Name: **Enrolment No:**	UPES UNIVERSITY WITH A PURPOSE

UNIVERSITY OF PETROLEUM AND ENERGY STUDIES
Online End Semester Examination, December 2021

Course: GPU Programming **Semester: VII**
Program: B. Tech. G&G **Time 03 hrs.**
Course Code: CSGG4008 **Max. Marks: 100**

SECTION A

1. **Each Question will carry 4 Marks**
2. **Instruction: Write short answers**

S. No.	**Question**	**CO**
Q 1	Explain CPU, GPU and GPGPU.	**CO1**
Q2	1. CUDA is a parallel computing platform and programming model a. True b. False 2. Host codes in a CUDA application can Transfer data to and from the device a. True b. False 3. The kernel code is only callable by the host a. True b. False 4. The kernel code is executable on the device and host a. True b. False	**CO2**
Q3	1. _____ is a form of parallelization which relies on splitting the computation by subdividing data across multiple processors. a. Data parallelism b. Task parallelism c. Function parallelism d. Object parallelism 2. Data parallelism performed ____, Task parallelism performed _____ a. Synchronous, Asynchronous Computation b. Synchronous, Synchronous Computation c. Asynchronous, Synchronous Computation d. Asynchronous, Asynchronous Computation. 3. CUDA is developed by _____ a. AMD b. NVIDIA c. Intel d. RAPIDS	**CO2**
Q4	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. d. None of the above 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	**CO3**

	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. a. True b. 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. a. True b. False	
Q5	1. For a vector addition, assume that the vector length is 4000, each thread calculates one output element, and the thread block size is 1024 threads. How many threads will be in the grid? A. 2000 B. 3000 C. 1024 D. 4096 2. If a CUDA device's SM (streaming multiprocessor) can take up to 1536 threads and up to 6 thread blocks. Which of the following block configuration would result in the most number of threads in the SM? A. 128 threads per block B. 256 threads per block C. 512 threads per block D. 1024 threads per block	**CO4**

SECTION B

1. **Each question will carry 10 marks**
2. **Instruction: Write brief notes**

Q 6	Explain with examples, how do you achieve computations in terms of (i) Concurrency (ii) Parallelism with CPU's.	**CO1**
Q 7	Explain CUDA and how CUDA is used to achieve parallelism?	**CO2**
Q 8	Explain host device architecture in GPU programming using block diagram and elaborating each component.	**CO3**
Q 9	Write a code to ~~dem~~onstrate concurrency using threads with que OR Write a code to demonstrate multiprocessing with execution time	**CO4**

SECTION C

1. **Each Question carries 20 Marks.**
2. **Instruction: Write long answer.**

Q 10	Explain Open CL memory model with pictorial diagram and explain each component.	**CO3**
Q11	Explain memory architecture of host and kernel with pictorial diagram and write a code snippet to demonstrate data transfer between host memory and kernel memory using CUDA. Or Explain threads, blocks and relation between them in CUDA perspective of execution. Write a code snippet to demonstrate the same.	**CO4**