

# Peterson Guo

647-606-9486 | petersonguo@gmail.com | petersonguo.com |  PetersonGuo |  PetersonGuo

## EDUCATION

### University of Waterloo

BASc. in Honours Electrical Engineering

Waterloo, ON

09/2023 - 04/2028

## SKILLS

**Languages:** C/C++, Python, MLIR, Bash

**Software:** LLVM, CUDA, ROCm, Pytorch, Git

**Others:** Compilers, GPU Drivers, Linux, Operating Systems, Kernel Debugging, LLMs

## EXPERIENCE

### Systems Software Engineer Intern

Nvidia

05/2025 – 12/2025

Toronto, ON · Santa Clara, CA

- Built and owned an internal **Python-C++** profiler for CUTLASS latency decomposition with up to **21x** lower overhead than torch.profiler and adopted by **15+** systems developers
- Reduced end-to-end **JIT** compilation time by **20%** by designing a deterministic Merkle-tree hashing system **5x** faster than SHA256 for multi-GB binaries, improving kernel generation **latency** and iteration speed
- Reduced kernel launch overhead by **60%** by eliminating **Python→C++→CUDA** dispatch hotspots, achieving performance within **10%** of tvm-ffi and over **2x** faster than Triton
- Core contributor of an **MLIR-based CUDA** dialect, adding **20+** ops and lowering patterns to enable **Python** DSL access to graphs, streams, async memops, and PDL
- Migrated low latency **Blackwell**-optimized kernels to CuTe DSL, enabling dynamic JIT integration and cross-architecture kernel specialization
- Delivered up to **2.67x** faster **LLM inference** at a pre-acquisition startup by integrating custom kernels and improving fusion heuristics of a **Python-based ML Graph**-compiler

### Software Engineer Intern

AMD

09/2024 – 12/2024

Markham, ON

- Delivered AMD's most stable GPU software release, by resolving **25+** kernel-level crashes, hangs, and performance regressions, using crash dumps, **firmware** logs, and **register-level** analysis
- Shaped display-pipeline research, shown by my **ML upscaling** PoC being adopted and later expanded into an internal technical conference paper by my successor, by building a lightweight **PyTorch/ROCM** prototype
- Strengthened display-pipeline stability on next-gen AMD GPUs/APUs by building **C/C++ kernel** drivers, ensuring robust framebuffer→display behavior for Ryzen AI and Radeon hardware
- Improved bring-up reliability by fixing **5+** Microsoft OS-GPU issues, analyzing **firmware**/memory dumps and eliminating early-boot bottlenecks
- Enhanced **Linux** display pipeline correctness, by resolving defects in color calibration, frame-sync logic, DSC decoding, and corruption, contributing patches to AMD's **open-source** Linux display driver

### Security Developer Co-op

eSentire

01/2024 – 5/2024

Waterloo, ON

- Performed **LLM security research** on production backends, analyzing attack surfaces including prompt injection, data exfiltration, and privilege escalation in deployed pipelines
- Cut **ingestion latency** by **50%+** by **optimizing** JSON parsing and Snowflake execution paths, reducing pipeline stalls in **real-time analytics**
- Improved log-processing throughput by **4x** through **algorithmic optimizations** in **Python**'s data pipeline
- Built an AI threat analytics dashboard end-to-end (Snowflake, Python, Vue3) used by enterprise clients to investigate live security events, earning executive visibility
- Strengthened analyst workflows by adding packet-processing and correctness enhancements to an open-source PCAP scrubber

## PROJECTS

### ML Upscaling

- | ROCm, CUDA, Machine Learning, CNNs, PyTorch, Python
- Prototyped a real-time **ML upscaling** model for **AMD graphics cards**, leveraging **transformer** based **super sampling** to enhance visual fidelity and performance for potential display pipeline adoption

### Chess Bot

- | PyTorch, Python, C++, CUDA, Alpha Beta Pruning, NNUE
- Built a **GPU-accelerated** chess engine using custom **CUDA** kernels for evaluation/search, achieving major speedups over CPU baselines

### Bionic Evo

- | C/C++, Assembly, Neural Networks, TensorFlow, CUDA, STM32
- Engineered a humanoid arm prototype for amputees by utilizing **STM32** and **EMG** sensors, integrating **pattern recognition** to achieve **precise gesture classification** and seamless arm control

### InvestIQ

- | Python, LSTMs, IBKR API, CUDA
- Built a multi-source data pipeline supporting fundamentals, and event feeds, enabling feature generation for backtesting, pairs trading, and LSTM-based volatility models