🎓 Google Scholar

 github.com/YueFei0403

✉ yuefei9943@gmail.com

📞 (+1) 647-866-7311

Yue Fei SHE/HER

Machine-Learning & Hardware-Aware Optimization • Parallel Programming/HPC • Data Pipelines & CI/CD

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 and Channel Estimation in 5G Network"

University of Toronto

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

- Capstone Project: Convolutional neural network NPU Overlay (MobileNetV1) 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 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

2. Abnash Bassi, **Yue Fei**, Gilead Posluns, Mark C. Jeffrey. Optimized Priority Scheduling for Faster Scalable Belief Propagation. *The Association for the Advancement of Artificial Intelligence (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

Demystifying Machine Learning (QWomen San Diego)　　　*Mar. 2025*

Qualcomm Internal Panel Discussion

TECHNICAL SKILLS

Programming: Python (**Pandas** for SQL-style joins, group-by aggregation, filtering to process CSV files), C/C++, MATLAB, Julia, Arm Assembly, Perl

Data Processing & Automation: Built CI/CD (Jenkins) pipelines to organize large IC/IP regression suites, **enabling same-day dashboards vs. 1–2-day manual consolidation**; Makefile-driven build/test flows;

ML / Optimization: PyTorch, GRU-RNN, MLP, Attention Mechanism, Convex Optimization (fractional & quadratic-transform methods), Sampling-based Source Coding

Parallel Programming: Experience with **multithreading** (OpenMP-style parallel loops, SIMD intrinsics, custom thread management) and **multi-queue scheduling** strategies for scalable workload distribution.

Tools & Environments: Git, Linux/Unix shells, Vim/GVim, MurΦ Model Checker, SimpleScalar simulators

INDUSTRY EXPERIENCE	**Qualcomm** ∣ Markham, Canada *Jun. 2024 – Jul. 2025*

INDUSTRY EXPERIENCE

Qualcomm ∣ Markham, Canada *Jun. 2024 – Jul. 2025*

- Verified **UWB** receiver path and improved startup performance by reducing LNA charging delay **90%** (20ns → 2ns)."
- Validated **WLAN CP-PLL synthesizer** loop and built **UVM**-compatible test plans spanning 500+ channel indices, ensuring robust coverage across 2G, 5G, and emerging 5G alternative bands.
- Developed **multi-head GRU-based RNN** for receiver gain line-up optimization, where each head learns one analog block (LNA, GM, TIA, BQ, PGA). Transformed a complex combinatorial tuning problem into a scalable learning-based approach, easing designer effort.
- Built a physics-inspired **MLP** that predicts **VCO** capacitance from control inputs, removing the need for RF/analog designers to manually tune capacitors for 1000+ frequency targets.

Alphawave Semi ∣ Toronto, Canada *May 2020 – Jun. 2021*

- Developed UVM testbenches to verify **SerDes** (clocking, datapath, **SRAM**), expanding functional coverage across **50+ scenarios**.
- Enhanced CI/CD automation to support **15× growth** in regression testing (scaling from **4 to 60+ projects**), improving efficiency and reliability as the company expanded.

SELECTED PROJECTS

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

ML & Optimization
Modeling and optimization for wireless networks and semantic coding

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

- Applied **fractional-programming** and **quadratic-transform** optimization to improve multi-cell **MIMO beamforming**, boosting convergence and power-constrained sum-rate performance.

Sampling-Based Semantic Source Coding (One-Shot Info Theory) *Jan. 2023 – May. 2023*

- Implemented **Poisson functional representation, rejection sampling, importance sampling,** and hybrid Poisson + dithered-quantization for 6G semantic source-coding in MATLAB.

Transformer & Embedding Visualization (Research Assistant) *Jul. 2021 – Sep. 2021*

- Explored **RNN-based Transformer models** and **Attention mechanisms** for NLP tasks.
- Applied **PCA-based embedding visualization**—as used in **GloVe**—to project high-dimensional embeddings into **2D/3D space** using **Python** (**NumPy, Matplotlib**), enabling intuitive inspection of semantic clusters.

Signal Processing
Estimation and detection for modern wireless systems

Angle of Arrival (AoA) & Channel Estimation *Jan. 2023 – May. 2023*

- Implemented **MUSIC, DFT,** and **Matrix-Pencil** eigenvalue methods in MATLAB for AoA estimation under large-scale and Rayleigh-fading channels; **Matrix-Pencil delivered ≈2 dB gain in low-SNR regimes with 16-antenna arrays**, outperforming MUSIC and DFT for highly-correlated signals.

LTE Signal Processing *Jan. 2024 – Apr. 2024*

- Processed captured LTE signals for **time/frequency sync, OFDM demodulation,** and **pilot-power analysis** in MATLAB; resampled 40 MHz front-end data to 30.72 MHz LTE rate and validated **PSS/SSS** detection.

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

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

- Accelerated medical-imaging by **7×** (17 s ⊠ 2.5 s) via **16-thread data-parallel** ultrasound beamforming with SIMD intrinsics and memory optimizations (`restrict`, single-write); validated correctness (RMS error < 1e-16) and scalability (1–16 threads) — a paradigm relevant to **data-parallel LLM training**.
- Designed and verified a **3-hop directory cache-coherence protocol** (MSI/MESI) in Mur⊠; optimized with **Exclusive (E)** state to eliminate bus transactions (**0 MB** overhead vs 80 MB baseline), formalized **7 invariants** (e.g., single-writer ownership), and handled FSM edge-case scenarios for large-scale shared-memory systems.

Coding Theory
Error-correcting codes for reliable communications

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

- Implemented **LDPC**, fountain/LT, and **Polar encoders/decoders** over binary-erasure channel; developed custom simulators in Julia and MATLAB.

Convolutional Codes & Viterbi Decoder   *Jan. 2023 – May. 2023*

- Built a rate-½ **convolutional encoder** and **Viterbi decoder** in MATLAB with custom trellis structures; validated against built-in tools and demonstrated 2-bit-error correction on a noisy BSC channel.

Computer Architecture
Speculation mechanisms relevant to LLM decoding

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

- Developed a **5-stage pipelined CPU** with **hazard detection & forwarding**, implemented a **perceptron-based branch predictor, Tomasulo out-of-order execution, Bouquet prefetcher**, and **MSI-directory cache-coherence protocol**; these experiences deepened my understanding of **pipeline parallelism for distributed LLM training** and how **speculative execution** in CPUs parallels **LLM speculative decoding** for faster inference.