Question Paper

Exam Date & Time: 18-Jun-2024 (02:30 PM - 05:30 PM)

MANIPAL ACADEMY OF HIGHER EDUCATION

SIXTH SEMESTER B.TECH. (COMPUTER SCIENCE AND ENGINEERING) DEGREE EXAMINATIONS - JUNE 2024
SUBJECT: CSE 3252/CSE_3252 - PARALLEL COMPUTER ARCHITECTURE AND PROGRAMMING

Marks: 50 **Duration: 180 mins.**

Answer all the questions.

1A)	Differentiate between the CPU and GPU design philosophies with neat diagram.	(5)
1B)	Using CUDA API functions for managing data in the device memory develop Cuda program for vector-vector addition.	(3)
1C)	How does data inconsistency issue is handled in CUDA programming? Illustrate with example code.	(2)
2A)	Differentiate between following communication routines with the help of example code: i) MPI_Ssend and MPI_Bsend ii) MPI_Bcast and MPI_Reduce	(4)
2B)	Write a MPI program to read a 3 X 3 matrix. Enter an element to be searched in the root process. Find the number of occurrences of this element in the matrix using three processes. Also, handle different errors using error handling routines.	(3)
2C)	What is 1D sequential convolution? Find out output array P[] after performing 1D sequential convolution using N=[3,5,2,4,6,1], M=[2,1,–1]	(3)
3A)	Solve the problem of reversing a string by writing an OpenCL program which with the help of kernel function reverses the string. Also, find the time taken for the GPU to execute the kernel function.	(5)
3B)	Define ***compute to global memory access (CGMA)*** ratio. Assume that a GPU has a global memory bandwidth of 160 GB/s. For a single-precision (4 byte) floating-point value and a CGMA ratio of 1, calculate the number FLOPS carried out by a kernel in this GPU. Calculate the CGMA if the number floating-point-operations per second of this GPU is 1500. Explain any one method of increasing CGMA ratio.	(3)
3C)	Write the CSR format for the following sparse matrix along with the explanation. 5 0 0 0 0 8 0 0 0 0 3 0 0 6 0 0	(2)
4A)	Compare the shared memory and registers used in the CUDA programming by mapping to the memory of the von Neumann model. Clearly show how the operations are performed in case of "fadd" instruction.	(4)
4B)	Develop a CUDA kernel to perform tiled 1D parallel convolution which uses shared memory.	(4)
4C)	Following function prototype represents 1D parallel convolution kernel. Mention the best candidate for constant memory and caching? Justify your answer. Also mention the code snippet required in the host code to change the candidate to constant memory. __global__ void convolution_1D_basic_kernel(float *N, float *M, float *P, int Mask_Width, int Width);	(2)
5A)	Explain how constant memory can be used to increase the performance of the massively parallel processors? Write the kernel function for 1D convolution using constant memory in Cuda.	(4)

5B) Illustrate the importance of the shared memory and __syncthreads() in tiled matrix multiplication by writing complete tiled matrix multiplication kernel code. (4)

5C) Assume that a grid has 128 blocks arranged in 2D and grid length (x direction) is 32. Threads in a block are arranged in 2D with block height(y direction) is 5. Each block contains 30 threads. Fill the following table with appropriate values. Show the calculations along with the required formulae for the last column. (2)

Note: Global and local thread indexing and block indexing shown in the table starts with 0. For blocks and threads(x,y,z) notation is used.

gridDim.x	gridDim.y	blockDim.x	blockDim.y	Global thread id of a Thread (2,3) in block (1,3)

-----End-----