Exam Date & Time: 04-May-2024 (02:30 PM - 05:30 PM)

MANIPAL ACADEMY OF HIGHER EDUCATION

SIXTH SEMESTER B.TECH END SEMESTER EXAMINATIONS, MAY 2024

PARALLEL PROGRAMMING [DSE 3254]

Marks: 50 **Duration: 180 mins.**

A

Answer all the questions. Section Duration: 180 mins

Instructions to Candidates:

Answer ALL questions Missing data may be suitably assumed

1) Consider a scenario where multiple threads update a shared global array, in parallel. Each thread performs some computation and then writes its result to a specific location in the array. Discuss with appropriate examples the usage of the OpenMP Flush directive in this scenario. (4)

A)

B) Discuss with the appropriate MPI program the usage of Buffered Send for a scenario where the root process sends a message to process with rank 1. (3)

C) Analyze and evaluate the similarities and differences between MPI_REDUCE and MPI_ALLREDUCE by providing suitable examples to illustrate their usage. (3)

2) Analyze and evaluate the differences between collective and point-to-point communication in MPI within the context of a parallel program designed to toggle each letter of a given word. (4)

A)

B) Examine the role of memory coalescing in facilitating efficient data transfer from global memory to shared memory, supported by relevant examples. (3)

C) Discuss with an appropriate example, how memory can be a limiting factor to parallelism in the context of CUDA. (3)

3) Assume you are working on a real-time gray-scale image processing application where you are tasked with modeling the background of a given scene. You have chosen to do this task by performing background subtraction (Image_1 - Image_2) where every pixel in image one is subtracted from the corresponding pixel in image two. You have also chosen to do this task parallelly by using CUDA. Assume that the dimension of the input image size is 1200x900 pixels. Your GPU hardware supports a streaming multi-processor with a maximum of up to 2048 threads and it allows up to 256 threads in each block. Design a CUDA program to do the above-mentioned task efficiently. (5)

A)

B) Analyze and apply optimization strategies for barrier utilization in OpenMP for the below-given code: (3)

```
{
#pragma omp parallel for
for(i=0;i< n;i++)
a[i] += b[i];
#pragma omp parallel for
for(i=0;i< n;i++)
c[i] += d[i];
#pragma omp parallel for reduction (+:sum)
for(i=0;i< n;i++)
sum += a[i] + c[i];
}
```

C) Consider the following statement: “A variable stored in register memory decreases the CGMA ratio”. State true or false and justify your answer. (2)

4) Discuss how "loop fission" and "loop unroll and jam" contribute to loop optimization, and how can these techniques be applied to enhance the optimization of the below-provided code. (4)

A)

```
void main()
{
int i=0;
int j=0;
for (i=0;i< n;i++)
{
c[i]=c[i]*3;
for(j=0;j< n;j++)
a[j][i]=b[j][i]+d[i];
}
}
```

B) Given an initial serial program that executes in 9.30 seconds, where 90% of its execution time has been parallelized, and each additional processor introduces a 4% overhead to the total CPU time, calculate the speedup and efficiency achieved by the parallel program when utilizing 6 processors. Additionally, estimate the total CPU time and elapsed time of the program. (3)

C) Evaluate the optimal choice of thread block size (16x16 or 32x32) for matrix-matrix multiplication on a CUDA device, considering the hardware constraints of 8 blocks and 2,048 threads per Streaming Multiprocessor (SM). Justify your answer. (3)

5) Assume you are working on a physics simulation that involves calculating the maximum velocity of a large number of particles (N) in a closed environment. Each particle has a distance associated with it and the time taken to travel the distance. Each particle's

A) distance traveled, and time taken are stored in two different arrays. Velocity is estimated as distance traveled by time taken. Write a CUDA program that efficiently calculates and stores the velocity of all the individual particles in an array V. Also, write a reduction algorithm to estimate the maximum velocity from V parallelly. Assume that, N=1200 and a block can have 256 threads. (5)

B) Discuss with appropriate examples the benefits of using tiling techniques to reduce global memory traffic. (3)

C) Evaluate the limitations of fixed resource partitioning and analyze how dynamic partitioning effectively addresses these constraints. (2)

-----End-----