Reg. No. | | | | | | | | | | |

MANIPAL INSTITUTE OF TECHNOLOGY

MANIPAL

A Constituent Institution of Manipal University

V SEMESTER B.TECH. (COMPUTER AND COMMUNICATION ENGINEERING)

MAKEUP EXAMINATIONS, DEC. 2016

SUBJECT: PARALLEL PROGRAMMING [ICT 355]

REVISED CREDIT SYSTEM
(31/12/2016)

Time: 3 Hours MAX. MARKS: 50

Instructions to Candidates:

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

1A. Write efficient kernel function calls to perform multiplication of two vectors assuming the function name as VectorMultiply. Include the execution configuration parameters and necessary variables/parameters definitions assuming that there is only one connected device.
i) If the number of elements in the input array is same as the maximum number of blocks along X direction of a grid that is supported by the hardware.
ii) If the number of elements in the input array is same as maximum number of threads per block supported by the hardware.
iii) If the shared memory need to be dynamically allocated and is equal to the number of threads within the block.
vi) If the input arrays are assumed to be in constant memory of the device and the output need to be stored in device's global memory. **5**

1B. With a neat diagram explain the OOEE of Nehalem micro-architecture. **3**

1C. Write the efficient CUDA C kernel function to find dot product of two 1D input vectors. Assume the input data can be handled by a multiple blocks of threads. **2**

2A. With suitable diagrams explain the different types of graphic shaders. Explain the need of unified shader architecture. **5**

2B. With necessary code snippets explain how Thrust kernels can be replaced with CUDA kernel and vice versa. **3**

2C. With suitable example, explain the effect of serial code on the speedup using Amdahls law. **2**

3A. Write an efficient CUDA C program to perform matrix multiplication using multiple blocks and shared memory. Assuming 4 x 4 input with the block size of 2 x 2, write the execution phases for the thread10 and thread01 of block11. **5**

3B. With example code snippets, explain how GPU memory can be a limiting factor to parallelism. 3

3C. Write the CUDA C program to retrieve the total amount of constant memory available on the device and shared memory available on each SM. 2

4A. With code snippets explain the CUDA device memory model. Highlight the methods to allocate and copy the data from host to various device memories. 5

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 an efficient reduction approach. The kernel is launched with <<< 4,4 >>> execution configuration parameters. 3

4C. Write the complete Thrust program to find the sum of two matrices A, B of dimension N x N and store the result in C. 2

5A. 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: 12, 11, 13, 11, 10, 14, 11, 12, 10, 13, 11, 12, 15, 13, 11, 12. 5

5B. What is snooping? How can you classify the snooping protocols? Explain. 3

5C. With an example for each, explain any two optimization techniques to handle memory transfers and bandwidth. 2

6A. Write complete CUDA program to perform convolution operation on an input 1D array of size N with the Mask of M element. Use multiple blocks and shared memory to make efficient use of hardware. Assume that shared memory size is same as the block size. 5

6B. With an example, explain how warp divergence affects the performance. Also suggest the methods to overcome the divergence. 3

6C. Explain with suitable diagrams the working of Inclusive and Exclusive caches with respect to cache miss and cache hit. 2