Aaron Wang

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

Education

University of Waterloo - Bachelor of Computer Engineering 2022 – 2027

- GPA: 3.9 - Relevant Coursework: Operating Systems, System Programming and Concurrency, Discrete Logic

Skills

- **Languages:** C, C++, CUDA, Verilog, Python, Bash
- **Frameworks and Tools:** Linux, CMake, Docker, Kubernetes, Slurm, NumPy, Pandas, PyTorch

Work Experience

NVIDIA Santa Clara, CA

Software Engineer Intern - Python, C++, CUDA *May 2025 - Sep 2025*

- Performance engineering for **PyTorch**: Working on GPU kernels, deep learning compilers, and low precision data types
- Implemented fused kernels to eliminate communication overhead for RMSNorm operation and reduce memory footprint by bypassing autograd, improved throughput up to 9x
- Software bringup and performance optimizations for next-gen hardware architectures

NVIDIA Toronto, ON

Software Engineer Intern - Python *Sep 2024 - Dec 2024*

- Worked on various **vLLM** performance optimizations (CudaGraphs, speculative decoding, multi-node inferencing, TPUs)
- Built distributed model runner using **gRPC** for multi-node deployments, reducing inter-process communication time and resulted in **15% higher throughput** compared to Ray runtime
- Spearheaded integration of inference engine with Hidet graph compiler for quantized kernel generation
- Benchmarked and profiled inference performance characteristics, identified source of bottleneck (memory access, scheduling overhead) at various workloads

Huawei Toronto, ON

Research Engineer Intern - C++ *Jan 2024 - Apr 2024*

- Researched novel algorithms for **improving compute/communication** efficiency in large distributed AI-Training systems
- Trained ML models of various architecture in PyTorch, collected and analyzed network communication data
- Developed a simulator in **C++** and **Python** to model networking communication during distributed parallel training
- Co-authored a **research paper** on **load-balancing** techniques that leverages workload characterization and preemption
  - Symphony: Collective Scheduling in Multi-Tenant GPU Clusters

Manulife Program - Skillify Waterloo, ON

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

- Built web pages using React, Node, and GraphQL that generates feedback on user documents with OpenAI APIs

Extracurricular

Waterloo Aerial Robotics Group Waterloo, CA

Software Team Lead - Autonomy - C, C++, Python *Sept 2022 - Current*

- Led team of 30 to develop **autonomous control software** for drones, achieving **first place** at AEAC 2024 competition
- Architected and software bring-up of multiple projects: Perception-Decision Control Systems, Computer Vision Model, Ground station GUI, Path Planning, Data Telemetry

Projects

I-BERT Transformer on FPGA – RTL Design, System Verilog

- Designed, synthesized, and deployed a BERT model on PYNQ board, performed validation, timing and LUT analysis
- Implemented systolic arrays for efficient matrix multiplication, leveraged data buffering to maximize throughput