1. **How many times is each element of the input matrices loaded during the execution of the kernel?**

Each element of the input matrices, A is loaded n times and B is loaded m times during the execution of the kernel. In my implementation the thread (row, col) calculates the output value in matrix C. To do such calculation, this thread requires to access global memory and load 16 values from each matrix.

1. **What is the memory-access to floating-point computation ratio in each thread? Consider multiplication and addition as separate operations and ignore the global memory store at the end. Only count global memory loads towards your offchip bandwidth?**

The thread C performs 32 memory accesses to load the values of A and B. Along the computation it does 16 multiplications and 16 sums. Therefore, the total number of memory accesses and floating-point computations (per thread) are both 32. Then, the memory-access to floating-point computation ratio in each thread is 1.