Performance Engineering for a Tall & Skinny Matrix Multiplication Kernel on GPUs

Dominik Ernst[1], Georg Hager[1], Gerhard Wellein[2], Jonas Thies[1]
[1]Erlangen Regional Computing Center (RRZE), Erlangen, Germany
[2]German Aerospace Center (DLR), Simulation and Software Technology
dominik.ernst@fau.de

General matrix-matrix multiplications (GEMM) in vendor-supplied BLAS libraries are best optimized for square matrices but often show bad performance for tall & skinny matrices, which are much taller than wide. Nvidia's current CUBLAS implementation delivers only a fraction of the potential performance (as given by the roofline model) in this case. We describe the challenges and key properties of an implementation that can achieve perfect performance. We further evaluate different approaches of parallelization and thread distribution, and devise a flexible, configurable mapping scheme. A code generation approach enables a simultaneously flexible and specialized implementation with autotuning. This results in perfect performance for a large range of matrix sizes in the domain of interest, and at least 2/3 of maximum performance for the rest on an Nvidia Volta GPGPU.

Keywords: tall & skinny, matrix matrix multiplication, GPU.