Farshid Varno

Intro

Seasoned AI/Hardware Co-Design Engineer with 15+ years of experience in hardware-aware software development, hardware design, and AI research. I leverage expertise in systems programming, computer architecture, and system simulation to design and optimize solutions for large-scale machine learning infrastructure. My track record includes patented AI accelerator architectures, impactful research in distributed machine learning, development of FPGA-based telecom and encryption modules, and optimizing hardware-software integration for AI systems.

Experience

Research & Industry

Research Scientist (AI/HW co-design), Rain

August 2023–present, San Francisco, USA

- Driving the architecture of LUT-based non-linear functions.
- Leading the design of a scalable C++ numerics library to support arithmetic operations and quantization for OCP Microscaling formats.
- Conducting AI/HW co-optimization with mixed-precision quantization & RISC-V custom instructions.
- Integrated custom operations into PyTorch's FX/TorchDynamo tracing pipeline.
- Designing high-performance Compute-in-Memory (CIM) hardware units with SystemC and QEMU, alongside developing high-level scalable simulation systems and AI performance models for quantization and attention mechanisms.
- Led the design of a sparse VMM unit and an online FP Softmax unit with Base-2 conversion.
- Led on-device training and domain adaptation efforts.
- Inventor on four patents for Softmax HW and core CIM architecture, three of which as the lead inventor.

Research Scientist, Imagia

May 2018–March 2022 (internship period included), Montreal, Canada

- Led research on Federated Learning optimization, achieving SoTA performance via drift elimination.
- Researched Transfer Learning, out-of-distribution generalization, Meta-Learning, and Few-Shot Learning.
- Contributed to the design of an AI library to enhance Imagia's research efforts.
- Published a patent on Transfer Learning as the lead inventor.
- Scaled and optimized AI training workloads by porting and configuring the Polyaxon MLOps platform on distributed DGX GPU clusters, enhancing resource efficiency and researcher productivity.

Research Assistant, Institute for Big Data Analytics

May 2017–May 2018, Halifax, Canada

- Engineered and optimized a custom CUDA kernel for high-throughput streaming data processing, demonstrating expertise in accelerator programming and low-level performance tuning led to ~1000x speedup compared to the original CPU-based approach.
- Developed a CNN framework for detecting aircraft corrosion with D-Sight technology (DAIS).
- Researched sparsity, activation functions, and normalization for efficient machine learning models.
- Collaborated with Harvard University to research human behavior prediction from fMRI data.

Data Scientist (part-time), Cognitive Health and Recovery Research Lab

Mar 2020–Jun 2020, Halifax, Canada

- Integrated and visualized clinical data to support cognitive health research.
- Investigated post-operative cognitive dysfunction in elderly patients through data analysis.
- Analyzed surgical time series data (e.g., anesthesia depth, patient vitals) to identify patterns and insights.

FPGA Engineer, Kara Telephone Co.

Jun 2013–Jun 2014, Tehran, Iran

- Designed, implemented, and integrated TDM switches on FPGAs, supporting up to 16k x 16k channels.
- Developed a multi-channel I2C master controller for 16 modules with error checking and correction.
- Designed and implemented SPI and USART peripheral interfaces, ensuring seamless system integration.
- Worked with embedded processors and RTOS, optimizing hardware-software interaction.
- Led speed optimization efforts for FPGA designs on Altera Cyclone series.

RTL Designer Intern, SarvNet Telecommunication Inc.

March 2012–Sep 2012, Isfahan, Iran

- Designed and implemented AES modules for encryption in STM4 lines, ensuring efficient performance.
- Developed resource-sharing mechanisms to support both 128-key and 256-key AES modes, adapting dynamically based on the selected encryption mode.
- Area optimization for FPGA designs, targeting Xilinx Virtex 4 and 6 series to minimize resource usage.

Teaching

- Adjunct Professor, Computer Architecture, Chehelsotoon Institute for Higher Education, Fall 2015
- Adjunct Professor, System Programming, Chehelsotoon Institute for Higher Education, Fall 2015
- Co-instructor, Machine Learning for Big Data (CSCI-6515), Dalhousie University, Fall 2020
- Teaching Assistant, Machine Learning for Big Data (CSCI-6515), Dalhousie University, Fall 2018
- Teaching Assistant, Digital Circuits (ECED-2200), Dalhousie University, Winter 2016
- Teaching Assistant, System Analysis (ECED-3401), Dalhousie University, Fall 2016
- Teaching Assistant, Java Programming, University of Guilan, Winter 2009
- Teaching Assistant, Algorithms, University of Guilan, Winter 2010

Background

Education

- Ph.D., Computer Science. Dalhousie University. 2016–2023, CGPA: 4.19
- M.Sc., Computer Architecture. University of Isfahan. 2012–2015, CGPA: 4.02
- B.Sc., Comuter Engineering, Guilan University. 2008–2012.

Skills

Programming Languages: Python, C++, CUDA

Systems Programming & Compiler Technologies:

- **Proficient In:** Andes RISC-V, PyTorch Dynamo Integration, Low-Level Performance Optimization
- **Actively Learning:** Rust, MLIR, LLVM, Triton

AI & Machine Learning: Distributed Machine Learning (Federated Learning), Transfer Learning, Quantization & Compression, On-Device Training, LLM & Attention Architecture Performance Modeling

ML Infrastructure & HPC: System Simulation (SystemC, QEMU), AI Performance Modeling

Hardware Design & Verification: VHDL, Verilog, SystemC

Developer & MLOps Tools: Git, GitHub Actions, Bazel, Polyaxon, MLflow

Selected Publications

Papers

- Varno, Farshid, Marzie Saghayi, Laya Rafiee, Sharut Gupta, Stan Matwin, and Mohammad Havaei. "Minimizing Client Drift in Federated Learning via Adaptive Bias Estimation." *European Conference on Computer Vision.* – **ECCV** (2022).
- Varno, Farshid, Lucas May Petry, Lisa Di Jorio, and Stan Matwin. "Learn Faster and Forget Slower via Fast and Stable Task Adaptation." *arXiv preprint arXiv:2007.01388* (2020).
- Varno, Farshid, Behrouz Haji Soleimani, Marzie Saghayi, Lisa Di Jorio, and Stan Matwin. "Efficient neural task adaptation by maximum entropy initialization." *arXiv preprint arXiv:1905.10698* (2019).
- Jiang, Xiang, Mohammad Havaei, Farshid Varno, Gabriel Chartrand, Nicolas Chapados, and Stan Matwin. "Learning to learn with conditional class dependencies." In *international conference on learning representations.* – **ICLR** (2018).
- Saghayi, Marzie, Jonathan Greenberg, Christopher O'Grady, Farshid Varno, Muhammad Ali Hashmi, Bethany Bracken, Stan Matwin, Sara W. Lazar, and Javeria Ali Hashmi. "Brain network topology predicts participant adherence to mental training programs." ***Network Neuroscience*** 4, no. 3 (2020): 528-555.

Patent

- Varno, Farsheed, Behrouz Haji Soleimani, Marzie Saghayi, Lisa Di Jorio, and Stan Matwin. Method and system for initializing a neural network. https://patents.google.com/patent/WO2020225772A1 (2020).
- Four provisional patent applications filed on the following topics:
  - Softmax Hardware with Base-2 transformation (lead inventor).
  - Block quantization and online Softmax resource sharing (lead inventor).
  - Efficient sequential VMM unit with bit-level sparsity (lead inventor).
  - A new compute-tile topology with close-to-logic HBM stack (co-inventor).

Recognition

- **Vice-president of Public Relations**, Toastmasters International, Dal Toastmasters, 2020.
- **Mitacs Accelerate Award**, 56k CAD, 2021–2022.
- **Scotia Scholar Award**, 45k CAD, Research Nova Scotia, 2019–2021.
- **Best Graduate Student Research Award**, Big Data Congress, Sep 2017.
- Selected **Conference Program Committee Member** for ICLR (2020), KDD (2017), and Confoo (2023).
- **Reviewer** for top AI conferences: CVPR 2023/24/25, ECCV 2022/24, ICCV 2023, FedVision 2023.
- **1st Rank Student Recognition**, University of Isfahan, Mar 2015.
- Mentored post-secondary students through various programs and occasions (e.g., AI4ALL 2024).

This resume is compiled using LaTeX and is also available in HTML format at https://farshid.varnio.com/resume/farshid_cv/index.html.