103

RAJARATA UNIVERSITY OF SRI LANKA
FACULTY OF APPLIED SCIENCES

B.Sc. (Honors) Degree in Information Technology
Fourth Year - Semester I Examination – January/February 2021

ICT 4305 – Parallel and Cluster Computing

Time: THREE (03) hours

- Answer **ANY FIVE (05)** questions.
- This is a closed book examination.
- This paper includes **FOUR (04)** pages.

1. In the modern age of computing, different types of computing models, ranging from classical to more advanced, can be leveraged.

a) Explain the following terminologies using a short description and a diagram(s) if necessary.
   i. Serial computing
   ii. Parallel computing
   iii. Distributed computing
   iv. Cluster computing (60 Marks)

b) Parallel computing is possible since it is greatly supported by the Central Processing Unit (CPU). Explain how modern CPUs are capable of doing parallel computing. (20 marks)

c) The use of Graphics Processing Unit (GPU) helps in massive parallel systems in several computing areas. Explain a few areas where GPUs in action/are used to solve parallel computing. (20 marks)

2. Parallel computing can be done using CPU or GPU. Each has its own implementation based on the programming language/platform used.

a) Using Java as the programming language answer the followings;
   i. Explain, with code, what are the supporting classes and methods to achieve a simple summing algorithm.
   ii. Explain the pitfalls that data is easily corrupted when a shared memory method is used and explain how to avoid them.
   iii. Explain how autoboxing and unboxing can affect algorithm performance.
   iv. **Stream.iterate()** and **LongStream.rangeClosed()** are two ways to create a data stream. Which is faster to use in parallel computing and what makes the difference?

(50 Marks)

104

b) Using the kernel "**kernel <<<4, 4>>> (a)**", write the output of the array "**a**" in each of the following code.

i.
```
__global__ void kernel (*int a) {
        int i = threadIdx.x + blockId.x * blockDim.x
        a[i] = blockDim.x;
}
```

ii.
```
__global__ void kernel (*int a) {
        int i = threadIdx.x + blockId.x * blockDim.x
        a[i] = threadIdx.x;
}
```

iii.
```
__global__ void kernel (*int a) {
        int i = threadIdx.x + blockId.x * blockDim.x
        a[i] = blockId.x;
}
```

iv.
```
__global__ void kernel (*int a) {
        int i = threadIdx.x + blockId.x * blockDim.x
        a[i] = i;
}
```

(50 Marks)

3. Flynn's taxonomy is a classification of computer architectures that can be used to show both serial and parallel architectures in simple form. Parallel architectures use combination of memory models to solve the memory issues in parallel computing. These memory architectures can be used to explore the kind of challenges faced in the implementation of a true parallel system.

a) Explain the following concepts using diagrams and its practical applications.

i. Single Instruction, Single Data (SISD)

ii. Single Instruction, Multiple Data (SIMD)

iii. Multiple Instruction, Single Data (MISD)

iv. Multiple Instruction, Multiple Data (MIMD) (60 Marks)

b) Explain the following memory architectures including the pros and cons of each.

i. Shared Memory (Uniform Memory Access and Non-Uniform Memory Access)

ii. Distributed Memory

iii. Hybrid Distributed-Shared Memory (40 Marks)

4. Graphics Processing Units (GPUs) are used heavily in parallel computing. One of the most popular GPU platforms is Compute Unified Device Architecture (CUDA).

a) Explain CUDA programming architecture using diagrams. (30 marks)

b) Explain CUDA memory model using diagrams. (30 marks)

c) Explain the following terms used in CUDA:

i. Host code

ii. Grid

iii. Blocks

iv. Warps

v. Thread

vi. Kernel Function

vii. Device Function

viii. Compute Capability (CC) (40 marks)

5. a) Explain the following CUDA code snippet. Make sure to cover all the important lines. The same or similar lines appearing later can be ignored.

```
#include <stdio.h>

__global__ void saxpy(int n, float a, float *x, float *y)
{
        int i = blockIdx.x*blockDim.x + threadIdx.x;
        if (i < n) y[i] = a*x[i] + y[i];
}

int main(void)
{
        int N = 1<<20;
        float *x, *y, *d_x, *d_y;
        x = (float*)malloc(N*sizeof(float));
        y = (float*)malloc(N*sizeof(float));

        cudaMalloc(&d_x, N*sizeof(float));
        cudaMalloc(&d_y, N*sizeof(float));

        for (int i = 0; i < N; i++) {
            x[i] = 1.0f;
            y[i] = 2.0f;
        }

        cudaMemcpy(d_x, x, N*sizeof(float), cudaMemcpyHostToDevice);
        cudaMemcpy(d_y, y, N*sizeof(float), cudaMemcpyHostToDevice);

        // Perform SAXPY on 1M elements
        saxpy<<<(N+255)/256, 256>>>(N, 2.0f, d_x, d_y);

        cudaMemcpy(y, d_y, N*sizeof(float), cudaMemcpyDeviceToHost);

        float maxError = 0.0f;
        for (int i = 0; i < N; i++)
            maxError = max(maxError, abs(y[i]-4.0f));
        printf("Max error: %f\n", maxError);

        cudaFree(d_x);
        cudaFree(d_y);
        free(x);
        free(y);
}
```

(60 Marks)

b) By referring the above code snippet or some other, state the high level program flow of a CUDA application. Writing the code is not needed, but important steps should be included.

(40 Marks)

106

6. In modern days, cluster computing has become easy with software and cloud platforms. But there are trades to clustering. As users we have to determine what type of cluster we need to form and manage.

a) Name three (3) cluster classifications and provide a short description on each.

(30 marks)

b) Explain the CAP theorem related to clusters. Why we can have all the capabilities in one system?

(20 marks)

c) Docker can help you to set up a small cluster in your local machine or in a cloud. State five (5) docker commands and explain their functionality.

(25 marks)

d) Sending messages in a cluster is a key feature. Discuss how we send efficient messages in a cluster. What sort of hardware, software and protocols can be utilized to manage effective networks?

(25 marks)

--- **END** ---