Elijah Baraw

ebaraw@andrew.cmu.edu | (203) 731-9535 | Pittsburgh, PA | github.com/elijah-bae-raw

EDUCATION

Carnegie Mellon University, School of Computer Science **Aug 2021 – May 2025**
Bachelor of Science in Computer Science. Concentration in Computer Systems
GPA: 3.94
Relevant Coursework: Functional Programming, Machine Learning, Distributed Systems, Cloud Computing, Parallel Computer Architecture, Operating System Design, Deep Learning

EXPERIENCE

Epic Systems **Madison, WI**
Software Engineer **July 2025 – Present**

- Built LLM gateway service integrating LiteLLM to provide unified interface for multiple AI model providers
- Configured Grafana dashboards and Prometheus metrics and alerts for monitoring AI platform microservices
- Managed cloud infra using Terraform, debugged production issues across distributed services

PROJECTS

Strassen-Winograd Matrix Multiply for GPUs *CUDA, C++, Nsight* **Jan 2025 – May 2025**

- Implemented $O(n^{2.8})$ matmul on NVIDIA RTX 2080 and V100 GPUs, outperforming cuBLAS for $n \geq 4096$
- Designed custom CUDA kernels, tuning via programatic search over parameter space
- Reduced memory overhead from 2x to 1.5x through strategic intermediate result reuse and pointer swapping

x86 IA-32 Kernel from Scratch *C, ASM, Simics* **Aug 2024 – Dec 2024**

- Built a complete i386 kernel from scratch, solo, for CMU 15410, implementing preemptive multitasking
- Engineered hardware interfaces, memory management, and I/O system for concurrent ELF binary execution

BET Programming Language & Compiler *Rust, ASM* **Aug 2024 – Dec 2024**

- Designed type- and memory-safe language with structs, mutually-recursive types, and function pointers
- Built compiler in Rust (lexer, parser, type checker, x86 codegen) targeting MacOS and Linux
- Re-architected from recursive to stack-based to compile million-line programs without a stack overflow

Poker-Bots Hackathon Dev Team *GCP, K8s, GitHub Actions* **Mar 2024; Mar 2025**

- Helped CMU Data Science Club run their first AI Poker-Bot competition, with $6,000 in prizes and 63 teams.
- Used GitHub Actions to automatically build Docker images of user-submitted Python bots, allowing competitors to use custom dependencies and machine learning libraries of their choice, running containers on GCP.
- Helped build the second iteration of the competition in 2025 using AWS ECS for bots, Lambda for matches

Distributed Backend (Golang) *Replication, Actor Model, Mailbox/Message Passing* **Nov 2023**

- Designed and executed a concurrent server to manage the state for a multiplayer game, accessible via API.
- Handled client requests about and updates to the game state using RPCs and a message-passing model.
- Implemented node launching and server groups, ensuring replication and enforcing consistency within groups.

TECHNICAL SKILLS

Languages: C, Python, Rust, Go, SQL, Java, HCL
Frameworks & Libraries: CUDA, NumPy, PyTorch, FastAPI, Spark
Tools & Platforms: AWS, Docker, K8s, Terraform, Git, GCP, Azure, Grafana, Prometheus