香港浸會大學理學院
HKBU Faculty of Science

DEPARTMENT OF COMPUTER SCIENCE

PhD Degree Oral Presentation

PhD Candidate:	Mr. WANG Yuxin
Date	24 July 2024 (Wednesday)
Time:	10:00 am – 12:00 noon (35 mins presentation and 15 mins Q & A)
Venue:	1) DLB 637, 6/F, David C Lam Building, Shaw Campus 2) ZOOM (Meeting ID: 973 5413 9693) (The password and direct link will only be provided to registrants)
Registration:	https://bit.ly/bucs-reg (Deadline: 12:00 nn, 23 July 2024)

Towards Efficient and Reliable DNN Training and Inference on HPC Clusters

Abstract

The size of Deep Neural Networks (DNNs), especially large DNNs like Large Language Models (LLMs), is rapidly increasing, driving advanced AI applications in recent years. However, as DNNs grow in complexity, training and inference on High-Performance Computing (HPC) clusters become increasingly challenging. Efficiency and reliability are paramount, particularly as model and system scales expand. This thesis addresses these challenges through four interconnected research works that collectively aim to optimize DNN training and inference on HPC clusters, spanning from training to inference, and from single-machine efficiency optimizations to distributed system efficiency and reliability improvements.

Regarding DNN training, the thesis progresses from single-device evaluations to distributed system enhancements. First, to understand the performance and energy efficiency of DNN training on HPC accelerators, we provide a comprehensive empirical study that benchmarks various accelerators on different deep learning workloads. The findings guide researchers in accelerator selection and suggest efficiency improvements for hardware vendors. Then, to enhance system efficiency and reliability when scaling distributed large DNN training, we present REFT, an innovative in-memory fault-tolerant training framework. REFT reduces the cost of fault tolerance against frequent system failures by minimizing checkpoint saving and loading overheads, while integrating more frequent checkpoint saving to improve training reliability.

Regarding DNN inference, the thesis also provides insights into single-device improvements and cloud-serving insights. For single-device inference, we establish a performance and energy efficiency model through empirical benchmarking to understand and improve the performance and energy efficiency of DNN inference on GPUs. We then propose a Mixed Aligned Scheduling (MAS) scheme that significantly enhances the throughput and energy efficiency of LLM inference under specific service-level agreements (SLA). Then, to efficiently and reliably scale large DNN inference services on HPC clusters, we introduce BurstGPT, a real-world workload dataset to evaluate and optimize LLM serving. BurstGPT captures user, system, and model characteristics, highlighting scheduling inefficiencies in current serving systems due to workload burstiness. This evaluation underscores the critical need for dynamic hardware resource management to maintain service quality under varying workload conditions.

In conclusion, this thesis presents comprehensive evaluations and optimizations for improving the efficiency and reliability of DNN training and inference on HPC clusters. Integrating empirical studies with practical system designs, we contribute insights into optimizing hardware and software frameworks during the evolution of AI.

***** ALL INTERESTED ARE WELCOME *****