

Course Mini-Project #1: Matrix-Matrix Multiplication with SIMD Instructions

Due date: Feb. 15

1. Introduction

The objective of this mini-project is to implement a C/C++ module that can carry out high-speed matrix-matrix multiplication by explicitly utilizing x86 SIMD instructions. Matrix-matrix multiplication is one of the most important data processing kernels in numerous real-life applications, e.g., machine learning, computer vision, signal processing, and scientific computing. This mini-project aims to help you gain hands-on experience of (1) SIMD programming, and (2) cache access optimization. It will help you develop a deeper understanding of the importance of exploiting data-level parallelism and minimizing cache miss.

2. Requirement

Your implementation should be able to support (1) configurable matrix size that can be much larger than the on-chip cache capacity, and (2) both fixed-point and floating-point data. Each group should create a Github site that hosts the code/results of all the projects through this semester. Other than the source code, your Github site should contain

- (1) Readme that clearly explains the structure/usage of your code
- (2) Experimental results that show the performance of your code under different matrix size (at least including 1,000x1,000, 10,000x10,000, and 100,000x100,000) and data precision (4-byte floating-point, 2-byte fixed-point)
- (3) Comparison with native implementation of matrix-matrix multiplication (i.e., without any cache optimization and explicit use of SIMD instructions)
- (4) Analysis and conclusion

3. Additional Information

The easiest way to use SIMD instructions is to call the intrinsic functions in your C/C++ code. The complete reference of the intrinsic functions can be found at <https://software.intel.com/sites/landingpage/IntrinsicsGuide/>, and you can find many on-line materials about their usage. Moreover, matrix-matrix multiplication has been well studied in the industry, and the most well-known library is the Intel Math Kernel Library (MKL), which can be a good reference for you.