A **Reg. No. :** | | | | | | | | | | | | |

Question Paper Code: U8G63

B.E. / B.Tech. DEGREE EXAMINATION, APRIL / MAY 2025

One Credit Course

Artificial Intelligence & Machine learning

21UAM863 – GPU PROGRAMMING

(Regulations 2021)

Duration: 1.30 hours Maximum: 50 Marks

Answer ALL Questions

PART A - (10 x 1 = 10 Marks)

1. Which GPU architecture introduced CUDA programming? CO1- U

   a) Kepler b) Tesla c) Maxwell d) Turing

2. Threads in CUDA execute in parallel with in CO1- U

   a) Block b) Grid c) CPU d) Kernel

3. Which function is used to specify the number of blocks in a grid? CO1- U

   a) blockDim.x b) gridDim.x c) threadIdx.x d) warpSize

4. What is the primary challenge of Multi-GPU computing? CO1- U

   a) Lack of GPU memory b) Synchronization and data transfer overhead

   c) Slow CPU processing d) Inability to execute parallel tasks

5. Which technology allows multiple GPUs to communicate in NVIDIA systems? CO1- U

   a) Hyper-Threading b) NVLink c) DirectX d) OpenMP

6. Problem decomposition is most useful in which type of programming? CO1- U

   a) Use shared memory as a cache b) Use more CPU processing

   c) Increase the number of threads d) Use more warp divergence

7. The effect of poor load balancing in parallel programs is: CO1- U

a) Increased CPU usage

b) Decreased execution efficiency due to thread underutilization

c) Enhanced memory access speed

d) Better management of shared resources

8. An algorithm's performance is most affected by: CO1- U

a) The number of threads used

b) The algorithm's time and space complexity

c) The number of available memory registers

d) The processor's brand

9. Which bus interface is commonly used to connect multiple GPUs in a system? CO1- U

a) SATA b) PCIe (Peripheral Component Interconnect Express)

c) USB d) HDMI

10. To eliminate data dependencies, one approach is CO1- U

a) Use only sequential execution

b) Increase the number of threads

c) Apply synchronization mechanisms like locks and barriers

d) Reduce the number of CPU cores

PART – B (2 x 20= 40 Marks)

11. (a) (i) Given a problem that requires heavy computation (e.g., image processing, deep learning inference), determine whether a GPU or CPU is more suitable and justify your choice. CO2-App (10+10)

(ii) Implement a basic parallel reduction algorithm (sum reduction) and optimize it using warp-level primitives.

Or

(b) Implement CUDA-based parallelism with graphical processing unit, for solving the parallel matrix multiplication CO2-App (20)

12. (a) Describe the concept of inter-process contention in multi-core processors. How does contention for shared resources like cache memory or bus bandwidth impact application performance, and how can software optimizations alleviate these issues? CO2-App (20)

Or

(b) Identify the fault tolerance techniques help avoid errors in distributed computing systems? Discuss mechanisms such as redundancy, checkpoints, and replica consistency, and explain how they minimize the impact of system failures. CO2-App (20)