Reg. No.										

MANIPAL INSTITUTE OF TECHNOLOGY

MANIPAL

A Constituent Institution of Manipal University

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

MAKEUP EXAMINATIONS, DECEMBER 2017

SUBJECT: PARALLEL PROGRAMMING [ICT 3153]

REVISED CREDIT SYSTEM
(23/12/2017)

Time: 3 Hours MAX. MARKS: 50

Instructions to Candidates:

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

1A. Write the complete CUDA program to find the absolute minimum element in the input vector of dimension N. Assume multiple blocks are used to handle the large input and shared memory is used to reduce the global memory traffic. 5

1B. Explain Kepler's dynamic parallelism with quicksort example. 3

1C. With an example explain how synchronization is handled within CUDA blocks of threads. 2

2A. Write efficient CUDA kernel functions to perform below operations using parallel approach. Assume multiple blocks are launched.

i) Let A be the input matrix. Find transpose of matrix represented by A^{T}.

ii) Given the vectors X={x_1, x_2,... x_N} Y= {y_1, y_2, ..., y_N}, calculate covariance where m_x and m_y represents the mean of X and Y respectively.

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - m_x)(y_i - m_y)$$ 5

2B. Write the CUDA kernel to implement an efficient histogram algorithm for an input array of integers ranging between 0 to 255. Assume each integer will be mapped to a single bin. 3

2C. Differentiate between GPU and CPU micro-architectures. 2

3A. Differentiate between the two broad categories of cache coherency protocols. With the help of a neat diagram, explain the MESIF protocol adopted in Nehalem micro-architecture. 5

3B. Write the CUDA kernel to compute the cosine value for radian input using the below equation. Assume multiple blocks are launched.

$$\cos(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{(2n)!}$$

3

3C. Write the execution phases of the CUDA kernel that computes the prefix sum of input vector [-4, 1, -5, 8, 3, 1, -5, 3, 8, 6, 8, 7, 1, 3, 6, 4], that is launched using <<4,4>>> execution configuration parameters. 2

4A. Explain the various types of CUDA memories, their scope and life time with an example program. How to detect the memory limitations for any connected GPU? 5

4B. Write the CUDA kernel to perform 2D convolution on the input 2D grayscale image of (height x width) dimension using the mask width of (5x5) dimension. 3

4C. For the below kernel code snippet, if the block size is 768 and warp size is 32, how many warp/s will have divergence during the iteration where stride is equal to i)1 ii)16 iii) 64 iv) 512?

```
unsigned int t = threadIdx.x;
unsigned int start = 2*blockIdx.x*blockDim.x;
partialSum[t] = input[start + t];
partialSum[blockDim.x+t] = input[start+ blockDim.x+t];
for (unsigned int stride = 1; stride <= blockDim.x; stride *= 2)
{
__syncthreads();
if (t % stride == 0) {partialSum[2*t]+= partialSum[2*t+stride];}
}
```

2

5A. Write the complete and efficient CUDA program to perform tiled prefix sum on input 1D vector of dimension N. Assume that shared memory is dynamically allocated depending on the device capability. **5**

5B. Write the CUDA Thrust program to find the variance of real numbers vector given by $\sigma^2 = (1/n)\sum_{i=1}^{n}(x_i - \mu)^2$ **3**

5C. In a certain computation, 80% of the work is vectorizable. Of the remaining 20%, half is parallelizable for an MIMD machine. What are the speedups over sequential execution for a 10 processing elements MIMD machine on this computation? **2**