STATEMENT OF PURPOSE

Xingyu Wang (Tom)
BASC in Computer Engineering
University of British Columbia

+1 604-388-5164
tomxingyuwang@gmail.com
linkedin-profile

This Staement of Purpose is for the application to the graduate program in department of Electrical and Computer Engineering at University of British Columbia.

Interests of study

So far, the charm of Electrical and Computer Engineering in my opinion is how I can make computations faster and cheaper, whether it is through software optimization or hardware resources. Nowadays as the demand of computations grows since the evolution of AI, the need of faster and cheaper computations is more urgent than ever.

I am particularly interested in the field of computer architecture especially in accelerating ML workloads. So far, I have some experience with memory management in CPUs and GPUs which still attracts me, but also, I want to explore in interconnects, FPGA-based specialized hardware accelerators, etc.

Currently I have two plans in my mind:

- **Plan A:** I am interested in investigating the potential of FPGA-based accelerators for ML tasks (Let's break GPU Monopoly!). This involves understanding the design and implementation of custom hardware solutions to speed up inference or training processes for specific workloads. I think this is possible with FPGA if we know how to effectively program and utilize their parallel processing capabilities. After finding a few promising approaches on different workloads, I want to implement and seek the possibility to generalize these approaches for broader applications. (For each specific feature of the workload, pack up a block of FPGA IP made for it, and then combine these blocks to form a more complex accelerator for a broader range of workloads. HLS or FPGA IP programming in a sense?)
- **Plan B:** I want to dive deeper into memory management and scheduling techniques, especially in the context of ML workloads. This includes exploring advanced memory strategies for CPUs, GPUs or even NPUs. Nowadays, SOC gets so complex that efficient memory management is crucial for performance. I want to investigate how to optimize memory access patterns and improve data locality to enhance the overall efficiency of ML workloads.

Why University of British Columbia?

I have been in UBC for 4 years and I still love the school and the courses offered here.

I am familiar with the campus, the professors, the courses, and the research projects. I have been involved in a research project with Prof. Alexandra Fedorova for the past year and I would like to continue the research with the professors that I know, and the professors that know me.

I have already taken few grad level course in ECE and I really enjoyed the courses. I would like to take more grad level courses in the future.

The research experience I had with Prof. Alexandra Fedorova has been great. I have learned a lot from the project and I would like to continue the research in the future.

Motivation

When I first entered the university, I was not exposed of any details of how computations are done,

My Background

Academic Background

I took various courses in computer engineering, electrical engineering and computer science at UBC.

- For hardware, I took courses like Digital Logic Design, Computer Architecture, accelerator design and more.

- For software, I took courses like Operating Systems, Machine Learning and more.

All the above courses have given me a solid foundation in both hardware and software, which I believe is essential for the graduate study in Electrical and Computer Engineering at University of British Columbia.

Work Experience

ML Prefetching for Page Faults: May 2024 – April 2025

I worked as a research assistant under Prof. Alexandra Fedorova after receiving an NSERC award, focusing on ML-based cache and page prefetching. My main task was to collect memory traces and adapt an LSTM model for page fault prediction, implemented in PyTorch and trained on a UBC GPU server. Results showed cache models perform poorly on page faults, highlighting distinct access patterns. We began exploring transformer-based models, but I paused the project for a full-time internship.

vLLM Memory Management: Jan 2024 – April 2025

I also worked on memory management for vLLM, a framework for efficient large language model inference in my **Advanced Computer Architecture** course. Initially, my goal was to optimize GPU memory usage with the techniques I learned in ML Prefetching for Page Faults since the idea of vLLM is to manage GPU memory with virtual-memory-like techniques. However, I found that for GPU workloads, the memory access patterns are close to deterministic which is quite different from OS-level memory management.

FPGA Soft IP Engineer: May 2025 – Present

As an FPGA Soft IP Engineer in **Altera**, I am currently working on developing soft IP cores for FPGA platforms, more specifically focusing on the subsystems of HBM (High Bandwidth Memory). This role involves designing and optimizing hardware components for specific applications, leveraging my background in computer architecture and digital design. I am gaining hands-on experience with hardware description languages and FPGA development tools, further bridging the gap between software and hardware design.