Reg. No.										

MANIPAL INSTITUTE OF TECHNOLOGY
MANIPAL

VI SEMSTER B. TECH. (COMPUTER SCIENCE & ENGINEERING)
ENDSEM EXAMINATION (MAY-2023)

SUBJECT: PARALLEL COMPUTING ARCHITECTURE AND PROGRMMING-PCAP
[CSE 3252]

Time: 3 hrs. MAX. MARKS: 50M

Instructions to Candidates:

- Answer **ALL** the questions.
- Missing data may be suitably assumed.

Q. No.	Questions	Marks	CO	BL
Q1 a)	i) Explain two types of spatial parallelism with example. ii) Suppose process 0 has variable A and process 1 also has variable A. Design an MPI program to exchange these values between the processes.	5M	CO 1,2	B2 B6
Q1 b)	Identify Factors to be considered by the application developers to choose the processors for running their applications.	3M	CO1	B3
Q1 c)	Show the difference between pipelined and non-pipelined processor with space-time diagrams. Assume 3 instructions are to be processed and each instruction is processed in 4 stages.	2M	CO1	B2
Q2 a)	Using collective communication routines, implement an efficient MPI program which reads matrix *A* of size 4*x*4 and it produces a resultant matrix *RES* of size 4*x*4 such that every row elements of *A* are added with a key value. The key value for the first row is the minimum element of the last row and the key value for remaining row is the minimum element from the previous row of matrix *A*. Make use of 4 processes (including root) to perform this task. Example: A → RES 1 2 3 4 → **3 4 5 6** 5 6 7 8 → 6 7 8 9 2 4 3 5 → 7 9 8 10 **2** 3 4 6 → 4 5 6 8	5M	CO2	B3
Q2 b)	Which feature of CUDA support executing same application on different CUDA hardware? Explain in detail with neat diagram.	3M	CO5	B2
Q2 c)	Discuss the two advantages of message passing models.	2M	CO2	B6
Q3 a)	Design an OpenCL kernel code that takes a string ***S*** consisting of ***N*** number of words. It also takes another integer value ***pos***. It produces resultant string ***RS*** by taking one character from every word of ***S*** in the position ***pos***. If ***pos*** value is greater than any word length, then put a '+' in ***RS*** for the	5M	CO3	B6

<table>
<tr><td></td><td>corresponding position. The kernel code should produce every character of RS in parallel.
Eg: S: This is PCAP asSignment pos: 3 RS: i+AS
Write all the statements that are associated with buffer for the above kernel. What is the global work size? Assume the context context and command queue cmdQueue is already created.</td><td></td><td></td><td></td></tr>
<tr><td>Q3 b)</td><td>Create OpenCL kernels to sort N numbers using odd-even transposition sorting. How will you create the kernel objects for these kernels. Write the part of host code which will clearly show how many iterations are required for sorting and how the kernels are invoked. Write the global work size. Assume the context context, command queue cmdQueue and program object program is already created.</td><td>3M</td><td>CO3</td><td>B6</td></tr>
<tr><td>Q3 c)</td><td>How does OpenCL support heterogeneous computing and explain the implications for performance?</td><td>2M</td><td>CO3</td><td>B2</td></tr>
<tr><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>Q4 a)</td><td>Assume that a grid has 1024 blocks arranged in 2D and grid length in x direction is 32. Threads in a block are arranged in 3D with block length in x direction and block height in y direction are 4 and 2 respectively. Each block contains 32 threads. Fill the following table with appropriate values. Show the calculations along with the required formulae for the last two columns.
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.
<table>
<tr><td>gridDim.x</td><td>gridDim.y</td><td>blockDim.x</td><td>blockDim.y</td><td>blockDim.z</td><td>Global thread id of a thread (2,1,3) in block(28,7)</td><td>Global block id of a block (22,18)</td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
</table></td><td>4M</td><td>CO4</td><td>B2</td></tr>
<tr><td>Q4 b)</td><td>Illustrate the difference between a CUDA thread and a CPU thread? Explain how does CUDA differ from OpenCL, and in what situations might you choose one over the other?</td><td>4M</td><td>CO4</td><td>B2</td></tr>
<tr><td>Q4 c)</td><td>Illustrate barrier synchronization function in CUDA explaining usage and challenges.</td><td>2M</td><td>CO5</td><td>B2</td></tr>
<tr><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>Q5 a)</td><td>Create a CUDA program to check the validity of n dates provided as a single string where each date requires 10 characters in the DD-MM-YYYY format. Each thread checks one date and Stores the answer(valid/ or not) in an appropriate position in a boolean array and host code displays the output as shown below . Take leap year into consideration. Use 1D grid and 1D block. Year should be between 1900 to 2023 including 1900 and 2023.</td><td>5M</td><td>CO5</td><td>B6</td></tr>
<tr><td>Q5 b)</td><td>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.</td><td>3M</td><td>CO5</td><td>B6</td></tr>
<tr><td>Q5 c)</td><td>Explain how the Compute to Global Memory Access ratio affects the performance of a kernel.</td><td>2M</td><td>CO5</td><td>B5</td></tr>
</table>