

CMPS 224/324: GPU COMPUTING  
ASSIGNMENT 3

In this assignment, you will implement a matrix-matrix multiplication kernel that uses shared memory tiling. Your kernel is expected to work for any set of matrix dimensions so make sure to handle boundary conditions correctly.

### Instructions

1. Place the files provided with this assignment in a single directory. The files are:
  - `main.cu`: contains setup and sequential code
  - `kernel.cu`: where you will implement your code (you should only modify this file)
  - `common.h`: for shared declarations across `main.cu` and `kernel.cu`
  - `timer.h`: to assist with timing
  - `Makefile`: used for compilation
2. Edit `kernel.cu` where `TODO` is indicated to implement the following:
  - Allocate device memory
  - Copy data from the host to the device
  - Configure and invoke the CUDA kernel
  - Copy the results from the device to the host
  - Free device memory
  - Perform the computation in the kernel
3. Compile your code by running: `make`
4. Test your code by running: `./mm-tiled`
  - If you are using the HPC cluster, do not forget to use the submission system. Do not run on the head node!
  - For testing on different matrix sizes, you can provide your own values for matrix dimensions as follows: `./mm-tiled <M> <N> <K>`
5. You are also provided with a file called `questions.txt` which contains questions about the assignment. Answer the questions in the file after implementing your kernel.

### Submission

Submit your modified `kernel.cu` and `questions.txt` files via Moodle by the due date. Do not submit any other files or compressed folders.