Zhaoyang Han
http://www.coe.neu.edu/∼zhhan
Email : zhhan@coe.neu.edu
Mobile : +1-267-872-8272

Research Interests

I am interested in research projects related to computer architecture and computing accelerations, particularly in networking and machine learning models. I have experience with FPGAs and have completed a project enabling FPGAs in public clouds such as CloudLab and FABRIC. Currently, I am working on utilizing programmable hardware like FPGAs and SmartNICs for task offloading, particularly for network functions and Large Language Model tasks, to alleviate the burden on host CPUs/GPUs. Additionally, I am proficient in multiple languages including C/C++, Python, HDL, P4, and Bash/Tcl.

Education

Northeastern University — Boston, MA
PhD. in Computer Engineering; — *Aug. 2019 – Jun. 2025 (est.)*
Advisor: Miriam Leeser

University of Pennsylvania — Philadelphia, PA
Master of Science in Electrical Engineering; — *Aug. 2017 – Jun. 2019*
Advisor: Andre DeHon

The University of Manchester — Manchester, UK
Bachelor of Engineering(Hons) in Electrical Engineering; First Class; — *Aug. 2015 – Jun. 2017*

Jilin University — Changchun, China
Bachelor of Engineering in Electrical Information Engineering; — *Aug. 2013 – Jun. 2015*

Experience

Northeastern University — Boston, MA
Research Assistant at Reconfigurable Computing Lab — *Sep 2019 – Present*

- Research on RDMA-based distributed FPGA clustering and virtualization.
- Reserach on P4-based network security functions.
- Research on in-network computing, including in-network telemetry, sketches and machine learning tasks on FPGA-based SmartNIC and NVIDIA Bluefield SmartNIC through DPDK/DOCA.
- Research on Accelerating LLM inference on High-Bandwidth Memory based FPGA with quantization techniques.
- Main developer of the Open-Cloud Testbed in Cloudlab. Research on P4-FPGA toolchain for the Open Cloud Testbed on the public cloud Cloudlab and FABRIC. Supporting users from multiple institutes in conducting their own reasearches.

Mathworks — Boston, MA
Software Engineer Intern at HDL Verifier Team — *Jun 2020 – Aug 2020, Jun 2021 – Aug 2021*

- Improved HDL cosimulation between Matlab Simulink and EDA simulators through Matlab C++ backend APIs. Created auto-fill functionality for parsing HDL designs automatically.
- Implemented AXI Bus Function Model (BFM) in SystemVerilog and linked to Matlab Simulink for co-simulation.

University of Pennsylvania — Philadelphia, CA
Research Assistant at Implementation of Computation Group — *Jun 2018 - Jun 2019*

- Implemented in-network computing functions (key-value cache and header compression) on FPGA with 10Gb/s throughput on Xilinx ZCU102.
- Research in reducing FPGA compilation time on SoC like Zynq.

Programming Skills

- **Languages**: C++, C, Verilog, SystemVerilog, VitisHLS, P4, Python, Bash
- **Tools/Frameworks**: AMD Vivado/Vitis, DPDK, Mininet, Modelsim, CMake, Make, PyTorch, Git, LaTeX

Preprint

- **(Poster)** Xuan Shen, Zhaoyang Han, et al.: FOTA-Quant: FGPA-Oriented Token Adaptive Quantization Framework for the Acceleration of LLMs (Design Automation Conference 2024 DAC'24)
- Xuan Shen, Zhaoyang Han, et al.: EdgeQAT: Entropy and Distribution Guided Quantization-Aware Training for the Acceleration of Lightweight LLMs on the Edge

Publications

- Zhaoyang Han, Michael Zink, Miriam Leeser: Memory-efficient Sketch Acceleration for Handling Large Network Flows on FPGAs (FPT'24)
- **(Innovation Award)** Zhaoyang Han, Andrew Briasco-Stewart, Michael Zink; Miriam Leeser: Extracting TCPIP Headers at High Speed for the Anonymized Network Traffic Graph Challenge (IEEE HPEC'2024)
- **(Best Paper)** Zhaoyang Han, Suranga Handagala; Kalyani Patle; Michael Zink; Miriam Leeser: A Framework to Enable Runtime Programmable P4-enabled FPGAs in the Open Cloud Testbed (IEEE INFOCOM WKSHPS'2023)
- Xuan Shen, Zhaoyang Han, et al.: HotaQ: Hardware Oriented Token Adaptive Quantization for Large Language Models (IEEE TCAD)
- Zhaoyang Han, Yiyue Jiang, Mushini Rahul, John Dooley, Miriam Leeser: Hardware Software Codesign of Applications on the Edge: Accelerating Digital PreDistortion for Wireless Communications (IEEE HPEC'2022)
- Zhaoyang Han, Meabh Loughman, Yiyue Jiang, Mushini Rahul, John Dooley, Miriam Leeser: Computationally Efficient Look-up-Tables for Behavioral Modelling and Digital Pre-distortion of Multi-Standard Wireless System (EAI CROWNCOM'21)
- Nik Sultana, John Sonchack, Hans Giesen, Isaac Pedisich, Zhaoyang Han, André DeHon, Boon Thau Loo: Flightplan: Dataplane Disaggregation and Placement for P4 Programs (NSDI'21)
- Yuanlong Xiao, Dongjoon Park, Andrew Butt, Hans Giesen, Zhaoyang Han, Rui Ding, Raphael Rubin, André DeHon: Reducing FPGA Compile Time with Separate Compilation for FPGA building blocks (FPT'19)
- Miriam Leeser, Suranga Handagala, Dana Diaconu, Zhaoyang Han, Michael Zink: Realizing Network-Attached FPGAs in the Cloud (Computing in Science & Engineering (CiSE, 2024))
- Sandeep Bal, Zhaoyang Han, et al.: P4-based In-Network Telemetry for FPGAs in the Open Cloud Testbed and FABRIC (IEEE INFOCOM WKSHPS'2024)

Teaching Activity

- **Fall 2019, UPenn**: Teaching Assistant of ESE532: System-on-a-Chip Architecture

Honors & Awards

- **Sep 2023**: Travel Grant for KNIT7: A FABRIC Community Workshop, Columbus OH
- **April 2022**: Travel Grant for KNIT6: A FABRIC Community Workshop, Austin, Texas
- **Fall 2015, University of Manchester**: International Excellence Undergraduate Scholarships
- **Summer 2015, Jilin University**: Second-class Scholarship for Outstanding Students
- **Summer 2014, Jilin University**: Second-class Scholarship for Outstanding Students

Invited Talks

- **P4-based FPGA Framework**: Talk at **KNIT7**: A FABRIC Community Workshop, Columbus OH, Sep 28, 2023