1. What is the floating point to global memory access ratio (in OP/B) of each of the following matrix-matrix multiplication kernels?
2. The simple kernel described in Chapter 3, Multidimensional Grids and Data, without any optimizations applied.

**answer for a:**

Each thread computes one output element and fetches all its required data directly from global memory.

A thread performs a dot product to calculate one element of the output matrix. N multiplications and N additions, for a total of 2N FLOPs.

For each of the N steps in the dot product, the thread loads one float (4 bytes) from matrix A and one float (4 bytes) from matrix B.

Bytes per output element = N \* (4 bytes+ 4 bytes) = 8N bytes.

Ratio = 2N FLOPs / 8N bytes = **0.25 FLOP/B**

1. The kernel described in Chapter 5, Memory Architecture and Data Locality, with shared memory tiling applied using a tile size of 32 32.

**answer for b:**

The block computes 32 \* 32 = 1024 elements, each requiring 2N FLOPs.

FLOPs per block = 32 \* 32 \* 2N.

The block divides the computation into N/32 phases. In each phase, the 32x32 threads load one 32x32 tile from matrix A and one 32x32 tile from matrix B into shared memory.

Bytes loaded from A per phase = 32 \* 32 \* 4 bytes.

Bytes loaded from B per phase = 32 \* 32 \* 4 bytes.

Total bytes loaded per phase = 2 \* 32 \* 32 \* 4 bytes.

Total bytes loaded for the entire block = (N/32) \* (2 \* 32 \* 32 \* 4) = 256N bytes.

Ratio = 32 \* 32 \* 2N FLOPs / 256N bytes = **8 FLOP/B**

1. The kernel described in this chapter with shared memory tiling applied using a tile size of 32 32 and thread coarsening applied using a coarsening factor of 4.

**answer for c:**

The 32x32 thread block computes 32 \* 32 \* 4 output elements.

FLOPs per block = 32 \* 32 \* 4 \* 2N.

The block still iterates through N/32 phases along the k dimension.

It loads one 32x32 tile from matrix A. This tile is reused for all 4 output tiles. Bytes = 32\*32\*4 bytes.

It then loads four separate 32x32 tiles from matrix B, one for each of the 4 output tiles. Bytes = 4 \* 32\*32\*4 bytes.

Total bytes loaded per phase = (1 + 4) \* (32\*32\*4) = 5 \* 32 \* 32 \* 4 bytes.

Total bytes for the entire block = (N/32)\* (5 \* 32 \* 32 \* 4) = 640N bytes.

Ratio = 32 \* 32 \* 4 \* 2N FLOPs / 640N bytes = **12.8 FLOP/B**