CUDA Quiz Name _____

1. Which of the following statements are true with regard to compute capability in CUDA
   a. Code compiled for hardware of one compute capability will not need to be re-compiled to run on hardware of another
   b. Different compute capabilities may imply a different amount of local memory per thread
   c. Compute capability is measured by the number of FLOPS a GPU accelerator can compute.

2. Which of the following correctly describes a GPU kernel
   a. A kernel may contain a mix of host and GPU code
   b. All thread blocks involved in the same computation use the same kernel
   c. A kernel is part of the GPU's internal micro-operating system, allowing it to act as in independent host

3. True or false: Functions annotated with the `__global__` qualifier may be executed on the host or the device
   True
   False

4. True or False: The threads in a thread block are distributed across SM units so that each thread is executed by one SM unit.
   True
   False

5. Which of the following is *not* a form of parallelism supported by CUDA
   a. Vector parallelism - Floating point computations are executed in parallel on wide vector units
   b. Thread level task parallelism - Different threads execute a different tasks
   c. Block and grid level parallelism - Different blocks or grids execute different tasks
   d. Data parallelism - Different threads and blocks process different parts of data in memory

6. The style of parallelism supported on GPUs is best described as
   a. SISD - Single Instruction Single Data
   b. MISD - Multiple Instruction Single Data
   c. SIMT - Single Instruction Multiple Thread

7. Which of the following correctly describes the relationship between Warps, thread blocks, and CUDA cores?
   a. A warp is divided into a number of thread blocks, and each thread block executes on a single CUDA core
   b. A thread block may be divided into a number of warps, and each warp may execute on a single CUDA core
   c. A thread block is assigned to a warp, and each thread in the warp is executed on a separate CUDA core

8. Shared memory in CUDA is accessible to:
   a. All threads in a single block
   b. Both the host and GPU
   c. All threads associated with a single kernel

9. What strategy does the GPU employ if the threads within a warp diverge in their execution?
   a. Threads are moved to different warps so that divergence does not occur within a single warp
   b. Threads are allowed to diverge
   c. All possible execution paths are run by all threads in a warp serially so that thread instructions do not diverge

10. Which of the following does *not* result in uncoalesced (i.e. serialized) memory access on the K20 GPUs installed on Stampede
    a. Aligned, but non-sequential access
    b. Misaligned data access
    c. Sparse memory access