Bakhtiar Zadeh

bakhtiar.mzd@gmail.com 07780116124

A motivated PhD student at Imperial College London. Knowledgeable in multiple programming languages with a background in machine learning, low-level systems, FPGAs, and computer architecture.

EDUCATION

Imperial College London, London — *PhD Student, Machine Learning / Computer Architecture*

In Progress

Conducting research in efficient machine learning inference within the Custom Computing Group at Imperial College London. Exploring novel architectures and software–hardware co-design methods for high-performance inference systems.

Imperial College London — *MEng Electronic and Information Engineering*

September 2020 - July 2024

Graduated with First Class Honours / 4.0 GPA. Focused on machine learning, financial signal processing, and high-performance computing (FPGA, C++). Master's Thesis: *Accelerating Latent Diffusion Model Inference using Novel FPGA Architectures.*

EXPERIENCE

Bloomberg, London — C++ *Software Engineer*

July 2024 - April 2026

Developed high-performance market data feed handlers for real-time exchange data. Key projects involved onboarding new exchanges, and employing asynchronous programming paradigms to improve performance. Developed my C++ knowledge into more complicated topics such as template metaprogramming and compile time programming.

JP Morgan, London — C++ *Software Engineer*

April 2023 - October 2023

Built low-latency software features for fixed-income market-making systems, focusing on network performance and code quality. Designed and implemented a single-threaded HTTP server and supporting library to handle high-volume pricing requests.

Imagination Technologies, London— Hardware Engineer

July 2022 - October 2022

Contributed to GPU hardware verification using Python and SystemVerilog. Optimised testbench performance, reducing simulation times and improving verification efficiency

SKILLS

C++, FPGA, Machine Learning , Python, GPU, Generative AI, SystemVerilog, Computer Architecture, Heterogeneous Systems

PROJECTS

Automatic allocation of heterogeneous resources for efficient neural network inference

- This project develops methods to allocate heterogeneous resources automatically from a model definition.
- By taking advantage of various heterogeneous resources, larger networks can have higher performance with limited hardware resources.

Intelligent Constant Matrix-Vector Multiplication resource reuse

- This project aims to introduce scalable and automatic parallelism at the software level to neural networks by implementing Constant Matrix Vector Multiplication (CMVM) reuse.
- The work aims to decouple network definitions from available hardware resources.
- In combination with previous work, this is aimed to be used in CERN for high energy physics experiments.

High Granularity Quantization (HGQ) methods for networks with mixed compute strategies

- Collaborated with co-authors to develop a framework which performs Quantization Aware Training (QAT).
- Allows QAT to be done on networks with mixed compute strategies (Look-Up based and arithmetic based).
- Work will be used at CERN for high energy physics experiments.

Latent diffusion model accelerator, hardware and software

- Developed a highly parameterisable dataflow accelerator for latent diffusion model inference.
- Implemented layer wise Post-Training-Quantisation with measures of generative quality against hardware usage
- Results outperformed various graphics cards and matched a RTX4080 on a single U250 FPGA.