Reg. No.										

MANIPAL INSTITUTE OF TECHNOLOGY

MANIPAL

A Constituent Institution of Manipal University

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

END SEMESTER EXAMINATIONS, NOVEMBER 2017

SUBJECT: PARALLEL PROGRAMMING [ICT 3153]

REVISED CREDIT SYSTEM
(20/11/2017)

Time: 3 Hours MAX. MARKS: 50

Instructions to Candidates:

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

1A. With a suitable example, explain the five stages instruction pipelining that is implemented w ithin a modern CPU architecture. A lso dis cuss how t his i s handled by Intel Nehalem core pipeline hardware. 5

1B. Explain how CUDA threads can be mapped to multi-dimensional data by converting the serial code snippet given below to efficient parallel CUDA kernel.

```
for i := 2 step 1 until N
for j := 2 step 1 until N
X[i, j] := (X[i, j-1] + X[i, j+1])/2;
```

3

1C. Differentiate between task parallelism and data parallelism with an example code snippet. 2

2A. Write the efficient CUDA kernel functions to perform the operations given below using parallel approach. Shared memory is used to reduce the global memory traffic and multiple blocks are launched.

i) $S=\sum_{i=0}^{i=N} P_i * Q_i$

ii) Given the vectors $X=\{x_1, x_2, \ldots x_N\}$ $Y=\{y_1, y_2, \ldots, y_N\}$, calculate $D=\sum d_i$ where d_i is computed using $d_i=((X_i - Y_i) / 2)^2 \; \forall \; i \in 0,1\ldots N$ 5

2B. With a suitable diagram, explain any three key features of Kepler GPU architecture. 3

2C. For the kernel code snippet given below, if the block size is 1024 and warp size is 32, how many warp/s will have divergence during the iteration where stride is equal to i)0 ii)16 iii) 32 iv)1024? 2

```
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 = blockDim.x; stride > 0; stride /= 2) {
__syncthreads();
if (t < stride) {partialSum[t] += partialSum[t+stride];} }
```

3A. Write the complete CUDA program to perform inclusive prefix minimum. Assume multiple blocks are used to handle the large input and shared memory is used to reduce the global memory traffic. 5

3B. Write the efficient CUDA kernel to implement histogram algorithm for an input array of ASCII characters. There are 128 ASCII characters and each character will map into its own bin for a fixed total of 128 bins. Assume shared memory is used to reduce the global memory traffic. 3

3C. In a certain computation, 90% of the work is vectorizable. Of the remaining 10%, half is parallelizable for an MIMD machine. What is the speedup over sequential execution for a 10 processing elements SIMD machine? 2

4A. State true/false and justify your answer with suitable example/s. 5

i) Usage of shared memory is mandatory for CUDA programs.
ii) Lesser the CGMA ratio, greater the performance of CUDA programs.
iii) Memory is the only limiting factor for parallelism.
iv) Device synchronization(__syncthreads__) can be used to synchronize between multiple kernel executions.
v) CUDA constant variables should be declared outside the scope of main function.

4B. Write the CUDA kernel to perform 1D convolution on the input vector of dimension N using the mask width of (1x5) dimension. 3

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

5A. Write the complete and efficient CUDA program to perform tiled matrix multiplication to compute C = A * B. The code should work (without any modifications) for any connected GPU device and for any valid input dimensions. 5

5B. With an example, explain Thrust interoperability. 3

5C. What is memory coalescing? With an example, explain how it affects the performance of CUDA programs. 2