PS 703106
Exercise 10

Group 4:
• Jonas Boutelhik
• Michael Thöni
• Thomas Urban

Hardware:
1. GeForce GTX 860M
 1.1 Hardware version: OpenCL 1.2 CUDA
 1.2 Software version: 418.39
 1.3 OpenCL C version: OpenCL C 1.2
1.4 Parallel compute units: 5

Assignment “The Need for Speed”:
Maximize the performance of the matrix multiplication and the benchmark score for the hardware in RR 15

- Document all optimization steps and respective performance improvements/degradations
- For optimization parameters (e.g. work group sizes), explain how you chose them
- Any changes to the algorithm are permitted, as long as the given problem is not modified (e.g. changing memory storage is allowed, re-using pre-computed results is not allowed)
- When working with multiple kernels, the execution times must be summed up. For simplicity, queuing times may be ignored.

Measure the overall score on the hardware present in RR 15. The general goal and assignment is to beat the performance of the CPU version for all problem sizes.

For this Assignment „Tiling“ was used to split the resulting matrix into smaller blocks. For these blocks 2D arrays have been created in the local memory of the kernel. The array contains only a part of the resulting multipilication of a row of a matrix A and a column of a matrix B. All the related subtotals of the corresponding block array elements must be added together to get the right value in the resulting matirx. Figure 1 shows the principle of tiling graphically.

The program is based on the kernel of the lecture and the host will execute 4 kernel calls. The first two kernel calls adjust the size of the two matrices (A and B), since the matrices must be a multiple of the workgroup size. The third kernel call realizes the tiling and the last one finally resets the matrices to their original size.

Figures 2 shows the performance of the program. It can be seen, that the openCL implementation is 3 to 11 times faster than the CPU one. The results of the two tested workgroup sizes of 16 and 32 show that with a workgroup size of 32 the performance is almost 4% better.
According to “https://en.wikipedia.org/wiki/GeForce_800M_series” the GeForce GTX 860M has a maximum processing power of 1317,1 GFLOPS. With the maximum value of 176,3 GFLOPS (N = 4001 for a workgroup size of 32) about 13% of the maximum GFLOPS were reached. Better GFLOPS values can be reached with an more effective caching.

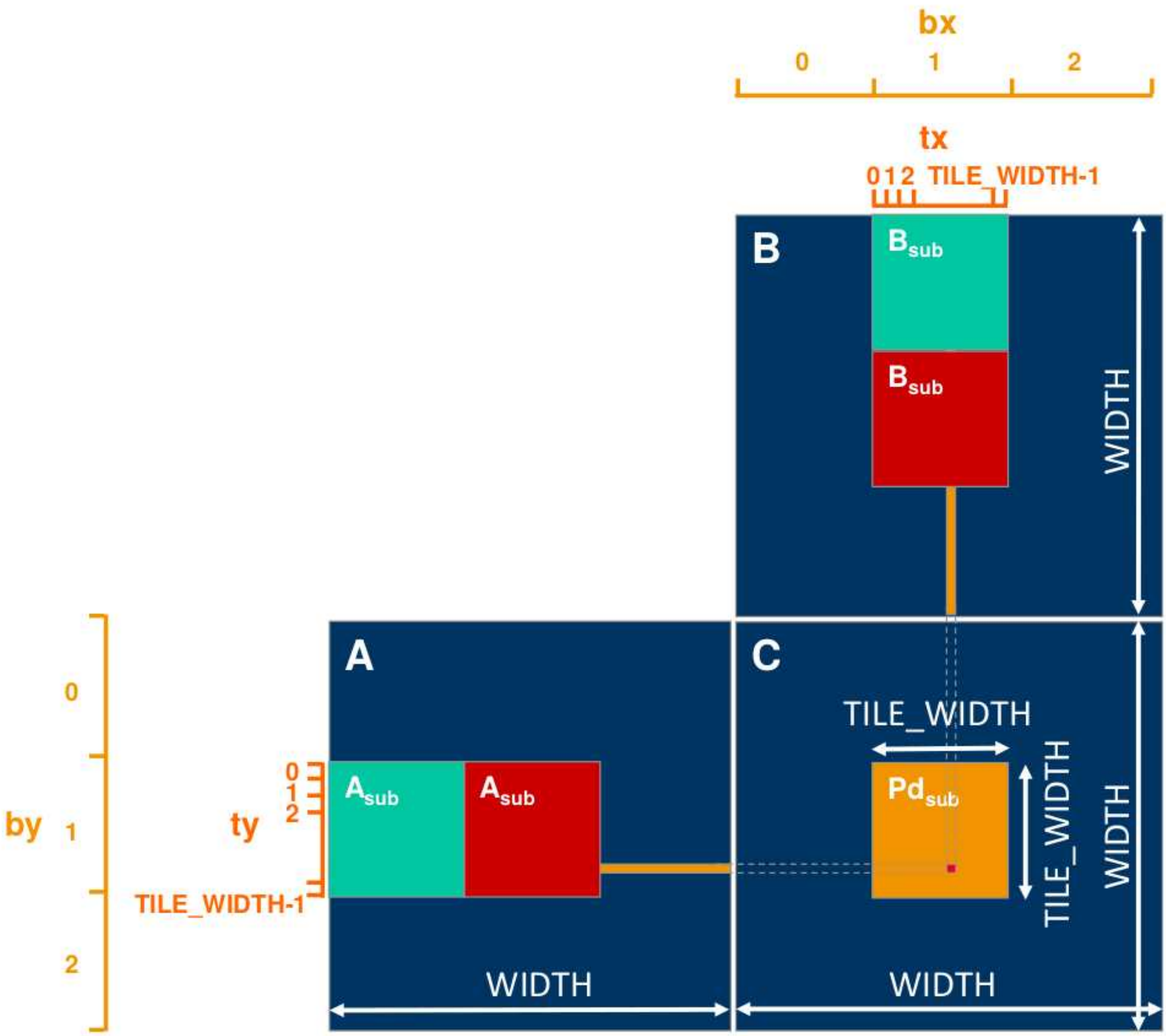

Figure 1: the principle of tiling

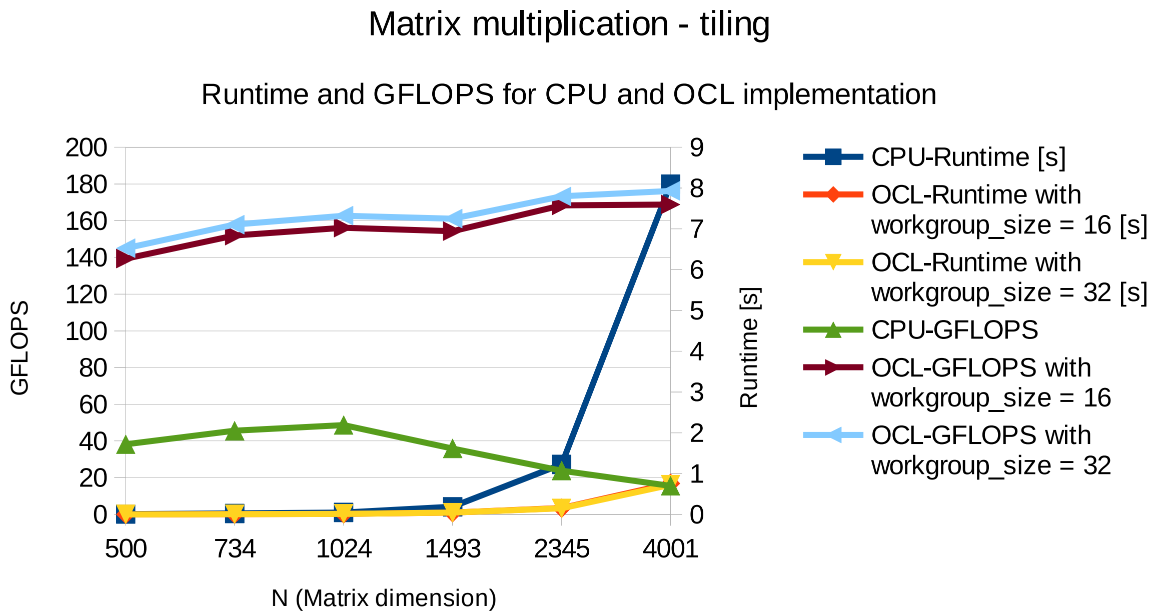

Figure 2: Runtime and GFLOPS of matrix multiplication for CPU and OCL (tiling) implementation