Google Scholar
github.com/YueFei0403
yuefei9943@gmail.com
(+1) 647-866-7311

Yue Fei she/her

Machine-Learning & Signal Processing • Parallel Programming • Data Analysis Pipelines

Education

University of Toronto
Master of Engineering, Electrical Engineering — *Sep. 2022 – Jun. 2024*

- Emphasis: Communications
- Advisor: Prof. Raviraj Adve
- MEng Thesis: "Pilot Training – Angle of Arrival & Channel Estimation in 5G Network"

University of Toronto
Bachelor of Applied Science, Electrical Engineering — *Sep. 2017 – Jun. 2022*

- Capstone Project: Convolutional Neural Network (CNN) NPU Overlay (MobileNet V1) for FPGA (Intel Stratix-10 NX 2100)
- Advisors: Prof. Vaughn Betz and Andrew Boutros

Publications (Peer-Reviewed Conference)

1. Arash Ahmadian, Louis S.P. Liu, **Yue Fei**, Konstantinos N. Plataniotis; Mahdi S. Hosseini. Pseudo-Inverted Bottleneck Convolution for DARTS Search Space. *IEEE ICASSP*, 2023.
2. Abnash Bassi, **Yue Fei**, Gilead Posluns, Mark C. Jeffrey. Optimized Priority Scheduling for Faster Scalable Belief Propagation. *AAAI [In Submission]*, 2026.

Awards and Honors

Dean's Honour List *2017 Fall, 2018 Winter, 2018 Fall, 2021 Fall, & 2022 Winter*
Edward S. Rogers Sr. Department Betz Entrance Scholarship ($5,000) *2017*

Certificate

Certificate in Engineering Business *Jun. 2022*

Invited Talks

Panel: *Demystifying Machine Learning* *Mar. 2025*

Talk: From Channels to States — Machine Learning in the Language of Communication and Control
QWomen San Diego — Qualcomm Internal Panel Discussion

Technical Skills

Programming: Python (Pandas, NumPy), C/C++, MATLAB, Julia, Arm Assembly, Verilog/SystemVerilog, Unix/Linux Shell, Perl
ML & Data: PyTorch, scikit-learn, GRU-RNN, Attention/Transformer, Q-Learning, Convex & fractional optimization, feature-engineering
MLOps & Data Eng.: CI/CD (Jenkins), Git, Makefile automation, REST/Flask APIs, Docker (basic), **HPC thread-pinning & NUMA-aware scheduling**
Parallel & Distributed: OpenMP-style multithreading, SIMD optimization, multi-queue schedulers, cache-coherence protocols
Tools/Cloud: Unix/Linux Shell, Vim/GVim, SimpleScalar; *quick to adopt distributed-data frameworks (e.g. Spark/Ray/K8s) given HPC background*

Industry Experience

Qualcomm – Design Verification / ML Engineer *Jun. 2024 – Jul. 2025*
Markham, Canada

- Built **multi-head GRU-RNN pipeline** to optimize receiver gain line-up; reduced manual tuning effort for analog designers and **operationalized ML in hardware workflow**.
- Designed **physics-inspired MLP** predicting VCO capacitance for **1,000+ frequency targets**, eliminating manual capacitor tuning and cutting parameter-search time by > 90%.
- Verified **UWB receiver path**, improving startup latency **20 ns → 2 ns (-90%)**; validated **WLAN CP-PLL loop** across **500+ channel indices** with automated **UVM test plans**.
- Integrated models & regression tests into **CI/CD (Jenkins)** for reproducible nightly runs and hardware-in-the-loop validation.

Alphawave Semi – Digital Verification Engineer *May 2020 – Jun. 2021*
Toronto, Canada

- Automated **SerDes datapath & clocking UVM testbenches**, expanding functional coverage by **50+ scenarios**.
- Enhanced **CI/CD pipelines** to support **15× growth** in regression jobs (4 → 60+ projects), improving release velocity and reliability.

Selected Projects

Highlighted academic, research, and technical projects spanning optimization, signal processing, and computer architecture.

Parallel & HPC
Multi-threaded acceleration and memory-coherence design

Parallel Beamforming & Cache-Coherence for Scalable Compute *Jan. 2024 – May 2024*

- Achieved **7× speed-up (17 s → 2.5 s)** on ultrasound beamforming via **SIMD intrinsics & memory-locality optimization**; validated scaling (1–16 threads) — relevant to **data-parallel training/inference**.

Optimized Priority Scheduling for Belief Propagation *Summer 2024*

- Developed a scalable **Stealing Multi-Queue (SMQ)** scheduler with **lazy priority updates** for asynchronous belief propagation, cutting locking overhead and improving cache locality; achieved up to **1.9× faster** than traditional MQ and **32× geomean speed-up** over sequential BP at 48 threads, while preserving convergence and work-efficiency — relevant to **fraud-graph & transaction-risk ML models**.

Optimization
Modeling and constrained optimization for wireless networks

Convex & Fractional Programming for Multi-Cell MIMO Beamforming *Sep. 2023 – Dec. 2023*

- Applied **fractional-programming & quadratic-transform** to improve constrained sum-rate optimization — transferrable to **portfolio-style resource-allocation** problems.

Coding Theory
Error-correcting codes for reliable digital communications

Graph-Based Error-Correcting Codes *Sep. 2023 – Dec. 2023*

- Implemented **LDPC**, **Fountain/LT**, and **Polar** encoders/decoders over binary erasure channel (BEC), binary symmetric channel (BSC), and binary additive white Gaussian noise channel (BI-AWGN); built custom simulators in Julia and MATLAB to demonstrate **graph-structured, ML-style decoding**.

Computer Architecture
Systems-level C/C++ implementations demonstrating pipeline parallelism & speculation

Computer Architecture Coursework *Sep. 2021 – Dec. 2021*

- Implemented in **C/C++** a **5-stage pipelined CPU** with **hazard detection & forwarding**, a **perceptron-based branch predictor**, **Tomasulo out-of-order execution**[Code], **Bouquet prefetcher**, and an **MSI-directory cache-coherence protocol**.
- Gained hands-on experience with **pipeline parallelism**, **speculative execution**, and **cache-coherence** — concepts that parallel large-scale distributed ML inference & speculative decoding in LLMs.