GPU resource allocation in DBMS

Modern database systems increasingly rely on GPU for:

1. Learned (ML-based) query optimization: Using neural networks or advanced ML models to find the optimal query plan.
2. ML-powered UDFs: Running inference tasks (e.g., image classification, text summarization, or anomaly detection) within the query pipeline.
3. GPU-acceleration: Speed up query execution, particularly for analytical workloads

If we have more than one of the above components in a DBMS, such GPU-hungry workloads can cause resource contention if they share the same GPU(s) and may hence hinder the speedup. Therefore, in this project, we would like to develop a GPU resource management solution that extends current DBMS. The expected outcomes are prototypes supporting GPU workloads, reduced overhead, and improved throughput for concurrent SQL queries and their ML-based optimizer. The student is also expected to conduct experiments, collect and analyze results, and present main findings in written reports and oral presentations.

Desired Qualifications:

- Solid systems programming skills in C/C++;
- Solid foundation in database systems. Prior experience in developing database engine preferred;
- Good understanding of computer architecture. Experience with machine learning libraries and GPU programming will be a plus.

Application Process:

Interested applicants should email Dr. Zhengjie Miao (zhengjie@sfu.ca) with the most recent transcript and Curriculum Vitae