CO332 – Heterogenous Parallel Computing
A1 – CUDA introduction

Submitted by :
Sagar Bharadwaj KS, **15CO141**
Aneesh Aithal, **15CO107**

Q1

1. **Architecture** **:** TESLA microarchitecture
   Compute Capability : 3.5

2. Maximum Block Dimensions : (1024 X 1024 X 64)

3. Maximum threads = Maximum Grid Dim * Maximum Block Dim
= 65535 * 512
= **33553920**

4. A programmer might not want to launch the maximum permitted number of threads if each of the thread demands excessive resources from the GPU. Increasing the number of threads does not necessarily increase the efficiency or throuhgput of the program.

Launching a lot of resource intensive threads would not only reduce the number of parallel executions that can take place in the GPU, it also adds the overhead of context swithcing between numerous threads making it slower than sequential execution.

5. The ‘register pressure’ exerted by each thread is what limits a program from launching maximum number of threads on a GPU. The number of registers available per block is limited. So if all threads belonging to a single block attempt to use up a lot of registers, it becomes impossible to parallely execute threads more than a limiting value.

Every Nvidia CUDA toolkit manual includes the effect of ‘register pressure’ on thread dimensions. Higher the register pressure, lower the actual thread dimensions are from the theoritical value.

6. Shared Memory : Memory allocated for a block which is shared by all threads in a block is called shared memory. Shared memory speeds up inter thread communication. Shared memory of one block is not accessible by another block.

Shared memory in queried GPU : **49512 Bytes**

7. Global memory : Memory accessible by all the blocks and threads executing on a GPU is called Global memory. The input given by the host to the device and the final output to be transferred back to the host is stored in Global memory.

Global memory in queried GPU : **11995578368 Bytes**

8. Constant Memory : We use constant memory for data that will not change over the course of a kernel execution. NVIDIA hardware provides 64KB of constant memory that it treats differently than it treats standard global memory. In some situations, using constant memory rather than global memory will reduce the required memory bandwidth.

Constant memory in queried GPU : **65536 Bytes (64kB)**

9. Warp size is the number of threads in a warp, which is a sub-division used in the hardware implementation to coalesce memory access and instruction dispatch.

Warp size : **32**

10. Yes. Double precision is supported on queried GPU.

Q2

Code attached.

Concepts of synchronisation of threads and array reduction were used to calculate sum of an array.

Q3

Code attached.

Q4

Code attached.

Dimensions of Matrix : M * N

1. Floating Point Operations **= M * N** (One operation per addition)

2. Global memory reads = **2* M * N** (One read per element in I/P matrices)

3. Global memory writes = **M * N** (One write per element in O/P matrix)