Matrix multiplication

Assignment 3

Milan Cimbaljević

Matrix multiplication using tiling in local memory and persistent thread model.

Idea of this approach is to run as many workgroups as there are compute units on the device.
Gpu in my laptop has 3 compute unites. I will be launching 3 workgroups each containing 256 threads (16 x 16).
Each workgroup will be assigned to compute certain amount of adjacent tile blocks inside the resulting array.

Matrices A and B will be sectioned into tiles, each workgroup is going to cache tile block that it currently processes.
Depending on the tile size each thread inside the workgroup will be assigned to handle certain number of elements.
By tiling memory and caching it we reduce the number of requests to global memory.
By assigned more work to each thread we can also reduce number of requests to the local memory.