Julian Singkham

1. What is the compute capability of the NVIDIA Turing architecture?

7.5

2. Suppose you are launching a one-dimensional grid and block. If the hardware's maximum grid dimension is 65535 and the maximum block dimension is 512, what is the maximum number of threads that can be launched on the GPU?

$65535 * 512 = 33{,}553{,}920\ Threads$

3. Under what conditions might a programmer choose not to want to launch the maximum number of threads?

A programmer may not want to launch with the maximum amount of threads for high memory work and power efficiency. Threads have a minimum memory requirement that could interfere with high memory workloads.

4. What can limit a program from launching the maximum number of threads on a GPU?

Not having enough registers and memory for the threads could limit the maximum amount a program can launch.

5. What is shared memory?

Shared memory is memory that threads within the same block have access to. Shared memory is located on the CPU chip itself and is similar to cache memory.

6. What is global memory?

The total amount of DRAM memory on the GPU.

7. What is constant memory?

Super-fast memory (cache speed) that is 64KB in size. This memory is shared with all threads and allows for near instant fetching. Additionally, all threads can access constant memory at the same time.

8. What does warp size signify on a GPU?

Threads are grouped into blocks, and under the hood those blocks are subdivided by the warp size. The warp size dictates how many threads are in a sub-block.
32 threads (warp size) -> 16 sub-blocks -> 1 block

9. Is double-precision supported on GPUs with 7.5 compute capability?

Yes, all GPU's past 1.3 compute capability supports double-precision.