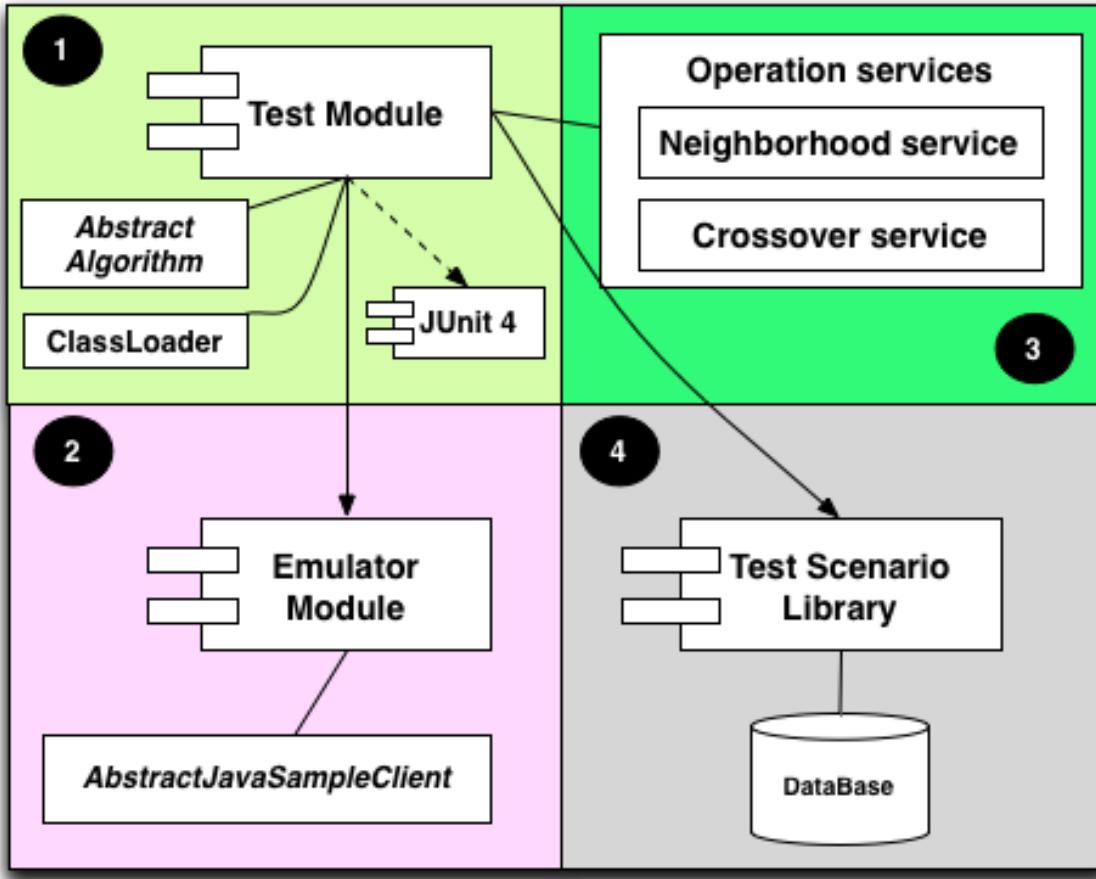


Figure 6-11: Testbed main architecture.

4 antipatterns test scenarios (Figure 6-14 -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 6-14 -❸).

6.3.6 Test Scenario Library

This modules provides a common representation for all workloads (Figure 6-11 -❷). Each workload is composed by a linear vector with 21 positions (Figure 1-2 -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 1-2 -❷). Each scenario is an atomic operation: the scenario must log into the application, run the task goal, and undo any changes performed, returning the application to its original state.

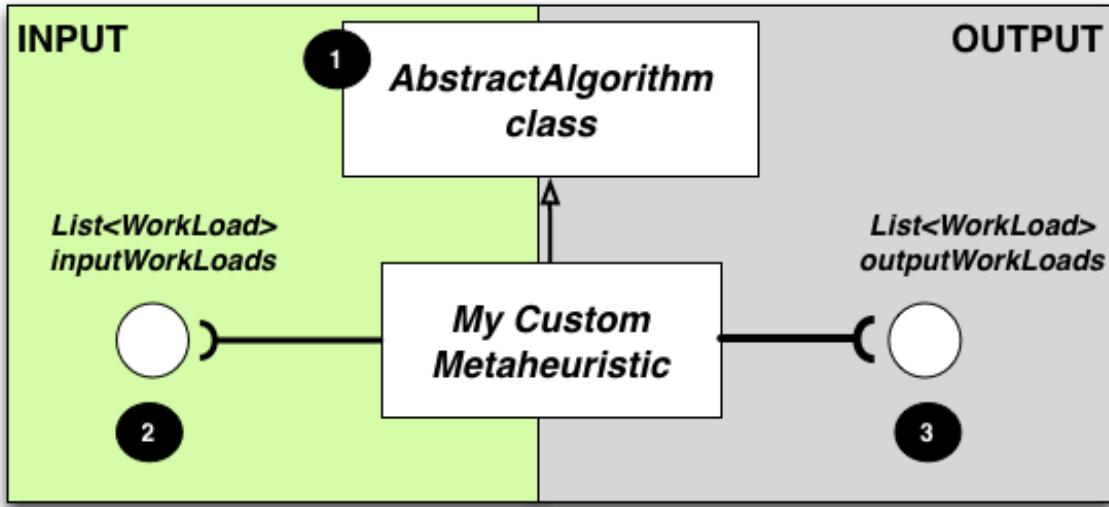


Figure 6-12: Test Module class diagram.

6.3.7 Operation services

The services are responsible for performing some operations performed by metaheuristics. Figure. 1-2 presents the solution representation and an example using the crossover operation. In the example, solution 1 (Figure 1-2 -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 1-2 -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 1-2 -❺). Figure. 1-2 -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.

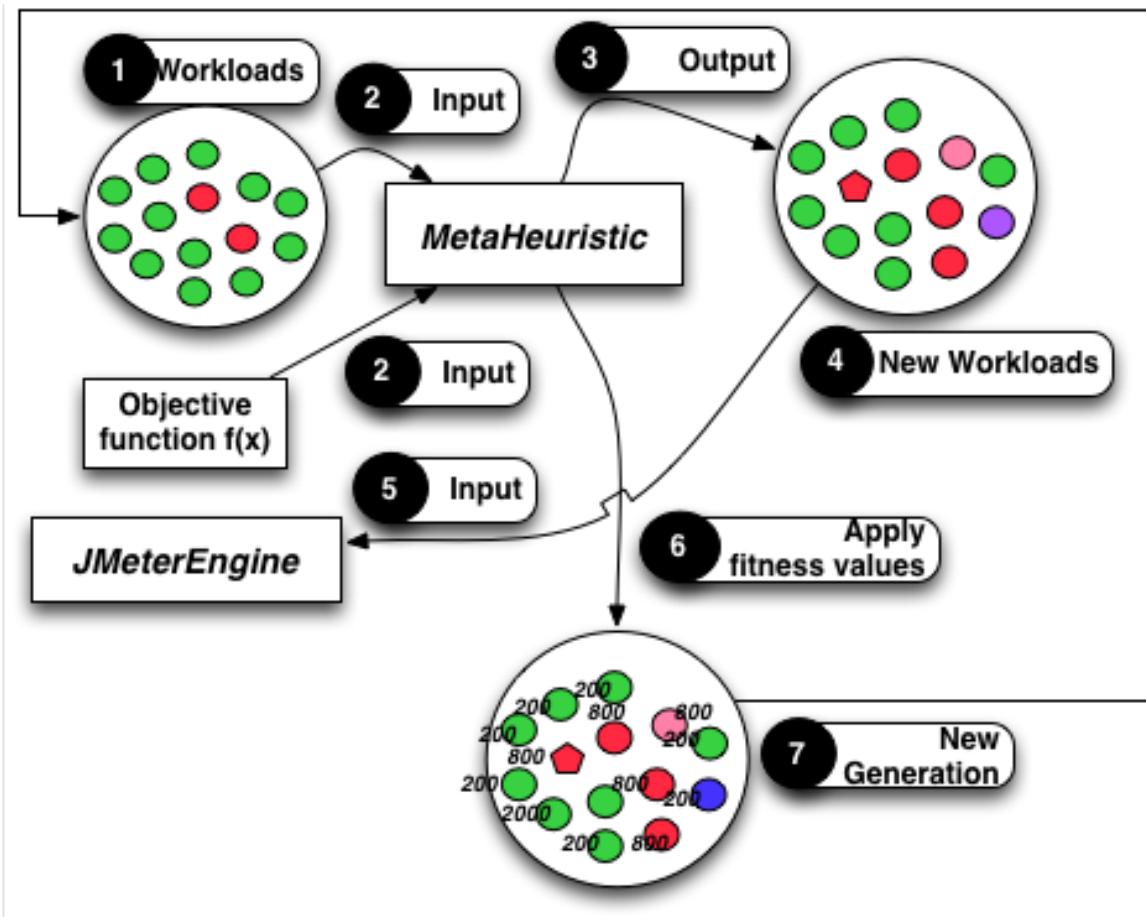


Figure 6-13: Test module life cycle.

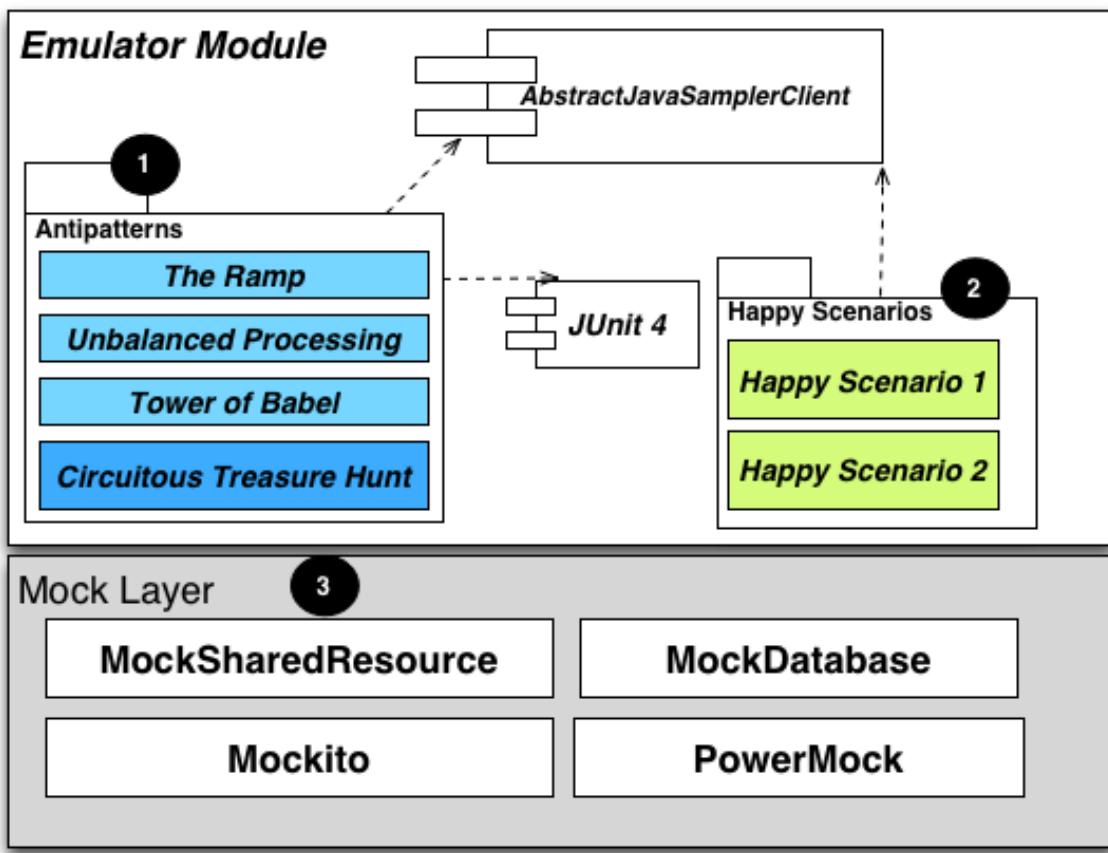


Figure 6-14: Emulator module

Chapter 7

Experiments

In this chapter, We present the results of experiments which we carried out to verify the hybrid and q-learning algorithm approaches ,the antipatterns implementation, the fitness objective function and the metaheuristics used by the testbed tool. 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 application. The fourth experiment was performed using an installed Moodle application. The fifth and sixth experiments are performned using an installed JPetStore application.

7.1 Emulated Class Test Experiment

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 7.1). The experiment was executed using the JMeter Java Request Sampler Component with IAdapter.

The experiment used the following fitness function:

Listing 7.1 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:    @Test  
12:    public void secondScenario() {  
13:        synchronized (StaticClass.class) {  
14:            for (int i = 0; i <= 2000; i++) {  
15:                StaticClass.x += i;  
16:            }  
17:            StaticClass.x = 0;  
18:        }  
19:    }  
20: }
```

$$\begin{aligned} fitness = & 0.9 * 90percentiletime \\ & + 0.1 * 80percentiletime \\ & + 0.1 * 70percentiletime + \\ & 0.1 * maxResponseTime + \\ & 0.2 * numberOfUsers - penalty \end{aligned} \tag{7.1}$$

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.

Fig.7-1 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.

Figure 7-1: Best results obtained in 27 generations

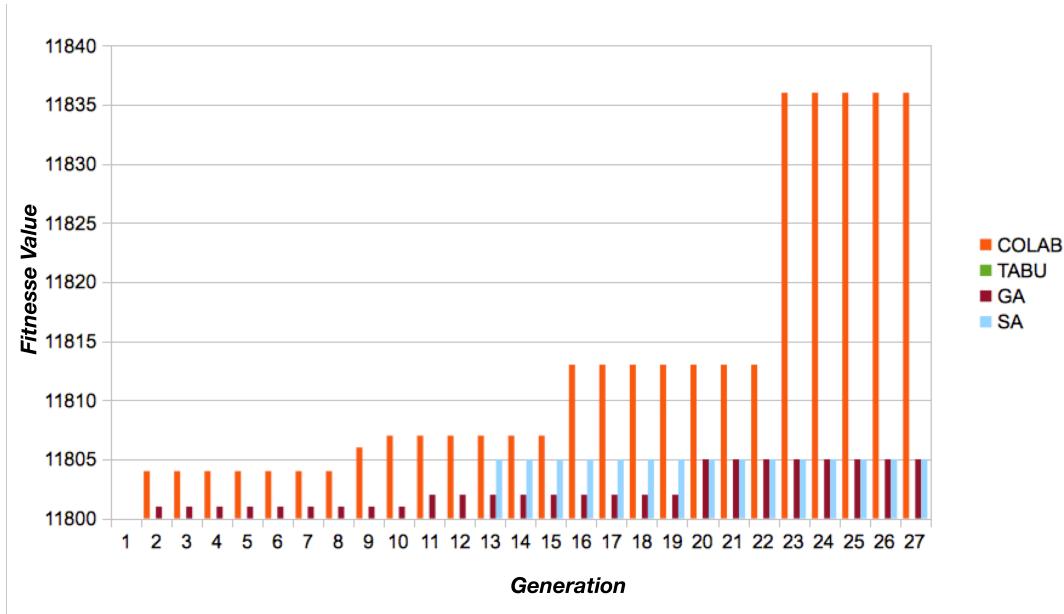


Table 7.1 presents the results obtained by the hybrid metaheuristic (HM) approach, genetic algorithm (GA), simulated annealing (SA), and Tabu search (TS) from 27 generations in the first experiment. The values are the maximum fitness value obtained by each algorithm.

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 $p \leq 0.05$ was 89. The result was significant at $p \leq 0.05$. The procedure showed that there was a significant improvement in the results with the collaborative approach.

Table 7.1: 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

7.2 Testbed Tool Experiments

We conducted two experiments in order to verify the effectiveness of the testbed tool. The experiments 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. [60].

The objective function applied is intended to maximize the number of users and minimize the response time of the scenarios being tested. In this experiments, better fitness values meaning to find scenarios with more users and a low values of response time. A penalty is applied when the response time is greater than the maximum response time expected. The experiments used the following fitness (goal) function. :

$$\begin{aligned}
 \textit{fitness} = & 3000 * \textit{numberOfUsers} \\
 & - 20 * 90\textit{percentiletime} \\
 & - 20 * 80\textit{percentiletime} \\
 & - 20 * 70\textit{percentiletime} \\
 & - 20 * \textit{maxResponseTime} \\
 & - \textit{penalty}
 \end{aligned} \tag{7.2}$$

The experiments addresses:

- Validate the operation of the testbed tool.
- Find the maximum number of users and the minimal response time.
- Analyze and verify the best heuristics among those chosen to the experiments.

7.2.1 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. The Fig. 7-2 presents the fitness value obtained by each metaheuristic. HybridQ metaheuristic obtained the better fitness values.

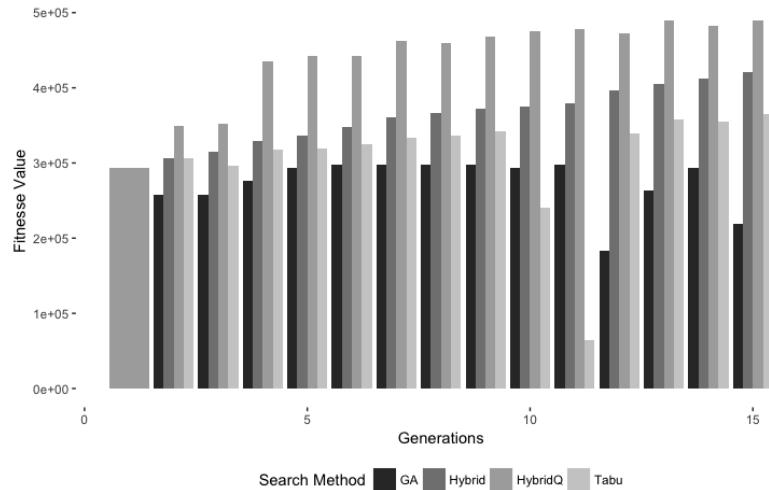


Figure 7-2: fitness value obtained by Search Method

Despite having obtained the best fitness value in each generation, the Hybrid algorithm performs twice as many requests as the tabu search (Fig. 7-3). The HybridQ algorithm obtained the best fitness value. The Fig. 7-4 shows the average, minimal e maximum value by search method.



Figure 7-3: Number of requests by Search Method

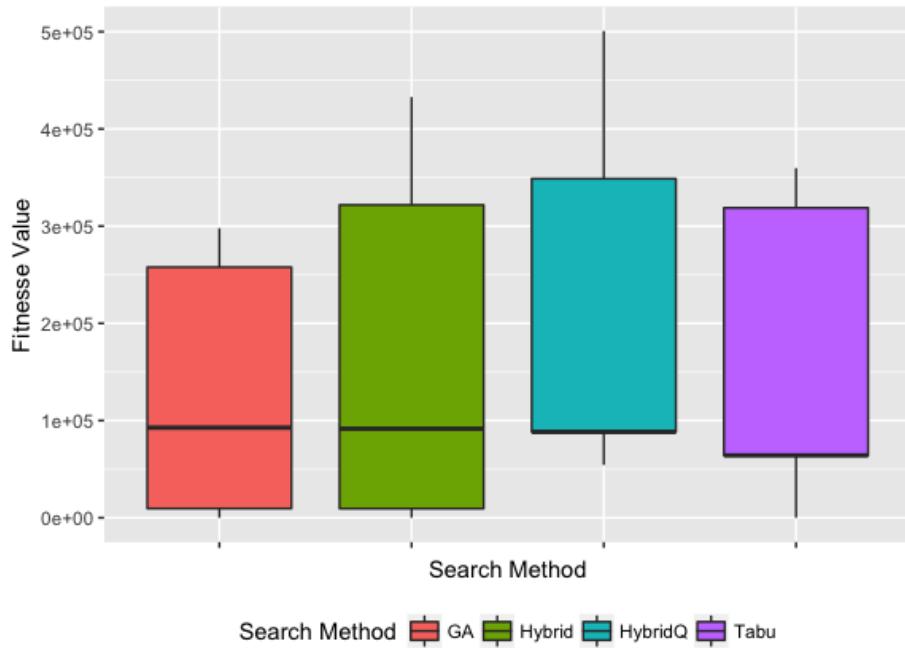


Figure 7-4: Average, median, maximum and minimal fitness value by Search Method

The Fig. 7-5 presents the maximum, average, median and minimum fitness value by generation. The maximum fitness value increases at each generation. The Fig. 7-6 presents the density graph of number of users by fitness value. The range between 100 and 150 users has the highest number of individuals found with higher fitness value.

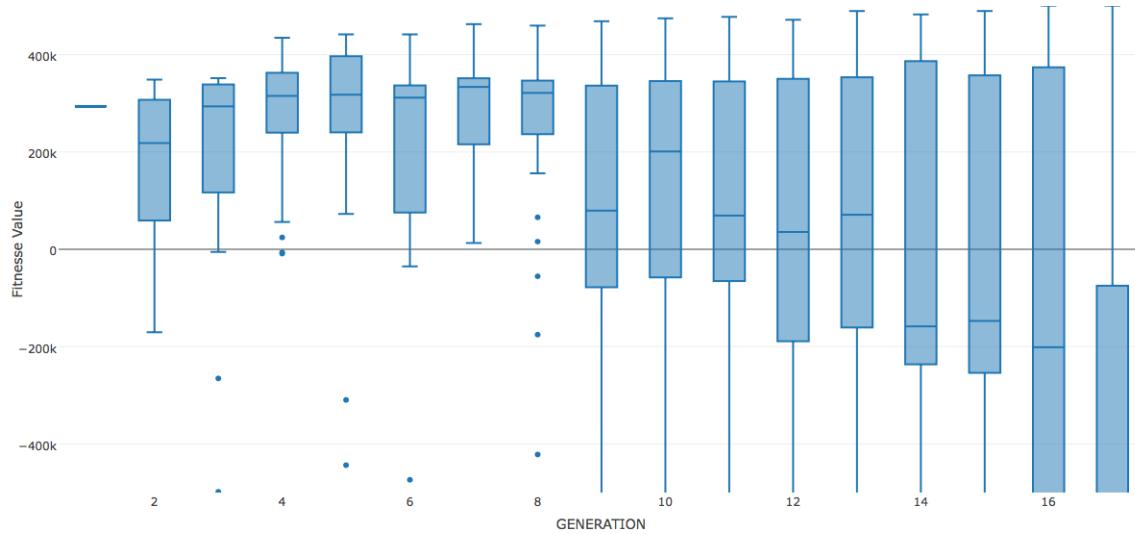


Figure 7-5: fitness value by generation

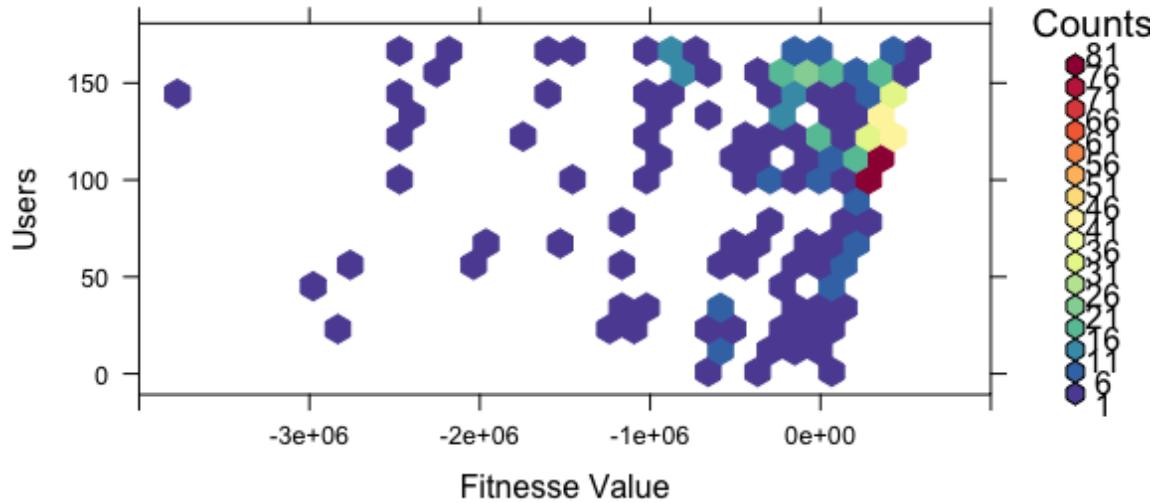


Figure 7-6: Density graph of number of users by fitness value

Table 7.2 shows 4 individuals with 164 to 169 users. These are the scenarios with the maximum number of users found with the best response time. 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 best individuals has one of the antipatterns used in the experiment.

Table 7.2: Best individuals found in the first experiment

Search Method	Generation	Users	fitness Value	Happy 2	Happy 1	Resp. Time
HybridQ	17	169	500740	153	16	13
HybridQ	16	169	500700	153	16	15
HybridQ	13	164	489740	149	15	13
HybridQ	15	164	489740	149	15	13

Fig. 7-7 presents the response time by number of users of individuals with Happy Scenario 1 and Happy Scenario 2. The Figure illustrates that the individuals with best fitness value has more users and minor response time. The Fig. 7-8 presents the response time by number of users of individuals with the Ramp and Circuitous Treasure antipatterns

scenarios. The Figure illustrates the smallest number of individuals with the antipatterns when compared to individuals who use the happy scenarios.

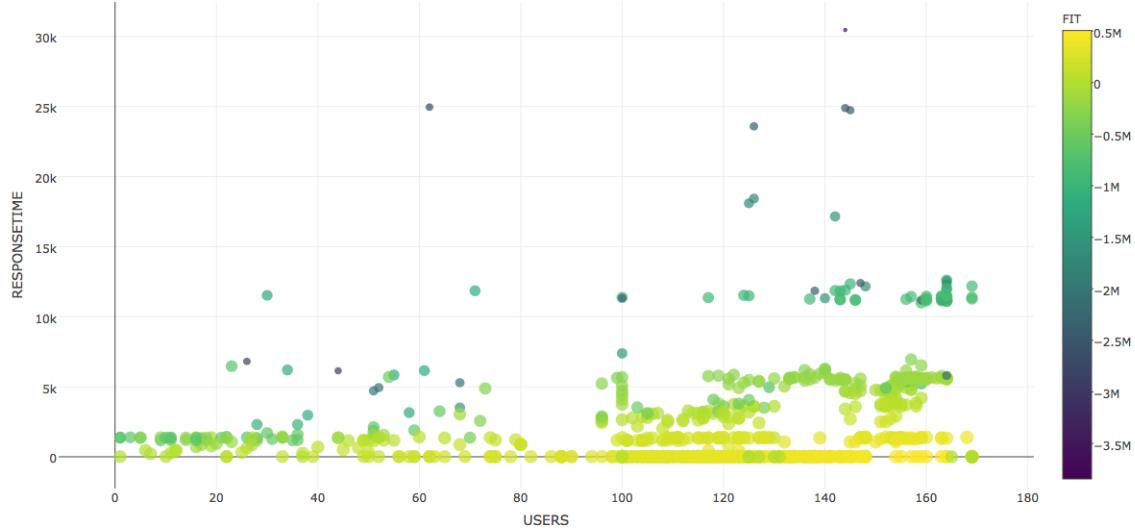


Figure 7-7: Response time by number of users of individuals with Happy Scenario 1 and Happy Scenario 2

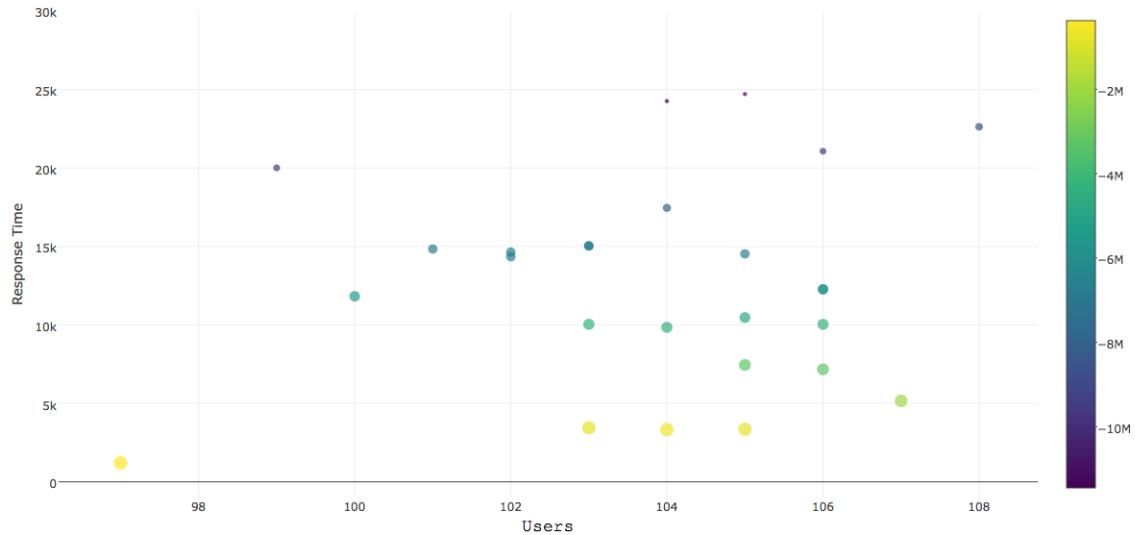


Figure 7-8: Response time by number of users of individuals with the Ramp and Circuitous Treasure antipatterns

Fig. 7-9 presents the markov decision process for the experiment. When the response time it is bellow or equal the service level, the action with major reward it is increase the number of users and include more positions with the Happy Scenario 2 (Happy 2).

When the response time is above the service level, the action with major reward value it is decrease the number of users and include more positions with the Happy Scenario 2. The actions with minor value of reward contains the both antipatterns Circuitous Treasure (CTH) and The Ramp antipatterns (Ramp).

In the first experiment, We conclude that the metaheuristics converged to scenarios with an happy path, excluding the scenarios with antipatterns. The hybridQ and hybrid metaheuristic returned individuals with higher fitness scores. However, the Hybrid metaheuristic made twice as many requests than Tabu Search to overcome it.



Figure 7-9: Markov decision process of experiment with Circuitous Treasure and The Ramp antipatterns

7.2.2 The Tower Babel and Unbalanced Processing experiment

The experiment was carried out for 6 continuous hours. All tests in the experiment were conducted without the need of a tester. 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. The Fig. 7-10 presents the fitness value obtained by each metaheuristic. The SA algorithm obtained the worst fitness values. Hybrid metaheuristic obtained the better fitness values.

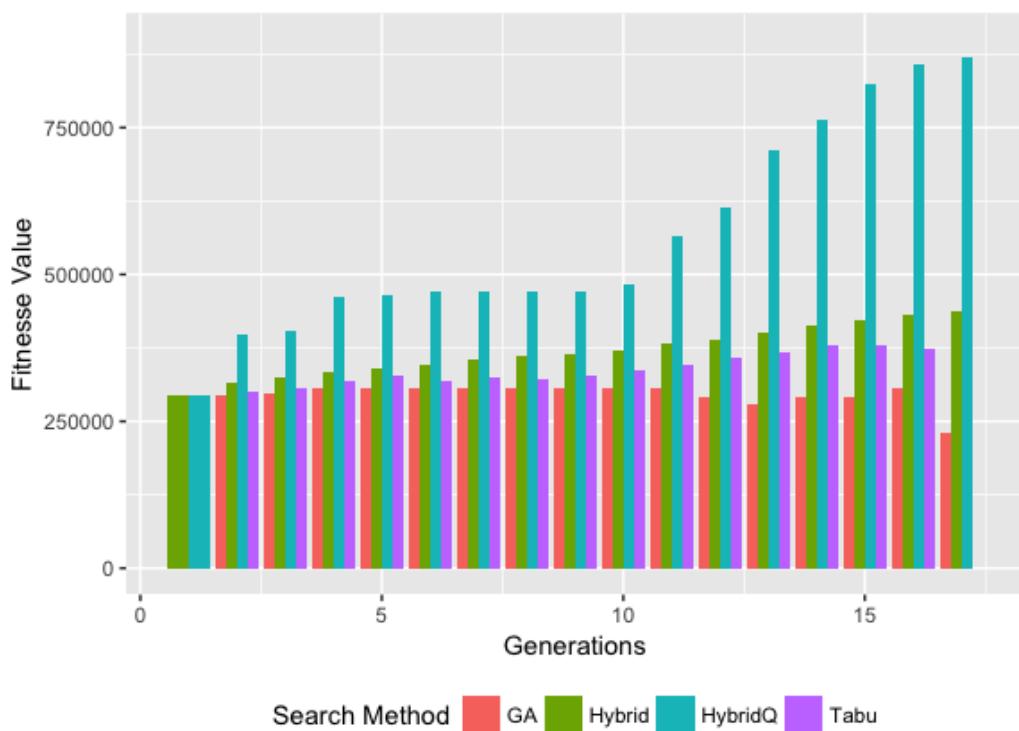


Figure 7-10: Fitness value obtained by Search Method

As in the first experiment, the Hybrid algorithm performs twice as many requests as the second one, the tabu search (Fig. 7-11). The Fig. 7-12 shows the average, minimal e maximum value by search method. The Fig. 7-13 presents the maximum, average, median and minimum fitness value by generation. The maximun fitness value increases at each generation. The Fig. 7-14 presents the density graph of number of users by fitness value. The range between 100 and 150 users has the highest number of individuals found with higher fitness value.

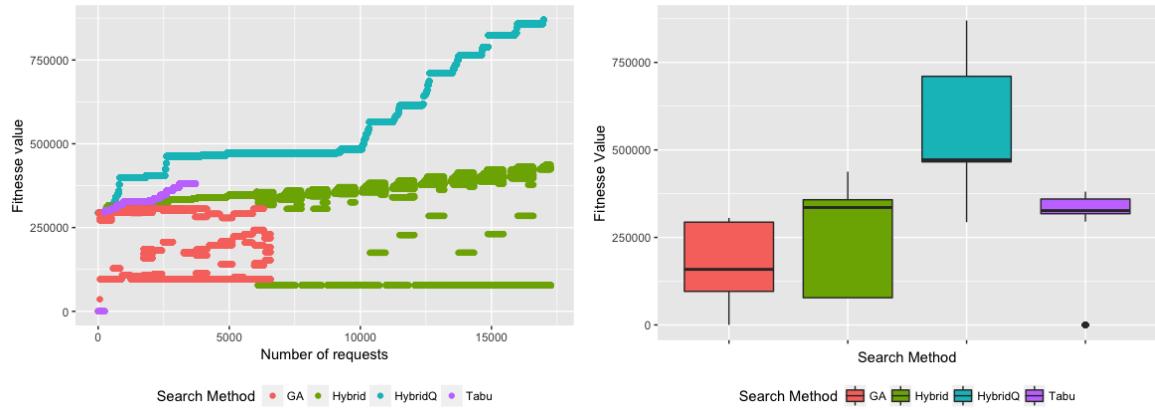


Figure 7-11: Number of requests by Search Method

Figure 7-12: Fitness value by generation in all tests

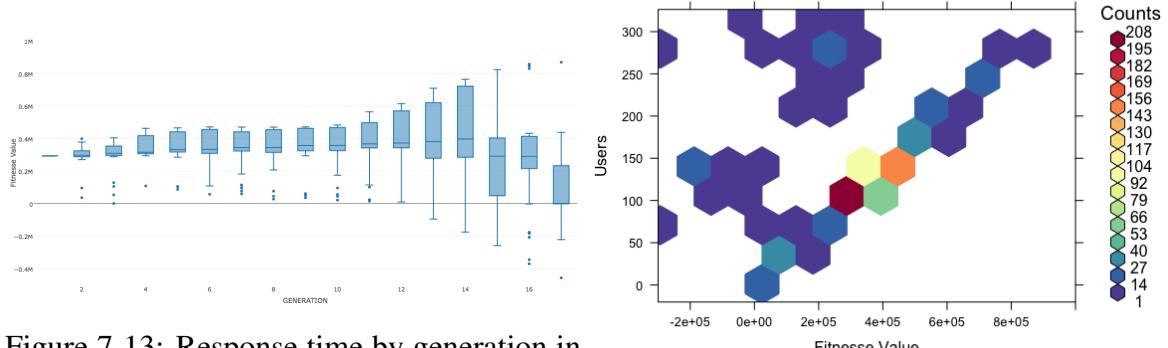


Table 7.3 shows 4 individuals with 279 to 292 users. The first individual has 121 users on Happy Scenario 2, 171 users on Happy Scenario 1 and a response time of 11 seconds. None of the best individuals found implements the Unbalanced Processing or Tower Babel antipattern.

Fig. 7-15 presents the response time by number of users of individuals with Happy Scenario 1 and Happy Scenario 2. The Figure illustrates that the individuals with best fitness value has more users and minor response time. The Fig. 7-16 presents the response time by number of users of individuals with Unbalanced Processing and Tower Babel antipatterns scenarios. The Figure illustrates the smallest number of individuals with the Unbalanced Processing and Tower Babel antipattern when compared to individuals who

use the happy scenarios and the Tower Babel antipattern.

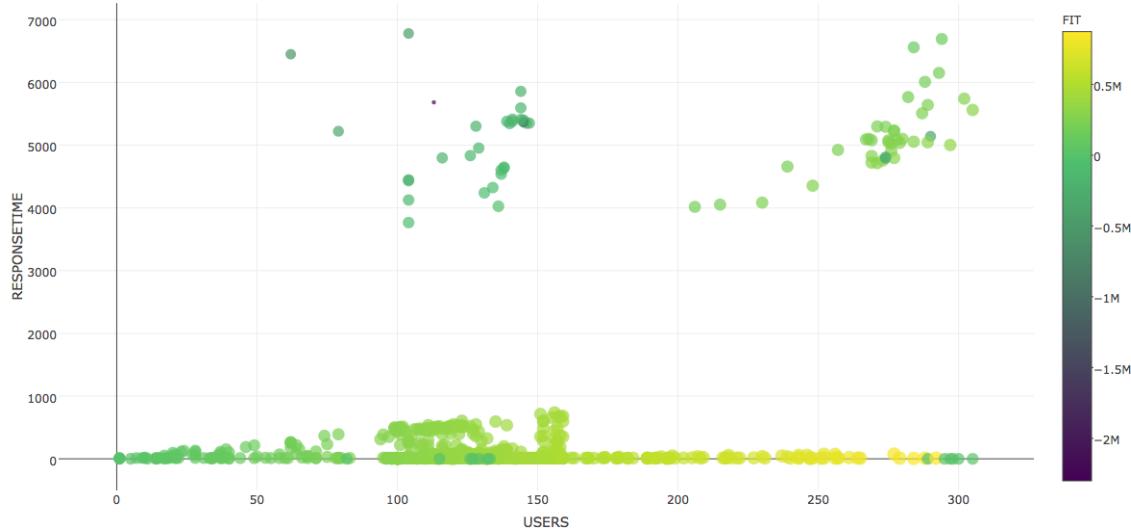


Figure 7-15: Response time by number of users of individuals with Happy Scenario 1, Happy Scenario 2 and Tower Babel antipattern

We conclude that the metaheuristics converged to scenarios with an happy path, excluding the scenarios with Unbalanced Processing and Tower Babel antipattern. The hybrid metaheuristic with Q-Learning (HybridQ) returned individuals with higher fitness scores. The Hybrid metaheuristic (Hybrid) 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 an antipattern over the 17 generations of the experiment. 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.

Table 7.3: Best individuals found in the second experiment

Search Method	Generation	Users	fitness Value	Happy 2	Happy 1	Time
HybridQ	17	292	869780	121	171	11
HybridQ	16	288	857780	103	185	11
HybridQ	16	284	845880	167	117	10
HybridQ	16	279	830780	144	135	11

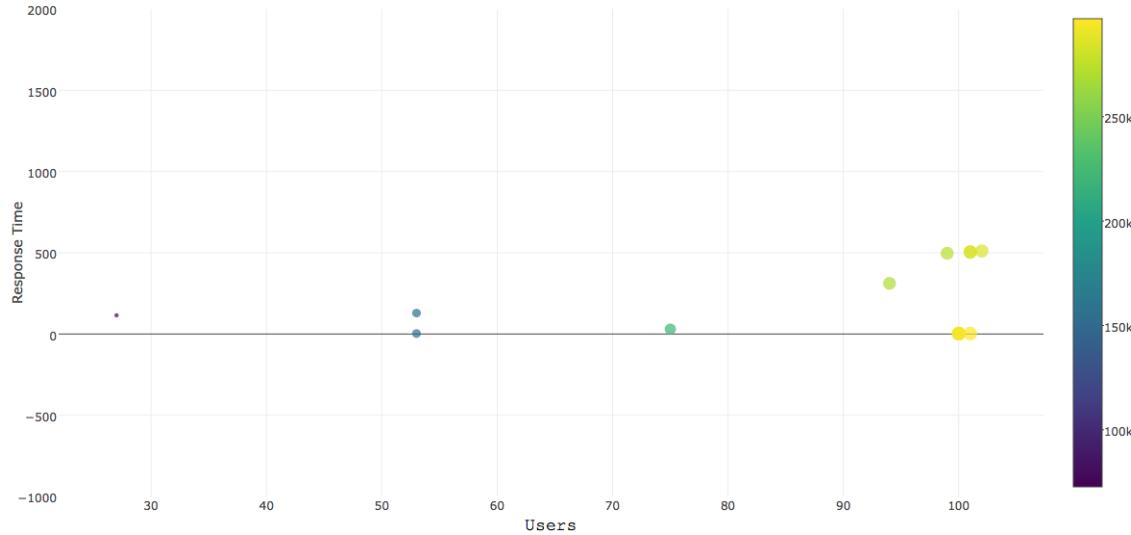


Figure 7-16: Response time by number of users of individuals with Unbalanced Processing antipattern

7.3 Moodle Application Experiment

This 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 experiment used the following fitness function:

$$\begin{aligned}
fitness = & 0.9 * 90percentiletime \\
& + 0.1 * 80percentiletime \\
& + 0.1 * 70percentiletime + \\
& 0.1 * maxResponseTime + \\
& 0.2 * numberOfUsers - penalty
\end{aligned} \tag{7.3}$$

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 7.4 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.

Table 7.4: 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 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. Table 7.5 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 7.6 presents the percentage of genes in all test scenarios by generation with and

Table 7.5: 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			

without collaboration. Most of the genes converged to the MyHome feature, which had the highest application response time.

Table 7.6: 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
Notifications	25	10	1	1	2	1	3
Post	8	2	1	1	1	2	1
Login	17	8	3	3	4	5	3

7.4 JPetStore Application Experiments

One experiment was conducted to test the use of the HybridQ algorithm in 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 seconds. Any individuals who obtained a time longer than the stipulated maximum time suffered penalties. The whole process of stress and performance tests, which run for 2 days and with about 1.800 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:

- Enter in the Catalog: the application presents the catalog of pets.
- Fish: The application shows the recorded fish items.
- Register: a new user is registered into the system.
- Dogs: The application shows the recorded dogs supplies.
- Shopping Cart: the application displays the shopping cart.
- Add or Remove in Shopping Cart: the application adds and removes items from shopping cart.

Table 7.7 presents the maximum number of users found in each scenario who obtained a time lower than the stipulated maximum time of 500 seconds.

Table 7.7: Maximum number of users of isolated test scenarios

Scenario	Max number of Users
Fish	85
Enter in The Catalog	60
Dogs	75
Cart	70
Register	90

The Fig. 7-17 and 7-18 present the response time by generation and by the number of users.

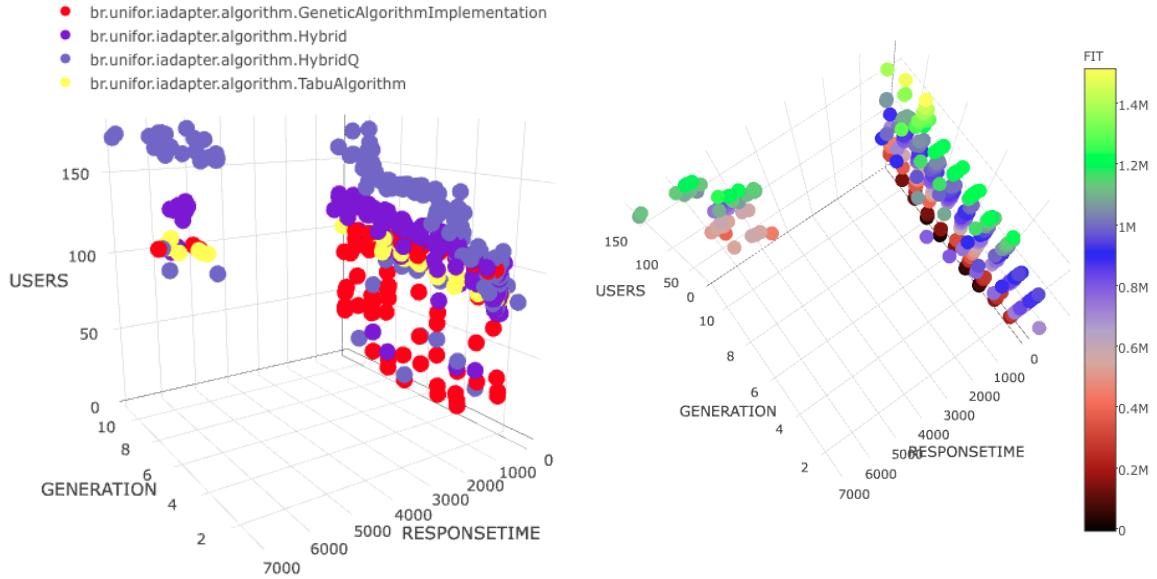


Figure 7-18: Search space of individuals

Figure 7-17: Search space of individuals found in the experiment by fitness value
found in the experiment by search method

The experiment used the following fitness function:

$$fitness = \begin{cases} n * \{3000 * \text{numberOfUsers} - 20 * 90\text{percentiletime} - 20 * 80\text{percentiletime} \\ - 20 * 70\text{percentiletime} - 20 * \text{maxResponseTime} - \text{penalty}\}, \text{ where } n \text{ is the} \\ \text{number of features used by the test in a set of previous selected application} \\ \text{features} \end{cases} \quad (7.4)$$

The purpose of the fitness function is to maximize the number of users and minimize the response time in the tests containing a selected n functionalities. For example, it is possible double the fitness value for tests that have the fish and user registration scenario.

This experiment tries to find the scenarios with maximal number of users and best response time tests that contains the Cart and Register features. The Fig. 7-19 and 7-20 show the fitness value by generation. The HybridQ obtained the best fitness values in all generations.

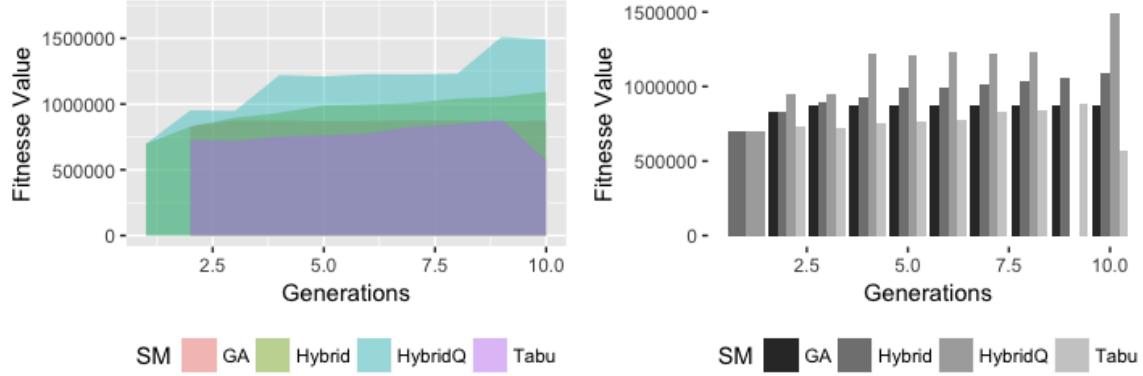


Figure 7-19: Fitness value by generation on JPetStore First experiment

Figure 7-20: Fitness value by generation on JPetStore First experiment

The Fig. 7-21 shows the fitness value by number of request by each Search Method. In the Figure, it is possible to observe that HybridQ obtained the best fitness value to same number of requests of the other algorithms.

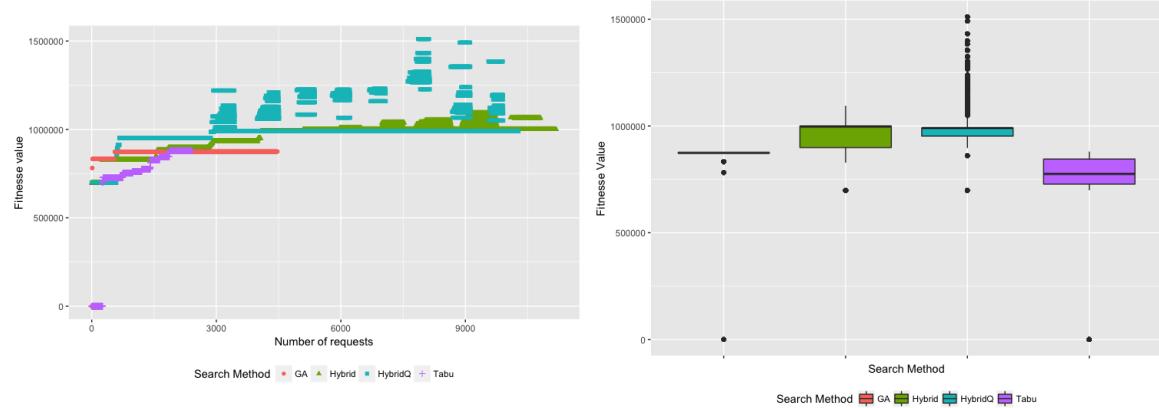


Table 7.8 shows 4 individuals with 233 to 398 users. The first individual has 73 users on fish scenario, 17 users on Dogs scenario, 50 users on Cart scenario, 33 Users on register scenario and a response time of 357 seconds.

We conclude that HybridQ found the individuals with major number of uses. The scenario with major number of users is the Fish search feature. The hybrid metaheuristic with Q-Learning (HybridQ) returned individuals with higher fitness scores. The individual with best fitness value has 73 users on fish scenario, 17 users on Dogs scenario, 50 users on Cart

Table 7.8: Best individuals found in JPetStore first experiment

Search Method	Response	Users	Gen	Fitness	fish	Dogs	Cart	Register
HybridQ	357	173	9	44773	73	17	50	33
HybridQ	398	171	10	44831	57	33	48	33
HybridQ	331	164	9	44774	71	14	51	28
HybridQ	233	159	9	44783	63	31	32	33

scenario, 33 Users on register scenario and a response time of 357 seconds.

7.5 Experimental Results

In this section, We present the result of the experiment which we carried out to verify the multiobjective 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 JPeteStore, available at <https://hub.docker.com/r/pocking/jpetstore/>. The maximum tolerated response time in the test was 500 seconds. 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 generations previously established by algorithm.

The experiments use the follow application features:

- Enter in the Catalog: the application presents the catalog of pets.
- Fish: The application shows the recorded fish items.
- Register: a new user is registered into the system.
- Dogs: The application shows the recorded dogs supplies.
- Shopping Cart: the application displays the shopping cart.
- Add or Remove in Shopping Cart: the application adds and removes items from shopping cart.

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 mutation operation was applied to 10% of the population on each generation. The objective functions 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 maximum response time expected.

The experiment used the following objective equations:

$$\begin{aligned} \text{objective function 1} = & -3 * \text{number of Users} \\ & -\text{penalty} \end{aligned} \quad (7.5)$$

$$\begin{aligned} \text{objective function 2} = & \text{response time} \\ & -\text{penalty} \end{aligned} \quad (7.6)$$

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

$$\begin{aligned} \text{penalty} &= 100 * \Delta \\ \Delta &= (t_{\text{CurrentResponseTime}} - t_{\text{MaximumResponseTimeExpected}}) \end{aligned} \quad (7.7)$$

7.5.1 Experiment Research Questions

The following research question is addressed:

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

7.5.2 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.

7.5.3 Hypotheses

- With regard to multi-objectives applied in the experiment:

- H_0 (A null hypothesis) :The NSGA-II doesn't found 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.

7.5.4 Results

Fig. 7-23 and Table 7.9 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. The workload number 1 with a single user accessing the dog scenario presented the response time of 245 seconds. The workload number 2 with a single user accessing the dog scenario, 7 users accessing the Enter Catalog feature and 3 users in register functionality presented the response time of 400 seconds.

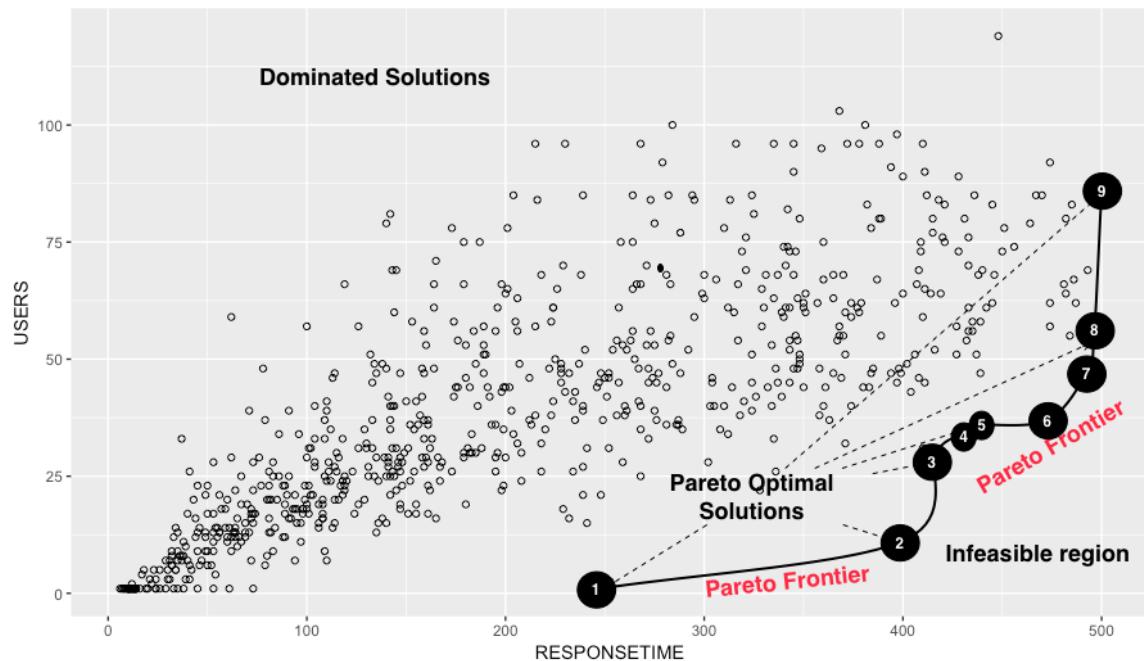


Figure 7-23: Experiment Pareto Frontier

Table 7.9: Pareto Frontier workloads 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

7.5.5 Threats to validity

- Construct Validity: In this work, we just evaluate the use of one multiobjective algorithm. However, several multiobjective 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.
- Conclusion Validity: The workloads found by NSGA-II may be the result of the elitist strategy used in the algorithm that can be applied in the genetic algorithm.

Chapter 8

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 application. 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.

8.1 Achievements

Two experiments were performed to validate the hybrid metaheuristic, two experiments were conducted to validate the testbed tool and two experiments were 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.

8.2 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.

Appendix A

Tables

Table A.1: Best individuals found in the second Testbed Tool experiments

Method	Resp.Time	Users	Fitness	Happy 1	Tower	Happy 2	Unbalanced
HybridQ	11	292	869780	171	0	121	0
HybridQ	11	288	857780	185	0	103	0
HybridQ	10	284	845880	117	0	167	0
HybridQ	11	279	830780	135	0	144	0
HybridQ	75	277	823500	147	35	77	18
HybridQ	12	265	788760	127	0	138	0
HybridQ	13	264	785740	170	0	94	0
HybridQ	20	261	776600	160	0	83	18
HybridQ	11	257	764780	181	0	76	0
HybridQ	16	257	764680	135	29	79	14
HybridQ	12	256	761760	150	0	106	0
HybridQ	72	256	760560	157	33	49	17
HybridQ	12	252	749760	191	0	61	0
HybridQ	74	252	748520	130	32	77	13
HybridQ	10	251	746800	130	30	77	14
HybridQ	12	248	737760	130	29	76	13
HybridQ	10	247	734940	155	0	92	0

Table A.4: My caption

Year	Type	Researches	Classification
1999	Book	[45]	Automated Software Testing
2004	Book	[99]	Method for Software Testing

2005	Book	[80]	Software Testing
2006	Book	[68]	Formal Tests
2007	Book	[37]	Load Testing
2008	Book	[65]	Load Test Tool
2009	Book	[92]	Load Testing
2011	Book	[97]	Complex Systems
2012	Book	[77]	Unit Test
2013	Book	[47]	Load Test Tool
2013	Book	[48]	WorkLoad Modeling
2014	Book	[56]	Performance Testing
2012	Book Section	[16]	Load Testing
2013	Book Section	[87]	Analysis of load testing results
1993	Conference Paper	[17]	Load Testing
1998	Conference Paper	[9]	Stress Search-Based Test
1998	Conference Paper	[103]	Stress Search-Based Test
1998	Conference Paper	[93]	Stress Search-Based Test
1998	Conference Paper	[116]	Stress Search-Based Test
1999	Conference Paper	[131]	Stress Search-Based Test
2005	Conference Paper	[36]	Stress Search-Based Test
2006	Conference Paper	[43]	Load Testing
2006	Conference Paper	[3]	Stress Testing
2007	Conference Paper	[98]	Stress Search-Based Test
2007	Conference Paper	[24]	Load Testing
2008	Conference Paper	[75]	Analysis of load testing results
2010	Conference Paper	[140]	Testing as Service
2011	Conference Paper	[96]	Performance Testing
2012	Conference Paper	[14]	Anti-patterns
2013	Conference Paper	[27]	Performance Testing
2014	Conference Paper	[58]	Performance Testing
2015	Conference Paper	[11]	Anti-patterns

2016	Conference Paper	[60]	Stress Search-Based Test
1994	Journal	[18]	Load Testing
1995	Journal	[15]	Load Testing
1995	Journal	[124]	Load Testing
1996	Journal	[139]	Load Test Tool
2000	Journal	[136]	Performance Testing
2000	Journal	[64]	Stress Search-Based Test
2001	Journal	[46]	Model-Based Test
2002	Journal	[90]	Performance Model
2002	Journal	[115]	Anti-patterns
2003	Journal	[33]	Mock Objects
2003	Journal	[8]	Performance Monitoring
2004	Journal	[44]	Analysis of load testing results
2004	Journal	[21]	Performance Testing
2004	Journal	[107]	Software Testing
2004	Journal	[88]	Survey
2005	Journal	[32]	Stress Search-Based Test
2005	Journal	[79]	Load Test Tool
2005	Journal	[101]	Stress Search-Based Test
2006	Journal	[52]	Stress Search-Based Test
2006	Journal	[52][7]	Stress Search-Based Test
2007	Journal	[35]	Performance Testing
2007	Journal	[94]	Load Test Tool
2007	Journal	[49]	Performance Model
2008	Journal	[26]	Stress Search-Based Test
2008	Journal	[5]	Systematic Map
2008	Journal	[1]	Testbed
2008	Journal	[82]	Performance Model
2009	Journal	[4]	Performance Monitoring
2009	Journal	[74]	Analysis of load testing results

2009	Journal	[6]	Survey
2009	Journal	[69]	Formal Tests
2010	Journal	[128]	Load Testing
2010	Journal	[83]	Analysis of load testing results
2010	Journal	[54]	Stress Search-Based Test
2010	Journal	[137]	Survey
2011	Journal	[20]	Analysis of load testing results
2011	Journal	[19]	Stress Search-Based Test
2011	Journal	[22]	Load Testing
2011	Journal	[95]	Load Testing
2011	Journal	[110]	Performance Model
2011	Journal	[108]	Load Testing
2011	Journal	[141]	Load Testing
2011	Journal	[2]	Test Coverage
2012	Journal	[138]	Load Testing
2012	Journal	[62]	Stress Testing
2013	Journal	[126]	Test Monitoring
2013	Journal	[84]	Analysis of load testing results
2013	Journal	[129]	Load Testing
2013	Journal	[134]	Anti-patterns
2013	Journal	[23]	Stress Testing
2013	Journal	[40]	Stress Search-Based Test
2014	Journal	[135]	Anti-patterns
2014	Journal	[41]	Stress Search-Based Test
2015	Journal	[73]	Survey
2015	Journal	[81]	Stress Testing
2015	Journal	[66]	Survey
2015	Journal	[12]	Stress Search-Based Test
2016	Journal	[127]	WorkLoad Prediction
2003	Thesis	[112]	Stress Search-Based Test

2006	Thesis	[52]	Stress Search-Based Test
2010	Thesis	[76]	Analysis of load testing results
2009	Thesis	[51]	Performance Model
2011	Thesis	[125]	Analysis of load testing results

Table A.2: Researches found using keyword 'Load Testing'

Year	Type	Researches
1999	Book	[45]
2004	Book	[99]
2005	Book	[80]
2007	Book	[37]
2008	Book	[65]
2009	Book	[92]
2012	Book	[77]
2013	Book	[47]
2012	Book Section	[16]
1993	Conference Paper	[17]
1998	Conference Paper	[9]
2006	Conference Paper	[43][3]
2008	Conference Paper	[75]
2010	Conference Paper	[140]
1994	Journal	[18]
1995	Journal	[15][124]
1996	Journal	[139]
2000	Journal	[136]
2001	Journal	[46]
2002	Journal	[90]
2003	Journal	[33]
2004	Journal	[44][107]
2005	Journal	[32]
2006	Journal	[52]
2007	Journal	[35][94][1]
2008	Journal	[26]
2009	Journal	[4]
2010	Journal	[128][83]
2011	Journal	[20][22][95][110][141]
2012	Journal	[138][62][126]
2013	Journal	[84][129][134][121][23]
2014	Journal	[135]
2015	Journal	[73][81]
2016	Journal	[127]
2010	Thesis	[76]

Table A.3: Researches found using keyword 'Stress Testing'

Year	Type	Researches
1999	Book	[45]
2004	Book	[99]
2005	Book	[80]
2006	Book	[68]
2007	Book	[37]
2008	Book	[65]
2009	Book	[92]
2012	Book	[77]
2013	Book	[47][121]
2012	Book Section	[16]
2013	Book Section	[87]
1993	Conference Paper	[17]
1998	Conference Paper	[9]
2006	Conference Paper	[43][3]
2008	Conference Paper	[75]
2010	Conference Paper	[140]
1994	Journal	[18]
1995	Journal	[15]
1996	Journal	[139]
2000	Journal	[136]
2001	Journal	[46]
2002	Journal	[90]
2003	Journal	[33]
2004	Journal	[44][21][107][88]
2005	Journal	[32]
2006	Journal	[52][7]
2007	Journal	[35][94]
2008	Journal	[26][5][1]
2009	Journal	[4][74][6][69]
2010	Journal	[128][83] [54]
2011	Journal	[20][23][95][110][141][108][2]
2012	Journal	[138][62] [126]
2013	Journal	[84][129][134]
2014	Journal	[135][128]
2015	Journal	[73][81][66]
2016	Journal	[127]
2003	thesis	[112]
2006	thesis	[52]
2010	thesis	[76]
2011	thesis	[125]

Table A.5: Comparison study presented by Pen~a-Ortiz on the book Modeling and Simulation of Computer Networks and Systems Methodologies [F- Fully attended feature P-Partial attended feature] [91]

Feature	Webstone	SPECweb	SURGE	Web polygraph	TPC-W
Analytical-Based Architecture	F	F	F	F	F
Distributed Architecture	F	F	F	F	
Business-Based Architecture		P		P	F
Client Parameterization	F	F		P	F
Workload Types		F			F
Multi-platform	F	F	F	F	F
Open Source	F		F	P	F
Search-based Testing or Learning features					
Feature	LoadRunner	WebLoad	JMeter	S-Clients	WebJammma
Analytical-Based Architecture	P	P	P		
Distributed Architecture	F	F			
Business-Based Architecture	F	F	F		
Client Parameterization	F	F	F		F
Workload Types	F	F	P		
Multi-platform	F	F	F	F	F
Open Source			F	F	F
Search-based Testing or Learning features					
Feature	Deluge	HAMMER HEAD 2	PTester	Siege	Locust
Analytical-Based Architecture					P
Distributed Architecture					
Business-Based Architecture					F
Client Parameterization					F
Workload Types					P
Multi-platform	F	P	F	P	F
Open Source	F	F	F	F	F
Search-based Testing or Learning features					

Appendix B

Figures

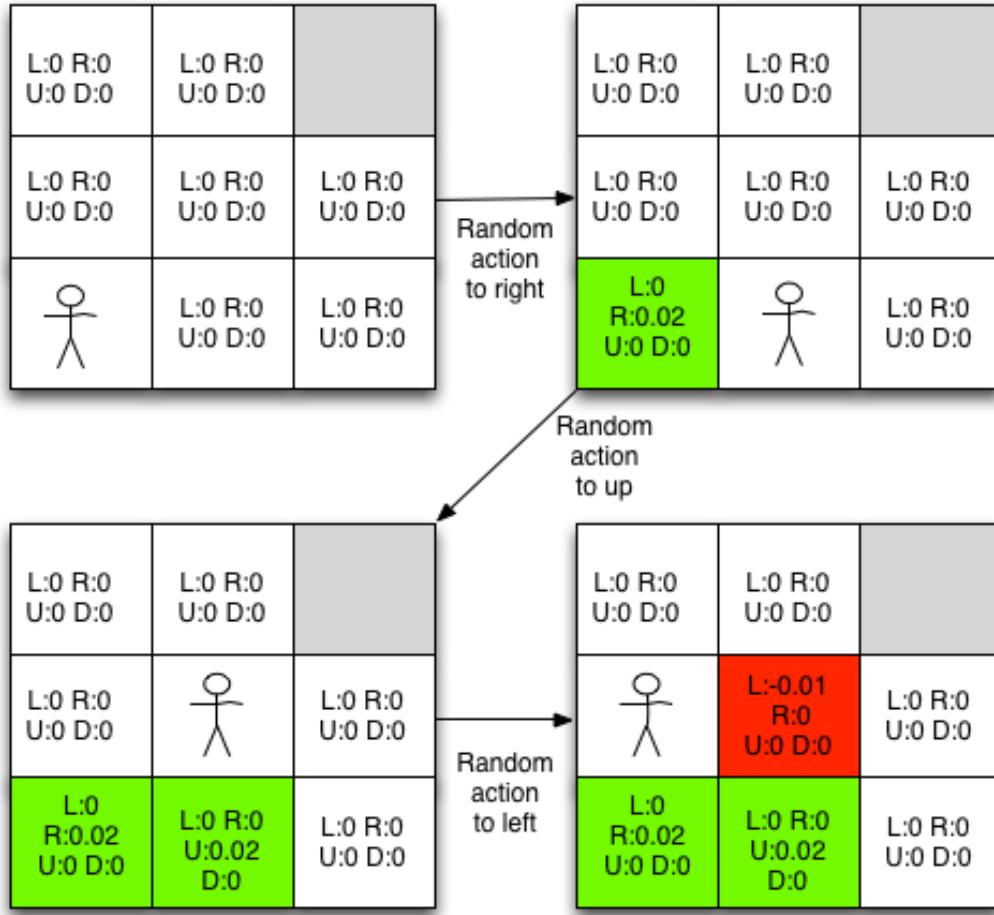


Figure B-1: GridWorld initial states

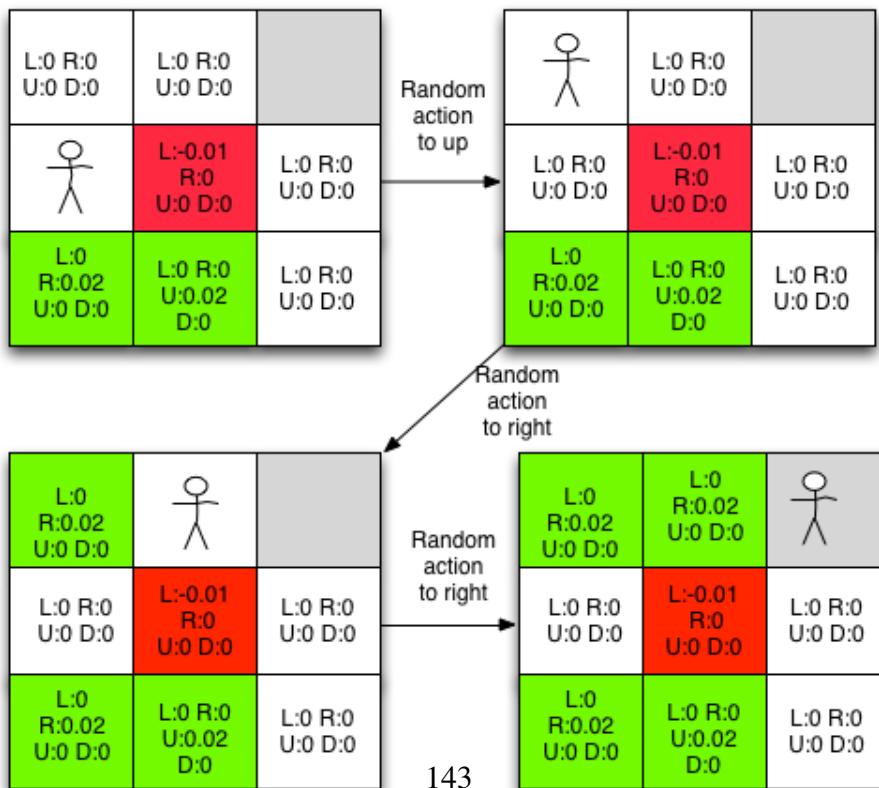


Figure B-2: GridWorld after four initial actions

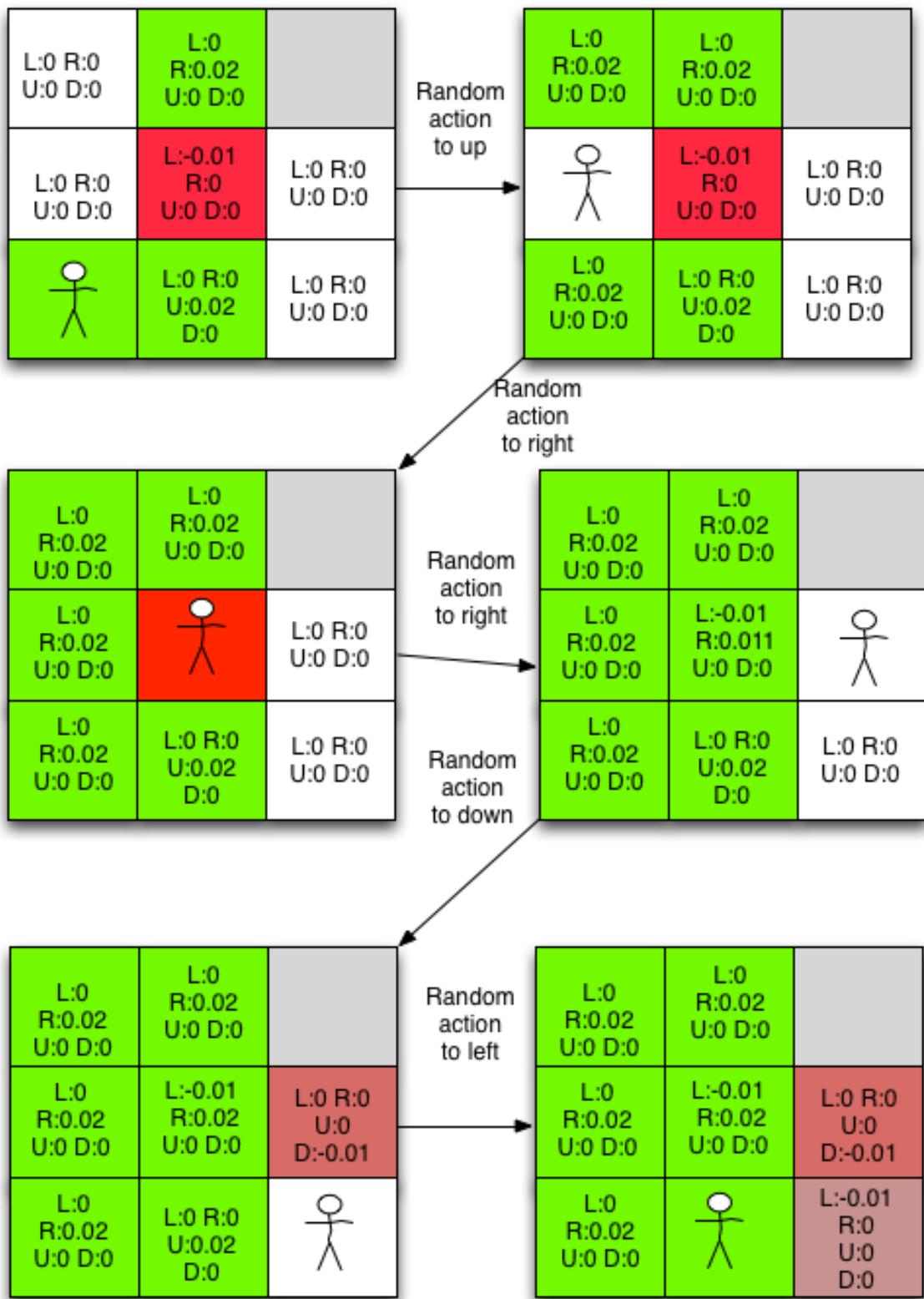


Figure B-3: GridWorld - exploration phase (Step 5 to 10)

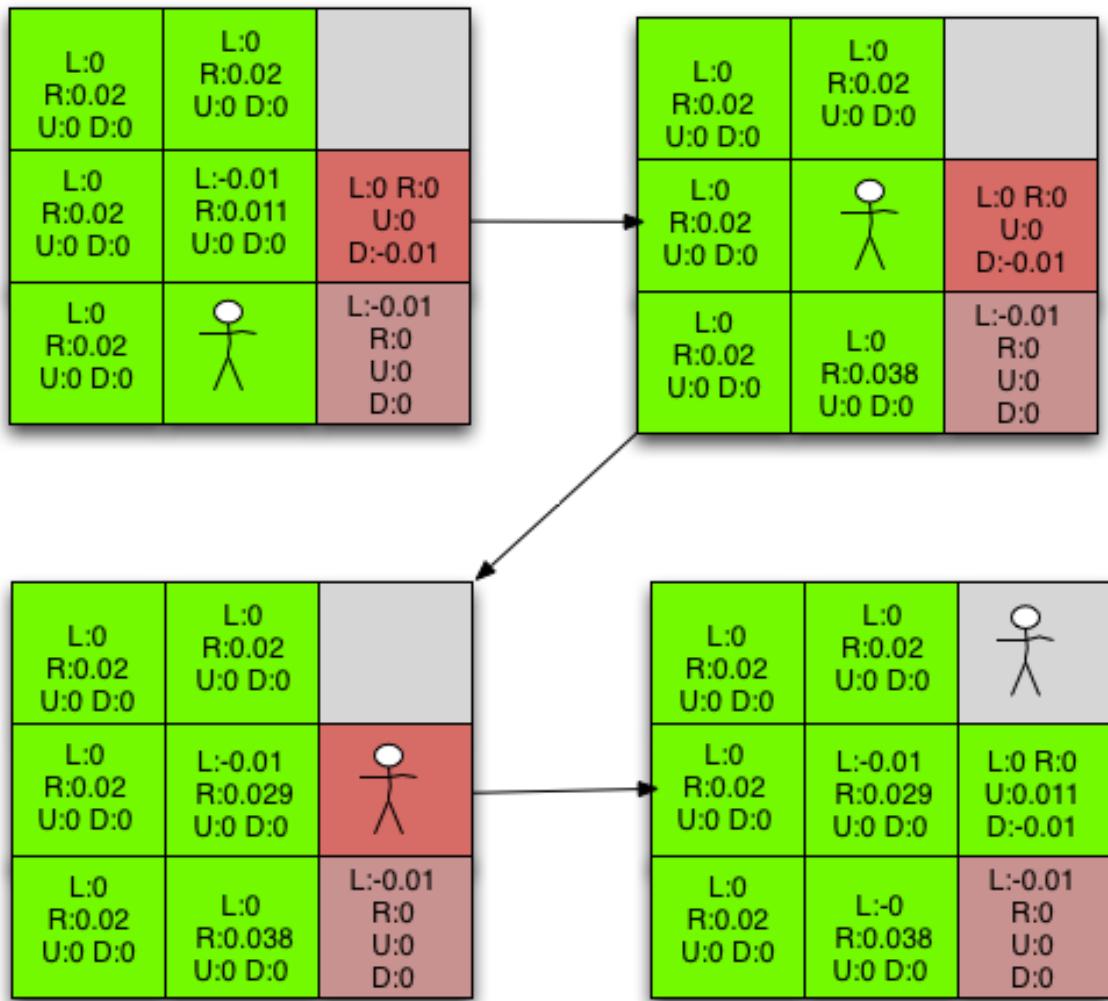


Figure B-4: GridWorld - exploration phase (Step 10 to 13)

Bibliography

- [1] Model-based generation of testbeds for web services. *Testing of Software and ...*, pages 266–282, 2008.
- [2] Software automated testing: A solution to maximize the test plan coverage and to increase software reliability and quality in use. *Computer Standards and Interfaces*, 33(2):152–158, 2011.
- [3] Saeed Abu-nimeh, Suku Nair, and Marco Marchetti. Avoiding Denial of Service via Stress Testing. 2001.
- [4] Mithun Acharya and Vamshidhar Kommineni. Mining health models for performance monitoring of services. *ASE2009 - 24th IEEE/ACM International Conference on Automated Software Engineering*, pages 409–420, 2009.
- [5] Wasif Afzal, Richard Torkar, and Robert Feldt. A systematic mapping study on non-functional search-based software testing. *Proceedings of the Twentieth International Conference on Software Engineering & Knowledge Engineering (SEKE'2008)*, (October 2015):488–493, 2008.
- [6] Wasif Afzal, Richard Torkar, and Robert Feldt. A systematic review of search-based testing for non-functional system properties. *Information and Software Technology*, 51(6):957–976, 2009.
- [7] Daimler Chrysler Ag, D Berlin, and Stefan Wappler. Improving Evolutionary Real-Time Testing Categories and Subject Descriptors. *Proceedings of the 8th annual conference on Genetic and evolutionary computation*, pages 1917–1924, 2006.
- [8] Marcos K. Aguilera, Jeffrey C. Mogul, Janet L. Wiener, Patrick Reynolds, and Athicha Muthitacharoen. Performance debugging for distributed systems of black boxes. *ACM SIGOPS Operating Systems Review*, 37:74, 2003.
- [9] Jarmo T. JT Alander, Timo Mantere, and Pekka Turunen. Genetic Algorithm Based Software Testing. In *Neural Nets and Genetic Algorithms*, 1998.
- [10] Enrique Alba and Francisco Chicano. Observations in using parallel and sequential evolutionary algorithms for automatic software testing. *Computers and Operations Research*, 35(10):3161–3183, 2008.

- [11] Azadeh Alebrahim and Maritta Heisel. Applying performance patterns for requirements analysis. In *ACM International Conference Proceeding Series*, volume 08-12-July, 2015.
- [12] Stefano D I Alesio, Lionel C Briand, Shiva Nejati, and Arnaud Gotlieb. Combining Genetic Algorithms and Constraint Programming. *ACM Transactions on Software Engineering and Methodology*, 25(1), 2015.
- [13] Aldeida Aleti, I. Moser, and Lars Grunske. Analysing the fitness landscape of search-based software testing problems. *Automated Software Engineering*, pages 1–19, 2016.
- [14] Davide Arcelli, Vittorio Cortellessa, and Catia Trubiani. Antipattern-Based Model Refactoring for Software Performance Improvement. *Proceedings of the 8th international ACM SIGSOFT conference on Quality of Software Architectures (QoSA '12)*, pages 33–42, 2012.
- [15] A. Avritzer and E.J. Weyuker. The automatic generation of load test suites and the assessment of the resulting software. *Software Engineering, IEEE ...*, 21(9):705–716, 1995.
- [16] Alberto Avritzer and Andre B. Bondi. Resilience Assessment Based on Performance Testing. In *Resilience Assessment and Evaluation of Computing Systems*, pages 305–322. 2012.
- [17] Alberto Avritzer and Brian Larson. Load Testing Software Using Deterministic State Testing. In *ACM SIGSOFT Software Engineering Notes, ISSTA '93*, pages 82–88, New York, NY, USA. ACM.
- [18] Alberto Avritzer and EJ Weyuker. Generating test suites for software load testing. ... *international symposium on Software testing ...*, pages 44–57, 1994.
- [19] Arthur I Baars, Kiran Lakhota, Tanja E J Vos, and Joachim Wegener. Search-based testing, the underlying engine of Future Internet testing. *Federated Conference on Computer Science and Information Systems (FedCSIS 2011)*, pages 917–923, 2011.
- [20] C Babbar, N Bajpai, and Dk Sarmah. Web Application Performance Analysis based on Component Load Testing. *International Journal of Technology*, 2011.
- [21] S Barber. Creating effective load models for performance testing with incomplete empirical data. ... *Energy Conference, 2004. INTELEC 2004. 26th ...*, pages 1–13, 2004.
- [22] Cornel Barna, M Litoiu, and H Ghanbari. Autonomic load-testing framework. *International conference on Autonomi*, pages 91–100, 2011.
- [23] Cornel Barna, Mark Shtern, Michael Smit, Vassilios Tzerpos, and Marin Litoiu. Mitigating DoS Attacks using Performance Model-driven Adaptive Algorithms. X(February), 2013.

- [24] Marcelo De Barros and Jing Shiau. Web services wind tunnel: On performance testing large-scale stateful web services. *... and Networks, 2007.* ..., 2007.
- [25] Roberto Battiti, Mauro Brunato, and Franco Mascia. *Reactive search and intelligent optimization*, volume 45. 2009.
- [26] 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.
- [27] Pavlo Bazilinskyy and Markus Brunner. Performance engineering and testing: The challenges on mobile platforms. *Student Conference on Optimisation of Software*, 2013.
- [28] Kristin P Bennett. The Interplay of Optimization and Machine Learning Research. *Journal of Machine Learning Research*, 7(November):1265–1281, 2006.
- [29] C. Blum and A. Roli. Metaheuristics in combinatorial optimization: overview and conceptual comparison. *ACM Computing Surveys*, 35(3):189–213, 2003.
- [30] Christian Blum. Hybrid metaheuristics in combinatorial optimization: A tutorial. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7505 LNCS(6):1–10, 2012.
- [31] Justin A. Boyan and Andrew W. Moore. Learning Evaluation Functions to Improve Local Search. *Journal of Machine Learning Research*, 1:77–112, 2000.
- [32] Lionel C. Briand, Yvan Labiche, and Marwa Shousha. Stress testing real-time systems with genetic algorithms. *Proceedings of the 2005 conference on Genetic and evolutionary computation - GECCO '05*, page 1021, 2005.
- [33] Matthew a Brown and Eli Tapolcsanyi. Mock Object Patterns. *Matrix*, pages 1–17, 2003.
- [34] 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.
- [35] 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.
- [36] Gerardo Canfora, Massimiliano Di Penta, Raffaele Esposito, and Maria Luisa Viliani. 2005., Canfora, G., An approach for QoS-aware service composition based on genetic algorithms.
- [37] Microsoft Corporation. Performance Testing Guidance for Web Applications, November 2007.

- [38] Vittorio Cortellessa and Laurento Frittella. A Framework for Automated Generation of Architectural Feedback from Software Performance Analysis. pages 171–185, 2007.
- [39] Leffingwell Dean and Widrig Don. Managing software requirements: A use case approach, 2003.
- [40] S Di Alesio, S Nejati, L Briand, and A Gotlieb. Stress testing of task deadlines: A constraint programming approach. *IEEE Xplore*, pages 158–167, 2013.
- [41] Stefano Di Alesio, Shiva Nejati, Lionel Briand, and Arnaud Gotlieb. Worst-Case Scheduling of Software Tasks – A Constraint Optimization Model to Support Performance Testing. *Principles and Practice of Constraint Programming*, pages 813–830.
- [42] Giuseppe a. Di Lucca and Anna Rita Fasolino. Testing Web-based applications: The state of the art and future trends. *Information and Software Technology*, 48:1172–1186, 2006.
- [43] 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.
- [44] Catalin Dumitrescu, Ioan Raicu, Matei Ripeanu, and Ian Foster. DiPerF: An automated distributed Performance testing framework. *Proceedings - IEEE/ACM International Workshop on Grid Computing*, pages 289–296, 2004.
- [45] Elfriede Dustin, Jeff Rashka, and John Paul. *Automated Software Testing: Introduction, Management, and Performance*. 1999.
- [46] Ibrahim K El-far and James a Whittaker. Model based software testing. *Encyclopedia of Software Engineering*, pages 1–22, 2001.
- [47] Bayo Erinle. *Performance Testing With JMeter 2.9*. 2013.
- [48] Dror G Feitelson. *Workload Modeling for Computer Systems Performance Evaluation*. Cambridge University Press, 2013.
- [49] P.C. Fritzsche, D. Rexachs, and E. Luque. TSP Performance Prediction Using Data Mining. *2007 4th IEEE Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications*, 2007.
- [50] Luca M Gambardella and Marco Dorigo. Ant-Q : A Reinforcement Learning approach to the traveling salesman problem. 5625:252–260, 1995.
- [51] As Ganapathi. *Predicting and optimizing system utilization and performance via statistical machine learning*. PhD thesis, 2009.
- [52] Vahid Garousi. *Traffic-aware Stress Testing of Distributed Real-Time Systems based on UML Models using Genetic Algorithms*. PhD thesis, 2006.

- [53] Vahid Garousi. Empirical analysis of a genetic algorithm-based stress test technique. *Proceedings of the 10th annual conference on Genetic and evolutionary computation - GECCO '08*, page 1743, 2008.
- [54] Vahid Garousi. A Genetic Algorithm-Based Stress Test Requirements Generator Tool and Its Empirical Evaluation. *IEEE Transactions on Software Engineering*, 36(6):778–797, November 2010.
- [55] Gregory Gay. Challenges in Using Search-Based Test Generation to Identify Real Faults in Mockito. pages 1–6.
- [56] Chris Geiger. Performance Testing in Continuous Integration Environments. 2014.
- [57] Jean-Yves Gendreau, Michel and Potvin. *Handbook of Metaheuristics*, volume 157. 2010.
- [58] Alim Ul Gias and Kazi Sakib. An adaptive bayesian approach for URL selection to test performance of large scale web-based systems. In *Proceedings of the 28th international conference on Software engineering*, volume undefined, pages 608–609, 2014.
- [59] Fred Glover and Rafael Martí. Tabu Search. *Tabu Search*, pages 1–16, 1986.
- [60] 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.
- [61] 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.
- [62] Mark Grechanik, Chen Fu, and Qing Xie. Automatically finding performance problems with feedback-directed learning software testing. *2012 34th International Conference on Software Engineering (ICSE)*, pages 156–166, June 2012.
- [63] Amy Greenwald, Keith Hall, and R Serrano. Correlated Q-learning. *Icml*, (3):84–89, 2003.
- [64] Hg Gross, Bryan F Jones, and David E Eyes. Structural performance measure of evolutionary testing applied to worst-case timing of real-time systems. *Software, IEE Proceedings-*, 147(2):25–30, 2000.
- [65] Emily H. Halili. *Apache JMeter: A practical beginner's guide to automated testing and performance measurement for your websites*. 2008.
- [66] 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.

- [67] Mark Harman and Phil McMinn. A theoretical and empirical study of search-based testing: Local, global, and hybrid search. *IEEE Transactions on Software Engineering*, 36(2):226–247, 2010.
- [68] Klaus Havelund, Manuel Núñez, Grigore Roșu, and Burkhardt Wolff. *Formal Approaches to Software Testing and Runtime Verification*. 2006.
- [69] 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. *ACM Comput. Surv.*, 41(2):1–76, 2009.
- [70] Tzung-Pei Hong, Hong-Shung Wang, and Wei-Chou Chen. Simultaneously applying multiple mutation operators in genetic algorithms. *Journal of heuristics*, 6(4):439–455, 2000.
- [71] B. Jones J. Wegener, K. Grimm, M. Grochtmann, H. Sthamer. Systematic testing of real-time systems. *EuroSTAR’96: Proceedings of the Fourth International Conference on Software Testing Analysis and Review*, 1996.
- [72] Wassim Jaziri. *Local Search Techniques: Focus on Tabu Search*. 2008.
- [73] Zhen Ming Jiang and Ahmed Hassan. A Survey on Load Testing of Large-Scale Software Systems. *IEEE Transactions on Software Engineering*, 5589(2):1–1, 2015.
- [74] Zhen Ming Jiang, Ahmed E Hassan, Gilbert Hamann, and Parminder Flora. Automated performance analysis of load tests. *2009 IEEE International Conference on Software Maintenance*, pages 125–134, 2009.
- [75] Zhen Ming ZM Jiang, Ahmed E. AE Hassan, Gilbert Hamann, and Parminder Flora. Automatic identification of load testing problems. In *IEEE International Conference on Software Maintenance, ICSM*, pages 307–316, 2008.
- [76] ZM Jiang. *Automated analysis of load testing results*. PhD thesis, 2010.
- [77] Tomasz Kaczanowski. *Practical unit testing with testNG and Mockito*. 2012.
- [78] Kaivan Kamali, L. J. Jiang, John Yen, and K. W. Wang. Using Q-Learning and Genetic Algorithms to Improve the Efficiency of Weight Adjustments for Optimal Control and Design Problems. *Journal of Computing and Information Science in Engineering*, 7(December 2007):302, 2007.
- [79] 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.

- [80] William E. Lewis, David Dobbs, and Gunasekaran Veerapillai. *Software testing and continuous quality improvement*. 2005.
- [81] Qi Luo, Aswathy Nair, Mark Grechanik, and Denys Poshyvanyk. FOREPOST: finding performance problems automatically with feedback-directed learning software testing. *Empirical Software Engineering*, pages 1–51, 2015.
- [82] Christof Lutteroth and Gerald Weber. Modeling a Realistic Workload for Performance Testing. *2008 12th International IEEE Enterprise Distributed Object Computing Conference*, pages 149–158, sep 2008.
- [83] Haroon Malik. A methodology to support load test analysis. *Software Engineering, 2010 ACM/IEEE 32nd . . .*, 2:421–424, 2010.
- [84] Haroon Malik, Hadi Hemmati, and Ahmed E Hassan. Automatic Detection of Performance Deviations in the Load Testing of Large Scale Systems. pages 1012–1021, 2013.
- [85] Alexander Pretschner Mark Utting and Bruno Legeard. A taxonomy of model-based testing approaches. *Software Testing Verification and Reliability*, 24(8):297–312, 2012.
- [86] Jackson Matsuura and Reinaldo A C Bianchi. Heuristically Accelerated Q – Learning : a new approach to speed up Reinforcement Learning. (March), 2015.
- [87] Michael Mayo and Simon Spacey. Predicting Regression Test Failures Using Genetic Algorithm-Selected Dynamic Performance Analysis Metrics. *Lecture Notes in Computer Science*, pages 158–171. Springer Berlin Heidelberg, jan 2013.
- [88] Philip McMinn, Regent Court, Software Testing, and Portobello Street. Search-based software test data generation: a survey. *Software testing, Verification and reliability*, 14:1–58, 2004.
- [89] Karl Meinke and Fei Niu. A learning-based approach to unit testing of numerical software. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6435 LNCS:221–235, 2010.
- [90] Daniel A Menascé and George Mason. TPC-W : A Benchmark for E-commerce. (June):1–6, 2002.
- [91] Petros Nicopolitidis Mohammad S. Obaidat and Faouzi Zarai. *Modeling and Simulation of Computer Networks and Systems Methodologies and Applications*.
- [92] Ian Molyneaux. *The Art of Application Performance Testing: Help for Programmers and Quality Assurance*. "O'Reilly Media, Inc.", 1st edition, January 2009.
- [93] F. Mueller and J. Wegener. A comparison of static analysis and evolutionary testing for the verification of timing constraints. *Proceedings. Fourth IEEE Real-Time Technology and Applications Symposium (Cat. No.98TB100245)*, 1998.

- [94] Dmitri Nevedrov. Using JMeter to Performance Test Web Services. pages 1–11, 2007.
- [95] Thanh H D Nguyen, Bram Adams, Zhen Ming Jiang, Ahmed E. Hassan, Mohamed Nasser, and Parminder Flora. Automated verification of load tests using control charts. *Proceedings - Asia-Pacific Software Engineering Conference, APSEC*, pages 282–289, 2011.
- [96] Tuli Nivas and Christoph Csallner. Managing Performance Testing With Release Certification and Data Correlation. In *European Conference on Software Engineering / Foundations of Software Engineering*, 2011.
- [97] Adam Jamison Oliner. *Using Influence to Understand Complex Systems*. Stanford University, 2011.
- [98] 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.
- [99] William E. Perry. *Effective methods for software testing*. 2004.
- [100] Éric Piel, Alberto González-Sánchez, and Hans-Gerhard Groß. Testing Software and Systems. *Ictss*, 6435(October):79–94, 2010.
- [101] Hartmut Pohlheim, Mirko Conrad, and Arne Griep. Evolutionary Safety Testing of Embedded Control Software by Automatically Generating Compact Test Data Sequences. *Analysis*, (724):804—814, 2005.
- [102] Jakob Puchinger and R Raidl. Combining Metaheuristics and Exact Algorithms in Combinatorial Optimization : A Survey and Classification. *Artificial Intelligence and Knowledge Engineering Applications a Bioinspired Approach*, 3562:41–53, 2005.
- [103] P. Puschner and R. Nossal. Testing the results of static worst-case execution-time analysis. *Proceedings 19th IEEE Real-Time Systems Symposium (Cat. No.98CB36279)*, 1998.
- [104] Günther R Raidl, Jakob Puchinger, and Christian Blum. Metaheuristic hybrids. In *Handbook of metaheuristics*, pages 469–496. Springer, 2010.
- [105] R Raidl. A Unified View on Hybrid Metaheuristics. *Hybrid Metaheuristics (LNCS 4030)*, pages 1–12, 2006.
- [106] Christian Raidl, Gunther R and Puchinger, Jakob and Blum. *Hybrid Metaheuristics An Emerging Approach*, volume 53. 2013.
- [107] Corey Sandler, Tom Badgett, and TM Thomas. The Art of Software Testing. page 200, September 2004.

- [108] I de Sousa Santos, AR Santos, and PA dos Santos Neto. Reusing Functional Testing in order to Decrease Performance and Stress Testing Costs. *SEKE*, 2011.
- [109] Yuji Sato and Taku Sugihara. Automatic generation of specification-based test cases by applying genetic algorithms in reinforcement learning. In *International Workshop on Structured Object-Oriented Formal Language and Method*, pages 59–71. Springer, 2015.
- [110] Yasir Shoaib and Olivia Das. Web Application Performance Modeling Using Layered Queueing Networks. *Electronic Notes in Theoretical Computer Science*, 275:123–142, 2011.
- [111] Marwa Shousha. *Performance Stress Testing of Real-Time Systems Using Genetic Algorithms*. PhD thesis, Carleton University Ottawa, 2003.
- [112] Marwa Shousha. *Performance Stress Testing of Real-Time Systems Using Genetic-Algorithms*. PhD thesis, Carleton University, 2003.
- [113] Connie U. Smith and Lloyd G. Williams. Software performance antipatterns. *Proceedings of the second international workshop on Software and performance - WOSP '00*, pages 127–136, 2000.
- [114] Connie U Smith and Lloyd G Williams. More New Software Performance AntiPatterns: EvenMore Ways to Shoot Yourself in the Foot. *Computer Measurement Group Conference*, pages 717–725, 2003.
- [115] C.U. Smith and L.G. Williams. Software Performance AntiPatterns; Common Performance Problems and their Solutions. *Cmg-Conference-*, 2:797–806, 2002.
- [116] 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.
- [117] Richard S. Sutton and Andrew G. Barto. Reinforcement learning. *Learning*, 3(9):322, 2012.
- [118] Csaba Szepesvári and Gabor Bartok. Algorithms for Reinforcement Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 4(x):1–103, 2010.
- [119] El-Ghazali Talbi. *Metaheuristics: from design to implementation*, volume 74. John Wiley & Sons, 2009.
- [120] El-Ghazali Talbi. *Hybrid Metaheuristics*, volume 2. 2012.
- [121] El-Ghazali Talbi. *Metaheuristics: From Design to Implementation*, volume 53. 2013.
- [122] N J Tracey, J a Clark, and K C Mander. Automated Programme Flaw Finding using Simulated Annealing. 1998.

- [123] Nigel James Tracey. *A search-based automated test-data generation framework for safety-critical software*. PhD thesis, Citeseer, 2000.
- [124] G Trent and M Sake. WebSTONE: The first generation in {HTTP} server benchmarking. *WWW Conference'95*, 1995.
- [125] Catia Trubiani. PhD Thesis in Computer Science Automated generation of architectural feedback from software performance analysis results Catia Trubiani. *Language*, 2011.
- [126] Martti Vasar, Satish Narayana Srirama, and Marlon Dumas. Framework for monitoring and testing web application scalability on the cloud. *Proceedings of the WICSA/ECSA 2012 Companion Volume on - WICSA/ECSA '12*, page 53, 2012.
- [127] Christian Vogeles, André van Hoorn, Eike Schulz, Wilhelm Hasselbring, and Helmut Krcmar. WESSION: extraction of probabilistic workload specifications for load testing and performance prediction??a model-driven approach for session-based application systems. *Software and Systems Modeling*, (October):1–35, 2016.
- [128] Xingen Wang, Bo Zhou, and Wei Li. Model Based Load Testing of Web Applications. *International Symposium on Parallel and Distributed Processing with Applications*, pages 483–490, 2010.
- [129] Xingen Wang, Bo Zhou, and Wei Li. Model-based load testing of web applications. *Journal of the Chinese Institute of Engineers*, 36(1):74–86, 2013.
- [130] J Wegener and M Grochtmann. Verifying timing constraints of real-time systems by means of evolutionary testing. *Real-Time Systems*, 15(3):275–298, 1998.
- [131] J. Wegener, H. Stamer, and H. Pohlheim. Testing the temporal behavior of real-time tasks using extended evolutionary algorithms. *Proceedings 20th IEEE Real-Time Systems Symposium (Cat. No.99CB37054)*, 1999.
- [132] Joachim Wegener, Harmen Stamer, Bryan F Jones, and David E Eyres. Testing real-time systems using genetic algorithms. *Software Quality Journal*, 6(2):127–135, 1997.
- [133] Harmen Wegener, Joachim and Pitschinetz, Roman and Stamer. Automated Testing of Real-Time Tasks. *Proceedings of the 1st International Workshop on Automated Program Analysis, Testing and Verification (WAPATV'00)*, 2000.
- [134] Alexander Wert, Jens Happe, and Lucia Happe. Supporting swift reaction: Automatically uncovering performance problems by systematic experiments. *Proceedings - International Conference on Software Engineering*, (May):552–561, 2013.
- [135] Alexander Wert, Marius Oehler, Christoph Heger, and Roozbeh Farahbod. Automatic detection of performance anti-patterns in inter-component communications. *QoSA 2014 - Proceedings of the 10th International ACM SIGSOFT Conference on Quality of Software Architectures (Part of CompArch 2014)*, pages 3–12, 2014.

- [136] EJ Weyuker and FI Vokolos. Experience with performance testing of software systems: issues, an approach, and case study. *IEEE transactions on software engineering*, 26(12):1147–1156, 2000.
- [137] Joshua White and Adam Pilbeam. A Survey of Virtualization Technologies With Performance Testing. page 6, 2010.
- [138] Minzhi Yan, Hailong Sun, Xu Wang, and Xudong Liu. Building a TaaS Platform for Web Service Load Testing. *2012 IEEE International Conference on Cluster Computing*, (2):576–579, 2012.
- [139] Cheer-Sun D Yang and Lori L Pollock. Towards a Structural Load Testing Tool. *SIGSOFT Softw. Eng. Notes*, 21(3):201–208, 1996.
- [140] Lian Yu, Wei Tek Tsai, Xiangji Chen, Linqing Liu, Yan Zhao, Liangjie Tang, and Wei Zhao. Testing as a service over cloud. In *Proceedings - 5th IEEE International Symposium on Service-Oriented System Engineering, SOSE 2010*, pages 181–188, 2010.
- [141] Pingyu Zhang, Sebastian Elbaum, and Matthew B Dwyer. Automatic generation of load tests. *Proceedings of the 2011 26th IEEE/ACM International Conference on Automated Software Engineering*, pages 43–52, 2011.