香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

数据科学与分析学域
DATA SCIENCE AND ANALYTICS THRUST
信息枢纽
INFORMATION HUB

DSA Research Experiences for Undergraduates

Research Project

Section1: Faculty Information

Full Name	Guoming Tang	Tel	13618480331
Thrust/Hub	DSA/INFO	Office	W3-306
Email	guomingtang@hkust-gz.edu.cn		

Section2: Research Project Proposal

Project Title	Green GPU Computing for Large-Scale LLM Inferences
Project Description (max 800 words)	As the scale of large language models (LLMs) continues to grow, so does the demand for **more efficient and adaptive GPU-based inference.** Traditional GPU resource allocation often relies on static or heuristic-based methods, which can lead to **suboptimal energy consumption** and increased latency. These challenges become even more pronounced in large-scale online inference scenarios, where diverse tasks—each with distinct **computational complexities** —must be processed simultaneously. This project will explore the following questions: 1. How can we define **new metrics regarding the GPU computing potentials** in evaluating LLM performance and overhead? 2. How do we seamlessly **match new tasks to the most suitable hardware** config class in an online environment with multiple GPU nodes? 3. Can LLMs **facilitate dynamic adjustments** to maintain efficiency amid variable workloads? 4. Can LLMs **generalize the hardware-tuning methodology** to new architectures, tasks, or even different hardware platforms? You will gain the following opportunities: 1. **Hands-on HPC & GPU Experience** Work with real GPU environments and gather practical skills in hardware tuning, optimization algorithm and cluster-based analysis. Gain a deep understanding of balancing performance metrics such as energy consumption and inference latency. 2. **LLM Integration & Agent Development** Explore how large language models can be harnessed for scheduling, optimization, and interpretability. Experiment with prompt engineering, fine-tuning (e.g., LoRA), and LLM-based reasoning to push the boundaries of HPC scheduling methods.

	3. **Research & Publications** Identify real-world problems within GPU hardware resource management and transform these insights into publishable research. There is a strong opportunity to work on cutting-edge topics and produce high-quality academic papers (if possible). 4. **Professional Growth & Internships** Students in this project may access internship opportunities with industry partners, opening doors to broader collaborations and future career development.
Proposed Research Duration	Start Date: Now End Date: Aug. 31, 2025
Student/Researcher Duties	Students should read papers in related fields, actively engage in team discussions, and contribute ideas to the project. Additionally, they will be responsible for part of coding tasks. Exploring different cutting-edge directions to solve problems is encouraged.
Technical Skills Required	☑Python ☑Machine Learning ☐ Big Data ☐ R ☑ Deep Learning ☐ SQL ☐ C/C++ ☐ Other: _____
Preferred Student/Researcher Background	Programming, Academic Reading & Writing Skills, Teamwork Ability
Maximum Number of Students/Researchers	☐ 1 ☑ 2

Section3: Pre-Application Research Exposure Meeting

Faculty members are encouraged to schedule a Research Exposure Meeting to introduce students to their projects.

Preferred Date	Mar. 14-16, 2025
Preferred Time	Morning or afternoon time, 1 hour
Meeting Mode	☑ In-Person ☐ Online
Venue (if in-person)	W3-306
Meeting Link (if online)	