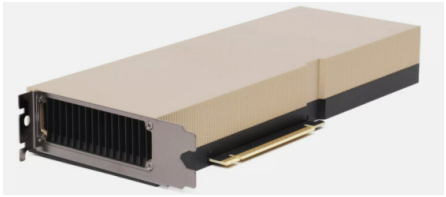
# Specialized QR Factorization for Tensor Algorithms on GPUs

**Keywords:** High-Performance Computing (HPC), Numerical Linear Algebra, GPU programming, Data Analytics, Tensor Decomposition

**Contact:** Jonas Thies (`j.thies@tudelft.nl`)


"Tensor-Train" decomposition of a 5-dimensional data tensor into five 2- and 3-dimensional tensors.



NVidia H100 ("Hopper") GPU

## Topic summary

In many modern data analysis algorithms, as well as simulation of high-dimensional problems like stochastic PDEs or quantum physics applications, tensors (the multi-dimensional generalization of matrices) are playing a central role [1]. Compared to classical (dense) linear algebra, tensor algorithms require a larger variety of "building blocks" that have to run efficiently on computing hardware. In this thesis, you will develop such building blocks for computing a low-rank approximation (tensor decomposition) on a modern "Graphics-Processing Unit" (GPU).

For decomposing large multi-dimensional data into a product of multiple much smaller tensors, a sequence of QR factorizations and/or Singular Value Decompositions (SVDs) of large rectangular or "tall & skinny" matrices are used. Common implementations for GPUs are not optimized for this case and either achieve poor performance or do not work robustly for possibly rank-deficient input matrices. We therefor suggest to implement and analyze a specialized QR factorization method for large data on current GPUs based on the TSQR algorithm introduced in [2]. Older GPUs provide insufficient high-bandwidth memory and cache (or shared memory) for this kind of algorithm. However, on newer systems such as, e.g., Nvidia GH200, implementing a special TSQR-like algorithm could be feasible and faster by orders of magnitude than conventional implementations.

**The thesis will be conducted in close collaboration with the German Aerospace Center (DLR) in Cologne**

## Required background

- The programming can be done in C++ (preferred) or Python
- CUDA programming (or at least keen interest in learning)
- Numerical linear algebra

## Literature

[1] S. Rabanser, O. Shchur, and S. Günnemann, "Introduction to Tensor Decompositions and their Applications in Machine Learning," 2017, doi: 10.48550/ARXIV.1711.10781.

[2] J. Demmel, L. Grigori, M. Hoemmen, and J. Langou, "Communication-optimal Parallel and Sequential QR and LU Factorizations," SIAM Journal on Scientific Computing, vol. 34, no. 1, pp. A206–A239, Jan. 2012, doi: 10.1137/080731992.