Started on	Sunday, 16 February 2020, 5:39 PM
State	Finished
Completed on	Sunday, 16 February 2020, 5:43 PM
Time taken	4 mins 2 secs
Marks	16.00/16.00
Grade	**100.00** out of 100.00

Question 1

Correct

Mark 1.00 out of 1.00

Which of the following is not a correct statement describing multicore (i.e. CPUs) and manycore (i.e. GPUs)

Select one:

- a. manycore architectures generally require thousands of parallel tasks to fully utilize the chip
- b. multicore architectures have better individual performance per processing unit than manycore
- c. manycore architectures don't always have hierarchical memory structures with smaller register memory close to the processors and larger slower caches and memories farther
- d. multicore architectures are tuned for aggregate throughput compared to manycore ✓

The correct answer is: multicore architectures are tuned for aggregate throughput compared to manycore

Question 2

Correct

Mark 1.00 out of 1.00

Which of the following is not a characteristic of SIMD instructions

Select one:

- a. expose parallelism to the programmer
- b. make use of data parallelism
- c. provide portable code across all SIMD chips ✓
- d. provide savings in area and energy

The correct answer is: provide portable code across all SIMD chips

Question 3

Correct

Mark 1.00 out of 1.00

A SIMD instruction on an NVIDIA GPU (also known as SIMT for "Single-Instruction, Multiple-Thread") because acts on a ...?

Select one:

- a. kernel
- b. warp ✓
- c. thread
- d. thread block

The correct answer is: warp

Question 4

Correct

Mark 1.00 out of 1.00

Thrust library copies data from CPU to GPU via

Select one:

- a. an explicit memcpy() call with same parameters as CUDA
- b. an explicit memcpy() call but with the parameters being the source and destination
- c. a simple assignment where the type of the pointers starts the copy operation ✓
- d. an explicit memcpy() call without the need for the direction parameter

The correct answer is: a simple assignment where the type of the pointers starts the copy operation

Question 5

Correct

Mark 1.00 out of 1.00

For transposing data structures in a thread block it is recommended you use

Select one:

- a. shared memory ✓
- b. global device memory
- c. texture memory
- d. thread registers

The correct answer is: shared memory

Question 6

Correct

Mark 1.00 out of 1.00

For which of the following OpenCL Runtime implementations would it be faster to not use float4 in the writing of the code?

Select one:

- a. AMD CPU Runtime
- b. Intel CPU Runtime
- c. AMD GPU Runtime
- d. NVIDIA GPU Runtime ✓

The correct answer is: NVIDIA GPU Runtime

Question 7

Correct

Mark 1.00 out of 1.00

Which of the following is not true about data stored in Device Global Memory?

Select one:

- a. is can be used by threads inside thread blocks to communicate and coordinate
- b. can be read by any function in CUDA ✓
- c. persists across different kernel invocations
- d. is used by thread blocks to communicate and coordinate

The correct answer is: can be read by any function in CUDA

Question 8

Correct

Mark 1.00 out of 1.00

Which of the following SIMD extensions has the largest width for SIMD

Select one:

- a. Larrabee
- b. Nvidia
- c. AVX
- d. ATI ✓

The correct answer is: ATI

Question 9

Correct

Mark 1.00 out of 1.00

What is the relation between the memory architectures on a manycore Fermi chip (Tesla C2050) where R is aggregate number of register memory, L1 is L1 cache size and L2 is L2 cache size

Select one:

- a. R > L1 > L2 ✓
- b. R < L2 < L1
- c. R > L2 > L1
- d. R < L1 < L2

The correct answer is: R > L1 > L2

Question 10

Correct

Mark 1.00 out of 1.00

```
__global__ void samplekernel(float *d_a)
       {
             int i = blockIdx.x * blockDim.x + ThreadIdx.x;
             ...
             if (i <16)
             {
                  d_a[i]=0;
                  __syncthreads();
             }
             else
             {
                  d_a[i]=1;
                  __syncthreads();
             }
              ...
       }
```

The code written above is a bad CUDA kernel example because

Select one:

- a. it contains a barrier in a divergent code and won't compile
- b. it has divergent code inside a CUDA kernel and won't compile
- c. it contains a barrier in a divergent code and could deadlock ✓
- d. it contains a barrier that is not defined for threads in thread blocks

The correct answer is: it contains a barrier in a divergent code and could deadlock

Question 11

Correct

Mark 1.00 out of 1.00

For efficient CUDA code one must consider all except which change mentioned below?

Select one:

- a. reducing execution divergence per warp
- b. reducing the number of thread blocks scheduled per core ✓
- c. reducing memory divergence per warp
- d. reducing the number of registers used by each thread

The correct answer is: reducing the number of thread blocks scheduled per core

Question 12

Correct

Mark 1.00 out of 1.00

Which of the following is a correct statement about the program CUDA threads must execute:

Select one:

- a. All threads must execute the same non-branching sequential program
- b. All threads must execute the same sequential program ✓
- c. All threads inside a thread block must execute the same non-branching sequential program but different thread blocks can have different programs
- d. All threads inside a thread block must execute the same sequential program but different thread blocks can have different programs

The correct answer is: All threads must execute the same sequential program

Question 13

Correct

Mark 1.00 out of 1.00

OpenCL compared to CUDA

Select one:

- a. has easier/faster coding of simple problems
- b. is less portable across architectures
- c. has a simpler and less performant task parallelism model
- d. has a similar data parallel execution model ✓

The correct answer is: has a similar data parallel execution model

Question 14

Correct

Mark 1.00 out of 1.00

Which of the following techniques does not waste memory bandwidth

Select one:

- a. random sparse accesses
- b. non-unit strided accesses
- c. padding cache lines with extra values ✓
- d. unaligned memory accesses

The correct answer is: padding cache lines with extra values

Question 15

Correct

Mark 1.00 out of 1.00

If a warp stalls because the required data is not in the registers, what are possible courses of action the GPU can take? Mark all that apply

Select one or more:

- a. Assuming there is another warp ready to execute, context switch to that warp ✓
- b. If there is no other warp ready to execute, then wait for the required data to arrive to registers ✓
- c. Even if context switching is expected to be too expensive, context switch to another warp

Your answer is correct.

The correct answers are: Assuming there is another warp ready to execute, context switch to that warp, If there is no other warp ready to execute, then wait for the required data to arrive to registers

Question 16

Correct

Mark 1.00 out of 1.00

Which of these statements about thread blocks within a kernel is not correct?

Select one:

- a. Any interleaving of the thread blocks by the GPU scheduler is a valid execution of the kernel
- b. Thread blocks within a kernel should have no dependencies among each other.
- c. Each thread block within a kernel can execute different codes
- d. All thread blocks within a kernel should execute the same code ✓

Your answer is correct.

The correct answer is: All thread blocks within a kernel should execute the same code