CSCI 423/523/723 Fall 2025

Mid term exam 1

The CUDA with C++ program multiplies two NxN integer matrices in parallel. N = 1 << 6.
Following functions are expected:

- kernel to fill the matrices on independent streams;
- kernel to implement the multiplication;
- host function to implement the multiplication;
- host function to verify the correctness of the result;
- main function.

The number of blocks should be equal to the warp size multiplied by the available streaming multiprocessors on the GPU. Measure running time of kernel and host versions of the multiplication functions.
Use and submit only one lastname.cu file.

Recommended steps:

- Initialize the matrices
- Implement sequential multiplication of the kernel
- Solve the task for the small matrix size (e.g. 4 x 4)
- Parallelize the kernel
- Verify the host and the kernel results
- Add streams
- Increase the matrix size.

Grading points (Total 20%):

- correct implementation following the task requirements 12%;
- grid stride loop 2%;
- optimal number of blocks and threads 2%;
- usage of the streams 2%;
- optimum data migration, clear memory 2%;