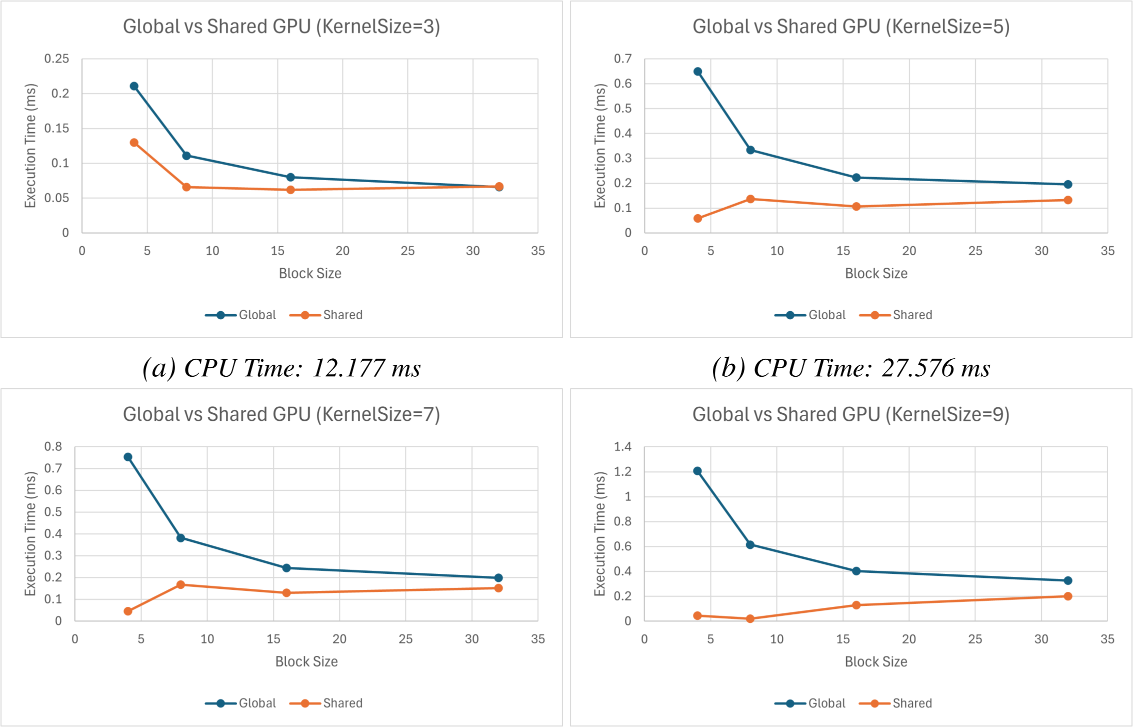
1. Assignment 2: CUDA Image Processing

1.1 Performance Analysis

Below are graphs of global and shared GPU execution times (ms) vs block size, with each graph having a fixed kernel size. Below each graph is its CPU time (independent of block size).



(a) CPU Time: 12.177 ms

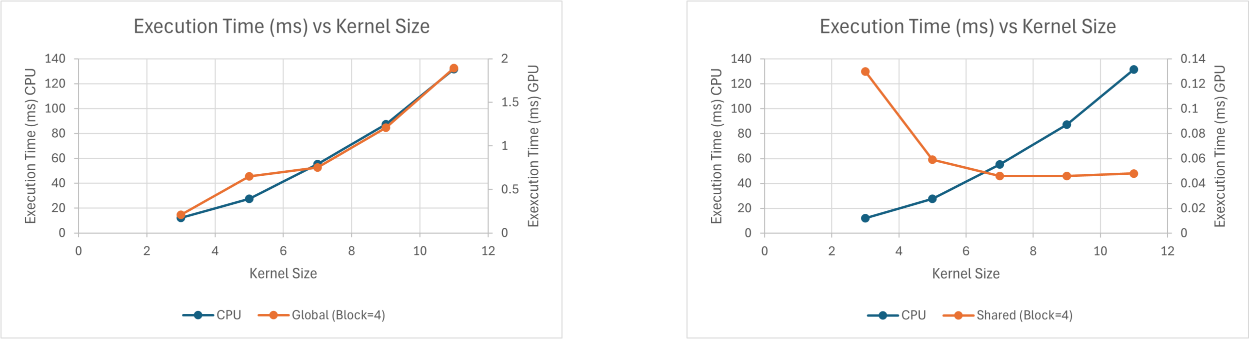
(b) CPU Time: 27.576 ms

(c) CPU Time: 87.273 ms

(d) CPU Time: 86.976 ms

Figure 1: Comparison of execution times vs block size for all implementations (different kernel sizes)

We observe from above, that shared memory implementation is consistently faster then the global memory implementation. Both are significantly faster then the CPU implementation. We observe that the shared memory implementation is rather consistent across different block sizes, while the global memory implementation shows exponential decrease based on block size. Note that the grid is directly effected by the block size as it is gridDim(ceil(width/block_size), ceil(height/block_size)).



(a) CPU vs Global Memory

(b) CPU vs Shared Memory

Figure 2: CPU and GPU execution times vs kernel size for fixed block size

From above, we first observe from figure 2(a) that the global and CPU implementations share very similar execution time growth patterns based on kernel size; following a quadratic shape. In contrast, figure 2(b) shows the shared memory implementation decreases execution time based on kernel size. One reason may be that larger kernels require more redundant memory access, meaning larger benefits for loading on shared memory. It is important to note, however, that the CPU execution times are still significantly longer then both global and shared memory implementations.