
Software Requirements Specification for Systematic Benchmarking Test Case Generation

Team 1

Emma Willson

Esha Pisharody

Grace Croome

Marie Hollington

Proyetei Akanda

Zainab Abdulsada

October 11th, 2024

Table of Contents

Table of Contents.....2

1. Versions, Roles and Contributions.....4

 1.1. Version History4

 1.2. Table of Contributions4

 1.3. Team Info4

2. Purpose of the Project5

 2.1 User Business.....5

 2.2 Goal of the Project5

 2.3. Stakeholders6

 2.3.1 Primary.....6

 2.3.2 Secondary6

 2.4. Naming Convention and Terminology6

 2.5 Relevant Facts and Assumptions6

 2.6 The Scope of the work6

 2.6.1 The Current Situation.....6

 2.6.2 Work Partitioning7

 2.7. Scope of the Product7

3. Constraints and Functional Requirements7

 3.1. Mandated Constraints.....7

 3.1.1 Solution Constraints.....7

 3.1.2 Partner or Collaborative Applications8

 3.1.3 Off-the-Shelf Software8

 3.1.4 Anticipated Workplace Environment.....9

 3.2 Functional Requirements9

4. Non-Functional Requirements11

 4.1 Usability and Human Requirements11

4.2. Performance Requirements	11
4.2.1 Precision or Accuracy Requirements	11
4.3. Operational and Environmental Requirements	11
4.3.1 Expected Physical Environment	11
4.3.2 Requirements for Interfacing with Adjacent Systems	11
4.4. Security Requirements	12
4.5. Legal Requirements	12
5. Risks and Predicted Issues.....	12
5.1 Risks	12
5.2 Predicted Issues	12

1. Versions, Roles and Contributions

1.1. Version History

Version #	Authors	Description	Date
0	All	Created first draft of document.	October 11, 2024
1	All	First major revision	April 4, 2025

1.2. Table of Contributions

Group Member	Contributions (Sections)
Emma Willson	2.3, 2.4, 2.5, 2.7, 4.1
Esha Pisharody	1, 2.1, 2.2, 2.6, 3.1, 4.3
Grace Croome	2.3, 2.4, 2.5, 2.7, 3.2, 4.1, 5
Marie Hollington	3.2, 4.1, 4.4
Proyetei Akanda	3.2, 4.1, 4.4
Zainab Abdulsada	2.1, 2.2 2.4, 2.6, 3.1, 4.3

1.3. Team Information

Group Members	Contact & ID	Roles
Emma Willson	willson@mcmaster.ca 400309856	Research Lead Developer
Esha Pisharody	pisharoe@mcmaster.ca 400325118	Developer
Grace Croome	croomeg@mcmaster.ca 400313932	Developer
Marie Hollington	hollim3@mcmaster.ca 400320562	Developer
Proyetei Akanda	akandap@mcmaster.ca 400327972	User Interface Lead Developer
Zainab Abdulsada	abduslaz@mcmaster.ca 400313736	Project Manager Developer

2. Purpose of the Project

2.1 User Business

Through this project we focus on modern interactive proof assistants (such as Lean¹, Idris², Agda³ and Rocq⁴), which are widely used in formal verification to ensure the correctness of mathematical proofs and software systems. Despite their importance, these systems have not been thoroughly tested for efficiency in terms of time and memory complexity. This project seeks to systematically test and compare these proof assistants to identify areas that can be optimized, highlighting potential areas for improving their performance. This initiative is driven not by a specific business problem, but by a significant opportunity to enhance the effectiveness of these tools. By pinpointing design inefficiencies and areas for improvement, this project aims to produce a comprehensive analysis, to guide future research in the optimization and development of more efficient proof-checking systems.

2.2 Goal of the Project

The goal of this project is to design and build an automated code generator that creates a series of tests of increasing size that will test the efficiency of modern interactive proof assistants, specifically Lean, Idris, Agda, and Rocq. Testing will be performed on these assistants as languages, outside of their specific capabilities as proof assistants. The deliverables will include a research paper outlining our findings through systematic testing, in addition to documentation on how to extend the test generator with new classes of test proofs, and the creation of test suites for the new class of proofs translated into each of the proof assistants. The project will also include an interactive GitHub workflow that produces web visualizations enabling users to view the time and memory complexity. Users will also be able to pull down the project locally using Docker, for ease of development.

By providing this analysis and method of performance evaluation of the various proof assistants, this project will enable researchers to identify key inefficiencies in order to optimize future designs of proof-checking systems. Furthermore, the webpages generated will enhance user exploration and analysis of the different proof assistants, allowing for easy comparison.

¹ <https://lean-lang.org/>

² <https://www.idris-lang.org/>

³ <https://agda.readthedocs.io/en/latest/getting-started/what-is-agda.html>

⁴ <https://Rocq.inria.fr/>

2.3. Stakeholders

2.3.1 Primary

Dr. Moradi and Mehrdad Eshraghi Dehaghani; Dr. Jacques Carette & research team; COMPSCI 4ZP6 Team 1; other programming language researchers.

2.3.2 Secondary

Developers of Lean, Idris, Agda, and Rocq, as well as users of these proof assistants.

2.4. Naming Convention and Terminology

- **CLI:** Command line interface, a text-based system that allows a user to interact with computer programs.
- **Proof assistant:** A proof assistant is a software tool used to assist a user with formalizing a logical or mathematical proof, ex. Agda, Lean, Idris, Rocq, or Isabelle.⁵
- **MHPG:** Mini Haskell Proof Grammar refers to the Haskell grammar we will be creating to format automated proofs. Located in Grammar.hs.
- **EP:** Elementary Proofs are the base set of test cases written in MHPG which will be the foundation from which tests of increasing size are generated.
- **PAL:** Proof assistant language, referring to the four languages we are working with: Agda, Lean, Idris, and Rocq.
- **UNIX:** Uniplexed Information Computing system, a multiuser and multitasking operating System.

2.5 Relevant Facts and Assumptions

- The user is proficient with the logic underlying the writing of proofs.
- Test cases are syntactically valid.

2.6 The Scope of the work

2.6.1 The Current Situation

Currently, modern interactive proof assistants have not been thoroughly tested for efficiency. In correlation, we do not find any tools to directly allow users to interactively compare various proof assistants.

⁵ <https://isabelle.in.tum.de/>

2.6.2 Work Partitioning

Step	Description
Create sample tests	Create a set of tests, in each PAL, to have a solid basis on how to create MHPG.
Set up Continuous Integration	Set up the project with proper Continuous Integration infrastructure (including the installation of Agda, Lean, Idris, and Rocq and the infrastructure for running the generator). This will make it easier for us to implement automated testing via Continuous Integration.
Create MHPG	Proceed with creating the Haskell Grammar used to write test cases.
Initialize EP	Create the elementary tests using MHPG (static tests).
Increase proof size	Given user input for type of proof and preset number of operations, corresponding algorithm for proof type will be called to increase EP size (dynamic tests).
Translate	Once a complex proof is generated, it will be translated into the four PALs.
Compile and measure time and memory	Then a script will be called to compile the proof in each PAL, timing compilation and recording memory used for each PAL.
Record and present	Once data is recorded, it is presented in graphs formatted through a webpage.

2.7. Scope of the Product

The product will test the time and space efficiency of the different PALs, through basic and proof-based tests, when provided with a proof type through the CLI. It will then provide a visual representation of the measured data through graphs.

3. Constraints and Functional Requirements

3.1. Mandated Constraints

3.1.1 Solution Constraints

Description: The project will utilize the proof assistants (Lean, Idris, Agda, and Rocq).

Rationale: These proof languages were chosen due to their prominence in research, and because they are all implemented in similar functional languages.

Fit Criterion: The solution must be able to translate the tests to each of these languages, and in turn analyze their time and memory performance.

Description: The code generator and translator will be implemented in Haskell.

Rationale: Haskell was chosen for its compatibility with the four proof assistants, due to its functional programming nature.

Fit Criterion: The solution must be able to generate test cases, run and analyze proof results.

Description: The end users shall interact with the final product through a GitHub workflow or local CLI through Docker.

Rationale: The final product must be accessible, through UNIX systems.

Fit Criterion: Users must be able to select test type and receive an output of the measured space and time utilization.

3.1.2 Partner or Collaborative Applications

The project will need to interface with external proof assistants: Lean, Idris, Agda, and Rocq.

3.1.3 Off-the-Shelf Software

Description: The project will use a main GitHub repository with version control, and GitHub workflow will be used to run tests.

Rationale: Using GitHub will allow all team members to seamlessly collaborate.

Fit Criterion: Each member must be able to commit and maintain versions effectively.

Description: The project will use the Python library matplotlib to create graphs for visualization.

Rationale: Using matplotlib will allow the data collected on the efficiency of proof assistants to be visualized.

Fit Criterion: The visualization library must be easy to use and produce graphs that can be easily interpreted.

Description: Docker will be used to create an environment where we can run the project. It will contain all the components of our project.

Rationale: We need one space to ensure fair testing on a shared environment.

Fit Criterion: Using Docker will enable easy pull on local machines to enable ease of future development.

Description: Flask is Web server setup, routing (@app.route), and HTML templating (render_template).

Rationale: Well documented, and high usage allows for ease of problem solving.

Fit Criterion: High customization will allow us to visualize things as we deem fit.

Description: Vercel will be used to host the Flask backend by running the app through a Python serverless function.

Commented [GC1]: also mention vercel/flask here

Rationale: The rationale behind using Vercel with Flask is to leverage Vercel's serverless deployment model for easy, scalable hosting without managing infrastructure, while still allowing dynamic backend logic like generating graphs on-the-fly. This setup ensures fast deployments, automatic previews, and seamless integration with frontend templates.

Fit Criterion: The Flask app must be able to deploy on Vercel and dynamically render graphs from test case data through serverless execution.

3.1.4 Anticipated Workplace Environment

- The system will be used in an academic and research-based setting.
- Users can work both in-person and remotely.
- Environments of such are shared by multiple researchers through collaborative efforts.
- Backend should be seamlessly installable for future research improvement.

3.2 Functional Requirements

The following functional requirements are ranked by priority from high to low.

PO (Minimum Viable Product)

With this collection of requirements, we will be able to complete our primary task of translating a variety of proofs into the base group of four PALs. After this, it will be possible to thoroughly compare the results we obtain from each PAL and answer our original questions about comparative efficiency.

- The system will contain a grammar MHPG.
- The system will contain a hard-coded initial set of EPs written in our proof grammar (MHPG).
- The system will be given a base set of EPs testing various language concepts (functions, records, datatypes etc.).
 - By covering a broad range of commonly used test types, we will be able to determine specific instances where certain languages are inefficient in comparison with other languages.
- The system, when given a number of operations and a test from the initial set, will produce a test case of the given size.
- The system will generate a set of tests of increasing sizes from the initial set.
- Post-test case generation, the system will translate them into the four PALs.
- The system will compile the test cases in each PAL.
- The system will record and display the compilation time and memory of each of these test cases in the four PALs.

- The system will provide documentation on how to extend the test generator with new classes of tests.
- The system will support installation on UNIX environments.

P1

With this additional collection of requirements, more detailed results will be generated, and comparisons can be more efficiently made.

- The system will measure the time and memory complexity of each test case as determined by comparison of the same base case across increasing sizes.
- The system will visualize the measured data in graphs on a webpage.
- The system will provide a description for each graph displayed on the webpage, explaining the data being visualized.
- The system will be available to pull down locally on Unix systems enabling users to install the complete system.

P2

With this additional collection of requirements, users will have a more customizable experience and will be able to interact more with the interface.

- The system will support installation on MacOS and Windows environments.
- The system will require users to select which type of test case is generated.
- The system will only allow users to select one type of test case per generation request.
- The system will allow users to select how many sizes of each test case is generated, with a lower and upper limit.
- The system will validate user input for all parameters and provide an error message if the input is invalid.
- The system will record the translated test cases in their respective PAL and store them in files.
- The system will record the compilation logs of each test proof and store them in a .txt file, for the purpose of supporting development (error capture), and future research.

P3

With this additional requirement, a comparison of efficiency across a broader range of proof assistants will be possible.

- The system will incorporate Isabelle as one of the PALs available for translation
 - While all four original PALs have similarities in their implementations, Isabelle is a popular proof assistant that differs more clearly from the others. Hence, this requires further research and may result in further issues with translations of the same proof being equivalent.

4. Non-Functional Requirements

4.1 Usability and Human Requirements

- The installation process should be straightforward, with minimal steps required to set up the system. A documentation should guide users through the installation, ensuring all necessary components are installed properly without requiring manual interference.
- The system should provide clear visualization of the resulting data for the time and memory complexity as we increase the size. It should provide graphs for the user to easily understand the benchmarking process.
- The installation process should be divided into distinct phases, with minimal steps required for each. A documentation should guide users through each phase of the installation process, ensuring all necessary components are installed properly without requiring manual interference. Separate phases of use will be defined, especially for time-consuming components like proof assistants.
- The system should be compatible with UNIX.
- Any webpages produced by the system must have a layout that is understandable and found to be clear by key stakeholders, with concise explanations of the graphs.
- All error messages should be clear, informative, and provide guidance to help users identify and resolve the problem.

4.2. Performance Requirements

4.2.1 Precision or Accuracy Requirements

- The translated code will have the same logical meaning as the input.
- The translation of accepted code will be able to be compiled with its associated proof assistant.

4.3. Operational and Environmental Requirements

4.3.1 Expected Physical Environment

- Our product will be primarily used in research environments, where it will be accessed on standard desktop/laptop computers.

4.3.2 Requirements for Interfacing with Adjacent Systems

- The MHPG must be translatable into all proof assistants available.
- The system shall support Chromium-based browsers.

Commented [GC2]: Web Application

All fields should be clearly labeled, ensuring users can easily identify where to input proofs and view results. The system must quickly process inputs and translations, minimizing delays to keep users engaged. Time complexity results must be clearly displayed with proper labels for easy comparison. The layout of the web application must be consistent and fully responsive across all devices of different screen sizes or resolution with a primary focus on desktop usability. The application will contain a three-panel layout: the left-hand side (LHS) will be a space where users can select from a dropdown of already written proofs. Then when they click "run", the right-hand side (RHS) will display the output of the translation based on the selected proof assistant, and a second RHS panel will show the output of running the translation. This layout should allow users to see the relationship between the input proof, the translation, and the execution results.

Commented [PA3]: "responsive and fast". Installing Agda alone takes 30 minutes... so you really need to divide each part of the system into phases of use, and have requirements for each separately, else this is never going to be achievable." - FIXED

Commented [PA4R3]: previously I had: "The system should be responsive and fast, with minimal delays in processing the visualized graphs."

Commented [PA5]: "user-friendly layout" is naturally desirable but amazingly hard to define and measure. You might want to adjust this to something like "is found to be understandable by a few key stakeholders" or some such." - FIXED IT

Commented [PA6R5]: previously I had: "Any webpages produced by the system must have a clear, user-friendly layout with concise explanations of the graphs/tables."

Commented [GC7]: another subsection is speed requirements- should we have one for the actual translation? I don't know if it makes sense given that our end result is determining efficiency

Commented [EW8R7]: we could add a compilation cutoff time in case an input is too large

Commented [EW9R7]: but yes i'm expecting the translation itself to be pretty fast

Commented [GC10]: necessary? doc specifies that this has to be quantifiable and I don't think we have anything that fits, just a general sense of accuracy

Commented [EW11R10]: i think a benchmark for this could be that all translations are reversible?

4.4. Security Requirements

- **Version Control Integrity:** While access to the GitHub repository can be open to the public for the benefit of transparency and collaboration, certain administrative actions (such as merging pull requests or modifying core functionality) should be limited to key stakeholders, including professors, teaching assistants (TAs), and authorized developers. This ensures proper oversight of contributions while still encouraging open access to the project.
- **Backup and Recovery:** Regular backups of program code must be maintained to ensure data recovery in the event of a system failure or breach. For example, team members must regularly commit their code to GitHub. In the event of a device failure, such as a laptop crash, the code can still be easily accessed from the GitHub repository, ensuring no work is lost. We will ensure that every group member commits their code consistently, allowing seamless access and recovery if local files become unavailable.

Commented [PA12]: "4.4: *output* should never be stored in a CI." -> I never wrote this? Idk what he is saying

Commented [PA13]: "there is no strong need for 'security' for this. In fact, having it open would be a strong benefit, so I'm not quite seeing why this will be closed?" - FIXED

Commented [PA14R13]: previously: "Access to the GitHub repository must be limited to key stakeholders, including the professors, teaching assistants (TAs), and authorized software developers. Only authorized individuals, such as professors, TAs, and the software development team, should have access to view or modify the code. This reduces the risk of unauthorized access or unintended changes. "

4.5. Legal Requirements

- Our software will use the open-source MIT license.

5. Risks and Predicted Issues

5.1 Risks

- As we must predetermine the base tests to work on, there is a risk of not being thorough enough and missing a case that has an unexpected result of time and memory complexity.
- Inefficiencies in how we create the translation could cause unnecessarily high times computed for certain test cases and incorrect conclusions. Potential shortcuts in languages must also be accounted for.

5.2 Predicted Issues

- We predict that it will be difficult to be certain that all the test cases are translated with exactly the same meaning across languages.