CPSC 524 Final Project Description

Rami Pellumbi*

December 11, 2023

*M.S., Statistics & Data Science

Introduction General matrix-matrix multiplication is an expensive operation to perform as the size of the matrices scales. The goal of this project is to aggressively optimize matrix-matrix multiplication on a single node via throughput analysis. From there, the goal will be to divvy up work between multiple nodes via MPI. The project has two (maybe three) phases:

1. Reproduce the results of the video motivating this project, optimizing for the specifications on a node of the Grace cluster.[1] This will utilize OpenMP for parallelization and AVX-512 instruction sets for performing dot products. The source provides a code in C++, uses SIMD instead of AVX-512, and optimizes for a consumer level processor.
2. Extend the result to also divvy up the work between multiple nodes via MPI.
3. (Maybe) Extend the result of the general matrix multiply to multiplication of lower and upper triangular matrices.

Throughput Analysis Given the maximum number of gigaflops and memory bandwidth on the node, and the fact that multiplying two $N \times N$ matrices is:

- $3N^2$ total memory access,
- $2N^3$ floating point operations,

the project will aim to reach both full memory and compute bandwidth during the matrix multiplies. Naive implementations are far from compute bound. The goal is to have the matrix multiply on a CPU be compute bound for a reasonably sized N on a single node.

[1] YouTube Source