Daniel Sanei

Professor Cheolhong An

GPU Programming

28 February 2025

Final Project Proposal

Project Topic

For my final project, I will implement high performance CUDA-accelerated matrix multiplication code within C++ using CUDA, and then utilize Pybind11 to bind these CUDA kernels from C++ to Python. This will allow using Python, a programming language with a more accessible syntax, while leveraging CUDA's parallel processing capabilities by abstracting the higher complexity of C++ from the driver code.

Description

This project will achieve the following goals:

1) Implement matrix multiplication code using CUDA kernels in C++
2) Use Pybind11 to bind C++ to Python
3) Create a Python script to initiate the driver code
4) Compare this result with a baseline CPU matrix multiplication
5) Analyze potential areas for memory optimization

Individual Role

This will be an individual project, and therefore I will be responsible for achieving the deliverables outlined in the section above.

Motivation

I am looking forward to a personal implementation of matrix multiplication using CUDA in C++, as I would like to pursue this project to become more proficient in CUDA programming and solidify my understanding of GPU architecture and memory optimization techniques. Additionally, I am very interested in binding C++ to Python, as I first heard of this technique during an extracurricular project team, and have since looked forward to an opportunity to accomplish this.

Justification

This project proposal aligns with the core focus of this course, which is understanding CUDA programming, GPU architecture, and memory optimizations. This project takes advantage of Python binding as an extension, though the core focus remains CUDA optimization on a matrix multiplication task. Considering the time constraints and the Professor and TA's advice to select a simple topic, I believe this project will be manageable within two weeks and I am looking forward to learning more about binding and CUDA programming.