Sonnet Salice

salicesonnet@gmail.com • sonnetsalice.com
Github: theonlysonnet • Linkedin: Sonnet Salice

EDUCATION

SEPTEMBER 2022 – JUNE 2027 (EXPECTED)
UNIVERSITY OF TORONTO
Bachelor's in Applied Science (B.A.S.c) in Computer Engineering + PEY Co-op
Awards: Dean's Honors List 2022-2023
Relevant courses: Algorithms & Data Structures, Applied Fundamental of Deep Learning, Intro to Machine Learning, Operating Systems, Digital Electronics, Computer Networks, Computer Architecture, Programming Fundamentals, Circuit Analysis, Digital Systems

MAY 2021 – JUNE 2022
ERIC HAMBER SECONDARY
Average: 96%

SEPTEMBER 2017 – MARCH 2020
CAMAS HIGH SCHOOL
GPA: 3.99 Unweighted, 4.74 Weighted,
AP Scholar with Distinction, Science Olympiad State Champion, Knowledge Bowl Regional Champion

SKILLS

Programming Languages: C++, Java, Python (NumPy, Matplotlib), ARM/Nios II Assembly
AI/ML: Pytorch, TensorFlow, HuggingFace, JAX, ONNX
Compiler: Torch, XLA, MLIR, StableHLO, Shardy
Hardware/Circuit Design: Verilog/VHDL, ModelSim, FPGAs, SoCs
Version Control: Git
Data science & Analysis: R/RStudio, Matlab, NumPy, Matplotlib
Web Development: JavaScript, CSS, HTML, REACT, Next.js, PostgreSQL, REST API
Languages: English, Japanese, Spanish

EXPERIENCE

AI/ML Compiler Engineering Intern, Tenstorrent -Torornto, ON   **MAY 2025 – PRESENT**
Working on the MLIR based compiler for running Deep learning workloads on custom processors and accelerators.
- Built LLM performance benchmarking infrastructure for tt-xla, the compiler pathway integrating XLA (OpenXLA/Google TPU) with Tenstorrent hardware.
- Created custom test suites to validate SOTA models (Llama 3, Qwen, Stable Diffusion, gpt-oss etc) through the tt-mlir compiler, gaining expertise in various LLM architectures and attention mechanisms.
- Contributed to the development of BFP8 quantization and conversion logic within the compiler to improve hardware utilization and model execution efficiency.
- Optimized graph compilation workflows across PyTorch, JAX, and ONNX by leveraging knowledge of IRs like StableHLO and integration strategies for torch-mlir and OpenXLA.
- Developed test suites for distributed execution, using Shardy for tensor partitioning and SPMD, and info from Hugging Face to optimize how the compiler shards SOTA models across multiple accelerators.
- Validated parallelism strategies like data, tensor, pipeline and expert (MoE) parallelism.
- Deployed to Japan as part of engineering team to establish Model-as-a-Service (MaaS) infra for Unsung Fields, configuring high-end accelerators for cloud-scale AI workloads.

Co-Founder and CTO, DayDream Education - Toronto, ON   **MAY 2024 – JAN 2025**
Ed-tech startup focused on reducing smartphone usage among high school students by rewarding healthy habits.
- Secured selection into the NEST incubator of the Entrepreneurship Hatchery at the University of Toronto as a co-founder.
- Developed technical infra of platform using React Native/Swift and Django/Django REST API with PostgreSQL, leading Agile development of a pilot program for local private schools to track phone usage of students and reward healthier habits.
- Created a business plan through market research, customer discovery, and financial forecasting, and forged partnerships with local businesses to offer real-world incentives for students.

INTERN, Repel Security Systems Ltd. - Vancouver, BC   **MAY 2023 – JULY 2023**
Led the digital transformation of accounting systems to meet business and client needs.
- Digitized the accounting infrastructure for a custom home automation firm by transitioning six months of financial records into a customized QuickBooks system.
- Analyzed ~$1M in financial data to identify historical accounting errors and design a tailored reporting structure.
- Consulted with leadership to audit business processes and provide necessary improvements, ensuring the new digital system aligned with specific client and operational needs.