STATEMENT OF PURPOSE

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

This Statement of Purpose is for the application to the graduate program in the department of Electrical and Computer Engineering at the University of Toronto.

Interests of Study

So far, the charm of studying in computers 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 for computations grows since the evolution of AI, the need for faster and cheaper computations is more urgent than ever.

My primary interest lies in **computer architecture**, with a particular focus on **accelerating ML workloads**. As models become more complex—moving from simple classifiers to deep neural networks and transformers—the computational requirements grow rapidly, demanding efficient processing of large-scale data and high-dimensional tensors.

Through my research experiences, I have come to appreciate the intricate relationship between ML model training/inference and the underlying hardware. Training deep neural networks involves massive parallel computations, frequent memory accesses, and complex data movement between different memory hierarchies. The efficiency of these operations is tightly coupled with hardware characteristics such as memory bandwidth, cache organization, and the availability of specialized compute units (e.g., GPUs, TPUs, or FPGAs).

For example, as a research assistant under Prof. Alexandra Fedorova, I worked on ML-based cache and page prefetching, where I used various models (LSTM, MLP, transformer, etc.) to predict memory access patterns and reduce latency. This project required analyzing memory traces, experimenting with various neural network architectures, and tuning hyperparameters to improve prediction accuracy. It gave me valuable insights into the challenges of applying ML to system-level problems and the importance of tailoring solutions to specific workloads.

In another project, I focused on optimizing GPU memory usage for large language models (vLLM). I explored how GPU memory access patterns differ from traditional OS-level management and applied techniques such as memory prefetching and scheduling to reduce latency during model inference. By experimenting with different scheduling algorithms, I was able to improve memory allocation efficiency and overall system performance. These experiences deepened my understanding of memory management and scheduling in the context of ML workloads and sparked my interest in further exploring this area.

From these projects, I saw that hardware limits often become the bottleneck for ML workloads. Improving memory access, reducing latency, and designing efficient data paths can make a real difference in performance. I want to work on practical solutions—like better memory scheduling or custom accelerators—that directly improve how ML models run. My goal is to build systems where hardware and software work together for faster, more efficient ML.

Inspired by Prof. Andrew Boutros's PhD thesis, I have identified several specific areas I would like to explore during my graduate studies:

1. I am interested in investigating the potential of **Customized accelerators for specific ML tasks** (Let's break the GPU Monopoly!). This involves understanding the design and implementation of custom hardware solutions to speed up inference or training processes for specific workloads. The prototype will start with FPGAs. FPGAs are powerful if we know how to effectively program and utilize their parallel processing capabilities and adaptability. Depending on the results, I would like to explore the possibility of transitioning to **ASIC design** or move on to **board-level reconfigurable computing**.

2. With the experience I have in memory management, I want to explore how to optimize memory access patterns and data locality for ML workloads on software-programmable customized accelerators. This includes techniques like customized-compiler, scheduling, interconnects, off-chip communication, and memory hierarchy design that is specifically tailored for specific ML models or tasks.

3. At the same time, I also want to explore flattening the model directly onto the hardware circuit, bypassing the traditional software stack. This could involve using **Domain Specific High-Level Synthesis (HLS)** tools to convert high-level descriptions of ML models directly into hardware implementations. (Accelerators which require complex software stacks like TPUs seem like another form of ... GPU to me.) 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 IP made for it, and then combine these blocks to form a more complex accelerator for a broader range of workloads. Modular design?)

While I am excited about hardware acceleration—such as designing customized accelerators and exploring FPGA-based solutions—I am equally fascinated by **software-level optimizations**, **CUDA acceleration** and **parallel computing**. Leveraging CUDA for GPU programming has shown me how efficient parallel algorithms and memory management can dramatically improve ML inference and training speed. I am eager to further explore how **hardware and software co-design**. Despite lack of experience in this area, I started learning CUDA and parallel computing on my own time after work, and I am excited to deepen my knowledge in this area during my graduate studies.

Why Graduate Study?

I have been working as an FPGA Soft IP Engineer in Altera for a few months now, and I have realized that the work I do in the industry is quite different from what I expect. I want to be able to develop new tools and technologies that can make a difference in the industry and society. I want to be able to work on innovative projects that can push the boundaries of what is possible.

I believe that graduate study will provide me with the opportunity to deepen my understanding of the field and to develop the skills necessary to conduct research and contribute to the advancement of knowledge. I want to be able to work on cutting-edge research projects and to collaborate with experts in the field. I believe that this will help me to achieve my long-term goal of contributing to the advancement of technology.

The faculty members such as Prof. Mark Jeffrey, Prof. Natalie Jerger and Prof. Vaughn Betz at the University of Toronto are leading experts in their respective fields, and I am excited about the opportunity to learn from them and to work with them on research projects.