Jack Cashman

cashman3@purdue.edu | +1-516-659-2087
LinkedIn Profile | jackfcashman@gmail.com
cashmanjack.github.io | GitHub: cashmanjack
Google Scholar

Education

•Purdue University *August 2026 - Present*
Ph.D. in Electrical and Computer Engineering West Lafayette, IN

•University of Rochester *August 2022 – May 2026*
B.S. in Computer Science (Honors Research) Rochester, NY

Research & Teaching Experience

•University of Utah *May 2025 – July 2025*
Research Assistant under Prof. Ponnuswamy Sadayappan Salt Lake City, UT

- Investigated sparse matrix reordering schema cache hierarchy utilization and its effects on Sparse Matrix Vector Multiplication (SpMV) performance.
- Reduced error of a machine learning model for pairwise ranking of data layouts based on predicted L1D, L2, and L3 cache misses, addressing the fragility of traditional reordering heuristics.
- Conducted controlled experiments by disabling hardware prefetching and varying compiler optimization levels to isolate performance effects and align empirical results with model predictions.

•University of Rochester, Dept. of Computer Science *August 2025 – April 2026*
Honors Thesis Researcher Rochester, NY

- Designing an analytical cache model that directly targets performance factors including cache capacity, associativity, latency, replacement policy, prefetchers, and thread count.
- Integrating analytical prediction with genetic search and heuristics to drive sparse matrix reordering decisions, departing from traditional graph theory approaches.
- Parallelizing reordering search and analytical modeling for improved speed of reordering, and validating results with hardware counters.

•University of Rochester *May 2024 – July 2024*
Research Assistant under Prof. Chen Ding Rochester, NY

- **Loop Asymptotic Locality Analysis (LALA):** Developed a symbolic tool to model the reuse distance distribution of affine programs, reducing analysis time from doubly exponential to linear relative to program size.
- **Zipfian Caching for Web Workloads:** Designed and evaluated caching policies to outperform LRU on dynamic Zipfian web workloads, achieving a faster convergence to the optimal hit rate.
- **LLM Text Detection:** Used locality measurements of data movement distance for varying N-gram text representation to classify text as human vs. LLM-generated.

•University of Rochester *August 2024 – April 2026*
Teaching Assistant, Computer Organization Rochester, NY

- Led study sessions and office hours to help undergraduate students understand core systems concepts, including memory hierarchy, instruction sets, and digital logic.
- Wrote, graded, and proctored exams, and provided feedback on assembly language and C programming assignments.

Selected Projects

•Tiny-QMoE LLM Compression System *Fall 2024*
Course Research Project, Machine Learning Systems

- Engineered a quantization and dictionary-based compression pipeline to enable inference of large language models (LLMs) on memory-constrained devices (e.g., mobile phones).
- Achieved **23–35x compression** on LLaMA models with minimal accuracy loss on MMLU and ARC benchmarks.
- Gained hands-on experience with GPU memory hierarchy, low-latency compression algorithms, and LLM benchmarking.

•Latency Based Tiling *Spring 2025*

Course Research Project, Honors Undergraduate Problem Seminar

- Built a triangle loop based cache latency micro benchmark for approximating cache capacities for L1D, L2, and L3 caches.
- Leveraged the cache approximation for polyhedral program tiling to produce a 21% performance speedup with effectively no compilation overhead unlike auto-tuning.

•Mid-End Compiler Optimization Framework *Spring 2025*

Course Project, Advanced Compilers (Software Analysis & Improvement)

- Built a compiler optimization framework from scratch, operating on an intermediate representation.
- Implemented classic and advanced optimizations including dead code elimination, partial redundancy elimination, and constant propagation using abstract interpretation.

Publications

- **Preprint:** Jack Cashman. *Latency Based Tiling.* arXiv preprint arXiv:2510.15912, 2025.
- **Preprint:** Jack Cashman, Jiaqi Nie. *Tiny-QMoE.* arXiv preprint arXiv:2509.22951, 2025.

Technical Skills and Interests

Languages: Rust, C, RISC-V Assembly, x86-64 Assembly, Coq, CUDA

Developer Tools: Git, Linux/Unix Shell, VTune, PAPI, LIKWID, Perf, GDB

Relevant Coursework: Advanced Compiler Optimizations, Machine Learning Systems, Computer Organization, Operating Systems, Programming Language Design, Data Structures & Algorithms, Computer Models & Limitations, Honors Research Problem Seminar, Computer Networks

Areas of Interest: Compiler Optimization, Locality, Cache Modeling, Computer Microarchitecture, Hardware Accelerators, Type Systems/Safety, Machine Verified Proofs