Aaron Wang

a23wang@uwaterloo.ca
github.com/AaronWang04
linkedin.com/in/aaron-wang-waterloo

Education

University of Waterloo - Bachelor of Computer Engineering 2022 – 2026

- Relevant Coursework – Operating Systems, System Programming and Concurrency, Formal Verification
- GPA: 3.9 - Recipient of the President's Scholarship of Distinction

Work Experience

CentML Toronto, CA

ML Systems Engineer Intern - Python, CUDA *Sept 2024 - Jan 2025*

- System level performance engineering for LLM Inference: Distributed runtime scheduling, metrics and logging
- Implemented model runner using gRPC and pickle serialization, reduced inter-process communication time and resulted in 20% lower host side execution overhead compared to Ray
- Added support for speculative decode metrics collection, updated Prometheus endpoint and Grafana dashboard
- Benchmark, profile, and collected inference performance characteristics, identified source of bottleneck (KV-Cache, NCCL overhead) at various workloads

Huawei Toronto, CA

Research Engineer Intern - Python, C++ *Jan 2024 - May 2024*

- Researched novel algorithms for improving compute/communication efficiency in large distributed AI-Training systems
- Trained ML models of various architecture in PyTorch, sampled and analyzed network communication data over NCCL
- Studied SoTA papers on architectures and solutions within modern distributed-training production systems
- Co-authored a research paper on scheduling techniques that leverages workload characterization and preemption
  - Symphony: Collective Scheduling in Multi-Tenant GPU Clusters (Under Submission)

Manulife Waterloo, CA

Student Developer Intern - TypeScript *May 2023 - Aug 2023*

- Built robust web pages using React, Node, and GraphQL that generates feedback on user documents with external APIs
- Implemented a modular React-Redux store structure, streamlining data flow and reducing re-renders by 2x
- Automated testing with Jest and Postman, devised test plan with 30+ unit tests to ensure code functionality and achieve full coverage

Extracurricular

Waterloo Aerial Robotics Group Waterloo, CA

Team Lead - Autonomy - C, Python, Dart *Sept 2022 - Current*

- Leading team of 50 to develop autonomous control software on custom drones for annual competition
- Led team in achieving first place at AEAC 2024 competition with Project Pegasus quadcopter and ground station
- Architected and software bring-up of multiple projects: Perception-Decision Control Systems, Computer Vision, Ground station GUI, Path Planning, Data Telemetry

Projects

Amber Tensor Library – C++, Cuda

- Built a lightweight PyTorch-style tensor template library with strided tensor representation, element-wise and matmul operations

Skills

- **Languages:** C, C++, CUDA, Python, Bash, JavaScript/TypeScript, Dart
- **Frameworks:** PyTorch, Jax, React, Node.js, GraphQL, Flutter
- **Tools:** Git, CI/CD with Github Actions, Docker, Kubernetes, GCP, AWS, CMake, Linux