First Essay

MohammadArman Soleimani, MIT EECS
soleimaniarman98@gmail.com

Introduction

The end of scaling laws is ushering in an exciting era in which computer architects must unleash their creativity to facilitate performance gains and energy savings. Building on my undergraduate research background, I am applying to MIT to pursue avant-garde architectures and circuits that can push the envelope of computing.

I aspire to study the employment of modern architectural paradigms such as processing-in-memory and neuromorphic computing to accelerate ML workloads and improve their energy efficiency. Groundbreaking ideas in hardware design can also lay the groundwork for breakthroughs in AI, as many advances in AI were made by virtue of cutting-edge computer architectures.

Background

Processing-in-Memory (PIM). I embarked on my research journey in PIM during my fifth semester under Prof. Rohbani and Prof. Sarbazi-Azad. Many designs enable bit-wise computation using DRAM; However, our study identified resilience against Process Variation (PV) as a critical concern, particularly for analog techniques. We presented 'PIPF-DRAM' (DAC'22) to address this concern and reduce energy dissipation. In this research, I contributed through various simulations and experiments. Notably, I ran Monte Carlo simulations to mimic the effect of PV and reduced the latency of XOR computation by optimizing the sequence of operations.

After our proposal was accepted, I was overjoyed at the feeling of creating value. With my maturing skills, I devised a novel circuit to enhance bit-wise computation speed and facilitate hyper-dimensional computing, a bio-inspired AI approach. This design can classify samples within DRAM sub-arrays and is currently under review for ISCA'24, which is my first time as a first author. Through this experience, I learned to identify gaps and make progress toward a solution independently; this is one of my critical assets in pursuing a Ph.D. Moreover, I plan to extend this line of research by devising architectures for more complex AI within DRAM, such as entirely in-DRAM regression.

Our endeavors also produced 'CoolDRAM' (ISLPED'23), a new DRAM structure that eliminates pre-charging to reduce energy dissipation, which received the best paper award in its respective track. I contributed to CoolDRAM by experimenting with our design and assessing it in terms of latency, energy dissipation, PV resilience, and noise immunity.

We decided to employ our experience by improving PIM for SRAM as well. Many SRAM PIM designs eliminate double-ended sensing, which deteriorates robustness and speed. We addressed this through a proposal currently under review for DAC'24. I proposed the usage of dual-word-line 6T SRAM and identified the need for a novel sense amplifier. Additionally, I ran the necessary simulations to assess our design and the state-of-the-art proposals. This proposal has gone through multiple submissions. While initially heart-breaking, I learned to embrace this as an inseparable part of scientific progress. Consequently, I fostered my critical analysis skills and became the first person to identify the shortcomings of my ideas. Moreover, I developed my tenacity in research - after each rejection, I asked myself how to strengthen our proposal.

My collaboration also included helping a Ph.D. student with her Networks-On-Chip research, 'OCRA' (NocArc'23). I analyzed OCRA's hardware overhead and prepared the paper and presentation.

FPGA Routing and Reinforcement Learning (RL). In parallel with our PIM research, I was humbled to get accepted for Summer@EPFL 2022, a selective summer internship at EPFL. Supervised by Prof. Stojilovic, I studied the employment of RL for FPGA routing. After analyzing state-of-the-art proposals, we concluded that complex objectives and the vast problem space of FPGA routing make it a challenging task for RL models, which failed to generalize. My internship did not end with a manuscript, but our research pruned future studies. This internship taught me the value of trimming proposals to pursue a solution.

Real-Time Scheduling and RL. I built on my EPFL experience earlier this year by joining a project at Sharif. Our goal was to leverage RL to allocate and schedule real-time tasks in a way that reduces energy consumption. Supervised by Prof. Ansari, I implemented and developed an algorithm to realize these goals. The outcome of this research is currently under review for the IEEE IoT Journal. In hindsight, a key catalyst in this research was my experience at EPFL, which shows how seemingly irrelevant experiences can become crucial assets later on.

Machine Learning. To propose architectures for ML, one must develop an understanding of ML. I engaged with ML in tandem with my research experiences. This semester, I joined Zista Gene Afarin as an intern to study efficient super-resolution imaging for biomedical samples using convolutional neural networks. Before this, I had learned about ML through online courses provided by Inria and Edge Impulse and our department's course on AI. These experiences and my research studies are vital assets in my pursuit of architectures for ML.

The Path Ahead

If given the extraordinary opportunity to pursue a Ph.D. at MIT, I aspire to leverage the resulting experience to engage in industrial research and turn research studies into real-life products. Multiple top-notch groups are involved in research in areas of my interest, which makes MIT a desirable destination for me.

Prof. Vivienne Sze has led influential research on designing circuits and architectures to accelerate emerging applications and reduce their energy dissipation, which strongly resonates with my interests. Her recent research in accelerator architectures at the Emze group includes studies on PIM, which comprises a key aspect of my background.

Similarly, I was fascinated by the recent contributions made by Prof. Daniel Sanchez in the field of accelerators. His previous studies also include data-centric computing, which aligns with my background in PIM.

My experience, particularly on PIM, can supplement the research led by Prof. Sze and Prof. Sanchez. Equipped with my background, I can contribute by devising circuits and architectures to accelerate ML workloads and save energy.

Additionally, Prof. Mohammad Alizadeh has conducted outstanding studies on hardware design for network analysis and RL, which overlap with my experience.

I am grateful to the faculty and committee for their time and consideration, and I hope my experience and interests align with their expectations.