Reese Levine

reeselevine@ucsc.edu
https://reeselevine.github.io

RESEARCH

My research centers on analyzing modern GPU abstractions in ways that both enable validation of their programming models, including security guarantees, and drive the rapid development of new GPU-based systems, e.g., LLM inference engines. My ongoing projects include empirically evaluating the correctness and security of GPU compilers and hardware and contributing to llama.cpp in order to bring performant and efficient LLM inference to all devices.

EDUCATION

- **University of California, Santa Cruz** — Santa Cruz, CA
  PhD in Computer Science; GPA: 4.0 — *September 2020 - June 2026 (expected)*
  - **Advisor**: Tyler Sorensen
  - **Thesis (Working Title):** Validating GPU Memory Consistency and Safety at Scale
  - **NDSEG Fellowship**

- **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

Under Submission

- Dennis Sprokholt, **Reese Levine**, Susmit Sarkar, Andrés Goens, Vijay Nagarajan, Soham Chakraborty. "Beyond the Fence: Sound Compilation for Homogeneous GPU and Heterogeneous CPU+GPU Systems". Under submission at top programming language conference.

- Rithik Sharma, **Reese Levine**, Ananthajit Srikanth, Yanwen Xu, Tyler Sorensen. "GPU Goldmines: Specifying, Executing, and Analyzing Tunable AI Shaders in the Browser using WebGPU". Under submission at systems/machine learning conference.

Conference

- Cheng Gu, **Reese Levine**, Zhenkai Zhang, Tyler Sorensen, Yanan Guo. "Behind Bars: A Side-Channel Attack on NVIDIA MIG Cache Partitioning Using Memory Barriers". To appear at USENIX Security 2026.

- **Reese Levine**, Ashley Lee, Neha Abbas, Kyle Little, and Tyler Sorensen. "Assessing and Addressing WebGPU Memory Safety in the Presence of Data Races". In *Proceedings of the ACM on Programming Languages (OOPSLA)*. 2025.

- **Reese Levine**, Mingun Cho, Devon McKee, Andi Quinn, and Tyler Sorensen. "GPUHarbor: Testing GPU Memory Consistency at Large (Experience Paper)". In *International Symposium on Software Testing and Analysis (ISSTA)*. 2023. **Distinguished Artifact**.

- **Reese Levine**, Tianhao Guo, Mingun Cho, Alan Baker, Raph Levien, David Neto, Andi 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

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

- **TEALS** — Velva, ND/Remote
  Teaching Assistant — *Fall 2018-Spring 2020*
  - Volunteer teacher for Technology Education and Learning Support (TEALS), a program supported by Microsoft
  - Taught high school students in a class in Velva, North Dakota introductory computer science topics using SNAP and Python

Mentoring

Students who I directly mentored/supervised, as well as (to the best of my knowledge) their next position afterwards:

- **Tianhao Guo (UCSC)**: Memory consistency testing framework using WebGPU in the browser.
  Outcome: Co-author on ASPLOS paper, Master's in Computer Engineering at NYU

- **Kyle Little (UCSC)**: Fuzzing tool for WebGPU memory safety.
  Outcome: Co-author on OOPSLA paper, started PhD at University of Utah

- **Ashley Lee (UCSC)**: Browser-based fuzzing framework for WebGPU memory safety.
  Outcome: Co-author on OOPSLA paper

- **Neha Abbas (UCSC)**: Fuzzing reduction for WGSL; llama.cpp shader development.
  Outcome: Co-author on OOPSLA paper

- **James Contini (UCSC)**: Parallel prefix sum optimization; llama.cpp shader development.
  Outcome: Undergrad thesis, applying to PhD programs

- **Kaushal Anbarasan (UCSC)**: Weak-memory analysis of shared-memory algorithms.
  Outcome: Software Engineer at Palo Alto Networks

- **Camille Gandotra (UCSC)**: Weak-memory analysis of shared-memory algorithms.
  Outcome: Software Engineer at Microsoft

Talks

- "SafeRace: Assessing and Addressing WebGPU Memory Safety in the Presence of Data Races", OOPSLA, Singapore, October 2025

- "Data Races: A Ticking Time Bomb for WebGPU Security", W3C GPU for the Web Working Group, August 2024

- "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"

  - January 2024: Imperial College London, University of Kent, Cambridge University, Bristol University
  - October 2024: Stanford University

- "GPUHarbor: Testing GPU Memory Consistency at Large", ISSTA, Seattle, WA, July 2023

- "MC Mutants: Evaluating and Improving Testing for Memory Consistency Specifications"

  - Khronos F2F, Phoenix, AZ, October 2022
  - ASPLOS, Vancouver, BC, March 2023

- "Testing Memory Models", Languages, Systems, and Data Seminar, UC Santa Cruz, January 2022

Industry

- **Apple**   Cupertino, CA
  Intern   *Summer 2022/2023*
  - 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   *Summer 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.

Grants and Funding

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

Awards, Press, and Impact

- "Cross-Process GPU Memory Disclosure on AMD GPUs" (CVE-2024-36353), 2024

    - https://nvd.nist.gov/vuln/detail/CVE-2024-36353

- Bounds-checking bug/vulnerability in Firefox pre-release

    - https://github.com/gfx-rs/wgpu/pull/6431

- **Distinguished Artifact Award**, ISSTA 2023

- MC Mutants work mentioned in UCSC press release/ACM TechNews

    - https://news.ucsc.edu/2023/03/sorensen-bugs/
    - https://technews.acm.org/archives.cfm?fo=2023-03-mar/mar-24-2023.html

- **Distinguished Paper Award, Distinguished Artifact Award**, ASPLOS 2023

- Memory model tests added to WebGPU Conformance Test Suite

    - https://github.com/gpuweb/cts/pull/1330

Projects

- **GPUHarbor: Memory Model Testing**

    - Public website: https://gpuharbor.ucsc.edu/webgpu-mem-testing/
    - Code: https://github.com/reeselevine/webgpu-litmus
    - Allows users to explore and tune parameters for weak memory model litmus tests, characterizing behaviors and finding bugs in conformance tests.

- **GPUHarbor: WebGPU Memory Safety Fuzzing**

    - Public website: https://gpuharbor.ucsc.edu/webgpu-race-testing/
    - Code: https://github.com/CPTforever/webgpu-race
    - Performs fuzzing that combines interesting program structures with data races to see if WebGPU implementations can accidentally optimize away bounds-checks necessary for memory safety.

- **Llama.cpp: LLM inference in WebGPU**

    - https://github.com/reeselevine/llama.cpp
    - Exploring new methods for portable and performant LLM inference using WebGPU, which also enables open-source models to be run in native and browser-based environments.

SERVICE

- **Languages, Systems, and Data (LSD) Seminar** — Santa Cruz, CA
  Co-organizer — *June 2024 - Present*
  - Invited speakers and led logistics planning for UCSC's LSD seminar, with a focus on inviting graduate students, including those from underrepresented backgrounds.

- **2024/2025 PLDI Artifact Evaluation Committee**
  - Reviewed 3 artifacts each year for accepted papers

- **2024 ASPLOS Artifact Evaluation Committee**
  - Reviewed 4 artifacts over the course of 3 submission cycles

- **Workshop on Irregular Applications: Algorithms and Architectures (IA^3) at SC**
  Sub-Reviewer — *2021-2025*
  - Helped review papers, focusing on one submission and holding a "mini-PC" meeting with my advisor and fellow PhD students

- **Advisory Committee for Campus Transportation and Parking** — Santa Cruz, CA
  Member — *September 2021 - June 2023*

- **SPLASH/OOPSLA** — Chicago, IL
  Student Volunteer — *October 2021*

- **Technology Education and Literacy in Schools (TEALS)** — Seattle, WA and Velva, ND
  Volunteer Teacher — *June 2018 - May 2020*
  - Co-taught introduction to programming class at a high school in North Dakota, mostly remotely but with several trips out per year to meet students in-person.

- **Computer Science Mentors** — Berkeley, CA
  Volunteer Teacher — *January 2015 - May 2016*
  - Led small group (3-4 students) supplementary discussion sections for data structures and algorithms classes

- **Vice Chancellor's Student Advisory Committee** — Berkeley, CA
  Member — *August 2014 - May 2015*

REFERENCES

- **Tyler Sorensen**
  Associate Professor, UC Santa Cruz
  Principal Researcher, RiSE group at Microsoft
  `tyler.sorensen@ucsc.edu`

- **Andi Quinn**
  Assistant Professor, UC Santa Cruz
  `aquinn1@ucsc.edu`

- **Vijay Nagarajan**
  Professor, University of Utah
  `vijay@cs.utah.edu`

- **Justin Hensley**
  Distinguished Engineer, Apple

Other references and contact information available upon request.