CARDIFF UNIVERSITY
EXAMINATION PAPER

Academic Year:	2017/2018
Examination Period:	Autumn
Examination Paper Number:	CMT106
Examination Paper Title:	High Perfomance Computing
Duration:	TWO hours

Do not turn this page over until instructed to do so by the Senior Invigilator.

Structure of Examination Paper:

There are 4 pages.
There are 4 questions in total.
There are no appendices.

The maximum mark for the examination paper is 60 and the mark obtainable for a question or part of a question is shown in brackets alongside the question.

Students to be provided with:

The following items of stationery are to be provided:
ONE answer book.

Instructions to Students:

Answer THREE questions.

Important note: if you answer more than the number of questions instructed, then answers will be marked in the order they appear only until the above instruction is met. Extra answers will be ignored. Clearly cancel any answers not intended for marking. Write clearly on the front of the answer book the numbers of the answers to be marked.

The use of translation dictionaries between English or Welsh and a foreign language bearing an appropriate school stamp is permitted in this examination.

Q1. (a) Two vectors of length 10000 are added together to produce a third vector. If CUDA is used to perform this operation on a GPU, and the thread block size is 1024, how many threads will be in the grid? [2]

(b) Consider the following CUDA kernel for vector addition, where **n** is the vector length:

```
__global__ void vecAddKernel(float *A, float *B, float *C, int n)
{
    int i = blockIdx.x*blockDim.x + threadIdx.x;
    if (i<n) C[i] = A[i] + B[i];
}
```

(i) What does threadIdx.x represent? [2]

(ii) Explain the purpose of the conditional statement. [2]

(c) Suppose you want to write a CUDA kernel that operates on an image of size 900x1500 pixels. You would like to assign one thread to each pixel. You would like your thread blocks to be square and to contain 1024 threads per block. Assuming a two-dimensional layout of blocks, what grid dimensions and block dimensions would you choose for your kernel? [4]

(d) Suppose the global memory bandwidth of a GPU is 320 Gbytes/s. Assuming a compute to global memory access ratio of 10.0, and that each floating-point value is of size 4 bytes, what is the maximum execution speed in Gflop/s? [4]

(e) Describe the global, shared, and register memory types for CUDA devices, paying attention to memory access efficiency and the relative size of each memory type. [6]

Q2. (a) Describe three ways in which the number of threads in a parallel section of an OpenMP program can be set. [3]

(b) Explain what is meant by task parallelism, and describe the use of a work-sharing construct for enabling task parallelism in an OpenMP program. [6]

(c) In the following fragment of OpenMP code, a, b, and c are one-dimensional arrays:

```
int chunk = 500, i;
#pragma omp parallel shared(a,b,c,chunk) private(i) num_threads(6)
{
        #pragma omp for schedule(static,chunk)
        for (i=0; i < 8000; i++) c[i] = a[i] + b[i];
}
```

(i) Describe how the iterations of the **for** loop above are partitioned for assignment to threads. [2]

(ii) Describe how the chunks of work created by the partitioning are scheduled on the threads. [2]

(iii) Assuming the threads are numbered 0, 1, 2, 3, 4 and 5, which thread will perform the loop iterations indexed by i from 3500 to 3999? [3]

(d) Describe two mechanisms for synchronising threads in OpenMP. [4]

Q3. (a) Define the following network metrics: bisection width, expansion increment, and narrowness. [3]

(b) Evaluate each of the metrics in part (a) of this question for a cubic mesh of k^3 nodes. [3]

(c) A regular 4×4 mesh is mapped onto a 4-dimensional hypercube so that neighbouring nodes in the mesh are also neighbours in the hypercube. Determine the node number of the hypercube node at location $(2, 3)$ in the mesh. [4]

(d) Define the speed-up of a parallel algorithm. What is meant by saying that a parallel algorithm is scalable? [4]

(e) In a data parallel application, particles move in a series of discrete time steps within a two-dimensional periodic space under the influence of the forces they exert on each other. This results in a non-uniform spatial distribution of particles. Describe a method for dynamically maintaining approximate load balance. [6]

Q4. (a) In a 6-dimensional hypercube, which nodes are directly connected to node number 43? [3]

(b) Explain what is meant by an asynchronous protocol in an MPI point-to-point communication. [2]

(c) Suppose the sequential version of an application runs in time T_{seq} on one node of a parallel computer. If α denotes the serial fraction of the application, give an expression for the time to run the parallel version of the application across N nodes of the parallel computer. [5]

(d) Use your answer to part (c) of this question to deduce Amdahl's Law, and explain why it appears to limit the usefulness of parallel computing. [6]

(e) Describe the two most common types of parallel computer architectures defined by Flynn's taxonomy. [4]

END OF EXAMINATION