Question Paper

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

MANIPAL ACADEMY OF HIGHER EDUCATION

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

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

Answer all the questions.

Missing data may be suitably assumed.

1A) Illustrate how digital computers are classified in Flynn's classification along with their applications? Also, mention on what basis they are classified. (5)

1B) Illustrate the device memory creation process and how to free the allocated device memory in CUDA with appropriate APIs along with their parameters. (3)

1C) Write an CUDA kernel to accept an integer array of size arr_size, that is multiples of 32. Each thread finds the average of consecutive 32 elements in the input array and stores it in an output array in the appropriate position. Also, write the block size used. (2)

2A) Write a MPI program to read integer matrix A and B of size 4x4. Modify matrix B by adding minimum element in every row with index say *r* of B matrix by the maximum element in the column having the same index *r* of A matrix. If principal diagonal element of B is not minimum element in row then replace it by 0. Implement this program using collective communications and without transpose using 4 processes (including root). (4)

A				B				Result (B)			
1	2	3	4	4	3	2	5	0	3	7	5
5	3	2	6	4	2	5	3	4	8	5	3
2	6	9	3	1	2	3	4	10	2	0	4
3	2	5	7	1	2	2	3	8	2	2	0

2B) Discuss the concept of collective communication in MPI and provide examples of collective operations along with syntax and detailed explanation of four different functions. (3)

2C) Develop a CUDA Kernel for 1D convolution with a condition that each output value is computed by a thread. Trace the thread wise computation for the following inputs:
Signal: [1, 2, 3, 4, 5, 6, 7, 8]
Kernel: [1, -1, 2] (3)

3A) Write an OpenCL program which with the help of a kernel function that takes a string S as input and one integer value N, produces string N times as follows in parallel:
Input : S = Hello N = 3
Output String: HelloHelloHello
Note : Each work item copies same character from the Input N times to the required position). Also find the execution time taken by this kernel. (5)

3B) (3)

Consider the kernel below

```
__global__ void Add(float* M) {
 int index = blockIdx.x * blockDim.x +
threadIdx.x;
        M[index] = M[index] + M[0];
}
```

What is CGMA for the above code? Justify your answer along with the definition of CGMA. Also, explain how we can improve the CGMA ratio in the above case?

3C) Given a sparse matrix of integers with m rows, n columns, and z non-zeros, how many integers are needed to represent the matrix in CSR format. If the information provided is not enough, indicate what information is missing. (2)

4A) Consider a CUDA program with a grid consisting of 1024 blocks arranged in a 2D configuration. The grid size in the x-direction is 32, and each block contains threads arranged in 2D with dimensions (4, 8) for x and y, respectively. Each block contains 32 threads.
Determine the global thread ID for the following scenarios:
I) For thread (2, 3) in block (15, 8).
II) For thread (1, 7) in block (20, 5).
III) For thread (3, 2) in block (29, 14).
IV) For Thread (3, 1) in block (31, 15)
Show your calculations for each case, considering the zero-based indexing convention. (4)

4B) The following function prototype is used for 1D parallel convolution kernel. Mention the method which changes this prototype to improve the performance by utilizing the cache memory? Explain it with the required code snippet including the kernel. (4)

```
__global__void
convolution_1D_basic_kernel(float *N, float
*M, float *P, int
Mask_Width, int Width);
```

4C) Illustrate with an example code demonstrating the usage and significance of the cudaMemcpyToSymbol() function in CUDA (2)

5A) Design a CUDA kernel for tiled parallel 1-dimensional convolution with halo elements. Also write and explain how to use constant memory for mask array. (4)

5B) Write a tiled matrix multiplication CUDA kernel to multiply two N×N matrices (4)

5C) There are 3D grids of size (3, 4, 2) and each block has 2D structure of size (2, 4). Draw this structure. Give a generalized formula for calculating global thread id. (2)

-----End-----