HW5: Array Reduction

1. The way the code is written, first call to syncthreads() is made when all threads read and add data from global memory and store the result into shared memory. This with jump = 512. Next, inside the loop, jump starts from 256 and goes down to 1. So 9 calls to syncthreads() are made. In total, block does syncthreads() **10 times.**

2. As long as jump >= 32, there is no thread divergence within a warp. So for all the iteration from jump = 16 to 1, there will be thread divergence. So all **5 warps** which are executed in last 5 iterations will suffer from it.