Institute of Software Technology
Reliable Software Systems

University of Stuttgart
Universitätsstraße 38
D–70569 Stuttgart

Bachelorarbeit

Controlled User Study: Usability and Efficiency Evaluation of the Parallel Performance Catalogue Extension for the Palladio-Bench

Denis Zahariev

Course of Study:	Softwaretechnik
Examiner:	Prof. Dr. Steffen Becker
Supervisor:	Markus Frank, M.Sc.
Commenced:	August 19, 2019
Completed:	Februray 19, 2020

Abstract

For the last two decades, the number of cores in modern CPUs has been steadily increasing. This enables a significant leap in the performance of modern software when the right parallel programming approaches and strategies are being used.

One tool used by software performance engineers to examine and evaluate the performance and reliability of pieces of software is the Palladio-Bench. This tool allows its users to analyse these and many more Quality of Software (QoS) properties such as sizing, scalability, and load balancing, based only on graphical models of the software architecture. After various pieces of research showed that the Palladio-Bench does not fully support parallelism, and the modelling of parallel programming strategies, a new extension for the tool was developed. This new extension for the Palladio-Bench incorporates fundamental parallel programming approaches into its already existing toolkit. The researchers that proposed the extension also claimed that it has higher usability and better time efficiency than the standard modelling toolkit of the Palladio-Bench. However, they were not able to prove it since the extension was not yet developed.

The purpose of this thesis is to put the supposed usability gains to the test. It compares the standard toolkit and the new extension in the context of the modelling of parallel behaviours. To support this study, a set of research questions was defined. The chosen research method was the conduction of a controlled empirical user study. Sixteen participants were recruited and split into two groups. Each group had to complete different modelling tasks with the standard toolkit and the extension. While they were working on the tasks, several metrics were recorded: task completion time, time spent in errors, number of errors, and usability evaluation. Afterwards, this data was statistically analysed and tested.

The results of the analysis prove that the extension increases the usability and the time efficiency of the Palladio-Bench. A reduction in the time spent in errors and the number of errors, however, could not be proved.

Kurzfassung

In den letzten zwei Jahrzehnten ist die Anzahl von Kernen in modernen Prozessoren stetig gestiegen. Dies ermöglicht einen signifikanten Anstieg in der Performanz moderner Software, wenn die richtigen Strategien und Ansätze für parallele Programmierung angewandt sind.

Ein Software-Werkzeug, das von Software-Performance-Ingenieuren verwendet wird, um die Performanz und Zuverlässigkeit zu analysieren und evaluieren, ist die Palladio-Bench. Dieses Werkzeug ermöglicht auch die Analyse von mehreren weiteren Softwarequalitätseigenschaften, wie Skalierbarkeit und Lastverteilung, basierend nur auf grafischen Modellen der Softwarearchitektur. Nachdem mehrere wissenschaftliche Arbeiten gezeigt haben, dass die Palladio-Bench Parallelität und die Modellierung von parallelen Programmieransätzen nicht völlig unterstützt, wurde eine neue Erweiterung für das Werkzeug entwickelt. Diese neue Erweiterung integriert fundamentale parallele Programmierstrategien in das schon existierende Toolkit von Palladio. Die Autoren, die die Erweiterung vorgeschlagen haben, behaupten, dass diese eine höhere Usability und Zeiteffizienz als das Standardtoolkit hat. Dies konnten sie aber nicht beweisen, da die Erweiterung noch nicht implementiert wurde.

Das Ziel dieser Thesis ist die Analyse der behaupteten Usability-Verbesserungen. Sie vergleicht das Standardtoolkit und die Erweiterung im Zusammenhang mit der Modellierung paralleler Verhalten. Um die Recherche zu unterstützen, wurden vier Forschungsfragen definiert. Die gewählte Forschungsmethode war die Durchführung einer kontrollierten empirischen Nutzerstudie. Sechzehn Teilnehmer wurden rekrutiert und in zwei Gruppen unterteilt. Während der Durchführung wurden mehrere Metriken gemessen: Zeit zur Aufgabenerfüllung, Zeit verbracht mit Fehlern, Fehleranzahl und Usability-Evaluation.

Die Resultate der Analyse haben die Verbesserungen in der Usability und der Zeiteffizienz erfolgreich bewiesen. Eine Reduzierung der Zeit, die mit Fehlern verbracht wurde und der Fehleranzahl konnte aber nicht bewiesen werden.

Contents

List of Figures

List of Tables

List of Acronyms

AT Architectural Template

GQM Goal-Question-Metric

PCM Palladio Component Model

PPC Parallel Performance Catalogue

QoS Quality of Software

SEFF Service Effect Specification

Chapter 1

Introduction

In the last 15 years, the number of cores in consumer and workstation CPUs has been steadily increasing. Even handheld devices such as a standard smartphone can have up to 8 cores. While these advances in the architecture of the CPUs allow faster parallel computation, the software must be written and optimised in a way that allows the utilisation of the growing number of cores. Such software often requires extensive reliability and performance analysis during its development. One tool that is especially useful for performance predictions and analysis is the Palladio-Bench.

Although the Palladio-Bench provides a rich modelling toolkit to its users, it has limited support for concurrency[BKR]. Furthermore, in the case when concurrency is supported, it requires a lot of manual modelling work even for simple parallel behaviours. Frank et al. examine Palladio's support of the modelling of two parallel programming approaches in their research[FSH17][FH18]. For one of the approaches, they conclude that although the defining characteristics of the approach are preserved when modelled in Palladio, a lot of time-consuming, error-prone, manual modelling work is required. To improve the modelling experience for this approach, they suggest an Architectural Template (AT) solution. For the other approach, they state that several abstractions have to be made in order to achieve a model of the approach. These abstractions, however, remove some of the defining characteristics of the approach. Therefore they conclude that an extension to the Palladio Component Model (PCM) is required to fully support the approach.

The research done by Frank et al. was also an inspiration for another study[ZWS19] conducted by Zahariev et al. The study gathered additional parallel programming approaches (such as Fork-Join and Pipes & Filters) and constructed a catalogue of important and fundamental approaches. Then all of the approaches in the catalogue were closely examined and modelled in Palladio. The authors received results similar to those of Frank et al., namely that the modelling process requires a lot of manual modelling work. The study also gave theoretical recommendations on how to incorporate the catalogue into the modelling toolkit of Palladio according to the AT method. This extension of the toolkit would result in an increase in the usability of Palladio and make the modelling process more time-efficient and less error-prone. However, these claims could not be tested during the study since the extension was not yet developed.

1.1. Problem Statement

Recently, the development of the extension proposed by Zahariev et al. has been finalised. The aim of this thesis is to examine the claimed improvements regarding the usability, error-proneness, and time-efficiency and to support the integration of the extension into the toolkit of Palladio with solid proofs. To gather such proofs, the extension has to be compared with the standard toolkit of Palladio in regard to the modelling of parallel behaviours. The chosen method for achieving this is the conduction of a controlled empirical user experiment, which allows a direct observation of the changes in the usability. In order for an empirical user experiment to be successful and to deliver reliable results, a detailed experimental design and its accompanying metrics and parameters have to be defined.

1.2. Research Objectives

In this section, the research questions of the thesis are defined, and a detailed research approach is outlined.

1.2.1. Research Question

The definition of the research questions follows a simplified variant of the Goal-Question-Metric (GQM) approach[BCR]. This approach allows the specification of a measurement system which can be especially suitable for empirical user studies. With this approach, each research question is accompanied by a goal that should be achieved and the metrics used for measurement of the answer to the question. Throughout the study, the following research questions, specified as GQM models, are going to be examined and answered:

Goal	Improve the usability of the Palladio-Bench
Question(RQ_1)	Does the Parallel Performance Catalogue improve the usability of the Palladio-Bench regarding the modelling of parallel behaviours?
Metric	Users' evaluation in the form of a survey

Table 1.1.: Research Question 1

Goal	Increase the time efficiency of the Palladio-Bench
Question(RQ_2)	Is the Parallel Performance Catalogue more time efficient than the standard modelling toolkit regarding the modelling of parallel behaviours?
Metric	Task completion time

Table 1.2.: Research Question 2

Goal	Reduce the error-proneness of the Palladio-Bench
Question(RQ_3)	Is the Parallel Performance Catalogue less error-prone than the standard modelling toolkit regarding the modelling of parallel behaviours?
Metric	Number of errors

Table 1.3.: Research Question 3

Goal	Reduce the time spent in errors while modelling parallel behaviours in the Palladio-Bench
Question(RQ_4)	Do users spend less time in errors while using the Parallel Performance Catalogue?
Metric	Time spent in errors

Table 1.4.: Research Question 4

1.2.2. Method

As already mentioned, the chosen method for answering the research questions of the thesis is a controlled empirical user experiment. A crucial part of the conduction of every user study is the creation of an objective and detailed experimental design, containing hypotheses, metrics, and a plan for the conduction. The experimental design will be constructed and improved in several iterations after being reviewed by a committee. After the completion of the design, suitable participants are going to be recruited, and participant groups will be created according to the experimental design. Also, during the recruitment, several workshop sessions will be planned for any inexperienced participants. Afterwards, a room suitable for the conduction of the experiment will be reserved, and machines will be equipped with the necessary software. The next step after the participants and rooms are prepared is the conduction of the user study. After the successful conduction of the experiment, the gathered data regarding the measured metrics will be presented. Then the measured data is going to be discussed and statistically analysed, supported by the usage of various tables, plots, and diagrams. The same data will then be used to test the hypotheses of the thesis. After the results from the hypothesis testing process are gathered, the research questions will be answered and evaluated. Afterwards, the thesis document will be finalised, and the final presentation of the thesis will be prepared. Flow chart 1.1 presents a graphical representation of the work packages that are a part of this thesis and their order of completion.

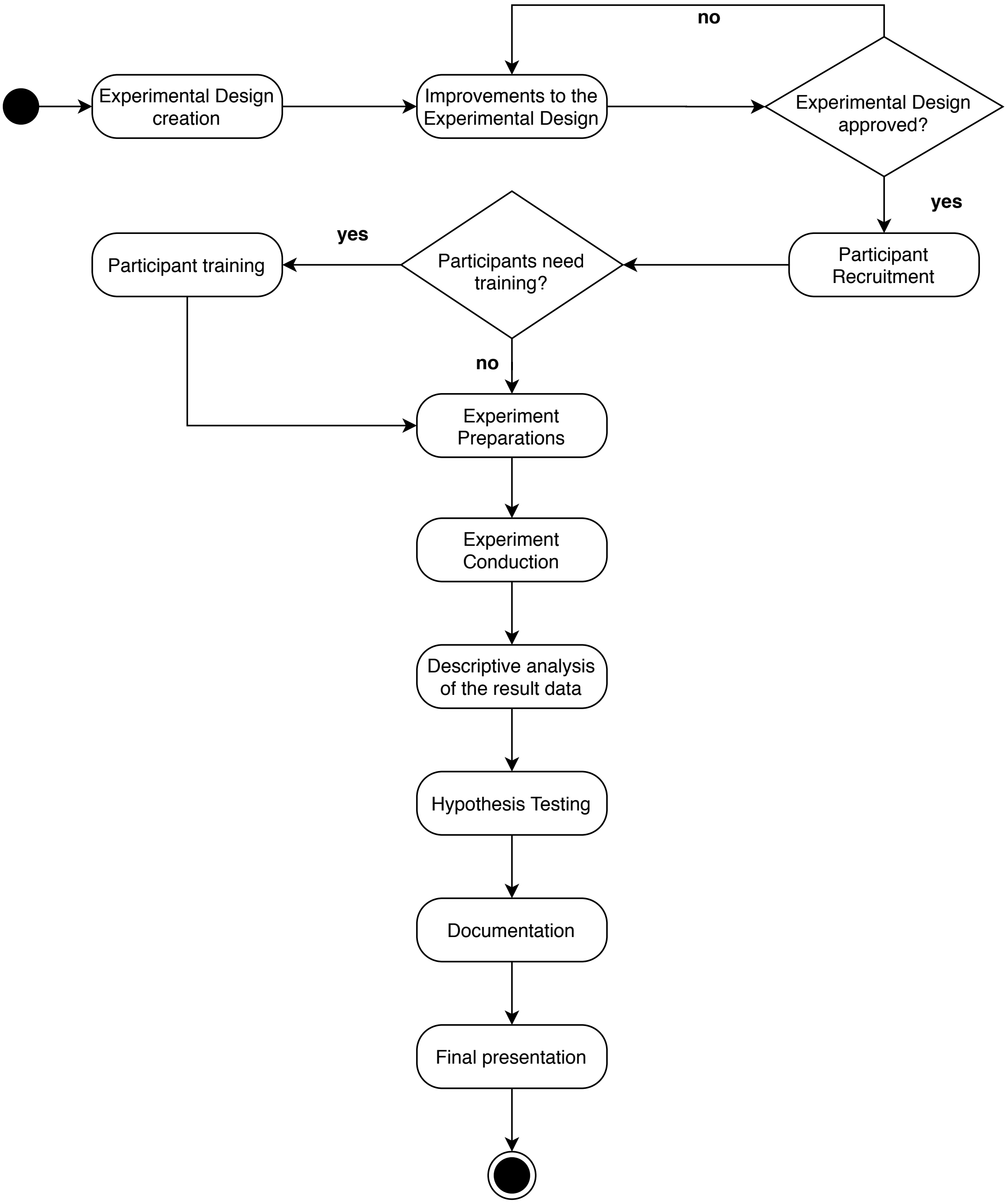

no
Experimental Design creation
Improvements to the Experimental Design
Experimental Design approved?
yes
yes
Participant training
Participants need training?
Participant Recruitment
no
Experiment Preparations
Experiment Conduction
Descriptive analysis of the result data
Hypothesis Testing
Documentation
Final presentation

Figure 1.1.: Research Methodology

Chapter 2

Technical Background and Foundations

2.1. Model-driven Software Development

The model-driven software development approach differs from other approaches in that models and modelling languages, such as the widely known UML, play a significantly more important role in the development process[VSB+]. The software development team constructs large and detailed models of the piece of software before the implementation process. Not only are the models more "user-friendly"[BBG05], but they also allow the communication and exchange of ideas and opinions of all the people that are a part of the team and not only the developers[VSB+]. The models also allow a better level of abstraction and the usage of predefined patterns for reoccurring behaviours and components[BBG05]. Afterwards, the actual code of the software could be generated automatically from the models, using a variety of different tools[VSB+], or it can be written manually.

2.2. Palladio

Palladio is a software architecture simulation approach that specialises in predicting various Quality of Software properties (QoS, e.g., performance and reliability)[BBB+16]. The Palladio-Bench is an Eclipse IDE-based supporting tool for the Palladio approach[Pal]. In the Palladio-Bench, users are able to graphically model the different components of a software architecture. These graphical models use the Palladio Component Model(PCM) meta-model and are respectively called PCM instances[BKR]. Furthermore, the Palladio-Bench enables users to model use-case scenarios, run various simulations on these, and provides visual representations of the results in the form of charts and diagrams[BKR].

The Palladio approach could be especially useful in model-driven development environments, allowing the development team to test, how reliable the architecture is, and to test the performance of the architecture on the planned hardware, on which it will be deployed, based only on the models and without the existence of any code[BKR].

2.3. Architectural Templates

In his Ph.D. Thesis[Leh18], Sebastian Lehrig introduces the Architectural Template (AT) method to Palladio. The main concept of this method is that software architects are able to apply already existing architectural knowledge in the form of patterns and templates to their architecture models. The application of such patterns and templates results in an easier and faster creation of a given software architecture. In order to incorporate the AT method in Palladio, the author creates a new Palladio extension that enables the application of architectural templates. Moreover, he provides an initial catalogue containing fundamental architectural templates that can be easily applied with the usage of his new extension.

2.4. Parallel Performance Catalogue

The following contents are based on research from this study[ZWS19].

Research done by Frank et al.[FSH17][FH18] regarding the modelling of parallel behaviours in Palladio states that some parallel programming approaches can not be modelled in Palladio and others can be modelled; however, the modelling process requires a lot of inefficient and error-prone manual modelling work. One of the papers concludes with a proposal for a new AT that should make the modelling of one of the approaches more efficient and less error-prone. These research results inspired another study[ZWS19] that gathered and researched additional parallel programming approaches. The main part of the research was to study which approaches can be modelled in Palladio, and when the construction of a model is possible if the defining properties of the approach are preserved in the model. After closely studying each pattern, the authors propose a new catalogue for the AT extension for Palladio, containing parallel programming approaches that can be incorporated in Palladio and also give recommendations on how to implement the respective architectural templates. The proposed Parallel Performance Catalogue contains the following approaches:

2.4.1. Fork-Join

The concept of this approach is that a task with a bigger workload can be forked, i.e., split into several tasks with a smaller workload, which are then executed in parallel. After the execution of all parallel threads completes, the execution can return to its original state, i.e., join[Eij17].

With the standard toolkit of Palladio, each of the forked threads has to be modelled manually. This process is doable for a small number of threads, e.g., four threads, without being too overwhelming. However, the modelling process becomes very overwhelming and inefficient when a higher number of parallel threads is introduced, for example, 16 or 32, which is a realistic scenario in modern times[ZWS19].

The Fork-Join AT in the Parallel Performance Catalogue solves the efficiency problems and eliminates the manual modelling work by allowing the users to model only a single thread and to create copies of it simply by applying the AT and specifying the number of copies.

2.4.2. Parallel Loops

The Parallel Loops pattern is an approach of achieving parallelism in loop constructs. It is applicable to loops that have a huge workload and require a lot of CPU work time, e.g., when there is a huge number of repetitions or the workload of a single iteration is very high. The parallelism is achieved by spawning a thread pool of a certain size and running a small number of iterations in each thread. To function properly, the operations in each iteration have to be independent of each other[MSM04].

With the standard toolkit of Palladio, all of the threads in the thread pool and the loops that they contain have to be modelled manually. Similarly to the already discussed Fork-Join approach, this is doable only for a small number of threads before the amount of required modelling work becomes too overwhelming[ZWS19].

The Parallel Loops AT in the Parallel Performance Catalogue enables the users to save a lot of manual modelling work by requiring them to model only a single parallel loop. Then after the application of the AT and the specification of the AT parameters, the rest of the parallel loops from the thread pool are created automatically.

2.4.3. Pipes & Filters

The Pipes & Filters is an approach that brings parallelism to data streams, where each stream consists of filters that represent computational steps and pipes which represent the flow between the filters. Defining characteristics of the approach are the expandability and interchangeability since filters can be easily added or removed[Mic17].

When modelling this approach with the standard toolkit of Palladio, every parallel pipeline has to be modelled manually. This includes the manual modelling of every filter and the respective pipes between the filters. Additionally, this must be completed in several diagrams. Similar to the already mentioned approaches, the manual modelling process for a large number of complex pipelines is very inefficient[ZWS19].

The Parallel Streams AT in the Parallel Performance Catalogue allows the users to model only a single parallel pipeline and to create multiple copies of it automatically. The AT creates all the necessary components in all diagrams, the connections between them, and the threads in which the pipelines will be executed.

2.5. Controlled Empirical User Experiments

The various empirical methods used in software engineering allow the direct observation of the human interaction with a given software, which could not be achieved via the conventional software testing methods, which primarily target qualities like functionality, performance, security, and reliability.

One of these methods is the controlled user experiment which is characterised by the ability to measure the effects of manipulating one variable on another variable[RH]. One of the advantages of this approach is the control of the subjects, objects, and instrumentation. Further advantages are the ability to perform statistical analysis and the possibility to replicate the experiment[WRH+12]. The main purpose of a controlled experiment is to prove or reject predefined hypotheses[DFAB03]. This is accomplished by observing, measuring and analysing the behaviour of the participants.

A vital part of the conception of a controlled user experiment is the definition of the research objectives and questions. One approach that can be used for this purpose is the Goal-Question-Metric (GQM) approach, which enables the definition of project goals in an operational and traceable manner[BRZ]. The goals are also used to derive questions, the answers to which are used to measure the accomplishment of the goal. The questions are also accompanied by metrics, which are derived from the questions. The purpose of these metrics is to define what information has to be gathered in order to answer the respective question. By following this approach, it is ensured that only the metrics relevant to the given goal are measured, and no unnecessary data is collected[RH].

The controlled user experiment is also characterised by having a fixed design[BRZ]. Therefore the creation of an experimental design is another crucial part of the conception of a controlled experiment. The careful creation of an experimental design ensures the reliability and generalisability of the results[DFAB03]. The first step of creating an experimental design is to specify the right hypotheses and the respective metrics and variables. The above-mentioned GQM approach can also be used in this step since it can allow a systematical and traceable grouping for each hypothesis and metric with a respective research goal. The next step is to select the experiment method, i.e. between-subject or within-subject, and to specify what each user has to do during the conduction of the experiment. The last step of the creation of the experimental design is the selection of the analysis procedures regarding the measured data and the hypotheses testing.

After the actual conduction of the controlled experiment ends, and the data is collected, the analysis of the data starts. The data is analysed with the help of descriptive statistics and presented with different plots, graphs, and diagrams in order to present and compare certain properties. The hypothesis testing process is also a part of the analysis. In this process, a hypothesis test, such as the widely used t-test[WRH+12], is used to either approve or reject the hypotheses defined in the experimental design. The results of the hypothesis testing process are then used to answer and evaluate the research questions.

Chapter 3

Related Work

This chapter presents relevant pieces of research and literature used throughout the controlled user experiment and the thesis document.

3.1. Design of Empirical User Studies in Model-based Software Development for Evaluating Modeling Languages

In his paper[Fra], Patrick Franczak discusses various empirical methods and techniques used for the purpose of conducting user studies. The paper also outlines the steps required to create a detailed experimental design for such studies. The author also examines if an empirical study is able to prove the supposed usability improvements of the AT method. He concludes that this is indeed possible and that the most suitable type of user study is a controlled experiment since it allows the researcher to take more control over the events during the conduction in comparison to the other empirical strategies. The author also proposes a basic example of an experimental design for such a controlled experiment.

3.2. An Efficiency Comparison Between Architectural Templates and SimuLizar

In his bachelor's thesis[Nüt], the author compares the efficiency of the Architectural Templates and SimuLizar extensions for the Palladio-Bench. For the purpose of the comparison, he conducts a controlled user study with two user groups. One of the groups has to complete tasks using the SimuLizar extension, and the other has to complete similar tasks using the Architectural Templates extension. After the conduction of the user study, the data measured during the execution of the tasks is statistically analysed, and the hypotheses introduced with the experimental design are tested. The results from the analysis state that the AT method is more time-efficient than the SimuLizar method.

3.3. Reporting Experiments in Software Engineering

In their work[JCP], Jedlitschka et al. provide a structuring proposal for the reporting of controlled experiments. They propose a specific outline that researchers can follow in order to present experiments and their results in a structured and systematic way. The structure of this thesis is based on their proposal to some extent.

3.4. Experimental Design

This thesis uses various pieces of detailed literature[BRZ][RH][WRH+12][DFAB03] about empirical methods used in the field of software engineering. The literature provides a deep insight into the process of creating detailed and objective experimental designs. Additionally, suitable metrics, data collection approaches, and data analysis guidelines are presented and discussed.

3.5. Hypothesis Testing

The book by Wohlin et al.[WRH+12] also provides extensive knowledge about different hypothesis testing techniques. This includes the definition of various hypothesis tests and practical examples. The book is used during the completion of the thesis in order to find the most suitable testing method.

Chapter 4

Experimental Design

The controlled user experiment is chosen as the research method in this thesis since as already mentioned, it allows the manipulation of variables[RH]. Another aspect contributing to the choice of this approach is the ability to perform statistical analysis of the data measured during the experiment[WRH+12]. The hypotheses of the experiment are introduced in accordance with the GQM models of the research questions they are regarding. This is done because of the traceability benefits introduced by the GQM approach[BRZ]. The independent variable of the experiment is the modelling method, i.e. the standard toolkit or the Parallel Performance Catalogue (PPC) extension. The dependent variables are the usability evaluations, task completion time, number of errors, and time spent in errors. These are also incorporated into the GQM models of the research questions in order to ensure that only relevant data is collected[RH]. Three of these metrics are also defined as suitable measurement criteria for usability measurements according to Dix et al.[DFAB03] (originally adapted from Whiteside et al.[WBH88]). The usability evaluation metric is gathered from a questionnaire that every user has to fill out. The usage of questionnaires or surveys during the conduction of controlled experiments is a common practice in order to gather more data[Fra]. The questionnaire is comprised of open text questions and questions, having a Likert scale as an answer. The Liker scale was chosen since it is a widely used, easy to develop, and engaging option[RM17]. The experiment features a within-subject design where every user performs under each condition. This method was chosen because it is less costly and requires fewer subjects than the between-subject design[DFAB03]. The within-subject design is also particularly effective when learning is involved[DFAB03] and is therefore suitable for the experiment since not all participants have experience with the Palladio-Bench. Additionally, possible transfers of learning effects are lessened by incorporating a technique suggested by Dix et al.[DFAB03], where each group does each condition in a different order. The approach chosen for hypothesis testing is a t-test, which is one of the most used parametric tests and is used to compare two sample means[WRH+12]. The t-test will be conducted with the conventional confidence level of 95%[RM17].

4.1. Goal

The goal of this experiment is to gather a sufficient amount of data that should allow the statistical analysis and the testing of the hypotheses. The analysis and the results of the hypothesis testing are then used to answer the research questions of the thesis.

4.2. Experimental Units

In order to observe how the extension affects a broad range of users with varying experience with Palladio, the study is conducted with two user groups, containing an equal number of experts, advanced users, and beginners in each group. As experts and advanced users are considered people with considerable or intermediate experience in the field of performance engineering or with Palladio. These are professors and research assistants recruited from the Reliable Software Systems Institute at the University of Stuttgart. As beginners are considered participants, having basic knowledge in the field of software engineering but still capable of completing the tasks they receive. The two participant groups are referred to as Group A and Group B. The number of participants in both groups is equal, and each group contains eight participants. Since it is expected that participants of all backgrounds are not familiar with Palladio, several workshop sessions are planned to take place in which the participants will be introduced to Palladio. If a participant is not able to attend any of these sessions, they are provided with the necessary materials and have to complete the workshop on their own.

4.3. Experimental Material

The conduction of the user study requires the creation of several documents. A workshop document, containing a detailed step by step Palladio tutorial, including an easy to understand realistic example, is created for the planned workshop sessions. A task document, containing the definition of the modelling tasks for each group, is used by the participants during the conduction of the user study. Also, a survey document is used during the user study in order to capture the usability evaluations of the users. Additionally, a consent form for participation in the user study has to be filled out by every participant before their participation. The survey document, task document, and consent form are combined together in a leaflet which each user receives during the conduction. The measured data for each user is also noted in a protocol by the supervisor conducting the user study. All of the documents are presented in the appendix of the paper.

4.4. Tasks

The conduction of the user study requires the creation of two use case scenarios. Both use case scenarios describe systems utilising parallel behaviours, which should be modelled in Palladio, and both use case scenarios contain two different tasks. In one of the tasks, the users are required to model the scenario using only the standard modelling toolkit of Palladio, and in the other, the users are required to model the scenario with the usage of the PPC extension, in particular, the Parallel Loops AT. Only the usage of one of the templates offered by the PPC is required, because a full comparison between each template and the standard toolkit requires a user study of a much larger scale. Furthermore, the tasks require only the modelling of the parallel behaviours in the Service Effect Specification (SEFF) diagram. Each user is provided with a Palladio project for each use case scenario, where every diagram is fully modelled except the SEFF diagram, which the user has to complete on their own.

4.5. Procedure

During the conduction of the user study, both user groups receive the two use case scenarios; however, each group has to complete a different task for each scenario. For example, Group A has to model the first scenario with the standard toolkit and the second scenario with the extension, while Group B has to do the tasks vice versa. This is done in order to lessen learning effects. During the modelling process, the task completion time, the number of errors, and the time spent in errors are recorded. All participants complete the modelling tasks in an individual session with the supervisor conducting the study, in order to precisely and accurately measure the mentioned metrics, which is achieved harder in group sessions. After a participant is done with the modelling, he has to fill out a questionnaire about his user experience.

4.6. Hypotheses, Parameters, and Variables

This is the section where the hypotheses and the metrics of the user study are defined and discussed. Each of the hypotheses concerns a respective research question, and for this purpose, the hypotheses are specified together with the GQM models they are referring to. Each of the hypotheses will be tested with the measured data regarding the metrics from the respective GQM model. The first research question is concerning the usability of the Palladio-Bench, and the respective hypothesis H_1 is the following:

Goal	Improve the usability of the Palladio-Bench
Question(RQ_1)	Does the Parallel Performance Catalogue improve the usability of the Palladio-Bench regarding the modelling of parallel behaviours?
Metric	Users' evaluation in the form of a survey
Hypothesis(H_1)	The questions regarding the usability of the Parallel Performance Catalogue have a higher mean score than the same regarding the usability of the standard toolkit.

Table 4.1.: Hypothesis 1

The metric that is going to be used to test the hypothesis is the user experience and usability evaluation of each user, which is obtained in the form of a survey at the end of the user study. The next hypothesis, H_2, is concerning the time efficiency and is specified as follows:

Goal	Increase the time efficiency of the Palladio-Bench
Question(RQ_2)	Is the Parallel Performance Catalogue more time-efficient than the standard modelling toolkit regarding the modelling of parallel behaviours?
Metric	Task completion time
Hypothesis(H_2)	The Parallel Performance Catalogue has a lower mean task completion time than the standard toolkit regarding the modelling of parallel behaviours.

Table 4.2.: Hypothesis 2

The same metric from the GQM model for the respective research question is going to be used for hypothesis testing, namely the task completion time, which is measured in seconds. This metric is recorded for every participant and every task. The next hypothesis, namely H_3, is regarding the error-proneness and is the following:

Goal	Reduce the error-proneness of the Palladio-Bench
Question(RQ_3)	Is the Parallel Performance Catalogue less error-prone than the standard modelling toolkit regarding the modelling of parallel behaviours?
Metric	Number of errors
Hypothesis(H_3)	The mean number of errors while using the Parallel Performance Catalogue is lower than the one while using the standard toolkit.

Table 4.3.: Hypothesis 3

The use cases used during the conduction of the experiment are constructed in such a way that there is only one possible solution, and any modelling action that does not lead to this solution is counted as an error. This number of errors is recorded for every participant and every task. Hypothesis H_4 is regarding the time spent in errors during the modelling process and reads as follows:

Goal	Reduce the time spent in errors while modelling parallel behaviours in the Palladio-Bench
Question(RQ_4)	Do users spend less time in errors while using the Parallel Performance Catalogue?
Metric	Time spent in errors
Hypothesis(H_4)	The mean time spent in errors while using the Parallel Performance Catalogue is lower than the one while using the standard toolkit.

Table 4.4.: Hypothesis 4

Having already defined what is considered as a modelling error, the time spent in an error is interpreted as the time a user spends in order to correct the error. In particular, this is the time interval between the occurrence and the removal of a given error. Again, this metric is measured for each participant and each task.

4.7. Analysis Procedure

First, the user evaluations from the user survey will be compared. The questions regarding the user experience are identical for all use cases, and since the answers are on a scale from one to seven, a mean score for each question will be calculated. The testing of H_1 will consist of comparing the mean evaluations for each usability regarding question. In order to test hypothesis H_2, the mean task completion time for each modelling method will be calculated and then compared. To test if hypothesis H_3 holds, the average number of errors for each method will be calculated and then compared. Similarly, to test hypothesis H_4, the mean time spent in errors for each method will be calculated and compared. Furthermore, a t-test with a confidence level of 95% will be run on the results regarding each hypothesis, which will allow the confident approval or rejectment of the respective hypothesis.

Chapter 5

Execution

This chapter covers the preparation steps and the actual conduction of the user study, which was thoroughly planned in the previous chapter.

5.1. Preparation

The first step of the user study conduction was participant recruitment. During this phase, 16 participants with different backgrounds were recruited. This selection consists of ten students, five research assistants, and one professor. They were then split into two groups of eight participants according to the experimental design. The number of beginners, advanced users, and experts was spread evenly across both groups as much as possible.

As already mentioned in the experimental design, it was expected that many of the participants did not have any experience with Palladio prior to the conduction of the experiment. Therefore, four workshop sessions were planned and took place. During these, the participants received the workshop document that contained a detailed introduction to Palladio, including detailed examples. The sessions took place in one of the computer laboratories of the RSS Institute, where all required computers were equipped with the necessary software. Some of the participants were not able to attend any of these sessions. Those participants were provided with the workshop document and were required to complete it on their own. In total, 13 participants had to complete the workshop. During the execution of the preparation phase, there were no deviations, and everything conformed to the initial plan.

5.2. User Study Conduction

All participants did the user study in private sessions, where only the participant and the supervisor were present. After receiving the user study leaflet, the participants were prompted to read it and to ask the supervisor if anything was unclear, in which case the supervisor provided them with answers. After answering the introductory questions from the questionnaire, the participants could begin with the completion of the task in the first use case scenario. At the start of the completion, the start time was noted by the supervisor, and a timer for 30 minutes was started. If the participants made modelling errors during the completion, they were informed, and the times of the error's occurrence and removal were noted. After the successful completion of the task, the end time was also noted. If a participant was not able to complete the task in the given time, the task's completion was interrupted, which was also noted. When done with the first task, the participants had to fill out a part of the questionnaire regarding their user experience during the completion of the task. Afterwards, the participants had to solve the task from the second use case scenario. The conditions under which the second task was to be completed and the data measured during this process were identical to the ones for the task in the first use case scenario. The second task was followed by the rest of the questionnaire. After the questionnaire was answered, the conduction of the experiment came to an end, and the participants were free to leave. During the execution of this phase, there were no deviations, and everything conformed to the initial plan.

5.2.1. Results

In the following sections, the results gathered from the user study are presented. All questionnaires and protocols can be found in the appendix of this thesis.

Group A

Table 5.1 and figure 5.1 show the task completion time of the participants in Group A. Group A is the group in which the participants completed their first task using the standard toolkit and the second with the PPC extension. Table 5.2 shows how many errors the participants made while solving the respective tasks and the total time spent in errors. As table 5.1 suggests, one of the participants was not able to finish one of the tasks in the 30 minute time window.

	Task completion time(in seconds)	
Participants	**Standard toolkit**	**PPC extension**
Participant 1 ◊	1582	339
Participant 2 △	1697	433
Participant 3 □	1372	343
Participant 4 △	1255	324
Participant 5 △	1447	425
Participant 6 △	1058	268
Participant 7 △	1344	327
Participant 8 □	not finished	455
Mean	**1393,57**	**364,25**
Standard Deviation	**210,22**	**65,48**

User Backgrounds: △ - Beginner; □ - Advanced; ◊ - Expert

Table 5.1.: Task completion time of Group A.

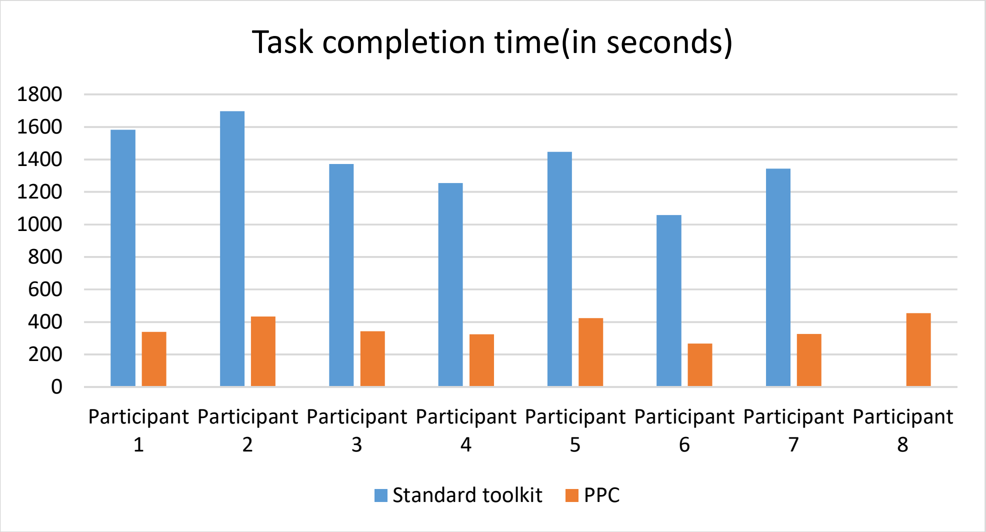


Figure 5.1.: Bar graph showing the task completion time of Group A.

	Standard toolkit		PPC extension	
Participants	**# of errors**	**Total time spent in errors**	**# of errors**	**Total time spent in errors**
Participant 1 ◊	0	0s	0	0s
Participant 2 △	2	43s	0	0s
Participant 3 □	1	16s	1	10s
Participant 4 △	1	14s	1	11s
Participant 5 △	0	0s	0	0s
Participant 6 △	0	0s	0	0s
Participant 7 △	1	52s	0	0s
Participant 8 □	0	0	0	0s
Total	5	125s	2	21s
Mean	0,63	15,63s	0,25	2,63s
Standard Deviation	0,74	20,88s	0,46	4,87s

User Backgrounds: △ - Beginner; □ - Advanced; ◊ - Expert

Table 5.2.: Number of errors and time spent in errors of Group A.

Group B

Table 5.3 and figure 5.2 show the task completion time of the participants in Group B. Group B is the group in which the participants completed their first task using the PPC extension and the second with the standard toolkit. Table 5.4 shows how many errors the participants made while solving the respective tasks and the total time spent in errors. Also in this group, one of the participants was not able to finish one of the tasks in the 30 minute time window.

	Task completion time(in seconds)	
Participants	**Standard toolkit**	**PPC extension**
Participant 9 ◊	not finished	240
Participant 10 □	1269	504
Participant 11 □	1172	505
Participant 12 □	1417	566
Participant 13 △	1472	504
Participant 14 △	1411	349
Participant 15 △	1680	493
Participant 16 △	1577	548
Mean	**1428,29**	**463,63**
Standard Deviation	**172,66**	**111,22**

User Backgrounds: △ - Beginner; □ - Advanced; ◊ - Expert

Table 5.3.: Task completion time of Group B.

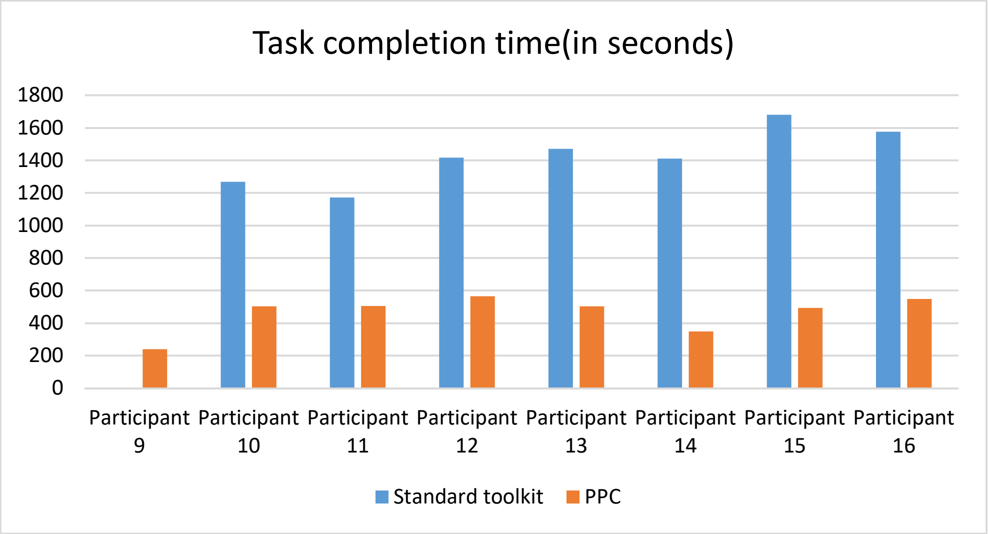


Figure 5.2.: Bar graph showing the task completion time of Group B.

	Standard toolkit		PPC extension	
Participants	**# of errors**	**Total time spent in errors**	**# of errors**	**Total time spent in errors**
Participant 9 ◊	0	0s	0	0s
Participant 10 □	0	0s	2	39s
Participant 11 □	1	5s	3	15s
Participant 12 □	2	25s	3	45s
Participant 13 △	3	97s	0	0s
Participant 14 △	0	0s	0	0s
Participant 15 △	1	32s	1	12s
Participant 16 △	0	0s	0	0s
Total	7	159s	9	111s
Mean	0,88	19,88s	1,13	13,88s
Standard Deviation	1,13	33,64s	1,36	18,42s

User Backgrounds: △ - Beginner; □ - Advanced; ◊ - Expert

Table 5.4.: Number of errors and time spent in errors of Group B.

Chapter 6

Analysis

In this chapter, the data recorded and presented in the previous chapter will be statistically analysed. The testing of the hypotheses introduced with the experimental design is also a part of this chapter.

6.1. Descriptive Statistics

The following sections discuss all measured metrics and provide useful statistics and diagrams regarding them.

6.1.1. Questionnaire answers

The questionnaire contained 17 questions; however, not all of them are going to be analysed in this section, since some of them serve as an introduction and others ask about overall feedback regarding the conduction of the user study. This section focuses only on the questions regarding the user experience and usability of the two modelling methods. As a result, only the answers of nine questions are compared.

Of these nine questions, there are three questions regarding the usability of the standard toolkit and three more regarding the usability of the PPC extension. The questions about each method are actually identical and inquire about the same information, and therefore for the purpose of the analysis, they will be summarised together into only three questions. The questions and their possible answers are the following:

Q_1: How would you rate your performance regarding the task in Use Case Scenario 1/2?

A: A scale from 1 to 7, where 1 denotes "very slow" and 7 denotes "very fast".

Q_2: How would you rate the amount of work required for completing the task in Use Case Scenario 1/2?

A: A scale from 1 to 7, where 1 denotes "too little" and 7 denotes "too much".

Q_3: How would you rate the usability of the standard toolkit/PPC extension regarding the modelling of parallel behaviours and your user experience with it?

A: A scale from 1 to 7, where 1 denotes "very bad" and 7 denotes "very good".

The numbering of the above-introduced questions does not correspond to the original numbering in the questionnaire in order to make the reading of the thesis easier.

Figures 6.1, 6.2, and 6.3 show the answers to Q_1, Q_2, and Q_3 of all participants for each method in the form of Likert plots. Boxplots 6.4, 6.5, and 6.6 provide an additional comparison between the values for each question.

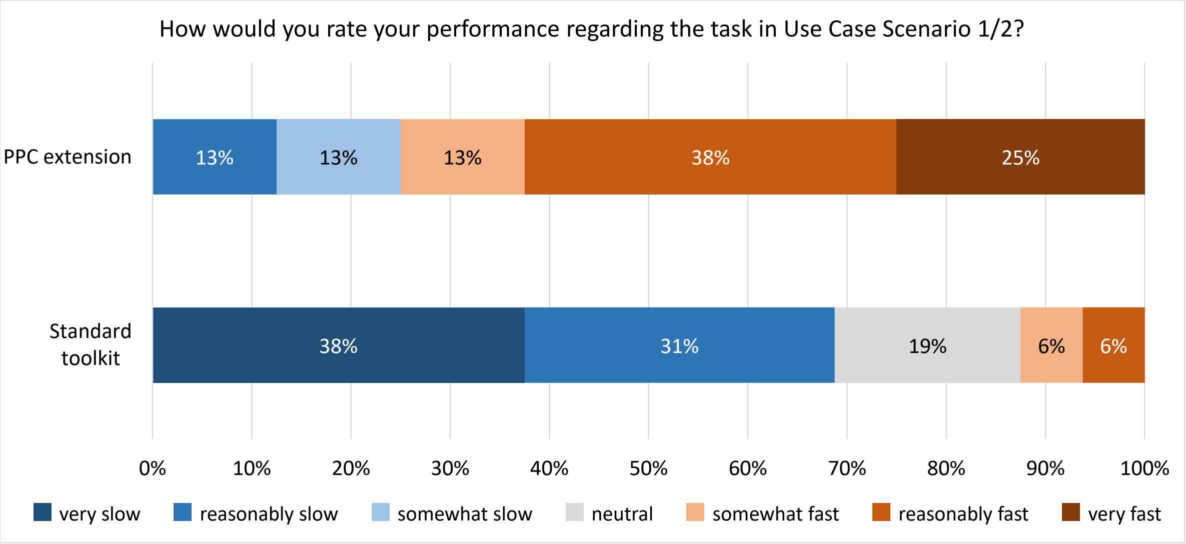


Figure 6.1.: Likert plot showing the answers to Q_1.

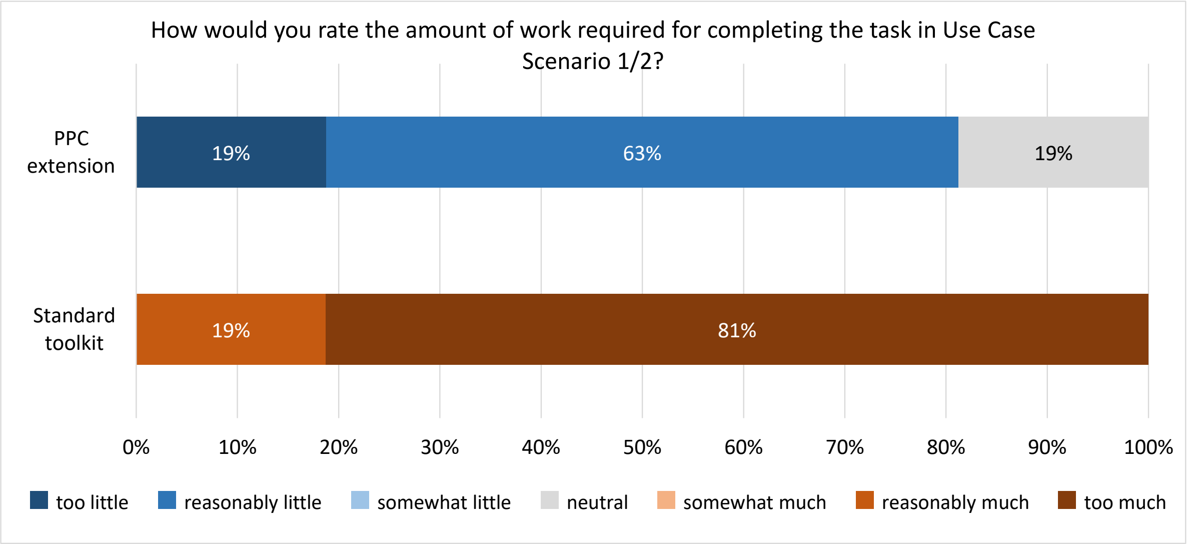


Figure 6.2.: Likert plot showing the answers to Q_2.

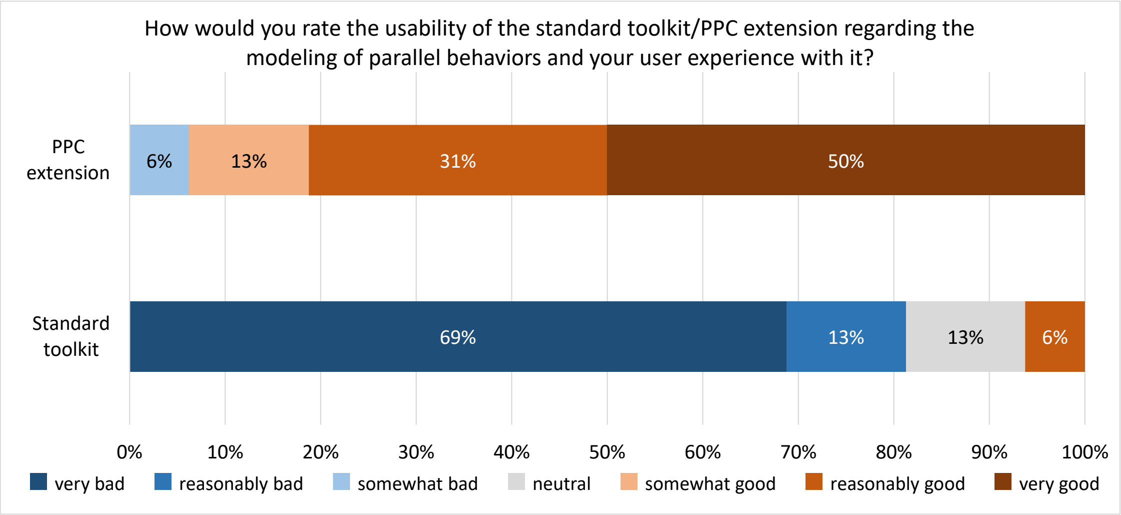


Figure 6.3.: Likert plot showing the answers to Q_3.

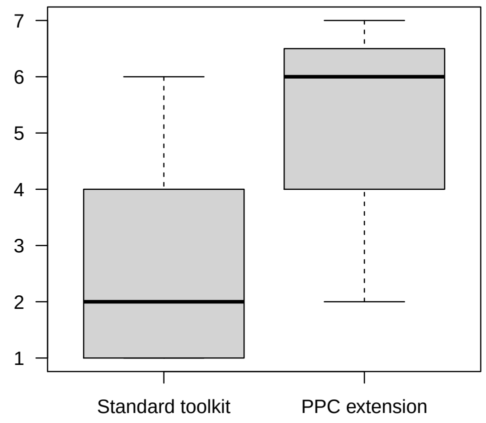


Figure 6.4.: Boxplot for the answers to Q_1.

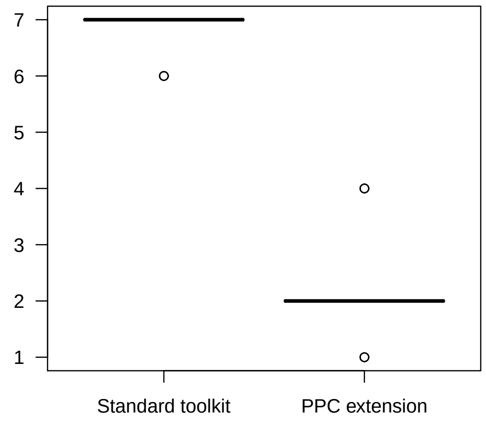


Figure 6.5.: Boxplot for the answers to Q_2.

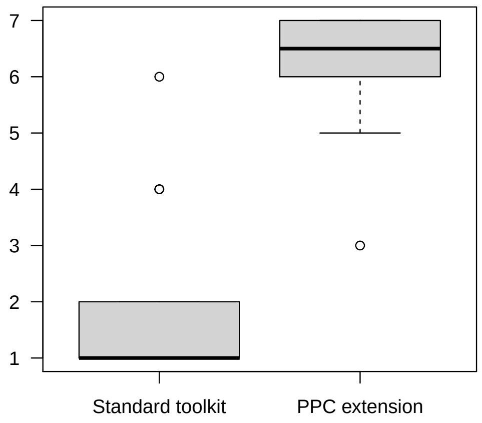


Figure 6.6.: Boxplot for the answers to Q_3.

When evaluating the mean scores for each question and method, it becomes clear that the users felt that they were faster when using the PPC extension. The users also rated the amount of work required with the standard toolkit significantly higher than the amount of work required with the PPC extension. Additionally, they rated the usability of the PPC extension significantly better than this of the standard toolkit.

The remaining three questions that have to be analysed are a comparison between the two modelling methods. The questions and their possible answers are defined as follows:

Q_4: How would you rate the usability of the Parallel Performance Catalogue in comparison to the standard toolkit?

A: A scale from 1 to 7, where 1 denotes "worse than the standard toolkit" and 7 denotes "significantly better and easier than the standard toolkit".

Q_5: How would you rate the following statement: "The Parallel Performance Catalogue introduces a very significant speed-up regarding the modelling of parallel behaviours."?

A: A scale from 1 to 7, where 1 denotes "false" and 7 denotes "true".

Q_6: Would you recommend the usage of the Parallel Performance Catalogue to other users of Palladio?

A: A scale from 1 to 7, where 1 denotes "definitely no" and 7 denotes "definitely yes".

Again, the numbering of these questions does not correspond with the original numbering in the questionnaire.

Since the answers given by the users to questions Q_4, Q_5, and Q_6 are not particularly suitable for a Likert plot representation, they are depicted in table 6.1. The answers of the users further reinforce their views that the PPC extension introduces a big speed-up and better usability.

Participants	Q_4	Q_5	Q_6
Participant 1	7	7	7
Participant 2	7	7	7
Participant 3	7	7	7
Participant 4	7	7	7
Participant 5	7	7	7
Participant 6	6	7	7
Participant 7	7	7	7
Participant 8	7	7	7
Participant 9	7	7	7
Participant 10	7	7	7
Participant 11	6	7	7
Participant 12	7	7	6
Participant 13	7	7	7
Participant 14	7	7	7
Participant 15	7	7	7
Participant 16	7	7	7
Mean	**6.875**	**7**	**6.9375**

Table 6.1.: Answers to questions Q_4, Q_5, and Q_6.

6.1.2. Task completion time

In order to provide a better overview of this metric, the boxplot in figure 6.7 is introduced. The figure provides a graphical comparison between the task completion time for the two methods. The figure uses the combined data from tables 5.1 and 5.3. When this data is put side by side, it becomes clear that there is a significant difference between the two methods regarding the time required to complete a model. It should also be noted that two participants were not able to complete the task using the standard toolkit within 30 minutes, which further emphasises the inefficiency of this method. Moreover, any reoccurring patterns and significant differences between the times of beginner, advanced, and expert users can not be identified. Hence, user backgrounds are not further discussed.

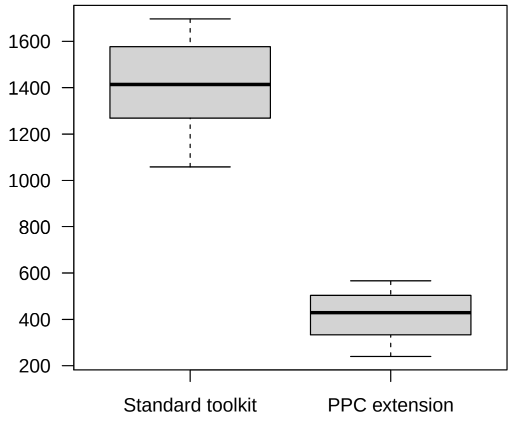


Figure 6.7.: Boxplot presenting the task completion time with each modelling method.

6.1.3. Number of errors

When the data regarding the number of errors in tables 5.2 and 5.4 is compared side by side, it is noticeable that each group made more errors during their first task. However, there is only a slight difference in the total number of errors for each method, namely 12 errors with the standard toolkit and 11 with the PPC extension. The same observation can also be made with the mean numbers for each method. Furthermore, when looking at each participant individually, clear patterns regarding the occurrence of errors can not be identified. Additionally, differences between the user backgrounds can not be found; therefore, they are not further discussed. Figure 6.8 introduces a boxplot constructed from the combined data regarding the number of errors in tables 5.2 and 5.4.

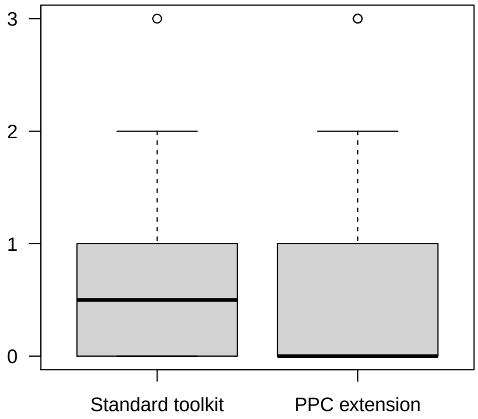


Figure 6.8.: Boxplot presenting the number of errors with each modelling method.

6.1.4. Time spent in errors

When the data regarding the time spent in errors of both groups in tables 5.2 and 5.4 is brought together, it is noticeable that the total time spent in errors with the standard toolkit is double the same with the PPC extension. Nonetheless, a comparison between the mean values reveals only a minor difference. Furthermore, when looking at each user individually and at the different user backgrounds, no reoccurring patterns regarding the time spent in errors can be identified. Figure 6.9 contains a boxplot constructed from the combined data in tables 5.2 and 5.4 regarding the time spent in errors.

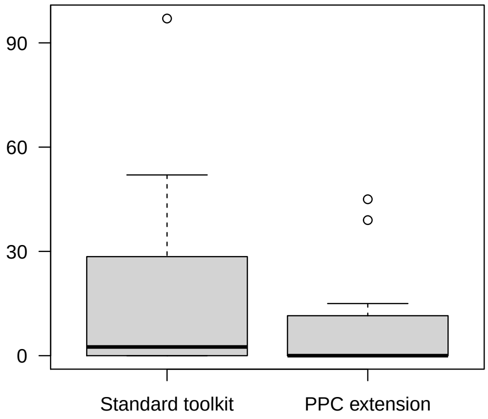


Figure 6.9.: Boxplot presenting the time spent in errors with each modelling method.

6.2. Hypothesis Testing

As mentioned in the experimental design, all hypotheses will be tested with a t-test with a confidence level of 95%. This thesis uses the t-test definition of Wohlin et al. [WRH+12]. Table 6.2 specifies the data and the calculations needed in order to conduct a t-test on a given hypothesis. In the following sections, each of the hypotheses is tested and respectively approved or rejected.

Input	Two independent samples: $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_m$
H_0	$\mu_x = \mu_y$ i.e. the expected mean values are the same
Calculations	Calculate $t_0 = \frac{\bar{x}-\bar{y}}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}}$ where $S_p = \sqrt{\frac{(n-1)S_x^2+(m-1)S_y^2}{n+m-2}}$ and, S_x^2 and S_y^2 are the individual sample variances
Criterion	Two-sided ($H_1 : \mu_x \neq \mu_y$): reject H_0 if $\lvert t_0 \rvert > t_{\alpha/2,n+m-2}$ One-sided ($H_1 : \mu_x > \mu_y$): reject H_0 if $t_0 > t_{\alpha,n+m-2}$ Here, $t_{\alpha,f}$ is the upper α percentage point of the t distribution with f degrees of freedom, which is equal to $n + m - 2$

Table 6.2.: T-test definition according to Wohlin et al. [WRH+12].

The following one-sided t-distribution table 6.3 for $\alpha = 0.05$, based on[Mas], is used during the hypothesis testing:

Degrees of freedom	Critical values of t	Degrees of freedom	Critical values of t
1	6,314	16	1,746
2	2,920	17	1,740
3	2,353	18	1,734
4	2,132	19	1,729
5	2,015	20	1,725
6	1,943	21	1,721
7	1,895	22	1,717
8	1,860	23	1,714
9	1,833	24	1,711
10	1,812	25	1,708
11	1,796	26	1,706
12	1,782	27	1,703
13	1,771	28	1,701
14	1,761	29	1,699
15	1,753	30	1,697

Table 6.3.: One sided t-distribution table for $\alpha = 0.05$, based on [Mas].

6.2.1. Hypothesis H_1

Before starting the hypothesis testing process, hypothesis H_1 is reintroduced in the following table 6.4:

Goal	Improve the usability of the Palladio-Bench
Question(RQ_1)	Does the Parallel Performance Catalogue improve the usability of the Palladio-Bench regarding the modelling of parallel behaviours?
Metric	Users' evaluation in the form of a survey
Hypothesis(H_1)	The questions regarding the usability of the Parallel Performance Catalogue have a higher mean score than the same regarding the usability of the standard toolkit

Table 6.4.: Hypothesis 1

The respective null hypothesis that is used for the testing reads as follows:

H_0: The mean scores of the usability regarding questions for the Parallel Performance Catalogue and the standard toolkit are equal.

Since the answers to questions Q_4, Q_5, and Q_6 are not suitable for comparison with a t-test, they are not tested in this section. The mean scores for each question, however, are strongly in favor of the Parallel Performance Catalogue, and therefore they will be regarded as contributing to the approval of hypothesis H_1.

The mean scores of questions Q_1, Q_2, and Q_3 are tested separately as follows:

Q_1: The individual sample variances are respectively equal to $2,6625$ for the standard toolkit and $3,1333$ for the Parallel Performance Catalogue. Having these values, S_p can be calculated and is equal to $1,7023$. Finally, t_0 is calculated and is equal to $4,6730$. When t_0 is compared to $t_{0.05,30}$ according to the one-sided criterion, it is clear that the difference in the means is significant.

Q_2: The individual sample variances are respectively equal to $0,1625$ for the standard toolkit and $0,9625$ for the Parallel Performance Catalogue. Having these values, S_p can be calculated and is equal to $0,75$. Finally, t_0 is calculated and is equal to $17,4420$. When t_0 is compared to $t_{0.05,30}$ according to the one-sided criterion, it is clear that the difference in the means is significant.

Q_3: The individual sample variances are respectively equal to $2,2958$ for the standard toolkit and $1,2292$ for the Parallel Performance Catalogue. Having these values, S_p can be calculated and is equal to $1,3276$. Finally, t_0 is calculated and is equal to

$9,3209$. When t_0 is compared to $t_{0.05,30}$ according to the one-sided criterion, it is clear that the difference in the means is significant.

After proving that the means for each question have a significant difference, hypothesis H_0 is confidently rejected, and respectively H_1 holds.

6.2.2. Hypothesis H_2

Before proceeding with the testing, hypothesis H_2 is presented once again in table 6.5. The null hypothesis used for the testing of H_2 is the following:

H_0: The mean task completion time of the Parallel Performance Catalogue and the standard toolkit are equal.

Goal	Increase the time efficiency of the Palladio-Bench
Question(RQ_2)	Is the Parallel Performance Catalogue more time-efficient than the standard modelling toolkit regarding the modelling of parallel behaviours?
Metric	Task completion time
Hypothesis(H_2)	The Parallel Performance Catalogue has a lower mean task completion time than the standard toolkit regarding the modelling of parallel behaviours.

Table 6.5.: Hypothesis 2

Since two users could not finish the task in the given time, they are not considered during the testing, and the sample size for the standard toolkit is regarded as 14. The individual sample variances are equal to $34480,5330$ for the standard toolkit and $10406,4625$ for the Parallel Performance Catalogue. Accordingly, S_p is equal to $146,9140$ and t_0 equals $18,5435$. After comparing t_0 and $t_{0.05,28}$ according to the one-sided criterion, the significant difference between the means is proven. As a result, the null hypothesis H_0 is rejected, and H_2 holds.

6.2.3. Hypothesis H_3

The null hypothesis used for the test of hypothesis H_3, shown once again in table 6.6 is the following:

H_0: The mean number of errors while using the Parallel Performance Catalogue is equal to the one while using the standard toolkit.

Goal	Reduce the error-proneness of the Palladio-Bench
Question(RQ_3)	Is the Parallel Performance Catalogue less error-prone than the standard modelling toolkit regarding the modelling of parallel behaviours?
Metric	Number of errors
Hypothesis(H_3)	The mean number of errors while using the Parallel Performance Catalogue is lower than the one while using the standard toolkit.

Table 6.6.: Hypothesis 3

The independent sample variances are equal to $0,8667$ for the standard toolkit and $1,1625$ for the Parallel Performance Catalogue. S_p equals $1,0073$ and t_0 equals $0,1755$. The comparison between t_0 and $t_{0.05,30}$ does not prove a significant difference between the means. Therefore hypothesis H_3 is rejected, and the null hypothesis holds.

6.2.4. Hypothesis H_4

Before specifying the null hypothesis used during the test, hypothesis H_4 is reintroduced in table 6.7.

Goal	Reduce the time spent in errors while modelling parallel behaviours in the Palladio-Bench
Question(RQ_4)	Do users spend less time in errors while using the Parallel Performance Catalogue?
Metric	Time spent in errors
Hypothesis(H_4)	The mean time spent in errors while using the Parallel Performance Catalogue is lower than the one while using the standard toolkit.

Table 6.7.: Hypothesis 4

The null hypothesis regarding H_4 reads as follows:

H_0: The mean time spent in errors while using the Parallel Performance Catalogue is equal to the one while using the standard toolkit.

The individual sample variances are equal to $736,4667$ for the standard toolkit and $203,1333$ for the Parallel Performance Catalogue. S_p equals $21,6749$ and t_0 equals $1,2397$. After comparing t_0 and $t_{0.05,30}$ according to the one-sided criterion, a significant difference between the means can not be proved. As a result, hypothesis H_4 is rejected, and the null hypothesis holds.

Chapter 7

Interpretation

In this chapter, the research questions of the thesis, which were defined in 1.2, are answered and evaluated. Additionally, limitations and possible threats to the validity of the thesis are presented.

7.1. Evaluation of Research Questions

After the testing of all hypotheses concluded, the answers to the research questions can now be presented.

RQ_1: The first research question RQ_1 is regarding the usability improvements that the PPC extension introduces. After showing that the participants in the user experiment evaluated the usability of the PPC extension better than this of the standard toolkit and after showing that hypothesis H_1 holds, the question from the GQM model associated with RQ_1 can be answered positively. This in respect, means that the goal from the GQM model, which is to increase the usability of the Palladio-Bench by introducing the new extension into its toolkit, is successfully achieved.

RQ_2: The next research question of the thesis, namely RQ_2, is concerning the time efficiency of the PPC extension. The comparison between the task completion times of both modelling methods and the test of hypothesis H_2 provide a positive answer to the question defined in the GQM model regarding RQ_2. Consequently, the goal to increase the time efficiency of the Palladio-Bench is successfully accomplished.

RQ_3: Research question RQ_3 is about the error-proneness of the modelling process. After observing the data regarding the number of errors made with each modelling method and the test of hypothesis H_3 which was rejected, the question from the GQM model for RQ_3 can not be answered positively. Therefore, the goal of reducing the error-proneness of the Palladio-Bench is not achieved.

RQ_4: The last research question of the thesis, namely RQ_4, is regarding the time spent in errors during the modelling process. The inability to show significant differences in the time spent in errors for each of the two modelling methods and the rejectment of hypothesis H_4 result in a negative answer to the question specified in the GQM model regarding RQ_4. As a result, the goal of reducing the time spent in errors is not accomplished.

7.2. Limitations and Threats to Validity

In the following two sections, first, the limitations of the thesis are presented, and then the threats to its validity.

7.2.1. Limitations

The thesis compares the standard toolkit of the Palladio-Bench and the PPC extension with regards to the modelling of parallel behaviours. The standard toolkit provides a wide variety of different modelling solutions in all areas of the Palladio-Bench; however, for the purpose of the thesis, only those used in Palladio's SEFF diagram were considered. Similarly, the PPC contains several ATs; however, only one of them was used during the conduction of the user study. Additionally, the PPC extension requires the creation of numerous additional files in order to run a simulation on a given model. These, however, were created beforehand and provided to the users since the user study focused only on the actual modelling process.

7.2.2. Threats to Validity

The following sections present possible threats to the validity of the research in this thesis.

Flawed Experimental Design

It is possible that a flawed experimental design was conceived and later accepted. In such a case, a different experimental design could have lead to different results. In order to mitigate this threat, the experimental design was reviewed by a committee before it was accepted.

Inexperienced Users

Even though the participants came from different backgrounds, the majority of them were new to Palladio. A selection of users, where everyone is an experienced Palladio user, could have resulted in other results. To mitigate this threat, a workshop took place where the inexperienced participants were trained on how to use the Palladio-Bench.

Insufficient Training

The contents of the workshop may not have been sufficient enough, resulting in insufficient participant training, which could eventually lead to the inability to solve the given tasks. To eliminate this threat, the contents of the workshop were reviewed repeatedly by the supervisor of the thesis.

Inadequate sample size

The sample size of the user experiment, i.e. the number of recruited participants, may have been too small. A larger sample size could lead to different results and outcomes.

Unsuitable Use Case Scenarios and Task Definitions

Flawed use case scenarios and task definitions that are either too easy or too difficult could have been created. To counter this, the use case scenarios and the task definitions were reviewed by the supervisor of the thesis.

Wrong Template Selection

The AT selected to represent the PPC could have been the wrong one for the purpose of the user study. The selection of another template could have also lead to different results. To mitigate this threat, the selection of the template that would represent the capabilities of the PPC at best was discussed with the supervisor of the thesis.

Inaccurate Data Measurement

It might have been that the metrics and data measured during the conduction of the user study were recorded inaccurately, which would lead to different results. To counter inaccurate measurements, the measurements were noted in protocols as they were recorded, and the times were recorded with the help of a stopwatch.

Chapter 8

Conclusion

The final chapter of the thesis contains a summary of the whole research process, presents valuable lessons learned during this process, and provides an insight into possible future work.

8.1. Summary

The purpose of this thesis was to evaluate the supposed quality improvements resulting from the introduction of the PPC extension to the Palladio-Bench. These quality improvements were summarised in the research questions of the thesis. The chosen method of research was the conduction of a controlled empirical user experiment. This required the creation of an extensive experimental design, where metrics, parameters, and hypotheses were specified. The experimental design also required the creation of the use case scenarios, the tasks, the questionnaires, and the protocols used during the conduction. Sixteen participants were recruited, and the majority of them had to be trained to work with the Palladio-Bench. After all of the above was completed, the actual user study took part, during which all of the metrics specified in the experimental design were measured. After the successful conduction of the user study, all of the measured data was presented and then analysed. As a result of the analysis and the hypothesis testing process, two of the hypotheses were proved to be true, namely that the PPC extension increases the usability and time efficiency of the Palladio-Bench with regards to the modelling of parallel behaviours. However, the remaining two hypotheses regarding the reduction of the error-proneness and time spent in errors were rejected. The results of the hypothesis testing were then interpreted into answers to the research questions. Lastly, the limitations and the threats to the validity of the thesis were discussed.

8.2. Lessons Learned

The first lesson learned after the conclusion of the user experiment and the analysis is that even though the PPC extension supposedly limits the extent to which errors can occur, the users are still able to make a number of errors, comparable to the one with the standard toolkit. The same also applies to the time spent in errors during the modelling process.

Another lesson learned after the recruitment process is that there is only a small number of people who have considerable expertise with the Palladio-Bench and can be considered as experts. As a result, the selection pool was limited, and the recruitment was harder.

The last lesson learned is that the planning and the conduction of a user experiment can be a long and time-consuming process. The creation of a detailed experimental design requires a lot of work and the organisational work, such as the planning of each user session, has to be done timely. These aspects should not be underestimated when planning a user experiment. Fortunately, these were considered during the conception of this thesis, and it was not affected negatively.

8.3. Future Work

After showing that the PPC extension successfully increases the usability and time efficiency of the Palladio-Bench, the current number of ATs can be further extended by the implementation of more templates for other parallel programming approaches and strategies. Additionally, new extensions based on the AT method, supporting various other approaches and strategies, could be implemented in order to increase the usability of other areas of the Palladio-Bench.

The reduction of the error-proneness and the time spent in errors, which this thesis could not prove, could be the basis for new studies and research. Also, other solutions for the reduction of the mentioned quality attributes can be researched and implemented. One possible approach to this could be to study the graphical user interface of the Palladio-Bench, and in particular, the PPC extension and to limit the points at which users can make unwanted errors.

Appendix A

Appendix

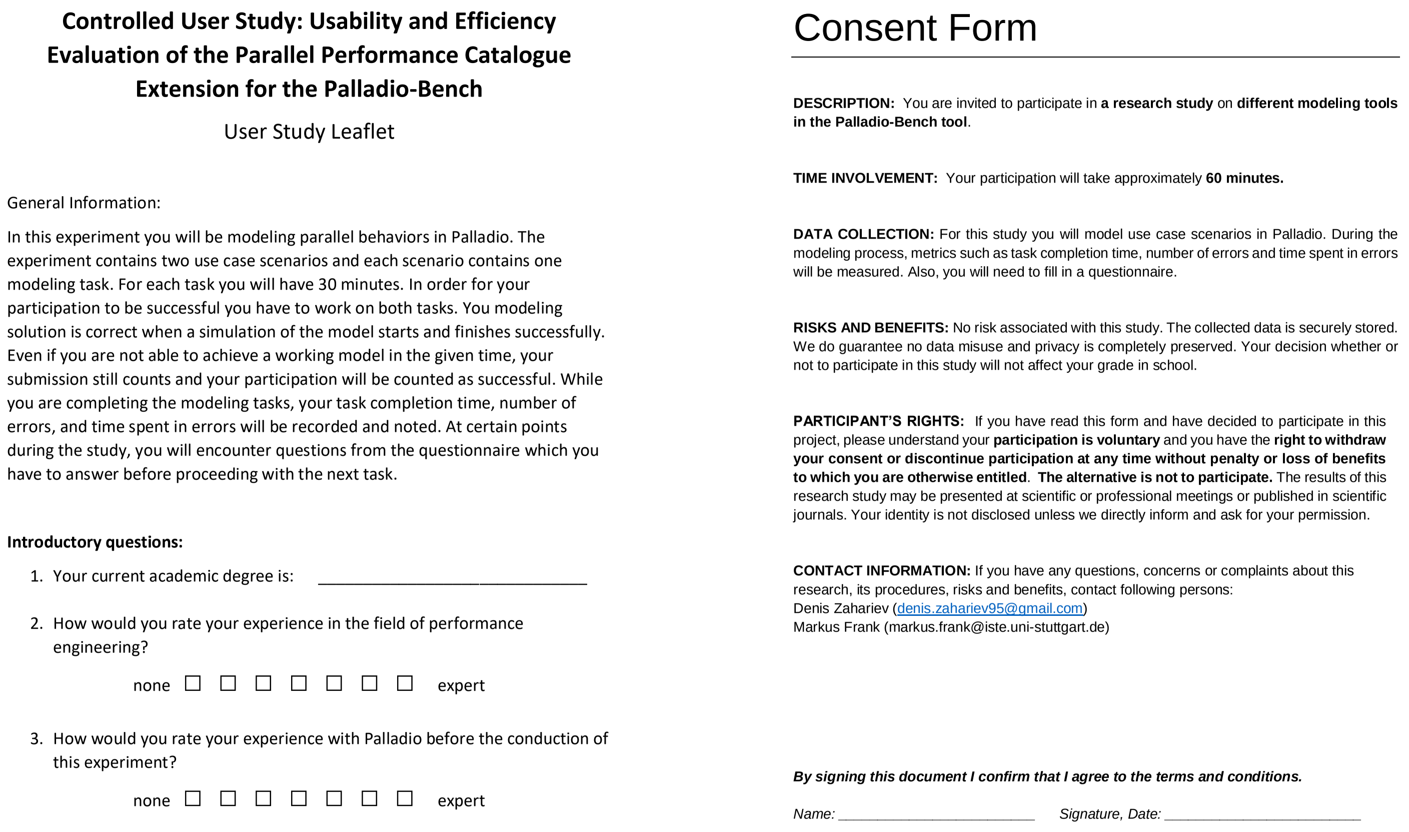
A.1. User Study Leaflet

The user study leaflet consists of several items:

1. General information about the conduction of the user experiment.
2. Consent form.
3. Use case scenario descriptions and task definitions.
4. Questionnaire

In the following two sections, the two versions of the leaflet distributed to each group are presented.

A.2. Blank User Study Leaflet - Group A



Controlled User Study: Usability and Efficiency Evaluation of the Parallel Performance Catalogue Extension for the Palladio-Bench

User Study Leaflet

General Information:

In this experiment you will be modeling parallel behaviors in Palladio. The experiment contains two use case scenarios and each scenario contains one modeling task. For each task you will have 30 minutes. In order for your participation to be successful you have to work on both tasks. You modeling solution is correct when a simulation of the model starts and finishes successfully. Even if you are not able to achieve a working model in the given time, your submission still counts and your participation will be counted as successful. While you are completing the modeling tasks, your task completion time, number of errors, and time spent in errors will be recorded and noted. At certain points during the study, you will encounter questions from the questionnaire which you have to answer before proceeding with the next task.

Introductory questions:

1. Your current academic degree is: ______________________________
2. How would you rate your experience in the field of performance engineering?

   none □ □ □ □ □ □ □ expert

3. How would you rate your experience with Palladio before the conduction of this experiment?

   none □ □ □ □ □ □ □ expert

Consent Form

DESCRIPTION: You are invited to participate in **a research study** on **different modeling tools in the Palladio-Bench tool**.

TIME INVOLVEMENT: Your participation will take approximately **60 minutes.**

DATA COLLECTION: For this study you will model use case scenarios in Palladio. During the modeling process, metrics such as task completion time, number of errors and time spent in errors will be measured. Also, you will need to fill in a questionnaire.

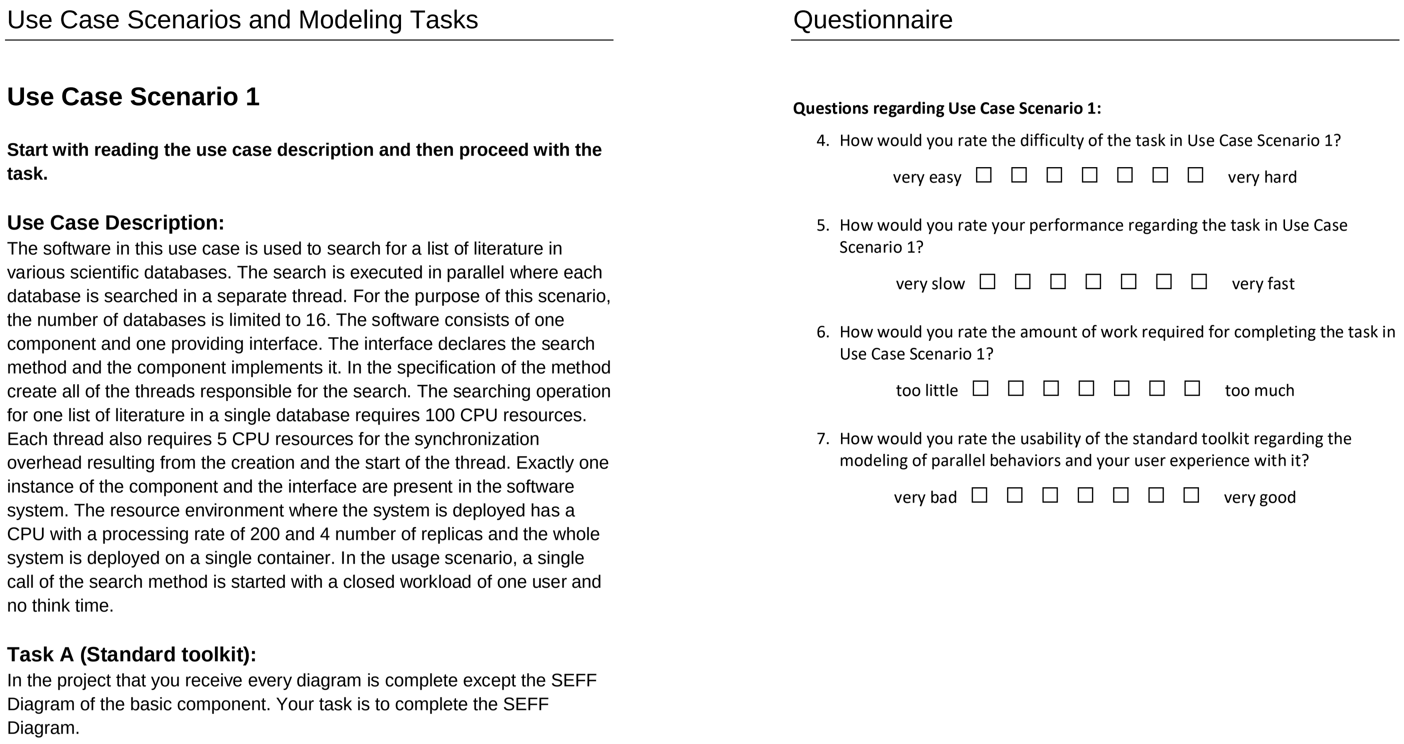
RISKS AND BENEFITS: No risk associated with this study. The collected data is securely stored. We do guarantee no data misuse and privacy is completely preserved. Your decision whether or not to participate in this study will not affect your grade in school.

PARTICIPANT'S RIGHTS: If you have read this form and have decided to participate in this project, please understand your **participation is voluntary** and you have the **right to withdraw your consent or discontinue participation at any time without penalty or loss of benefits to which you are otherwise entitled**. **The alternative is not to participate.** The results of this research study may be presented at scientific or professional meetings or published in scientific journals. Your identity is not disclosed unless we directly inform and ask for your permission.

CONTACT INFORMATION: If you have any questions, concerns or complaints about this research, its procedures, risks and benefits, contact following persons:
Denis Zahariev (denis.zahariev95@gmail.com)
Markus Frank (markus.frank@iste.uni-stuttgart.de)

By signing this document I confirm that I agree to the terms and conditions.

Name: ________________________ *Signature, Date:* ________________________



Use Case Scenarios and Modeling Tasks

Use Case Scenario 1

Start with reading the use case description and then proceed with the task.

Use Case Description:

The software in this use case is used to search for a list of literature in various scientific databases. The search is executed in parallel where each database is searched in a separate thread. For the purpose of this scenario, the number of databases is limited to 16. The software consists of one component and one providing interface. The interface declares the search method and the component implements it. In the specification of the method create all of the threads responsible for the search. The searching operation for one list of literature in a single database requires 100 CPU resources. Each thread also requires 5 CPU resources for the synchronization overhead resulting from the creation and the start of the thread. Exactly one instance of the component and the interface are present in the software system. The resource environment where the system is deployed has a CPU with a processing rate of 200 and 4 number of replicas and the whole system is deployed on a single container. In the usage scenario, a single call of the search method is started with a closed workload of one user and no think time.

Task A (Standard toolkit):

In the project that you receive every diagram is complete except the SEFF Diagram of the basic component. Your task is to complete the SEFF Diagram.

Questionnaire

Questions regarding Use Case Scenario 1:

4. How would you rate the difficulty of the task in Use Case Scenario 1?

   very easy □ □ □ □ □ □ □ very hard

5. How would you rate your performance regarding the task in Use Case Scenario 1?

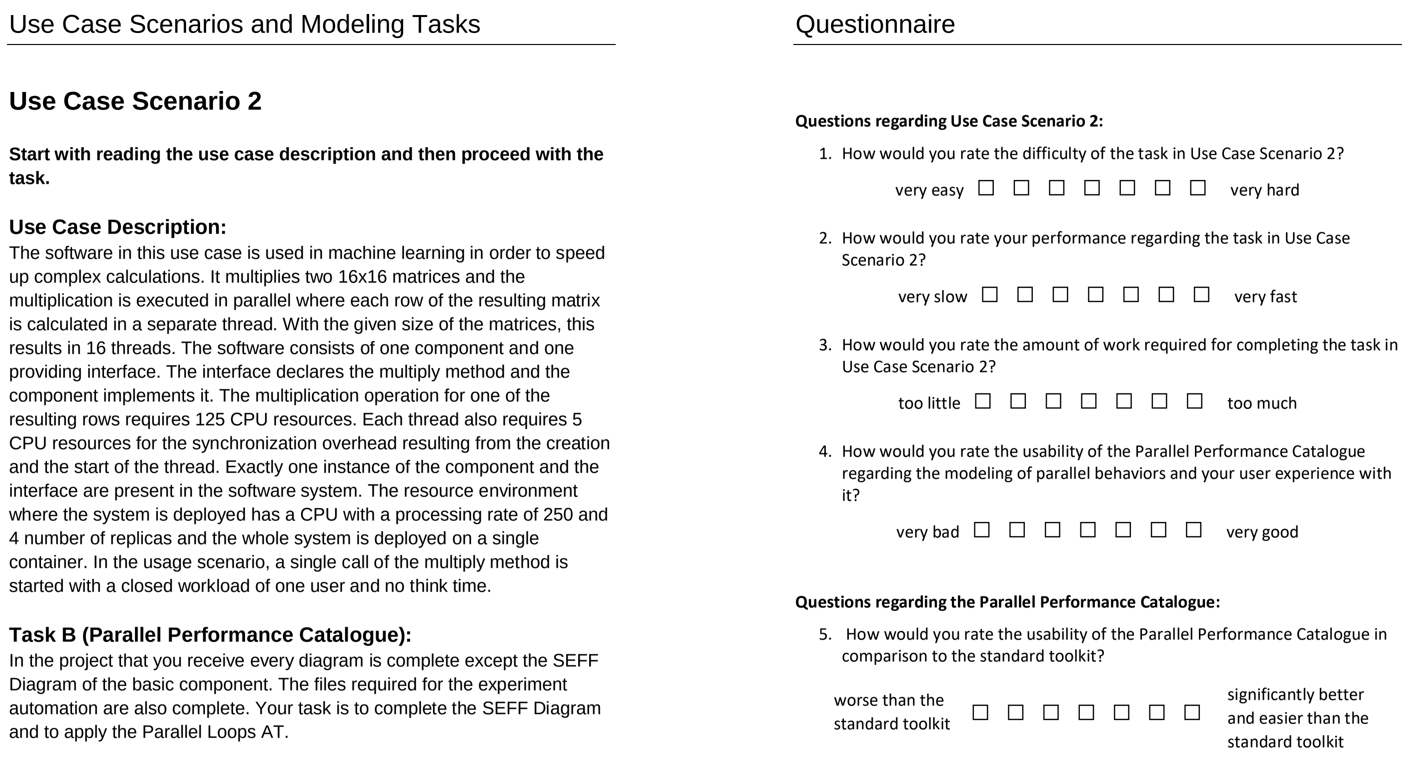
   very slow □ □ □ □ □ □ □ very fast

6. How would you rate the amount of work required for completing the task in Use Case Scenario 1?

   too little □ □ □ □ □ □ □ too much

7. How would you rate the usability of the standard toolkit regarding the modeling of parallel behaviors and your user experience with it?

   very bad □ □ □ □ □ □ □ very good



Use Case Scenarios and Modeling Tasks

Use Case Scenario 2

Start with reading the use case description and then proceed with the task.

Use Case Description:

The software in this use case is used in machine learning in order to speed up complex calculations. It multiplies two 16x16 matrices and the multiplication is executed in parallel where each row of the resulting matrix is calculated in a separate thread. With the given size of the matrices, this results in 16 threads. The software consists of one component and one providing interface. The interface declares the multiply method and the component implements it. The multiplication operation for one of the resulting rows requires 125 CPU resources. Each thread also requires 5 CPU resources for the synchronization overhead resulting from the creation and the start of the thread. Exactly one instance of the component and the interface are present in the software system. The resource environment where the system is deployed has a CPU with a processing rate of 250 and 4 number of replicas and the whole system is deployed on a single container. In the usage scenario, a single call of the multiply method is started with a closed workload of one user and no think time.

Task B (Parallel Performance Catalogue):

In the project that you receive every diagram is complete except the SEFF Diagram of the basic component. The files required for the experiment automation are also complete. Your task is to complete the SEFF Diagram and to apply the Parallel Loops AT.

Questionnaire

Questions regarding Use Case Scenario 2:

1. How would you rate the difficulty of the task in Use Case Scenario 2?

   very easy □ □ □ □ □ □ □ very hard

2. How would you rate your performance regarding the task in Use Case Scenario 2?

   very slow □ □ □ □ □ □ □ very fast

3. How would you rate the amount of work required for completing the task in Use Case Scenario 2?

   too little □ □ □ □ □ □ □ too much

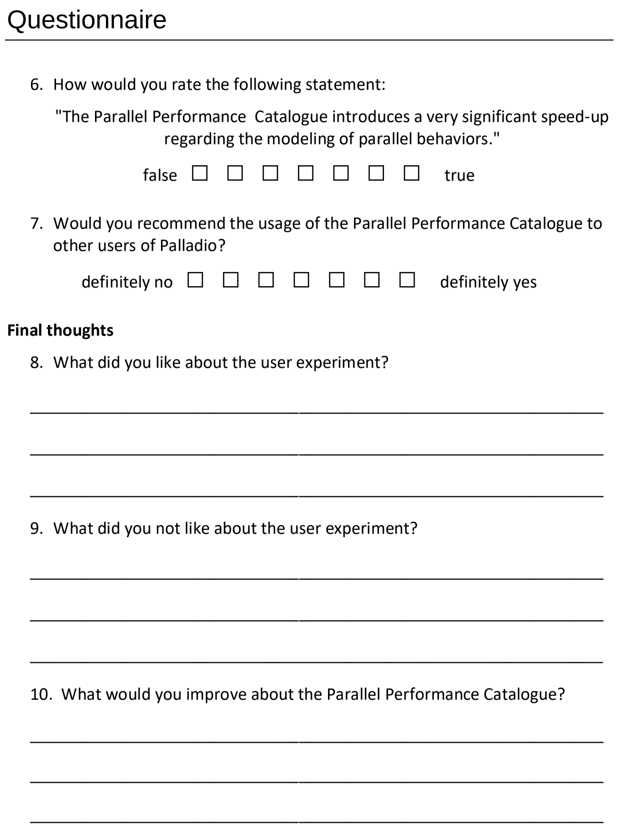
4. How would you rate the usability of the Parallel Performance Catalogue regarding the modeling of parallel behaviors and your user experience with it?

   very bad □ □ □ □ □ □ □ very good

Questions regarding the Parallel Performance Catalogue:

5. How would you rate the usability of the Parallel Performance Catalogue in comparison to the standard toolkit?

   worse than the standard toolkit □ □ □ □ □ □ □ significantly better and easier than the standard toolkit



Questionnaire

6. How would you rate the following statement:

   "The Parallel Performance Catalogue introduces a very significant speed-up regarding the modeling of parallel behaviors."

   false □ □ □ □ □ □ □ true

7. Would you recommend the usage of the Parallel Performance Catalogue to other users of Palladio?

   definitely no □ □ □ □ □ □ □ definitely yes

Final thoughts

8. What did you like about the user experiment?

   ____

   ____

   ____

9. What did you not like about the user experiment?

   ____

   ____

   ____

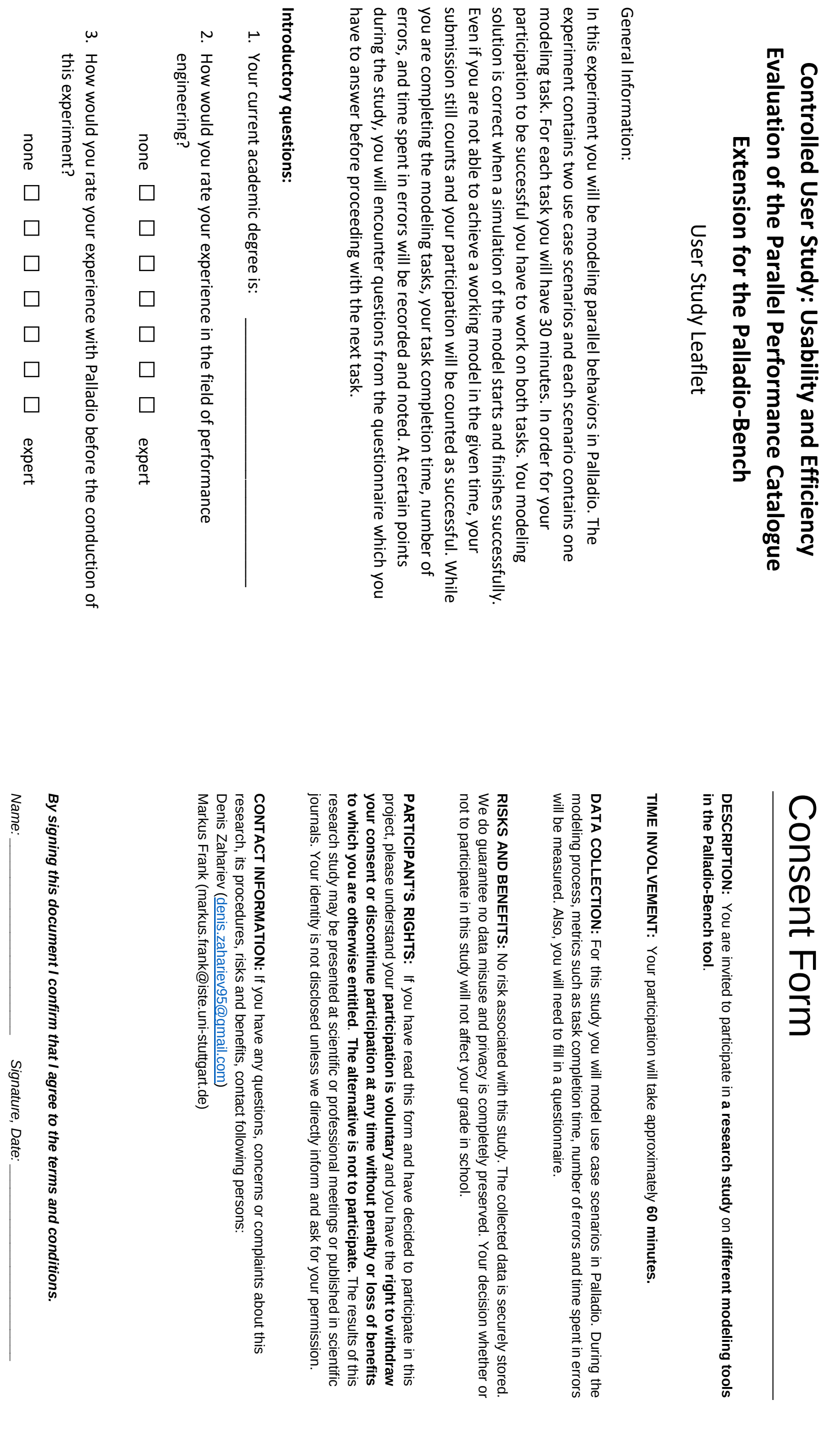
10. What would you improve about the Parallel Performance Catalogue?

   ____

   ____

   ____

A.3. Blank User Study Leaflet - Group B



Controlled User Study: Usability and Efficiency Evaluation of the Parallel Performance Catalogue Extension for the Palladio-Bench

User Study Leaflet

General Information:

In this experiment you will be modeling parallel behaviors in Palladio. The experiment contains two use case scenarios and each scenario contains one modeling task. For each task you will have 30 minutes. In order for your participation to be successful you have to work on both tasks. You modeling solution is correct when a simulation of the model starts and finishes successfully. Even if you are not able to achieve a working model in the given time, your submission still counts and your participation will be counted as successful. While you are completing the modeling tasks, your task completion time, number of errors, and time spent in errors will be recorded and noted. At certain points during the study, you will encounter questions from the questionnaire which you have to answer before proceeding with the next task.

Introductory questions:

1. Your current academic degree is: ________________
2. How would you rate your experience in the field of performance engineering?

   none □ □ □ □ □ □ □ expert

3. How would you rate your experience with Palladio before the conduction of this experiment?

   none □ □ □ □ □ □ □ expert

Consent Form

DESCRIPTION: You are invited to participate in **a research study** on **different modeling tools in the Palladio-Bench tool**.

TIME INVOLVEMENT: Your participation will take approximately **60 minutes.**

DATA COLLECTION: For this study you will model use case scenarios in Palladio. During the modeling process, metrics such as task completion time, number of errors and time spent in errors will be measured. Also, you will need to fill in a questionnaire.

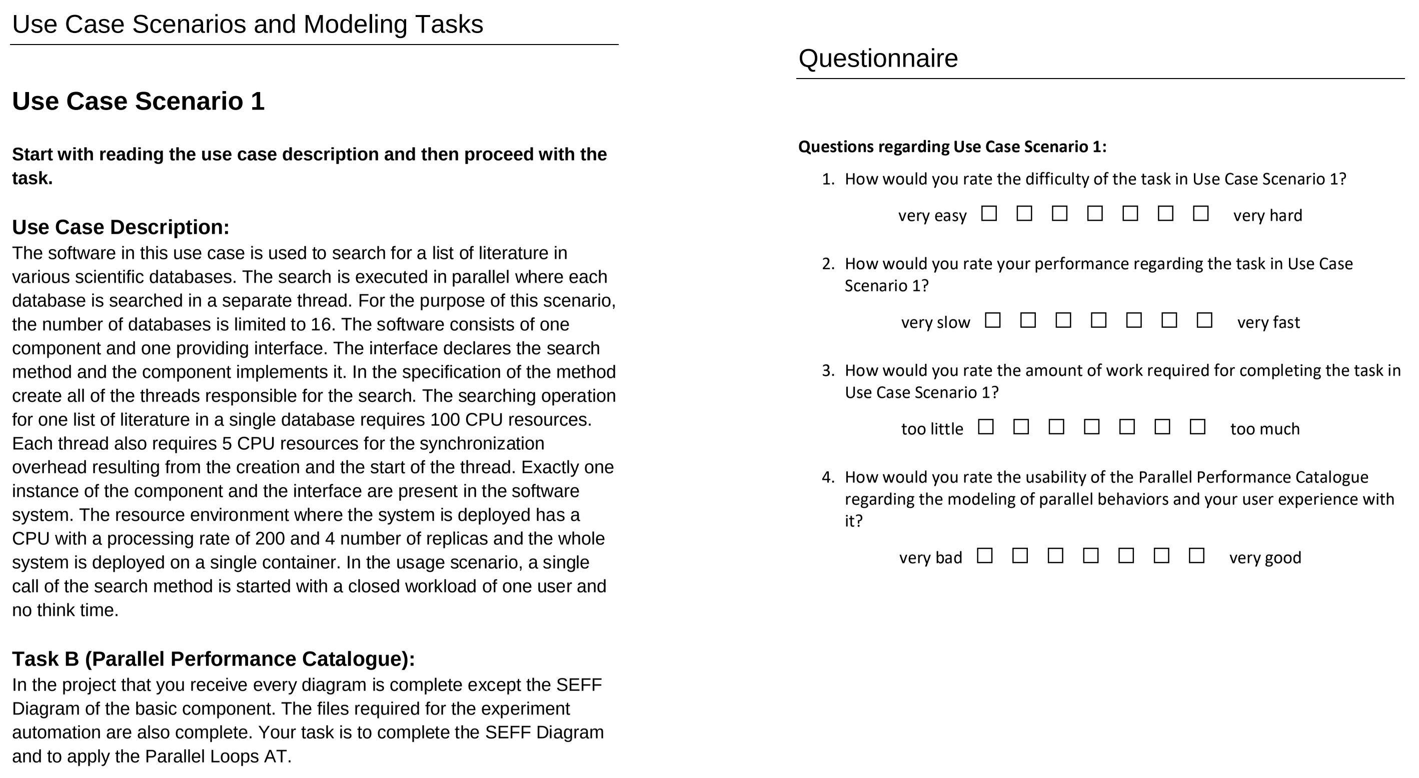
RISKS AND BENEFITS: No risk associated with this study. The collected data is securely stored. We do guarantee no data misuse and privacy is completely preserved. Your decision whether or not to participate in this study will not affect your grade in school.

PARTICIPANT'S RIGHTS: If you have read this form and have decided to participate in this project, please understand your **participation is voluntary** and you have the **right to withdraw your consent or discontinue participation at any time without penalty or loss of benefits to which you are otherwise entitled**. **The alternative is not to participate.** The results of this research study may be presented at scientific or professional meetings or published in scientific journals. Your identity is not disclosed unless we directly inform and ask for your permission.

CONTACT INFORMATION: If you have any questions, concerns or complaints about this research, its procedures, risks and benefits, contact following persons:
Denis Zahariev (denis.zahariev95@gmail.com)
Markus Frank (markus.frank@iste.uni-stuttgart.de)

By signing this document I confirm that I agree to the terms and conditions.

Name: ____________ *Signature, Date:* ____________



Use Case Scenarios and Modeling Tasks

Use Case Scenario 1

Start with reading the use case description and then proceed with the task.

Use Case Description:
The software in this use case is used to search for a list of literature in various scientific databases. The search is executed in parallel where each database is searched in a separate thread. For the purpose of this scenario, the number of databases is limited to 16. The software consists of one component and one providing interface. The interface declares the search method and the component implements it. In the specification of the method create all of the threads responsible for the search. The searching operation for one list of literature in a single database requires 100 CPU resources. Each thread also requires 5 CPU resources for the synchronization overhead resulting from the creation and the start of the thread. Exactly one instance of the component and the interface are present in the software system. The resource environment where the system is deployed has a CPU with a processing rate of 200 and 4 number of replicas and the whole system is deployed on a single container. In the usage scenario, a single call of the search method is started with a closed workload of one user and no think time.

Task B (Parallel Performance Catalogue):
In the project that you receive every diagram is complete except the SEFF Diagram of the basic component. The files required for the experiment automation are also complete. Your task is to complete the SEFF Diagram and to apply the Parallel Loops AT.

Questionnaire

Questions regarding Use Case Scenario 1:

1. How would you rate the difficulty of the task in Use Case Scenario 1?

   very easy □ □ □ □ □ □ □ very hard

2. How would you rate your performance regarding the task in Use Case Scenario 1?

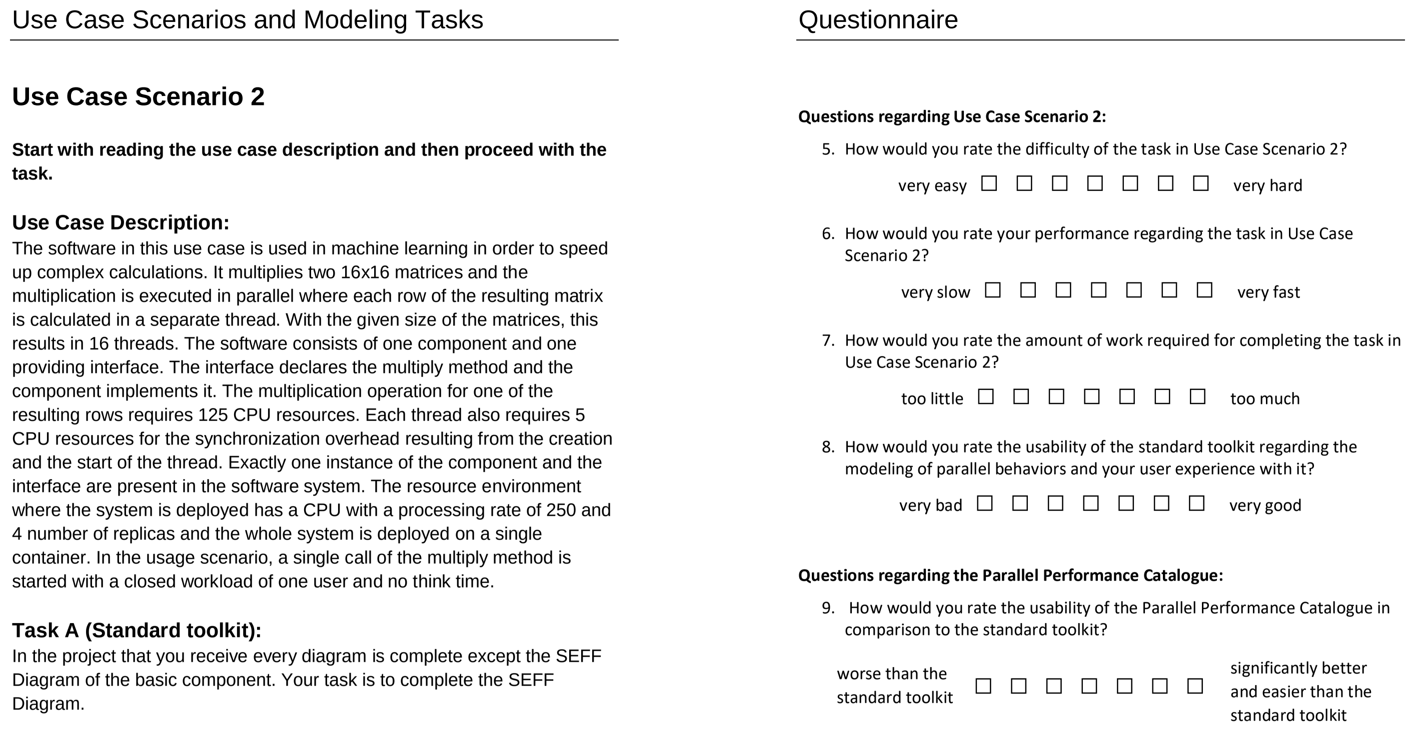
   very slow □ □ □ □ □ □ □ very fast

3. How would you rate the amount of work required for completing the task in Use Case Scenario 1?

   too little □ □ □ □ □ □ □ too much

4. How would you rate the usability of the Parallel Performance Catalogue regarding the modeling of parallel behaviors and your user experience with it?

   very bad □ □ □ □ □ □ □ very good



Use Case Scenarios and Modeling Tasks

Use Case Scenario 2

Start with reading the use case description and then proceed with the task.

Use Case Description:

The software in this use case is used in machine learning in order to speed up complex calculations. It multiplies two 16x16 matrices and the multiplication is executed in parallel where each row of the resulting matrix is calculated in a separate thread. With the given size of the matrices, this results in 16 threads. The software consists of one component and one providing interface. The interface declares the multiply method and the component implements it. The multiplication operation for one of the resulting rows requires 125 CPU resources. Each thread also requires 5 CPU resources for the synchronization overhead resulting from the creation and the start of the thread. Exactly one instance of the component and the interface are present in the software system. The resource environment where the system is deployed has a CPU with a processing rate of 250 and 4 number of replicas and the whole system is deployed on a single container. In the usage scenario, a single call of the multiply method is started with a closed workload of one user and no think time.

Task A (Standard toolkit):

In the project that you receive every diagram is complete except the SEFF Diagram of the basic component. Your task is to complete the SEFF Diagram.

Questionnaire

Questions regarding Use Case Scenario 2:

5. How would you rate the difficulty of the task in Use Case Scenario 2?

   very easy □ □ □ □ □ □ □ very hard

6. How would you rate your performance regarding the task in Use Case Scenario 2?

   very slow □ □ □ □ □ □ □ very fast

7. How would you rate the amount of work required for completing the task in Use Case Scenario 2?

   too little □ □ □ □ □ □ □ too much

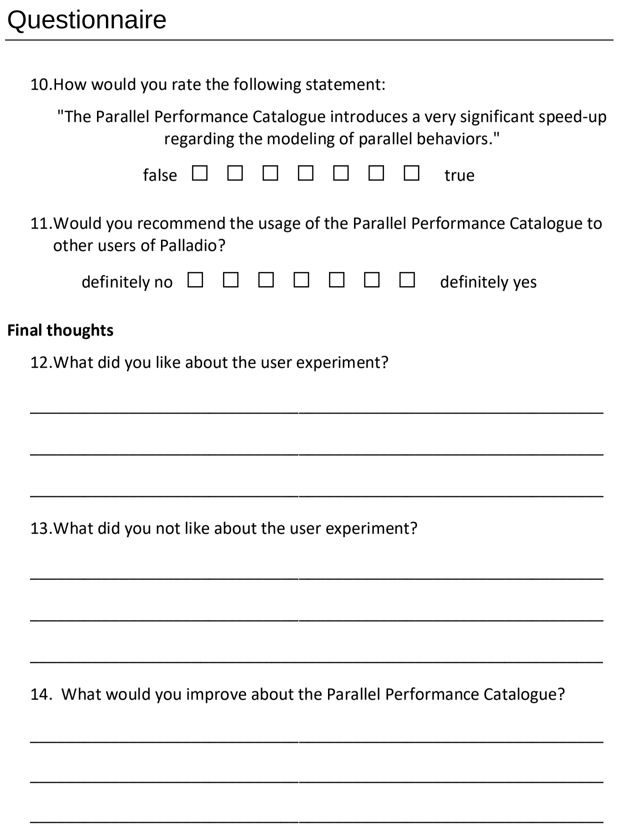
8. How would you rate the usability of the standard toolkit regarding the modeling of parallel behaviors and your user experience with it?

   very bad □ □ □ □ □ □ □ very good

Questions regarding the Parallel Performance Catalogue:

9. How would you rate the usability of the Parallel Performance Catalogue in comparison to the standard toolkit?

   worse than the standard toolkit □ □ □ □ □ □ □ significantly better and easier than the standard toolkit



Questionnaire

10. How would you rate the following statement:

"The Parallel Performance Catalogue introduces a very significant speed-up regarding the modeling of parallel behaviors."

false □ □ □ □ □ □ □ true

11. Would you recommend the usage of the Parallel Performance Catalogue to other users of Palladio?

definitely no □ □ □ □ □ □ □ definitely yes

Final thoughts

12. What did you like about the user experiment?

____

____

____

13. What did you not like about the user experiment?

____

____

____

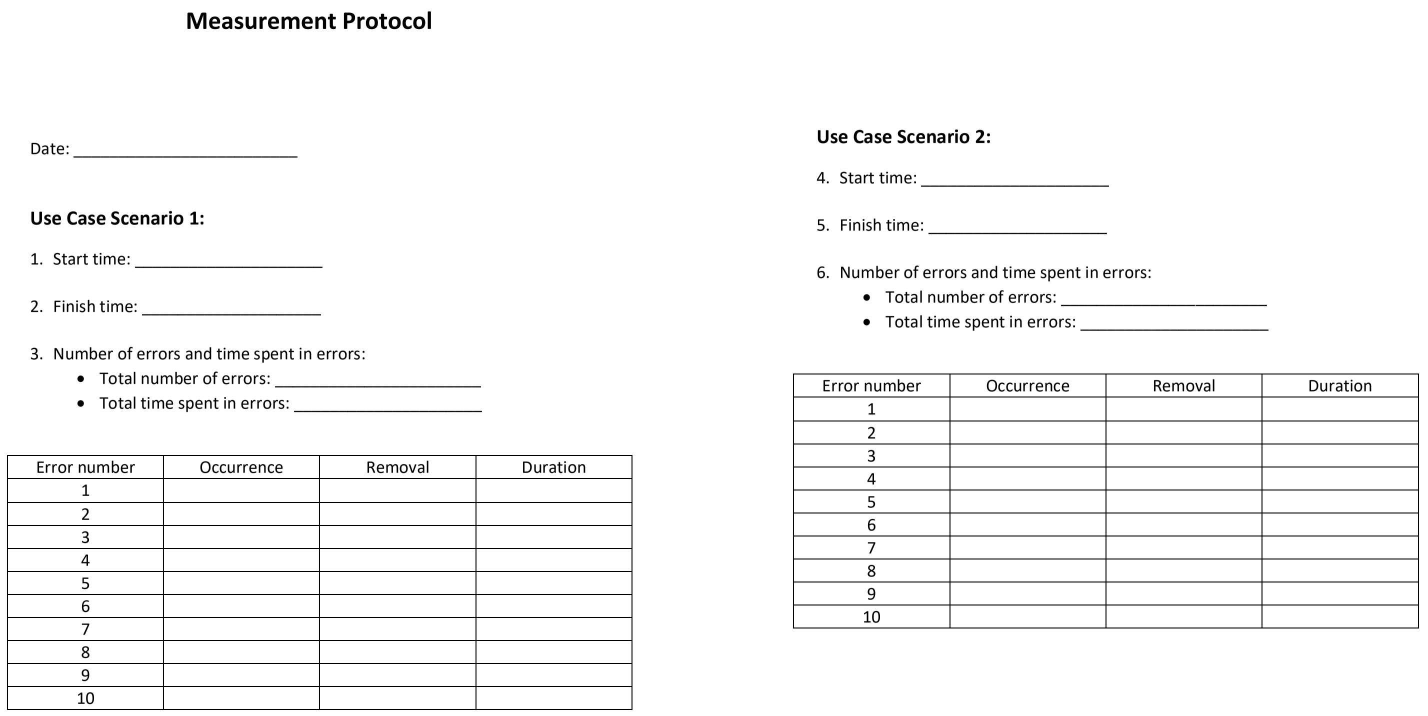
14. What would you improve about the Parallel Performance Catalogue?

____

____

____

A.4. Blank Measurement Protocol



Measurement Protocol

Date: ______________________

Use Case Scenario 1:

1. Start time: ____________________
2. Finish time: ____________________
3. Number of errors and time spent in errors:
   - Total number of errors: ______________________
   - Total time spent in errors: ____________________

Error number	Occurrence	Removal	Duration
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			

Use Case Scenario 2:

4. Start time: ____________________
5. Finish time: ____________________
6. Number of errors and time spent in errors:
   - Total number of errors: ______________________
   - Total time spent in errors: ____________________

Error number	Occurrence	Removal	Duration
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			

A.5. User Study Leaflets and Protocols

The following GitLab repository[RSS] contains all user study leaflets(including use case scenarios, task descriptions and questionnaire) filled out by the participants, and all measurement protocols:

RSS GitLab - Leaflets and Protocols

A.6. Palladio Workshop

The workshop document used in the workshop sessions can be found in the same GitLab repository[RSS]:

RSS GitLab - Workshop Document

Appendix A

Bibliography

[BBB+16] S. Becker, F. Brosig, E. Burger, A. Busch, Z. Durdik, J. Happe, L. Happe, C. Heger, R. Heinrich, J. Henss, N. Huber, O. Hummel, B. Klatt, A. Koziolek, H. Koziolek, M. Kramer, K. Krogmann, M. Küster, M. Langhammer, A. Wert. *Modeling and Simulating Software Architectures - The Palladio Approach*. Oct. 2016. ISBN: 9780262034760 (cit. on p. 7).

[BBG05] S. Beydeda, M. Book, V. Gruhn. *Model-Driven Software Development*. 2005 (cit. on p. 7).

[BCR] V. R. Basili, G. Caldiera, H. D. Rombach. *The Goal Question Metric Approach* (cit. on p. 2).

[BKR] S. Becker, H. Koziolek, R. Reussner. *The Palladio component model for model-driven performance prediction*. https://www.sciencedirect.com/science/article/pii/S0164121208001015. (Accessed on 06/24/2019). (Cit. on pp. 1, 7, 8).

[BRZ] B. Boehm, H. D. Rombach, M. V. Zelkowitz. *Foundations of Empirical Software Engineering* (cit. on pp. 10, 11, 14, 15).

[DFAB03] A. Dix, J. E. Finlay, G. D. Abowd, R. Beale. *Human-Computer Interaction (3rd Edition)*. USA: Prentice-Hall, Inc., 2003. ISBN: 0130461091 (cit. on pp. 10, 11, 14, 15).

[Eij17] V. Eijkhout. *Parallel Programming in MPI and OpenMP*. 2017 (cit. on p. 9).

[FH18] M. Frank, A. Hakamian. "An Architectural Template for Parallel Loops and Sections." In: *Proceedings of the Symposium on Software Performance 2018, 7-9 November 2018, Hildesheim, Germany*. 9th Symposium on Software Performance 2018. Hildesheim, Nov. 2018. URL: https://www.performance-symposium.org/fileadmin/user_upload/palladio-conference/2018/papers/FrankHakamian18.pdf (cit. on pp. 1, 8).

[Fra] P. Franczak. *Design of Empirical User Studies in Model-based Software Development for Evaluating Modeling Languages* (cit. on pp. 13, 15).

[FSH17] M. Frank, S. Staude, M. Hilbrich. "Is the PCM Ready for ACTORs and Multicore CPUs? - A Use Case-based Evaluation." In: *Proceedings of the Symposium on Software Performance 2017, 9-10 November 2017, Karlsruhe, Germany*. 8th Symposium on Software Performance 2017. Karlsruhe, Nov. 2017. URL: http://www.performance-symposium.org/fileadmin/user_upload/palladio-conference/2017/papers/Is_the_PCM_Ready_for_ACTORs_and_Multicore_CPUs_A_Use_Case-based_Evaluation.pdf (cit. on pp. 1, 8).

[JCP] A. Jedlitschka, M. Ciolkowski, D. Pfahl. *Reporting Experiments in Software Engineering. In: Shull F., Singer J., Sjøberg D.I.K. (eds) Guide to Advanced Empirical Software Engineering*. Springer, London. ISBN: 978-1-84800-044-5 (cit. on p. 14).

[Leh18] S. M. Lehrig. "Efficiently Conducting Quality-of-Service Analyses by Templating Architectural Knowledge." PhD thesis. Karlsruher Institut für Technologie (KIT), 2018. 514 pp. ISBN: 978-3-7315-0756-7. DOI: 10.5445/KSP/1000079766 (cit. on p. 8).

[Mas] Massachusetts Institute of Technology. *T-Distribution Table*. http://math.mit.edu/~vebrunel/Additional%20lecture%20notes/t%20(Student%27s)%20table.pdf. (Accessed on 06/14/2019) (cit. on p. 35).

[Mic17] Microsoft. *Pipes & Filters*. July 2017. URL: https://docs.microsoft.com/en-us/azure/architecture/patterns/pipes-and-filters (cit. on p. 10).

[MSM04] T. Mattson, B. Sanders, B. Massingill. *Patterns for Parallel Programming*. First. Addison-Wesley Professional, 2004. ISBN: 0321228111 (cit. on p. 9).

[Nüt] C. Nützel. *An Efficiency Comparison Between Architectural Templates and SimuLizar: A Controlled Experiment*. Chemnitz University of Technology (cit. on p. 13).

[Pal] Palladio. *Palladio Software Architecture Simulator: Tools*. https://www.palladio-simulator.com/tools/. (Accessed on 01/27/2020). (Cit. on p. 7).

[RH] P. Runeson, M. Höst. *Guidelines for conducting and reporting case study research in software engineering* (cit. on pp. 10, 11, 14, 15).

[RM17] C. Robson, K. McCartan. *Real World Research, 4th Edition*. Dec. 2017. ISBN: 978-1-118-74523-6 (cit. on p. 15).

[RSS] RSS. *RSS GitLab Repository*. URL: https://git.rss.iste.uni-stuttgart.de/theses/parallel-performance-pattern-controlled-experiment (cit. on p. 57).

[VSB+] M. Völter, T. Stahl, J. Bettin, A. Haase, S. Helsen, K. Czarnecki, B. von Stockfleth. *Model-Driven Software Development: Technology, Engineering, Management* (cit. on p. 7).

[WBH88] J. Whiteside, J. Bennett, K. Holtzblatt. “Chapter 36 - Usability Engineering: Our Experience and Evolution.” In: *Handbook of Human-Computer Interaction*. Ed. by M. HELANDER. Amsterdam: North-Holland, 1988, pp. 791–817. ISBN: 978-0-444-70536-5. DOI: https://doi.org/10.1016/B978-0-444-70536-5.50041-5. URL: http://www.sciencedirect.com/science/article/pii/B9780444705365500415 (cit. on p. 15).

[WRH+12] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, A. Wesslén. *Experimentation in Software Engineering*. Springer Berlin Heidelberg, 2012. ISBN: 978-3-642-29043-5 (cit. on pp. 10, 11, 14, 15, 34, 35).

[ZWS19] D. Zahariev, A. Weller, J. Schuder. *Identifying Reoccurring Parallel Design Patterns to Build a Parallel Pattern Catalogue for Palladio*. 2019 (cit. on pp. 1, 8–10).

All links were last followed on February 19, 2020.

Declaration

I hereby declare that the work presented in this thesis is entirely my own and that I did not use any other sources and references than the listed ones. I have marked all direct or indirect statements from other sources contained therein as quotations. Neither this work nor significant parts of it were part of another examination procedure. I have not published this work in whole or in part before. The electronic copy is consistent with all submitted copies.

---

place, date, signature