Annie Hu

✉ anniegracehu@gmail.com   📞 614-636-1232   ⚙ /anniehu17   in /in/annie-grace-hu

Experience

Jane Street | Software Engineering Intern   May 2025 – Aug 2025
- GPU Performance (ongoing): Implementing PyTorch baselines and profiling computational bottlenecks for 5 state-of-the-art sequence model architectures (Mamba2, RetNet, Hyena, RG-LRU, xLSTM). Developing CUDA/Triton kernels for accelerated inference. Baselines will serve as reference implementations for upcoming GPU MODE hackathon.
- Built an automated testing system for trading applications by enabling deterministic replay of external system interactions, preventing production incidents across 6+ critical apps.

The San Francisco Compute Company | Trading Systems Software Engineer   Feb 2025 – May 2025
- Architected and led the overhaul of the trading engine, resource allocator, and API to support customized on-demand GPU configurations, enabling customers to specify GPU requirements at purchase time.
- Developed a high-performance asynchronous Rust bid/ask matching engine to replace the mixed-integer programming solver, reducing order processing time from hours to seconds.

Five Rings | Software Engineering Intern   January 2025
- Implemented C++ structured binding support for custom trading types through tuple specialization, automating testing and boilerplate generation with Python, enabling cleaner and more maintainable code.

Roblox | Software Engineering Intern   May 2024 - Aug 2024
- Built a full-stack investigation tool for violative 3D content, consolidating data from multiple disparate sources into a single dashboard, eliminating the need to navigate between different tools.
- Led the full development lifecycle from stakeholder interviews and requirements gathering through technical design and implementation.
- Designed real-time APIs and React frontend with search, filtering, and rich media display capabilities.

Projects

Sieve: Learning-Based Prefetch Filter, Paper   C++, Python, gem5
- Developed Sieve, a flexible machine learning-based filter for hardware prefetching, which intelligently decides which memory prefetches to allow, reducing wasted bandwidth and cache pollution in modern CPUs.
- Engineered a modular system in gem5 that supports combining and evaluating multiple prefetching strategies at once, and introduced a novel way to combine filtering with adaptive prefetcher selection ("sandboxing") for even better performance.
- Achieved an improvement of up to 4.15% in processor performance on real-world GAP and PARSEC parallel workloads.

Operating System Kernel with Ext2 FS and Shell   C++, Rust, QEMU, GDB, mkfs.ext2
- Built a multi-threaded kernel with synchronization primitives (semaphores, spin locks, blocking locks) and system calls (fork, exec, pipe, file operations).
- Implemented an Ext2 file system driver with intelligent caching (92% I/O reduction) and symbolic link resolution.
- Created a POSIX-compliant cross-platform shell with user programs (ls, cat, echo) that support pipes, redirection, and command substitution.

Pythonic Language to Assembly Compiler   Rust, C, ARM/x86-64 Assembly, CPython bytecode
- Built a compiler from a custom Pythonic language to ARM/x86 assembly, using a Python disassembled bytecode inspired intermediate representation.
- Implemented tail call and constant folding optimizations, with support for first-class functions, pointer arithmetic, and hardware I/O operations.

Education

University of Texas at Austin (B.S. in Computer Science, Turing Scholars Honors)   GPA: 4.0/4.0

Technical Skills

Languages: Python, C/C++, CUDA, Triton, Rust, x86/ARM Assembly, OCaml, Java, TypeScript, SQL
Technologies: PyTorch, Perfetto, macOS Instruments, Linux, AWS, Docker, Grafana, Elasticsearch, Kibana, React, PostgreSQL