Reg. No.

MANIPAL INSTITUTE OF TECHNOLOGY

MANIPAL

A Constituent Institution of Manipal University

II SEMESTER M.TECH. (SOFTWARE ENGINEERING/COMPUTER NETWORKING AND ENGINEERING)

END SEMESTER EXAMINATIONS, APRIL/MAY 2017

SUBJECT: PROGRAM ELECTIVE III - PARALLEL COMPUTATION AND APPLICATIONS [ICT 5241]

REVISED CREDIT SYSTEM (29/04/2017)

Time: 3 Hours MAX. MARKS: 50

Instructions to Candidates:

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

1A. With suitable diagrams, explain the key features of Kepler GPU architecture. 5

1B. With suitable code snippets, explain how interoperability can be achieved between CUDA and Thrust. 3

1C. With suitable diagrams, explain the parallel computer classifications that are adopted in most of current supercomputers and modern GPUs. 2

2A. Write the CUDA kernel to reduce 1D input array of size N using shared memory with least divergence. Given the input array kernel [2, 1, 8, 1, 0, 4, 4, 2, 0, 3, 1, 2, 5, 3, 1, 2] and configuration parameter <<<4,4>>>, write the execution phases for the above kernel. 5

2B. With a neat diagram explain the OOEE of Nehalem micro-architecture. 3

2C. Can memory be the only limiting factor for parallelism on GPU? Justify your answer with an example. 2

3A. With suitable diagram, explain the cache coherency protocol that is adopted in Nehalem micro-architecture. 5

3B. With a suitable example CUDA program, explain how to allocate memory for any three CUDA variables. 3

3C. What is thread divergence? Explain how it can be minimized with an example. 2

4A. Write the CUDA C kernel to compute inclusive prefix scan for 1D array elements. Show the execution phases of the above kernel for the input: 2, 6, 30, -8, 0, -9, 1 2, 5, 3, 4, 8, 7, 3, 1, 2. Assume the kernel is launched with <<< 8, 2 >>> execution configuration parameters. 5

4B. With an example explain the difference between task and data parallelism. 3

4C. With suitable code snippets, explain how error handling can be done in CUDA C programs. 2

5A. Write the complete CUDA program to perform convolution on 1D input X of dimension N and mask K of dimension M (such that M<N) using the shared and constant memory. 5

5B. Write complete CUDA program using Thrust library to find the division of two input arrays A and B of length N and store the result in C on a GPU (i.e. $\forall$ a$\in$A and $\forall$ b$\in$B , c=a/b, only if b is non-zero). 3

5C. Is it possible to achieve synchronization within the CUDA blocks? Justify your answer with an example. 2