Project Deliverable 1: Project Proposal and Abstract

CSE 4059: Applied Parallel Programming using GPUs

Group Members:

- Member 1: **Bei Jin**

Project Abstract:

1. **Project Title:** Quantized Matrix Multiplication: An FP8/INT8 CUDA-Based Implementation and Performance Analysis
2. **Problem Statement:** With the increasing size and complexity of machine learning models, especially large language models (LLMs), the need for faster and more memory-efficient matrix operations has grown significantly. Traditional FP32 and even FP16 computations pose limitations in terms of memory bandwidth and energy efficiency. Low-precision formats such as FP8 and INT8 present promising alternatives, but their implementation and quantization effects on accuracy and performance are not fully explored in educational settings.
3. **Objective**: This project aims to implement and analyze low-precision matrix multiplication on GPUs using CUDA. Specifically, it will compare the performance and numerical accuracy of FP8 and INT8 matrix multiplication against FP16 baseline, demonstrating the impact of quantization on speed, memory usage, and output quality.
4. **Approach**: We will implement quantization-aware matrix multiplication kernels using CUDA. The project includes quantization (Q), dequantization (DQ), and fused Q-DQ-GEMM operations for both FP8 and INT8 formats. Data-parallel techniques such as thread-block tiling, memory coalescing, and shared memory usage will be applied to optimize performance. Baseline results from cuBLAS FP16 GEMM will be used for comparison.
5. **Expected Outcome**: We expect to observe significant speedup and memory savings with FP8 and INT8 formats while maintaining acceptable numerical accuracy. The project will provide insights into precision trade-offs in GPU computing and contribute a foundational low-bit GEMM implementation that can be extended to real-world deep learning applications.