Performance Engineering for Tall & Skinny Matrix Multiplication Kernels on GPUs

Dominik Ernst
Erlangen Regional Computing Center, Germany
dominik.ernst@fau.de

Abstract

Block Vector Algorithms, i.e. algorithms that are formulated to operate on a matrix consisting of several vectors, have been shown to be useful in the context of Eigenvalue Solvers [1], where they have both numerical and performance benefits. Block Vectors are so called Tall & Skinny Matrices (TSM), for which the standard GEMM implementation approaches that are used in the vendor libraries fail to deliver good performance. We therefore introduced specialized TSM matrix-matrix multiplication operations into GHOST (the General Heterogeneous Sparse Matrix Toolkit [2], a sparse linear algebra kernel library) to enable efficient Block Vector computations. In this work, we show several implementation approaches for TSM matrix multiplication on GPUs, differing mostly by the way work is mapped to the hardware. Extensive performance modeling is used to analyze and explain the approaches' performances by identifying key factors like amount of parallelism, access patterns, limiting data paths or reuse opportunities.

References

1. Melven Röhrig-Zöllner and Jonas Thies and Moritz Kreutzer and Andreas Alvermann and Andreas Pieper and Achim Basermann and Georg Hager and Gerhard Wellein and Holger Fehske. Increasing the Performance of the Jacobi–Davidson Method by Blocking. In: SIAM Journal on Scientific Computing 37.6 (2015), pp. C697–C722. doi: 10.1137/140976017.
2. Moritz Kreutzer et al. GHOST: Building Blocks for High Performance Sparse Linear Algebra on Heterogeneous Systems. In: International Journal of Parallel Programming (2016), pp. 1–27. issn: 1573-7640. doi: 10.1007/s10766-016-0464-z.