Reese Levine

reeselevine1@gmail.com
https://reeselevine.github.io

Research

My research interests are in parallel and concurrent programming. My work has focused on GPU programming models, including designing and evaluating techniques to test the conformance of compilers and hardware to memory model specifications, and testing and improving the safety properties of GPU programming languages in the face of data races. I am currently collaborating on projects related to GPU security and compiler correctness. I am also actively contributing to llama.cpp in order to bring GPU-accelerated inference to the browser using WebGPU, opening up research opportunities in portable performance, efficient LLM quantization, and improving GPU programming models.

Education

- **University of California, Santa Cruz** — Santa Cruz, CA
  PhD in Computer Science; GPA: 4.0 — *September 2020 - Present*
  - **Selected Classes:** Compiler Design, Computer Architecture, Advanced Programming Languages, Distributed Systems, Formal Methods
- **University of California, Berkeley** — Berkeley, CA
  Bachelor of Arts in Computer Science; GPA: 3.6 — *August 2013 – May 2017*
  - **Selected Classes:** Operating Systems, Efficient Algorithms, Computer Security, Introduction to Databases

Publications

Conference

- **Reese Levine**, Ashley Lee, Neha Abbas, Kyle Little, and Tyler Sorensen. "Assessing and Addressing WebGPU Memory Safety in the Presence of Data Races". In *In Proceedings of the ACM on Programming Languages (OOPSLA)*. 2025.
- **Reese Levine**, Mingun Cho, Devon McKee, Andrew Quinn, and Tyler Sorensen. "GPUHarbor: Testing GPU Memory Consistency at Large (Experience Paper)". In *In International Symposium on Software Testing and Analysis (ISSTA)*. 2023. **Distinguished Artifact**.
- **Reese Levine**, Tianhao Guo, Mingun Cho, Alan Baker, Raph Levien, David Neto, Andrew Quinn, and Tyler Sorensen. "MC Mutants: Evaluating and Improving Testing for Memory Consistency Specifications". In *Architectural Support for Programming Languages and Operating Systems (ASPLOS)*. 2023. **Distinguished Paper, Distinguished Artifact**.

Workshop

- **Reese Levine**, and Tyler Sorensen. "Probabilistic Memory Consistency Specifications". In *Young Architect Workshop*. 2023

Articles

- Arkaprava Basu, Natalia Gavrilenko, Keijo Heljanko, **Reese Levine**, Ajay Ashok Nayak, Hernan Luis Ponce de Leon, Tyler Sorensen and Haining Tong. "GPU Memory Consistency: Specifications, Testing, and Opportunities for Performance Tooling". In *Computer Architecture Today* (ACM SIGARCH Blog). 2025

Talks

- "Testing the Vulkan Memory Model", Vulkanised 2024, Sunnyvale, CA, February 2024
- "Evolving Weak Memory Models for Evolving Architectures", Future of Weak Memory at POPL 2024, London, January 2024
- "Testing GPU Memory Consistency at Large", Imperial College London, University of Kent, Cambridge University, Bristol University, Stanford University, January-October 2024
- "MC Mutants: Evaluating and Improving Testing for Memory Consistency Specifications", Khronos F2F, Phoenix, AZ, October 2022
- "Testing Memory Models", Languages, Systems, and Data Seminar, UC Santa Cruz, January 2022

Awards and Grants

- **National Defense Science and Engineering Graduate (NDSEG) Fellowship**, 2023

Teaching

- **UC Santa Cruz** — Santa Cruz, CA
  Teaching Assistant — *Spring 2021/Winter 2022/Winter 2023*
  - TA for undergraduate class on parallel programming: homework development, office hours, grading homeworks/exams
- **UC Berkeley** — Berkeley, CA
  Teaching Assistant — *Summer/Fall 2016*
  - Taught students concepts in computer architecture, updated lab exercises, developed and graded exams

Industry

- **Apple** — Cupertino, CA
  Intern — *July 2022 - September 2022/June 2023 - September 2023*
  - Worked on Apple's GPU Platform Architecture team doing new feature design/testing and performance optimization
- **Qualtrics** — Seattle, WA
  Software Engineer — *August 2017 - September 2020*
  - Developed new method of storing data in Elasticsearch indexes to reduce hardware usage by 10x while maintaining customer latency SLAs
  - Designed and implemented an improved ingestion pipeline using Scala and Akka Streams that increased data indexing rates by 40 percent while reducing operational load and providing fairness and prioritization
  - Built application using Scala and the Play framework to perform background tasks for Qualtrics' Analytics Engine like garbage collection and defragmentation of data in Elasticsearch
  - Contributed to incident remediation and operational hardening, including presenting analysis of severe incidents to engineering leadership
  - Mentored intern in summer long project involving new data analysis feature requested by key customers
- **Munchery** — San Francisco, CA
  Software Engineering Intern — *May 2015 - July 2015*

- Developed Ruby bot on Slack allowing customer care to communicate directly with delivery drivers through Twilio SMS.
- Contributed to open-source Jenkins plugin allowing provisioning of Docker containers on Amazon EC2.
- Wrote comprehensive QA tests for updated Munchery checkout page.

Service

- **Languages, Systems, and Data (LSD) Seminar** — Santa Cruz, CA
  Co-organizer — *June 2024 - Present*
- **2024/2025 PLDI Artifact Evaluation Committee**
- **2024 ASPLOS Artifact Evaluation Committee**
- **Advisory Committee for Campus Transportation and Parking** — Santa Cruz, CA
  Member — *September 2021 - June 2023*
- **SPLASH/OOPSLA** — Chicago, IL
  Student Volunteer — *October 2021*
- **TEALS** — Seattle, WA and Velva, ND
  Volunteer Teacher — *June 2018 - May 2020*
- **Computer Science Mentors** — Berkeley, CA
  Volunteer Teacher — *January 2015 - May 2016*
- **Vice Chancellor's Student Advisory Committee** — Berkeley, CA
  Member — *August 2014 - May 2015*