My research interest lies in the intersection of systems and machine learning. The increasing computational demands from exponentially growing models coupled with the rapid (3∼5 years) hardware life cycles in large-scale data centers intensify their environmental impact [1] as few systems are built with sustainability considerations. I aim to address this challenge in three directions: 1) to re-imagine the evaluation and design of data centers software hardware stacks to extend silicon life and minimize hardware waste; 2) to address the rapid evolution and diversity ML workloads by developing efficient dynamic resource management systems; and 3), to meet intense computational requisites via software-defined infrastructure optimization. Guided by these goals, **I aspire to build next-generation systems that promote social responsibility as a first-level design objective in large-scale ML and data-intensive systems alongside efficiency, performance, and reliability**. Backed by my research and industry experience, I am confident that pursuing a Ph.D. in Computer Science at UC Berkeley is the next step toward accomplishing my long-term research aspirations.

Carbon-aware scheduling: To introduce sustainability as an optimization metric for current computing systems, I am co-leading a research project at **CMU** to repurpose old yet functional servers in data centers, advised by **Prof. Akshitha Sriraman** and **Prof. Udit Gupta**. Based on my characterization of two microservice applications in DeathStarBench [2], I identified feasible computational paths on older servers without compromising the end-to-end performance. However, this optimization presents NP-hard challenges at industrial-scale deployments. To address this, I designed a new methodology to efficiently prune the criticality search space by analyzing the microservice-level tail latencies. To further generalize, I proposed a simple representation of the **carbon-performance correlation** with two constant values abstracting away hardware specifications during the microservice placement search. The offline empirical profiling result forms the basis for online heterogeneous scheduling. I built *a carbon-aware auto-scaling scheduler* using Docker Swarm, demonstrating ∼69% carbon-saving on production workload traces while maintaining tail latency requirements, enabling lifetime extension of older hardware. We have submitted this work to IEEE CAL as my co-first author publication [3]. Through this experience, I learned how to frame and conduct **end-to-end systems research** by formally defining a design, methodology, and evaluation framework. This work solidified my research passion for hyperscale systems and *responsible metrics*, and presented future directions along the *data center software hardware stacks* such as ML workload scheduling, heterogeneous processors, and load shifting, which I am excited to explore further during my Ph.D.

GPU collective optimization: With a curiosity for exploring ML systems and improving ML workload efficiency at data centers, I joined the ML Scaling team (legacy Google Research) at **Google** to gain industrial research experience under the mentorship of **Dr. Yu Feng**. The project initially aimed to reduce GPU idle time from all-reduce communication for DTensor, the distributed API for Tensorflow, targeting a large language model, BERT, to speed up training. Through fast prototyping, I soon identified that the given fixed-size clustering optimization approach was suboptimal due to its inability to capture the model architecture, resulting in a lack of predictable performance. I conducted a self-driven compute graph analysis across BERT and T5 with custom configurations and discovered a common **all-reduce topological pattern**, a new knowledge within the sibling teams. Over brainstorming sessions with my mentors backed by extensive documentation reviews, I proposed and built a new **clustering method** by utilizing all-reduces' *locality* in the compute graph. This addressed the challenge of inefficient grouping, replaced excessive hyperparameter tunings with a one-time compute graph analysis, and expanded beyond the original optimization scope from BERT to general transformer architectures. Extensive experiments across model configurations demonstrated that this optimization reduced GPU's idle time up to 78%, effectively eliminating wasted GPU compute during communication and speeding up training by up to 5%. It is currently integrated into DTensor and available for public use. This project highlighted the importance of efficiency in ML systems, a direction I seek to focus on in ML workload scheduling and resource management in my Ph.D.

Industrial Engineering: Before CMU, I worked as a software engineer at **IBM** on Db2 database engine development, where I gained technical skills across **data center stacks** and grew tremendously as an independent thinker and collaborator. As a new grad, my first project was constructing a new infrastructure from scratch for the highly available distributed database architecture, enabling four *parallelism* modes, including intra- and inter-database partitions, demanding substantial engineering effort and domain knowledge. I independently completed the project with minimal guidance via extensive open-source self-study, fast codebase (32GB) navigation, and proactive cross-team collaborations. I gained a unique experience spanning *automatic machine configurations, developing a client-server framework, and creating a machine health monitoring and recovery system*. The process of developing real-world large-scale systems that service enterprise data prepared me technically for my three subsequent projects at IBM and my current research. It instilled in me the mindset of *reliability* and *scalability* in system design, taking modular steps in research, and enhanced my professional communication skills, all of which I am eager to apply in my Ph.D. studies.

Driven by my strong passion for finding innovations in large-scale systems, I voluntarily initiated and led a team of 5 SDEs at **IBM** on **system research and design** outside of regular work, mentored by **Dr. Petr Novotny** and **Shaikh Quader**. Based on our previous project of enabling Db2 for federated learning (FL) on IBM Cloud, I identified inefficiency in client setup and latency from network conditions during aggregation, leading to a patent submission and publication for a tier-based *FL infrastructure and pipeline* with IBM. This experience provided me insights into distributed training infrastructure across physical data centers, and showed me meaningful headroom for more efficient and sustainable AI data centers. More importantly, it validated my strive for intellectual freedom in searching, scoping and solving a problem, ultimately led to my decision to pursue graduate study and a career in research.

Research Values: My core research value in building *socially responsible systems* evolved during my undergraduate study at **UofT** under **Prof. Steve Mann**'s guidance. To help underserved communities, I engineered a wearable system for prosopagnosia. This project presented multifaceted challenges: technically advancing towards lightweight computing devices, logistically evaluating the system in the absence of real patients, and internally coordinating the team for publications. I overcame each hurdle through rapid prototyping, applying *interdisciplinary* techniques from psychology, and effective team communication, all fueled by my research values and passion. This journey reinforced my *resilience* towards research challenges, my curiosity for new and innovative work, and my commitment to developing socially responsible systems. I presented and published this work to IEEE SMC2020 [4].

Career Plan: My academic and industry experiences have prepared me to conduct novel research with real-world applications in large-scale systems. After completing my Ph.D. I plan on pursuing a career as a professor in academia with strong industry collaborations. My previous teaching experience revealed my enthusiasm for mentorship and promoting STEM education for female, diverse, and underserved communities. I volunteered for three years at the STEM4Grils summer camp as a Python-related course instructor for girls aged 5 to 11 years old. As an engineer, I mentored student interns for 16 months, aiding them in delivering industrial products by turning ideas into features. As a TA at CMU, I hold office hours, assist the professor with course organization, and manage student questions, which reaffirmed my interest in college-level teaching. I hope my research, mentorship, and teaching can improve social responsibility that will sustain 10, 20, 50 years into the future. Pursuing a Ph.D. at Berkeley will enable me to grow my research while gaining further teaching and mentoring experience.

Why Berkeley: The intellectual freedom, unparalleled leadership in systems and ML research, and strong ties with the industry at Berkeley present an ideal environment for my continued research into efficient, and socially responsible ML systems and solutions for data centers. I am particularly inspired to work with **Prof. Matei Zaharia**, whose significant contributions to large-scale distributed systems and data processing have profoundly influenced my understanding of efficiency and scalability in my summer research system design. In particular, leveraging PTD-P and with pipelining schedule in 'Efficient Large-Scale...GPU Clusters Using Megatron-LM' inspired me to promote sustainability via computation and resource efficiency. Additionally, I am keen to work with **Prof. Ion Stoica**, whose innovative work in ML system design within data centers, such as Ray that provides a single dynamic execution engine, presents an exciting avenue for exploration in heterogeneity computes for large models with additional optimization metrics on top of efficiency and performance.

In conclusion, I aim to establish a new standard in large-scale data center and AI systems that balances societal impacts with efficiency and performance. And I am also excited to explore new directions. At Berkeley, I will bring a combination of research experience, software engineering expertise, professional soft skills, and a strong passion with an open mind and enduring curiosity for new directions to the CS Ph.D. program.

[1] U. Gupta, M. Elgamal, G. Hills, G.-Y. Wei, H.-H. S. Lee, D. Brooks, and C.-J. Wu, “Architectural co2 footprint tool: Designing sustainable computer systems with an architectural carbon modeling tool,” *IEEE Micro*, vol. 43, no. 4, pp. 107–117, 2023.

[2] Y. Gan, Y. Zhang, D. Cheng, A. Shetty, P. Rathi, N. Katarki, A. Bruno, J. Hu, B. Ritchken, B. Jackson, K. Hu, M. Pancholi, Y. He, B. Clancy, C. Colen, F. Wen, C. Leung, S. Wang, L. Zaruvinsky, M. Espinosa, R. Lin, Z. Liu, J. Padilla, and C. Delimitrou, “An open-source benchmark suite for microservices and their hardware-software implications for cloud & edge systems,” in *Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS ’19, (New York, NY, USA), p. 3–18, Association for Computing Machinery, 2019.

[3] W. Jaylan, **Z. Pan***, U. Gupta, and S. Akshitha, “Ecoscale: Giving old servers new life at hyperscale,” *IEEE Computer Architecture Letters*, dec 2023 (under submission).

[4] S. Mann, **Z. Pan**, Y. Tao, A. Gao, X. Tao, D. E. Garcia, D. Shi, and G. Kannan, “Face recognition and rehabilitation: A wearable assistive and training system for prosopagnosia,” in *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 731–737, 2020.