Improving SHA-2 Collisions Using Satisfiability Modulo Theory (SMT) Solvers

BSc Computer Science
Marcel Barlik
marcel.barlik@city.ac.uk

General Information

The project idea originated from a City academic (Nyx-Brain, 2024), posted both on Moodle, as well as outside their office. This makes Martin Nyx Brain a supervisor on this project.

The project does not involve any external clients, or create any arrangements involving outside help as of this time. Outside collaborators, such as field experts, may involve themselves by providing additional insight, information or potential resources, such as computing power like a Virtual Machine (VM), **but will not directly contribute to the codebase of the project**.

This document makes use of Open Sans font for better compatibility, due to the lack of out-of-box presence of Times New Roman on most Unix based distributions.

This document contains approximately 1800 words, excluding the cover sheet, references, ToC, ethics checklist and appendix. Excluding titles, figure captions and other aspects not part of general "content" yields about 1750 words, about +3%, which is within City's policy of word count deviation.

Document Version: 1.0

Version	Date
1.0	2025-02-02

Table of Contents

Solved Problem

This research project aims to utilise unexplored opportunities that have arisen from advances in *"New Records in Collision Attacks on SHA-2"* (Li, Y. Liu, F. And Wang, G, 2024). This research will expand on the novel concept of using a Satisfiability Modulo Theory (SMT) solver for practical SHA-2 collisions, using the principles and mathematics described, in addition to their code as a reference.

Project Objectives

The main purpose of this research is to investigate potential measurable quantified performance differences in SMT solvers and their parameters for SHA-2 collisions.

Research Questions

This project consists of two primary research questions (RQs):

1. **RQ1:** Does using a more effective SMT solver yield better SHA-256 collision results?
2. **RQ2**: Can encodings provided in the research (Li, Y. Liu, F. And Wang, G, 2024) be improved upon, aiming for better practical SHA-256 collisions?
   2.1. Using the *ESPRESSO logic minimizer* (Wikipedia, 2024) heuristic?
   2.2. Using better parallelism?

Justification

Li, Y. Liu, F. And Wang, G, 2024 research has proved that an SMT solver, a well-known NP-Hard problem (Wikipedia, 2024), can be utilised to solve SHA-2; thus implying SHA-2 collision is only as complex as SMT in complexity space. The research did however lack vast experimentation with different SMTs and respective parameters – which could in turn provide better results. **RQ1** shall provide benchmark insights and analysis of the differences and similarities noted.

Li, Y. Liu, F. And Wang, G, 2024 research encodings can potentially be expanded on to provide better efficiency. The longest running aspect of the code is the SMT solver. Number of clauses and variables directly exponentially influence the search space; creating more concise clauses would reduce the search space, and could allow for better throughput. One example of this is the *ESPRESSO logic minimizer* heuristic (Wikipedia, 2024), aiming to remove the "don't-care terms".

The Rust language is built targetting concurrent programming, as defined by one of its values - “Fearless Concurrency” (ch16-00, ‘The Rust Programming Language’, Rust Project Contributors, 2025). The emphasis on correct memory structure and power of concurrency could allow for a potential “*Cube-and-Conquer*” approach when interacting with the SMT solver. (Marijn, Heule, J., Kullmann, O., Wieringa, S. and Biere, A.)

These heuristics and encoding efficiencies can be combined, in addition to other methods, not mentioned here for brevity. **RQ2** shall provide findings on potential heuristics and encoding efficiencies, not limited or confined to these two examples. It shall describe what has been attempted, how it affected benchmarks and an explanation of proving/disproving any performance uplift.

Additional Notes

It is important to note, as this is research, either proving or disproving any of the RQs is seemed as a valid outcome. For example: failure to find an SMT as effective or better, could provide vital information as to explain what key characteristics in SMT are crucial specific to SHA-2 collisions. This knowledge would potentially open a path for future research around the basis of those key characteristics.

During research, additional questions may arise, but will not be the primary concern of this research.

Project Beneficiaries

“The SHA-2 hash function is implemented in some widely used security applications and protocols, including TLS and SSL, PGP, SSH, S/MIME, and IPsec.” (sec. *“Applications”* para. 1, Wikipedia, 2025)

This research will provide a formal verification; assurance that SHA-2 is still securely sound in the near-foreseeable future, while pushing the current field boundaries in SHA-2 SMT collisions, knowledge and benchmarks. In the very unlikely event of a breakthrough, the project can become a research publication, **outside of the scope of this project**, in order to create pressure and emphasis on moving away from SHA-2 onto more secure, quant-safe standards for the public.

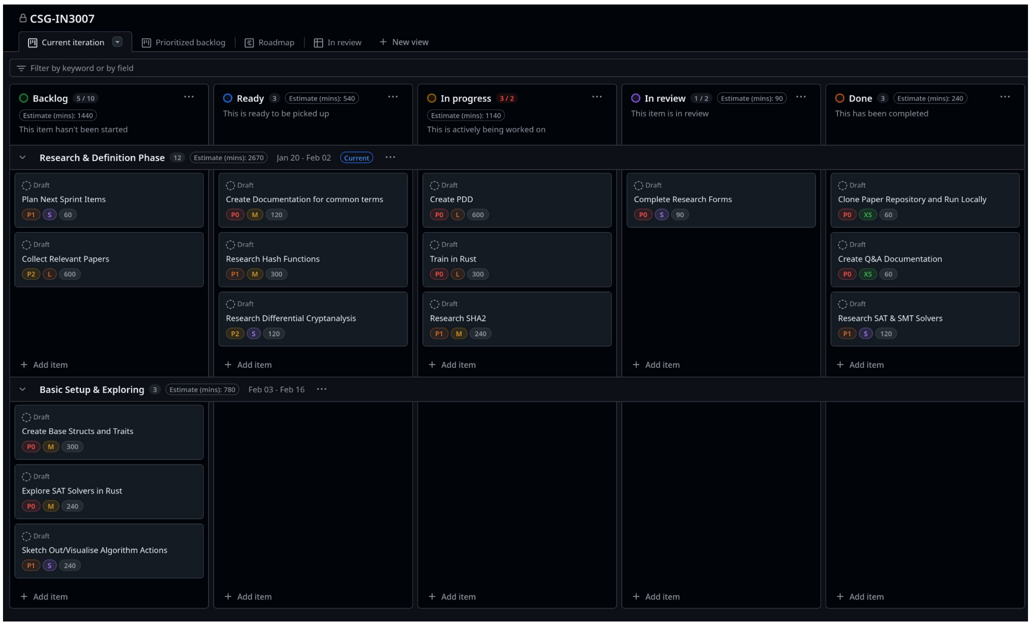
Project Plan

As for project methodology, a Kanban-styled board, split into **multiple sprints** is in use for this project. Downtime and work pressure is accounted for, with two 2-week sprints followed by a 1-week break. The project splits very effectively into these sprints; as an **agile-esque Kanban approach** allows out-of-order task completion – including, but not limited to documentation, development, research and testing. This ensures everything is accounted for as outlined by requirements.

Since this out-of-order task approach does not work very well with a gantt chart, one has not been used for this project. Instead, it is possible to define tasks as dependencies where necessary, but a sprint-planning approach should reduce the need for dependency nested tasks.

I have decided to use GitHub's in-built "project" boards, since they tie in directly to the repository, allowing for easy access and referencing throughout, where necessary.

The GitHub Kanban board (private) can be found here.



Backlog 5 / 10 Estimate (mins): 1440 This item hasn't been started	Ready 3 Estimate (mins): 540 This is ready to be picked up	In progress 3 / 2 Estimate (mins): 1140 This is actively being worked on	In review 1 / 2 Estimate (mins): 90 This item is in review	Done 3 Estimate (mins): 240 This has been completed
Research & Definition Phase 12 Estimate (mins): 2670 Jan 20 - Feb 02 Current				
Draft Plan Next Sprint Items P1 S 60	Draft Create Documentation for common terms P0 M 120	Draft Create PDD P0 L 600	Draft Complete Research Forms P0 S 90	Draft Clone Paper Repository and Run Locally P0 XS 60
Draft Collect Relevant Papers P2 L 600	Draft Research Hash Functions P1 M 300	Draft Train in Rust P0 L 300		Draft Create Q&A Documentation P0 XS 60
	Draft Research Differential Cryptanalysis P2 S 120	Draft Research SHA2 P1 M 240		Draft Research SAT & SMT Solvers P1 S 120
+ Add item	+ Add item	+ Add item	+ Add item	+ Add item
Basic Setup & Exploring 3 Estimate (mins): 780 Feb 03 - Feb 16				
Draft Create Base Structs and Traits P0 M 300				
Draft Explore SAT Solvers in Rust P0 M 240				
Draft Sketch Out/Visualise Algorithm Actions P1 S 240				
+ Add item	+ Add item	+ Add item	+ Add item	+ Add item

Figure 1: Current Project Board

Risks Affecting the Project

The **lack of supervisor time** could result in lack of deeper understanding in the topic, slowing down the research. RQ1 has been chosen to avoid extreme depth in SHA-2 and SMTs, in order to reduce the risk; my current knowledge and prior base research, would be sufficient to answer this RQ. However, as for RQ2, I would require depth of knowledge and would require weekly time with my supervisor. As an alternative, I can utilise public knowledge such as publications, contact other researchers in the field and find out more information, gaining help through other routes.

The project will use Rust as the primary language. As of my knowledge, at present, there is no natively written SMTs or SHA-2 collisions in Rust, only bindings exist. As this is research, anything can happen; it may occur that Rust is an incredibly **poor choice of programming language** for this use case. This in turn would add complications as to requiring a rewrite in a known and trusted language in this field, C++. This is unlikely, since Rust and C++ have direct out-of-box compatibility with linkers, and in fact as defined by '*The Rustonomicon*' (Rust Project Contributors, 2025) the "Repr(C)" trait aims to reproduce C code where possible. This in theory would allow Rust code to be interchangeable with C++, and as of my knowledge, some of it has already been achieved through open-source contributions. Additionally, the community for Rust is very open to answering questions. Specifically related to this project, the Discord community server has #cryptography-and-security and #dark-arts channels, these are very active and have a lot of Rust knowledgable users. In the worst case scenario, a direct fork of the Li, Y. Liu, F. And Wang, G, 2024 research, in C++ and Python, with alterations could be utilised as a base for the project.

A **high-performance HVM** will be required to match the performance benchmarks of the original paper. My aim is to obtain such HVM from outside collaborators potentially interested in this research. As an alternative, I am aware that City does have a HVM, but it does fall short of my needs for longer-term high-performance throughput. I own a X86 24-thread 128GB RAM home-lab, as well as have access to an always-free Oracle Cloud Infrastructure ARM Ampere A1 4-thread 24GB RAM VM. Either of these could be utilised to continue creation of benchmarks, working around the limitations and providing comparable benchmarks. There still would be a conceivable difference based on implementations; "algorithm progress, [...] is sometimes orders of magnitude more important than hardware." (Thompson, N.C., Ge, S. and Sherry, Y.M., 2021)

LSEP Issues

Prefix Notes

To note, it is **nearly impossible** a breakthrough occurs. It is **impossible** to determine the probability of advancements and how large they would be. Based on previous research in this field, which has been very slow-progressing, it is predictable that the performance will not be too dissimilar, to that of current records, set out by Li, Y., Liu, F. and Wang, G. (2024). Since this risk is **nearly impossible**, but **extremely consequential**, I believe this project does **not** pose any **probable** high-risk, similarly to how CVE categorises risk based on knowledge requirement and likely-hood.

Legal

General Data Protection Act (GDPR) does not directly apply to the scope of this project. However, a breakthrough in SHA-2 collisions could be used to break the security of many GDPR compliant companies. Therefore, ethical and responsible disclosure would be required in such scenario.

NCSC does not clearly define any legal requirement to disclose a vulnerability. It does make mention of "Vulnerability Reporting with a UK government online service" – which this project could apply to, but makes no legal obligations. (NCSC, 2018)

ENISA makes mention of the "Coordinated Vulnerability Disclosure" mentioning "the Cybersecurity Act (2019), the NIS2 Directive (2022), and the upcoming Cyber Resilience Act (2024)." (ENISA. n.d.)

As such, if the project were to require vulnerability disclosure, the national bodies responsible for the UK/EU would be contacted, ensuring an ethical, legal and responsible disclosure.

Social

This research will have benefits in progressing the cryptography field, ensuring the safety of public data. It will either create a formal verification that SHA-2 is still secure sound with present knowledge. Alternatively it could disprove the security of SHA-2, showing it is potentially near end-of-life; in this case putting pressure for the world to push for, ideally quant-safe, better alternatives to replace it.

In the **extremely unlikely** scenario that a breakthrough occurs, practical reproducible polynomial-time SHA-2 collisions would pose a real-time threat to many protocols, users and all electronically stored information behind encryption, requiring a mass-scale action of all international security bodies. It would also permanently break some cryptocurrencies, that are embedded in SHA-2 too deeply to change, potentially crashing the cryptocurrency market. It is more positive for a research to discover and disclose this ethically, rather than a user with malicious intent - where it may be too late causing international harm.

Ethical

The original codebase provided by Li, Y., Liu, F. and Wang, G. (2024) does not contain a licence. Their paper does make note that "The source code to search for the differential characteristics and verify the (SFS/FS) collisions for SHA-256 and SHA-512 is available [...]", but it does not set out if forking and working on-top of it would be permissible. By default copyright standards, it is assumed that such property is reserved for the copyright owners. The original research paper has the **CC BY 4.0** licence, which allows for sharing and adapting the principles enclosed as long as attribution is given (CreativeCommons, n.d.). However, it may be unethical to assume that this same licence applies to the code hyperlinked in the paper. This has been mitigated by the decision to create a new codebase in Rust, and use the principles of the paper.

Professional

One professional risk for me, is potential to release skewed benchmarks, misleading others. One way of preventing misleading benchmarks, is to ensure all variables remain the same, and only the SMT or its parameters change for **RQ1.** This would provide a like-for-like benchmark, and with enough runs for averages and standard deviations could prove to be an accurate representation of best algorithmically performing SMTs to answer **RQ1**.

Ethics Review

Part A: Ethics Checklist

A.1: If you answer YES to any of the questions in this block, your consultant/supervisor must have obtained approval for the project from an appropriate external ethics committee, and you need to have received written confirmation of this from him/her. Students cannot themselves apply for ethics approval in this case as the project is considered high risk". This type of research is not covered by City's process, and external approval from an appropriate institution is required.		Answer
1.1	Does your research require approval from the National Research Ethics Service (NRES)?	NO
1.2	Will you recruit participants who are covered by the Mental Capacity Act 2005?	NO
1.3	Will you recruit any participants who are covered by the Criminal Justice System, for example, people on remand, prisoners and those on probation?	NO

A.2: If you answer YES to any of the questions in this block your consultant/supervisor must have obtained appropriate ethics committee approval.		Answer
2.1	Does your research involve participants who are unable to give informed consent?	NO
2.2	Is there a risk that your research might lead to disclosures from participants concerning their involvement in illegal activities?	NO
2.3	Is there a risk that obscene and or illegal material may need to be accessed for your research study (including online content and other material)?	NO
2.4	Does your project involve participants disclosing information about protected characteristics (as identified by the Equality Act 2010)?	NO
2.5	Does your research involve you travelling to another country outside of the UK, where the Foreign & Commonwealth Office has issued a travel warning that affects the area in which you will study?	NO
2.6	Does your research involve invasive or intrusive procedures?	NO
2.7	Does your research involve animals?	NO
2.8	Does your research involve the administration of drugs, placebos or other substances to study participants?	NO

A.3: If you answer YES to any of the questions in this block, then unless you are applying to an external ethics committee or the Senate Research Ethics Committee (SREC), you must apply for approval from the Computer Science Research Ethics Committee (CSREC) through Research Ethics Online - https://researchmanager.city.ac.uk/. Depending on the level of risk associated with your application, it may be referred to the Senate Research Ethics Committee (SREC).		**Answer**
3.1	Does your research involve participants who are under the age of 18?	NO
3.2	Does your research involve adults who are vulnerable because of their social, psychological or medical circumstances (vulnerable adults)?	NO
3.3	Are participants recruited because they are staff or students of City, University of London?	NO
3.4	Does your research involve intentional deception of participants?	NO
3.5	Does your research involve participants taking part without their informed consent?	NO
3.5	Is the risk posed to participants greater than that in normal working life?	NO
3.7	Is the risk posed to you, the researcher(s), greater than that in normal working life?	NO
A.4 If you answer YES to the following question and your answers to all other questions in sections A1, A2 and A3 are NO, then your project is deemed to be of MINIMAL RISK. **If this is the case, then you can apply for approval through your supervisor under PROPORTIONATE REVIEW. You do so by completing PART B of this form.** **If you have answered NO to all questions on this form, then your project does not require ethical approval. You should submit and retain this form as evidence of this.**		**Answer**
4	Does your project involve human participants or their identifiable personal data?	NO

Part B: Ethics Proportionate Review Form

Omitted for brevity due to not being applicable. The answer to A.4 is “No”, since no human participants or their data will be involved. All testing and benchmarking will be done on generated SHA-2 pairs, unrelated to anyone or anything.

References

[1] Nyx-Brain, M. (2024) 'Improving Attacks on the SHA-2 Algorithms'. Available at: https://moodle4.city.ac.uk/pluginfile.php/1093061/mod_folder/content/0/%5BSecurity%5D%5BCryptography%5DImproving-Attacks-on-the-SHA-2-Algorithms.docx accessed on 2025-01-30.

[2] Li, Y., Liu, F. and Wang, G. (2024) 'New Records in Collision Attacks on SHA-2'. Available at: https://eprint.iacr.org/2024/349 accessed on 2025-01-30.

[3] Thompson, N.C., Ge, S. and Sherry, Y.M. (2021) 'Building the algorithm commons: Who discovered the algorithms that underpin computing in the modern enterprise?', Global Strategy Journal, 11(1), pp. 17–33. Available at https://onlinelibrary.wiley.com/doi/epdf/10.1002/gsj.1393 accessed on 2025-02-02.

[4] Rust Project Contributors (2025) 'The Rust Programming Language'. Available at: https://github.com/rust-lang/book/blob/main/src accessed on 2025-01-31, with latest commit SHA fa312a3.

[5] Rust Project Contributors (2025) 'The Rustonomicon'. Available at: https://github.com/rust-lang/nomicon accessed on 2025-01-31, with latest commit SHA bc22988.

[6] Wikipedia Contributors (2025) 'SHA-2'. Available at: https://en.wikipedia.org/w/index.php?title=SHA-2&oldid=1271918606 accessed on 2025-01-31.

[7] Wikipedia Contributors (2024) 'Satisfiability modulo theories'. Available at: https://en.wikipedia.org/w/index.php?title=Satisfiability_modulo_theories&oldid=1250965920 accessed on 2025-02-02.

[8] Wikipedia Contributors (2024) 'Espresso heuristic logic minimizer'. Available at: https://en.wikipedia.org/w/index.php?title=Espresso_heuristic_logic_minimizer&oldid=1251095784 accessed on 2025-02-02.

[9] Marijn, Heule, J., Kullmann, O., Wieringa, S. and Biere, A. (n.d.) 'Cube and Conquer: Guiding CDCL SAT Solvers by Lookaheads'. Available at: https://www.cs.utexas.edu/~marijn/publications/cube.pdf accessed on 2025-02-02.

[10] National Cyber Security Centre (NCSC) (2018) 'Vulnerability Reporting'. Available at: https://www.ncsc.gov.uk/information/vulnerability-reporting accessed on 2025-02-02.

[11] European Network and Information Security Agency (ENISA) (n.d.) 'Vulnerability Disclosure'. Available at: https://www.enisa.europa.eu/topics/vulnerability-disclosure accessed on 2025-02-02.

[12] CreativeCommons (n.d.) 'Attribution 4.0 International Deed (CC BY 4.0)'. Available at: https://creativecommons.org/licenses/by/4.0/ accessed on 2025-02-02.

Appendix

No client information sheet is provided, since the project does not involve a client.

No AI tools have been utilised as part of this PDD.

For category 3, section 4.5.7 "Legal, Social, Ethical and Professional Issues (LSEPI)" a "Prefix Notes" section has been used instead of an Appendix for better clarity.