Eric Blasko

Dr. Zemoudeh

CSE 525

29 May 2019

Homework #3

**3. If the SM of a CUDA device can take up to 1536 threads and up to 4 shared blocks. Which of the following block configuration would result in the largest number of threads in the SM?**

1. 128 threads per block
2. 256 threads per block
3. **512 threads per block**
4. 1024 threads per block

Explanation:

1. 128 threads per block \* 4 blocks = 512 threads
2. 256 threads per block \* 4 blocks = 1024 threads
3. 512 threads per block \* 3 blocks = 1536 threads
4. 1024 threads per block \* 1 block = 1024 threads. More than 1 block goes over limit

Solution: C: 512 threads per block only uses 3 blocks, but fills all of the available thread contexts:

**4. For a vector addition, assume that the vector length is 2000, each thread calculates one output element, and the thread block size is 512 threads. How many threads will be in the grid?**

1. 2000
2. 2024
3. **2048**
4. 2096

Explanation:

1. The number of thread blocks needed to cover the entire output array is ceiling(2000/512) = 4. This means that there is a total of 4 blocks with 512 threads per block, giving 4 \* 512 = 2048 threads that will be created

Solution: C: 2048

**6. Assume that a CUDA kernel is launched with 1,000 thread blocks, with each having 512 threads. If a variable is declared as a local variable in the kernel, how many versions of the variable will be created through the lifetime of the execution of the kernel?**

1. 1
2. 1000
3. 512
4. **512000**

Explanation:

Each thread will create its own local version of the variable that it only it has access to. In this case 1000 blocks \* 512 threads = 512000 versions created

Solution: d) 51200

**7. In the previous question, if a variable is declared as a shared memory variable, how many versions of the variable will be created throughout the lifetime of the execution of the kernel?**

1. 1
2. **1000**
3. 512
4. 51200

Explanation:

Shared memory variables are allocated to each thread block. Since there are 1000 thread blocks, the variables are allocated per each block. Thus, there are only 1000 version of the variable

Solution: b) 1000

**9. A kernel performs 36 floating-point operations and 7 32-bit word global memory accesses per thread. For each of the following device properties, indicate whether this kernel is compute- or memory bound.**

1. Peak FLOPS= 200 GFLOPS, Peak Memory Bandwidth= 100 GB/s

(36 ops)/(200\*109 ops/s) = 0.18 \* 10-9

(7\*4 Bytes)/\*100 GB/s) = 0.28 \* 10-9

0.28 \* 10-9  > 0.18 \* 10-9

Solution: Thus, kernel is memory bound

1. Peak FLOPS= 300 GFLOPS, Peak Memory Bandwidth= 250 GB/s

(36 ops)/(300\*109 ops/s) = 0.12 \* 10-9

(7\*4 Bytes)/\*250 GB/s) = 0.112 \* 10-9

0.112 \* 10-9  < 0.12 \* 10-9

Solution: Thus, kernel is compute bound