High Performing Matrix–Panel Multiplication Routine on Intel Processor with AVX-512 Instructions

Muhammad Rizwan*, and Jaeyoung Choi

School of Computer Science & Engineering, Soongsil University

369 Sangdo-ro, Dongjak-gu, Seoul 06978, Korea

Corresponding author (Electronic mail: mrizwan@soongsil.ac.kr)

Linear algebra equations are used to solve numerical and scientific problems in the field of computational science. Matrix–matrix multiplication is the fundamental operation in basic linear algebra subprograms (BLAS). General matrix–matrix multiplication (GEMM) kernels are widely available. These high-performing BLAS level 3 kernels performed well on modern processors for large square matrices. However, these GEMM kernels achieved sub-optimal performance when one of the dimensions is small. Matrix–matrix multiplication operations used in LU, QR, and Cholesky factorizations require optimal performance of tall and skinny matrices for panel–matrix, matrix–panel, or panel–panel multiplication operations. A matrix is termed as panel when one of its dimensions is small.

In this study, we analyzed the effects of matrix size, register blocking parameters, and thread distribution on the performance of our previously implemented high-performing blocked matrix–matrix multiplication routine [1, 2]. Despite the memory structure is same, the optimal blocking parameters and thread distribution are not the same for matrices with different sizes. As a result, the matrix–matrix multiplication optimal blocking parameters cannot produce good results for panel–matrix, matrix–panel, or panel–panel multiplication operations. Different algorithms may be optimal for different matrix dimensions depending on the shapes of the matrices involved. We modified our previously implemented column major variant of the matrix–matrix multiplication routine [2] and improved the performance of the matrix–panel multiplication operation using AVX-512 instructions. We also presented the performance of ScaLAPACK QR factorization by replacing double-precision general matrix–matrix multiplication routine (DGEMM) with our matrix multiplication routines on Intel Xeon Phi Knights Landing (KNL).

We named our previously implemented blocked matrix–matrix multiplication algorithm as USERDGEMM.

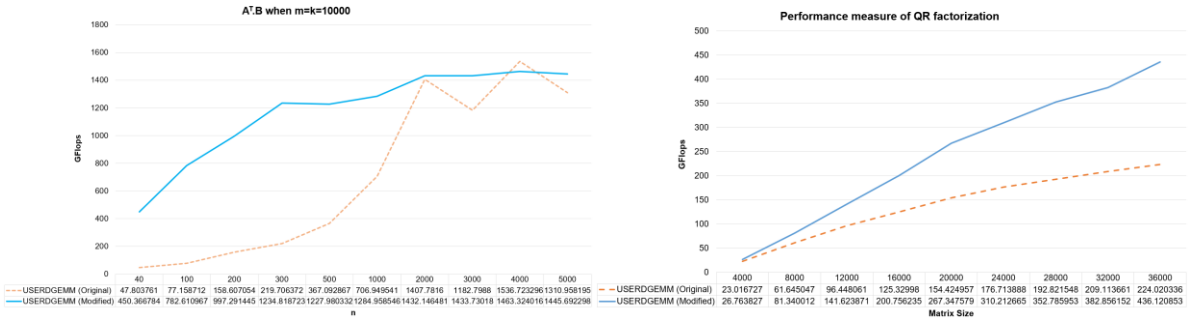

Figure 1. Performance measure of original vs modified routine on KNL: $A^T.B$ matrix multiplication (left) and then utilizing it in QR factorization (right)

Acknowledgments This work was supported by the Supercomputer Development Leading Program of the National Research Foundation of Korea (NRF) funded by the Korean government (MSIT) (No. 2020M3H6A1084984).

References

[1] R. Lim et al., “An implementation of matrix-matrix multiplication on the Intel KNL processor with AVX-512,” *Cluster Computing*, 21, 4, 1785–1795 (2018).

[2] Y. Park et al., “Improving blocked matrix-matrix multiplication routine by utilizing AVX-512 instructions on intel knights landing and xeon scalable processors,” *Cluster Computing*, 1-11 (2021).