Quiz Questions for Module 4

1. Assume that a kernel is launched with 1000 thread blocks each of which has 512 threads. If a variable is declared as a shared memory variable, how many versions of the variable will be created through the lifetime of the execution of the kernel?

    a. 1
    b. 1,000
    c. 512
    d. 512,000

    Answer: (B)
    Explanation: Shared memory variables are allocated to thread blocks. So, the number of versions is the number of thread blocks, 1,000.

2. For our tiled matrix-matrix multiplication kernel, if we use a 32X32 tile, what is the reduction of memory bandwidth usage for input matrices A and B?

    a. 1/8 of the original usage
    b. 1/16 of the original usage
    c. 1/32 of the original usage
    d. 1/64 of the original usage

    Answer: (C)
    Explanation: Each element in the tile is used 32 times, as explained in lecture 4.3.

3. For the tiled single-precision matrix multiplication kernel as shown in Lecture 4.4, assume that the tile size is 32X32 and the system has a DRAM burst size of 128 bytes. How many DRAM bursts will be delivered to the processor as a result of loading one A-matrix tile by a thread block?

    a. 16
    b. 32
    c. 64
    d. 128

    Answer: (B)
    Explanation. For a 32X32 A-tile, each row in the tile consists of 32 consecutive words and is accessed by a warp. The total amount of data in the row is just a single burst. We have 32 rows in a tile so there will be 32 bursts delivered to the processor.

4. Assume a tiled matrix multiplication that handles boundary conditions as explained in Lecture 4.5. Assume that we use 32X32 tiles to process square matrices of 1,000X1,000. Within EACH thread block, what is the maximal number of warps that will have control divergence due to handling boundary conditions for loading A tiles throughout the kernel execution?

   a. 32
   b. 24
   c. 16
   d. 8

   Answer: (A)
   Explanation: Control divergence happens due to the handling of the right edge. For thread blocks processing tiles that are totally within the valid range in the y-dimension, all 32 warps in a block will experience divergence at the right boundary. For the thread blocks that process the bottom A tiles on the right edge, only 8 warps will experience control divergence because all threads in the lower 24 warps will fail the boundary test.