STATEMENT OF PURPOSE

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

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

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

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 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**. While working on ML-based page-level prefetching under Prof. Alexandra Fedorova, I explored using LSTM models for page fault prediction, which gave me valuable insights into how ML can improve system performance. Additionally, I had the chance to explore on vLLM memory management and scheduling. It broadened my sight on the intricate relationship between hardware and software in optimizing ML workloads and the current challenges in this area. These projects have deepened my curiosity about efficient ML acceleration and motivated me to further explore related areas such as **interconnects** and **FPGA-based specialized hardware accelerators**.

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.

2. 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**.

3. 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 these customized accelerators. This includes techniques like scheduling, interconnects, off-chip communication, and memory hierarchy design that is specifically tailored for specific ML models or tasks.

4. 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.)

5. 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 at 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. I have read several papers published by the faculty members that I selected as my potential advisors, and I am impressed by their work. I have sent emails to a few of them listing my thoughts on their papers and expressing my interest in working with them. The email address is fortily@student.ubc.ca if you missed it. I am also impressed by the resources and facilities available at University of Toronto, which I believe will provide me with the tools necessary to succeed in my studies.

Why Me?

I believe I am a strong candidate for the graduate program in Electrical and Computer Engineering at University of Toronto for several reasons:

- **Strong Academic Background:** I have a solid foundation in computer engineering, with a focus on both hardware and software aspects. My coursework has provided me with a comprehensive understanding of the field (**Computer Architecture**, **Digital Logic Design**, **Operating Systems**, **ML**), and I have consistently performed well academically.

- **Research Experience:** I have gained substantial research experience through two significant projects.

  - As a research assistant under Prof. Alexandra Fedorova, I worked on **ML-based cache and page prefetching**, where I tried to predict the memory access patterns of applications to reduce latency. I started with training an LSTM model on collected memory traces of specific programs, to predict future memory accesses. While the LSTM model showed potential in relatively simple scenarios, it struggled with more complex patterns. To address this, I experimented with various architectures (Attention layer, MLP and transformer) and hyperparameters, ultimately improving the model's performance. This project provided me with valuable insights into the challenges of applying ML to system-level problems and how to customize models for specific tasks.

  - In another project focused on **vLLM memory management**, I explored optimizing GPU memory usage for large language models, discovering the unique characteristics of GPU memory access patterns compared to traditional OS-level management. This project is motivated by the nature of vLLM, where vLLM manages the memory of GPUs in blocks as virtual memory, which encouraged me to apply OS techniques learned from my previous project to optimize memory allocation and scheduling. The initial tackle was to prefetch memory pages before they are needed, reducing latency during model inference. However, different from traditional OS, ML workloads do not have the same level of randomness in memory access patterns, making it unnecessary to predict which pages to prefetch. While working on this project, I found that the scheduling of memory requests led to some additional memory swap-in-swap-out overhead, which is not ideal for latency-sensitive applications. To address this, I explored various scheduling algorithms to optimize the order of memory requests, ultimately improving the overall performance of the system. The report can be found on my GitHub This project has deepened my understanding of memory management and scheduling in the context of ML workloads and has sparked my interest in further exploring this area.

  - These experiences have not only enhanced my **technical skills** but also taught me how to approach **complex problems** and work collaboratively in a **research setting**.

- **Relevant Work Experience:** My current role as an FPGA Soft IP Engineer has given me practical experience in **hardware design** and **software stack for custom hardware**. I worked on large-scale traffic generator IP for testing efficiency of memory operations on the on-chip memory such as HBM and DDRRAM. I have gained knowledge with NoC interfaces, AXI protocol, and FPGA toolchains. This industry experience complements my academic background and research skills, making me well-prepared for the challenges of graduate study.

- **Passion for Learning and Innovation:** I am deeply passionate about the field of computer engineering and am eager to explore new ideas and technologies. I am motivated to contribute to the **advancement of knowledge** and to make a meaningful impact in the field.

- **Clear Research Interests:** I have a clear vision of my research interests, particularly in the areas of **computer architecture** and **ML acceleration**. This focus will allow me to make significant contributions to the field during my graduate studies.