CSCI-GA.3033-004

Graphics Processing Units (GPUs): Architecture and Programming

Fall 2015 – (90 minutes)

NAME: **ID:**

- The exam is open book/notes.
- If you have to make assumptions to continue solving a problem, state your assumptions clearly.
- You answer on the question sheet. You can use extra white papers if you want.

1. [1 point] We know in CUDA that commands in a stream (e.g. kernel launch, data movement between host and device, etc) are executed in order. Why is this restriction, given that it may lead to some performance loss?

2. [2 points] We have seen that if-else may lead to branch divergence in a warp due to lockstep execution of instructions. Now, suppose there is a thread that has and *if* without else. Can this also lead to performance loss in some cases? If yes, explain a scenario where there is performance loss. If no, explain why not. No need to write full code, just explain.

3. In OpenCL there is only one queue between the host and each device. So we cannot have several queues between the host and device like streams in CUDA.

a. [2 points] Does this restrict the performance of OpenCL? Justify.

b. [2 points] Will we gain any performance in OpenCL if we allow multiple queues between the host and the device? If yes, give a scenario where multiple queues give better performance. If no, explain why not.

4. [2 points] Beside overlapping data-transfer and computation, state two other scenarios where streams are useful.

-
-

5. [2 points] State two characteristics of a problem that makes it a good candidate for GPU instead of CPU.

-
-

6. [3 points] State three reasons you may want to have several kernels instead of one big kernel.

-
-
-

7. [4 points] Suppose NVIDIA decides to have larger warps in their future GPUs. Give two advantages of doing so, and two disadvantages.
Advantages:

-
-

Disadvantages

-
-

8. We have discussed a lot the importance of memory coalescing. Also we say that having an L2 cache (servicing all the SMs) helps in global memory coalescing.

a. [1 point] How does L2 helps in memory coalescing?

b. [1 point] Does the existence of L2 mean that the programmer does not need to pay attention to global memory access to be coalesced? Explain.