Assignment for the final exams

The exam consists in a discussion on the achieved results of parallel scaling following the implementations of two codes for distributed multi-GPUs parallel systems. For both codes a scaling analyses is expected using 1, 2, 4, 8, ... GPU nodes of the Boost partition hosted into the Leonardo supercomputer. The analysis is presented within a brief report including a short comment on the results of scaling, possibly plotted on a stacked bar chart that shows in details how the total time is spent between computation and communication. For both cases a 1D-distribution of processes among the *i-direction* on a 2D-domain of size *N x N* is expected. Plots are required for at least two different resolutions: *N* of the order of 10^3 and 10^4.

The two codes are described as follows:

1) Code a distributed matrix multiplication (C = A x B) using MPI and implementing the *All_Gather* algorithm presented during classes. The implementation should include three ways of computing the matrix multiplication local to all processes: a simple *naive* implementation for CPU only, use of the DGEMM (cblas_dgemm) for CPU only and use of the of the DGEMM for GPU (cubals_dgemm). Compare scaling results for the three versions.

Note: the cublas_dgemm requires initialization and expects the data in row-major order. On the GPU code the initialization is expected to happen on CPU where the OpenMP parallelization can be applied to speed up the procedure for large matrixes.

2) See the notes included into the repository for the Jacobi problem.