Annual Progress Report for Ph.D Students

Student Name: Lai Ruiqi

Advisor: Dmitrii Ustiugov

School: College of Computing and Data Science

A brief Introduction of my project

I am currently conducting research on AI-as-a-Service, aimed at deploying large-scale AI

models for inference services. My current research focuses on the deployment of LLM (Large Language Models) at the cluster level, addressing issues related to the large data size of LLM models and improving the throughput of LLM model inference services on GPU clusters.

Summary of research progress

During my second year, after the initial rejection of my first paper submission, I made several adjustments to my research methodology. Rather than focusing solely on parallelism, I broadened the scope to develop a more general framework for analyzing and comparing different autoscaling policies. This refinement has led to more robust and insightful results. I am currently finalizing the updated project and plan to submit it to ASPLOS in August or EuroSys in September.

A detailed completion plan

We plan to submit this work to ASPLOS’26 Oct or EuroSys’26 Oct.

Jul 2025 - Aug 2025 System development and evaluation experiment

Aug 2025 - Sep 2025 Paper writing and revising

Sep 2025 Submit to conference