Felix Zhang

felixfzhang@cs.toronto.edu | ff-zhang | felixfzhang

Education

University of Toronto Sept. 2025 – June 2027
Masters of Science in Computer Science

University of Toronto Sept. 2021 – June 2025
Honours Bachelor of Science in Computer Science; Major in Mathematics *3.96/4.0 cGPA*

Awards: Fields Undergraduate Summer Research Program ($3 800), Louis Savlov Scholarship in Sciences And Humanities ($1 000), Second Malcolm Wallace Scholarship ($5 000), University of Toronto Scholar ($7 500), Dean's List Scholar, B.C. Achievement Scholarship ($1 250), District/Authority Scholarship ($1 250)

Publications

PD3: Prefetching Data with DPUs for Disaggregated Memory May 2026
Sidharth Sankhe, Felix Zhang, Umayrah Chonee, Sherman Lim, Jason Hu, Jialin Li, Qizhen Zhang
23^{rd} USENIX Symposium on Networked Systems Design and Implementation

Research Experience

Research Assistant; Far Data Lab, University of Toronto Sept. 2024 – Aug. 2024
Supervisor: Prof. Qizhen Zhang

- Investigated offloading computations onto SmartNICs and data processing unit (DPUs)
- Parallelized the execution of *Monodepth2* in **Python** and **C++** with the latter achieving linear performance scaling with the number of threads
- Built a DPU-based prefetcher *PD3* with a team of 6 which intercepts network traffic to predict and prefetch data for tiered key-value stores
- Designed an external service for offloading shuffle operations from database management systems which supports either a memory or storage backend managed by DPUs

Research Assistant; University of Toronto March 2024 – Dec. 2024
Supervisor: Prof. Jack Sun

- Worked on a team of 11 to implement a pedagogical kernel *KidneyOS* in **Rust** to be used in an introductory operating systems course with **500+** students annually
- Enabled thread creation and destruction, multi-threading, pre-emptive scheduling within the thread system
- Led a team of 3 to implement POSIX-compatible syscalls and add support running user-space executables

Research Assistant; PRISM Lab, Bloorview Research Institute June 2024 – Aug. 2024
Supervisors: Erica Floreani and Prof. Tom Chau; Funded by: FUSRP

- Curated deep-learning models from the literature on denoising electroencephalogram (EEG) data in a team of 4 and benchmarked them on the *EEGDenoiseNet* dataset
- Investigated the applicability of end-to-end transformer models to denoise EEG signals and the impact of using signals' time-frequency representation as input on model performance

Industry Experience

ML Runtime Engineer; Cerebras Systems May 2024 – Aug. 2025

- Implemented a runtime virtual memory system in **C++** with a team of 3 which pre-emptively loads data before it is accessed, allowing off-chip memory to be used for the first time with only a **10%** performance penalty
- Added support for network storage in the paging system with remote direct memory access, providing **100** GB/s read and write speeds with **10** μs latency to multiple remote servers
- Enabled the ability log and replay the network operations, decreasing the time to recreate stalls and timeouts by over **80%**, and setup unit tests to automatically catch breakages and performance regressions in the network layer
- Improved the throughput of the network layer by **6%** when transferring data by implementing best practices for remote direct memory access and reducing setup overhead
- Determined the cable and port mapping for a one node cluster which will be used for all future deployments

Technical Skills

Languages	C++, Rust, Python, C, C#, Bash, Java
Frameworks	Catch2, PyTest, PyTorch, TensorFlow, OpenCV, scikit-learn, NumPy, Pandas, SciPy, Eigen3
Tools	Git, Linux, Unix, tmux, CMake, Docker, Anaconda, Google Colab, QEMU, Jupyter, WSL, Slurm