My decision to pursue graduate studies stems naturally from my experiences as a master's student at **[Your University]** and, more recently, as a research staff member in the **Systems Group** at **[Company/Institution]**. My primary interests lie in **distributed systems** and **data-intensive cloud computing**, with a focus on the system-level challenges of **scaling and deploying machine learning models**. These research efforts have led to publications at leading venues, including **OSDI [1]** and **ICDE [3]**. Graduate studies at Stanford will enable me to advance these interests and represent the first step toward my long-term goal of a research career.

In recent years, **Graph Neural Networks (GNNs)** have emerged as powerful models for learning on graph-structured data. Existing systems for **distributed GNN training** often extend designs from deep neural network (DNN) training or graph processing frameworks. While seemingly natural, such retrofitting inherits tradeoffs not suited for GNN workloads, leading to communication stalls, underutilized compute resources, and scalability bottlenecks.

To address these challenges, I collaborated with **[Your Collaborators]** to develop the **P³** [**1**] system, published at **OSDI 2021**. Unlike traditional partitioners (e.g., METIS) that focus solely on the structural dimension, P^3 introduces an independent partitioning across both feature and structural dimensions. Combined with intra-layer model parallelism and pipelining, this enables a novel push–pull distributed training strategy that achieves high resource utilization while minimizing communication and partitioning overheads. We are now extending this work by integrating pipelined push–pull parallelism into Microsoft's DeepGraph engine to scale training across thousands of GPUs.

More recently, I have been exploring **model serving systems**. Large-scale pre-trained language models such as BERT have significantly advanced NLP applications, but their computational demands make efficient serving challenging. A promising approach is the use of early-exit deep neural networks (EE-DNNs), which exploit differences in sample classification difficulty to achieve better accuracy–latency trade-offs by exiting inference early. However, we observe that coarse-grained batching—a common technique for throughput improvement—becomes suboptimal when samples dynamically exit at different stages, reducing hardware utilization. To address this, with **Your Research group members**, I proposed **SURGEON [2]**, a system that takes an EE-DNN model and SLA constraints as input, and generates an optimal partition and service assignment across heterogeneous resources using as few GPUs as possible. SURGEON consolidates batches at partition boundaries to improve hardware efficiency and employs dynamic programming to determine adaptive partitioning strategies. I am currently evaluating SURGEON across diverse EE-DNN architectures and service-level objectives.

Before joining MSR, I completed my master's (by research) in Computer Science at **[University, Country]**, advised by **[Supervisors]**. My thesis investigated system-level optimizations for **distributed temporal**[1] **graph analytics**. Existing frameworks often fail to scale due to redundant computation and excessive messaging across time points. To address this, I co-developed **GRAPHITE [3]**, which introduces the *time-interval* as the data-parallel unit. A novel *time-warp* operator automatically partitions a vertex's temporal state and temporally aligns messages, thereby reducing redundant execution and communication. GRAPHITE was published in **ICDE 2020** and is currently used in the **impact industry or Uni** for temporal analytics, including contact tracing. I also worked on **WAVE [4]**, an extension of GRAPHITE that incorporates *dependency-driven incremental processing*. By tracking dependencies, WAVE incrementally propagates changes across intermediate values, significantly reducing recomputation. This work was recognized at the **2nd ACM SRC (Graduate Category) @ SOSP 2019**, where it was a finalist and awarded the Bronze Medal.

I realize the need for a strong theoretical foundation to pursue advanced research. In this direction, I have always striven for academic excellence – I stood top of the class during both bachelor's and master's studies. My time at **YOUR UNIV** offered me opportunities to assist with two graduate courses, and to participate

[1] Graphs whose structure and attributes evolve

in the Artifact Evaluation Committee (AEC)[2] and the Shadow Program Committee[3] at several conferences. These engagements helped me learn how to organize and articulate ideas effectively; faux-PC discussions taught me to interpret reviewer subtext around rebuttals, while comparing submitted versus accepted papers at EuroSys 2021 gave me valuable insight into how feedback shapes stronger research contributions. The time I spent in industry, both before and after graduate school, helped me develop essential soft skills – time management, teamwork, and cooperation – which I believe are crucial for thriving in demanding academic environments.

I believe my experience with data-intensive systems at Microsoft Research has provided me with a unique perspective on the practical challenges faced by developers and cloud operators in deploying, operating, and monitoring large-scale computer systems. This background makes me well-prepared to address big-picture questions in the field. At Stanford, I aim to advance research on efficient **systems infrastructure and tooling for emerging data-intensive workloads**. As machine learning models grow in scale and complexity, particularly in safety- and performance-critical applications, the demand for compute, resiliency, resource-efficiency, and affordability grows in tandem. I find these challenges especially exciting to pursue. Stanford 's leadership in data-intensive systems research, world-class faculty, and interdisciplinary culture make it an ideal environment for my graduate study. I am particularly inspired by **Prof. A**, **Prof. B**, and **Prof. C**, whose work (e.g., Spark, Snorkel, Shinjuku, PipeDream, ROC, INFaaS, GNNAutoScale, GraphSAGE) has influenced my thesis and research papers, and with whom I would be eager to collaborate.

In summary, I believe I bring with me research experience, industry-sharpened programming and soft skills, and above all, an insatiable desire to learn and excel. I look forward to the next milestone in my life – a PhD in Computer Science from Stanford.

[1] First Author, Second Author, "P^3: KJHS", In Proceedings of the 15th USENIX Symposium on Operating Systems Design and Implementation (**OSDI 2021**), July 2021. yourjournalDOI

[2] First Author, Second Author, "HGFD" (**Ongoing Project**)

[3] First Author, Second Author, "ABG", In Proceedings of the 36th IEEE International Conference on Data Engineering (**ICDE 2020**), Dallas, Texas, April 2020. yourjournalDOI

[4] First Author, "ADC", 2nd ACM Student Research Competition (**SRC**) at the 27th Symposium on Operating Systems Principles (**SOSP 2019**), Ontario, Canada, Oct 2019. yourjournalDOI

[2] AEC SOSP 2019, OSDI 2020, and ASPLOS 2020
[3] Shadow PC EuroSys 2021 and 2022