Mohammad Nasirifar

LLM hacker keen on inference and low-level perf.

farnasirim@gmail.com
farnasirim.com
Canadian Citizen residing in SF Bay Area

Career Highlights

Software Engineer, Fireworks AI PyTorch, Cuda/GPU Sep 2024–Now

- LLM serving, performance, robustness, fine-tuning, and hardware compatibility
- Recent: Deepseek MOE NCCL sharding, ring attention, and speculative decoding

Software Engineer, Facebook Ads ML Inference PyTorch, Cuda, C++ Jan 2022–Aug 2024

- Led XFN enablement of first event-based model in Ads, as founding member of ML Inference team in Toronto
- Built an automated search frameworks for model splitting and distributed serving strategies based on min-cost flows, saving 15 engineer years on model iterations in one year
- Various perf improvements in PyTorch graph, Cuda, lowering, and op-fusion levels
- Led post-training model validation and tuning, achieving heterogeneous hardware parity

Software Engineer, Lorica C++, GPU, Number Theory, WASM Oct 2020–Jan 2022

- Led the invention of the first infinite-length IR system in BFV Fully-Homomorphic Encryption scheme (US Patent 20230229801)
- Achieved 60+% improvement over Microsoft SEAL with custom WASM SIMD optimization compiler passes

Visiting Researcher, Massachusetts Institute of Technology C, mTLS Summer 2019

- Optimized assembly code running on a small wearable, squeezing out 10% power efficiency

Systems Researcher, University of Toronto C++, RDMA/Infiniband, Linux 2018–2020

- Built Slope, an RDMA-based system for distributed model serving

Software Engineer, CafeBazaar Inc. TensorFlow, Go, Python, Kubernetes 2015–2018

- Led Ad Quality engineering and ops, reporting to CEO (team of 6 SWE/MLEs)
- Privileged to work on Blacksmith, a bare-metal cluster manager for Kubernetes, while learning from a top-tier Infrastructure team

Education

University of Toronto: M. Sc. in CS	GPA: 4.0	2020
Shahid Beheshti University: B. Sc. in CS	GPA: 18.55/20	2018

Misc

- Featured on isocpp.org for a 2021 article on compile-time precalculations in C++
- Bronze medalist in 2018 National Mathematics Competition
- ACM ICPC Regional contest champion, 2017 World Finalist (humbly placing 56th)