Module 6 (18)

1. Write a parallel formulation of bionic sort on a hypercube with n=2d processes also derive expression for parallel run time. 4

OR

1. Show communication during the last stage of a bitonic sort considering each wire is mapped to a hypercube processand each connection represents a compare-exchange between processes 4
2. In a parallel formulation of Prim’s Algorithm why it is not easy to perform different iterations of the while loop in parallel? 4
3. Parallel run time for Matrix-vector multiplication with block 2-D partitioning for using fewer than n2 processes is Tp=n2/p + ts logp + twnrootp. Calculate overall asymptotic isoefficiency function of scalability. 6 Marks
4. Prove that isoefficiency function of Cannon’s algorithm is big o(p)3/2 4

Module 4 ( 8)

1. What could be possible impact of finite buffers in message passing? Explain it by writing small code fragment showing communication between two processes………………..4 marks MPI)
2. Describe in detail various environment variables in OpenMP 4

Module 5 (12) Any 3

1. Draw a neat labelled diagram of CUDA Memory hierarchy. Illustrate the difference between texture memory and constant memory. 4 Marks
2. Write a CUDA kernel to illustrate the use of shared and global memory. 4 Marks
3. **How do memory operations in GPUs differ from those in CPUs? 4 marks**
4. **Why Must Cuda Divide Computation Twice: Into Grids And Then Blocks? 4**

Module 1 (8)

1. Differentiate between SIMD and MIMD architecture? 4 marks
2. What do you mean by diameter of static interconnection network? Mention Diameter of completely-connected, star and 2-D mesh wraparound mesh. 2

Module 3

1. What are the different sources of overhead during execution of parallel program? Derive the expression for Total Parallel Overhead 4
2. The sequential version of a program executes in 10000 hours. Executing on 1000 processors it executes in 20 hours. What is the speedup and efficiency? How long would the program take to execute if the efficiency when executing on 1000 processors was 100%?............4

Module 2… 8 marks

1. If the diameter of an interconnection network is 4, would you prefer a store-and-forward network with ts=10 and tw=0.1 or a cut through network with ts=10 and tw=0.5 . Justify your answer 4
2. What is speculative decomposition? Explain with proper example 4

**Why Must Cuda Divide Computation Twice: Into Grids And Then Blocks?**

**Answer :**

The hardware is based on maximizing throughput. This has been done by allowing a large number of running threads -- all with a live context. This implies that only a fixed number of threads can fit in the hardware. This in turn means that these threads cannot communicate with or depend on other thread that could not be fit and hence must wait for the first set of threads to complete execution. Hence, a two level decomposition. Further, even the set of threads running together may execute at different SMs, and synchronization across SMs would be slow and onerous and hence not supported.