1)

We initilized the kernel with 1000 blocks and 1000 threads per block, where each block is responsible for a certain column, and each thread is responsible for a certain index in that column. For each block we initilized a shared array with one element. All the threads in that block are adding the which is the term in the index that the thread is responsible of. In each block, after every thread added their needed term, there is one thread that is responsible to add the sum of all the terms in his column to C - the output variable that should contain the sum of all the collumns.

And as requested we are doing the final p-root(How to write it lol) in the function dist\_gpu which is executed on the cpu.

You can see here the screenshot from our execution on the server:

