Reg. No.										

प्रज्ञानं ब्रह्म
Manipal
INSPIRED BY LIFE

MANIPAL INSTITUTE OF TECHNOLOGY, MANIPAL 576104

(Constituent College of Manipal University)

KNOWLEDGE IS POWER

SECOND SEMESTER M.TECH. (Network Engg.) DEGREE END SEMESTER EXAMINATION, May -2016
SUBJECT: ELECTIVE -I PARALLEL COMPUTATION AND APPLICATIONS (ICT-574)
(REVISED CREDIT SYSTEM)

TIME: 3 HOURS **10/05/2016** **MAX. MARKS: 50**

Instructions to candidates

- Answer ANY **FIVE FULL** questions
- Missing data, if any, may be suitably assumed

1A. With an example, explain the CUDA C programming model. Highlight the necessary changes to be made to the regular C program. With the neat diagram explain the hardware, where the threads execute the integer/floating point operations.

1B. What is snooping? How can you classify the snooping protocols? Explain.

1C. Write the CUDA C code snippets to allocate 2KB of floating type data to be held onto the device's constant memory and shared memory.

[5+3+2]

2A. With suitable diagrams explain the key features of Kepler GPU architecture.

2B. With suitable code snippets, explain any three CUDA C variable type qualifiers. Write the CUDA C kernel to add two matrices and store the result in third matrix. Assume that multiple 1D blocks of threads are launched to handle the huge data input. Every thread should compute the sum of each row.

2C. Write the CUDA C kernel function to scale a black and white picture to the given scale factor.

[5+3+2]

3A. Write the CUDA C kernel to compute exclusive prefix maximum scan for 1D array elements. Assume that multiple blocks of threads are launched to handle the input data. Given the block size of 8, show the execution phases of the above kernel for the input: 2, 1, 3, 1, 0, 4, 1, 2, 0, 3, 1, 2, 5, 3, 1, 2.

3B. With the neat diagram, explain how Nehalem core pipeline unit performs fetching and pipe-lining of instructions.

3C. Write the CUDA C program to retrieve the maximum threads allowed in the block and total number of registers available in each SM.

[5+3+2]

4A. Write the CUDA C kernel to perform matrix multiplication using shared memory. Write the necessary comments highlighting each steps/phases. For the above kernel, explain how shared memory usage improves the performance by considering 4 x 4 2D input array using 2 x 2 blocks.

4B. Write the execution phases to find the sum of : 7, 6, 10, -18, 0, -9, 15 12, 5, 3, 4, 8, 7, 3, 1, 2 using two reduction approaches.

4C. List out the major differences between Sandybridge and Nehalem micro-architectures.

[5+3+2]

5A. With suitable code snippets, explain any five Thrust algorithms.

5B. With an example for each, explain loop fusion, loop unrolling and memory coalescing optimization techniques.

5C. With suitable examples, explain the two forms of parallel computing.

[5+3+2]

6A. Write the complete CUDA C program to perform convolution operation on 1D input data N with the mask M. Shared memory has to be used to reduce the global memory traffic, mask has to be placed in constant memory and each thread should access the global memory only once. Multiple blocks should be launched to handle the huge input array.

6B. Write the complete CUDA C program using CUDA libraries to compute mean and variance of a dataset X of cardinality N.

6C. With an example, explain how mapping of threads to 2D data can be performed in CUDA.

[5+3+2]

******** ALL THE BEST ********