NPTEL Online Certification Courses

Indian Institute of Technology Kharagpur

GPU Architectures and Programming
Assignment- Week 2
TYPE OF QUESTION: MCQ/MSQ

Number of questions: 10 **Total mark: 10 X 1 = 10**

MCQ Question

Question 1:
Which of the following statements about vector processors is TRUE?

A. Vector processors are designed to process multiple data elements using a single instruction.
B. Vector processors execute scalar operations more efficiently than vector operations.
C. Vector processors are primarily used for control-intensive tasks.
D. In vector processors, data dependencies between vector elements are common and unavoidable.

Answer: A
Solution:
A. TRUE. Vector processors are optimized for SIMD (Single Instruction, Multiple Data) operations. They can perform the same operation on multiple data elements (such as arrays or vectors) simultaneously
B. FALSE. This is the opposite of how vector processors are designed. They are specifically optimized for vector operations, not scalar operations.
C. FALSE. Vector processors are not primarily used for control-intensive tasks. They are designed for data-intensive tasks
D. FALSE. While data dependencies can occur, they are not unavoidable. Vector processors generally work best when vector elements are independent of each other

Question 2:
A vector processor can perform 8 floating-point operations per clock cycle. If the clock frequency of the processor is 2 GHz, how many floating-point operations per second (FLOPS) can the processor achieve?

A. 8×10^9
B. 32×10^9
C. 16×10^9
D. 4×10^9

Answer: C
Solution:
FLOPS = Floating-point operations per cycle × Clock frequency
FLOPS = $8\times2\times10^9$
FLOPS = 16×10^9

Question 3:
Which of the following correctly describes the memory hierarchy in NVIDIA's Fermi GPU architecture from fastest to slowest access speed?

A. Global Memory → Shared Memory → Registers → L1 Cache

B. Registers → L1 Cache → Shared Memory → Global Memory

C. L1 Cache → Registers → Shared Memory → Global Memory

D. Shared Memory → Registers → L1 Cache → Global Memory

Answer: B

Solution:
In the Fermi GPU architecture, registers provide the fastest memory access, followed by the L1 cache, shared memory, and finally, global memory, which is the slowest but has the highest capacity.

Question 4:
Match column 1 with column 2.

a) Vertex Shader	i) Operates on lines and triangles defined by multiple vertices, changing or generating additional primitives in graphics
b) Pixel Shader	ii) Maps the position of vertices onto the screen, altering their position color, or orientation.
c) Geometry Shader	iii) Fills the interior of primitives, including interpolating per fragment parameters, texturing, and colouring in graphics.

A. a-i,b-ii,c-iii
B. a-ii,b-iii,c-i
C. a-ii,b-i,c-iii
D. a-iii,b-i,c-ii

Answer:B

Question 5:
An NVIDIA Tesla GPU consists of 128 Scalar Processor (SP) cores and 16 KB of Shared Memory. There are a total 4 TPC and each TPC has 2 Streaming Multiprocessors (SMs). How many SPs are present in one SM and What is the total shared memory available across all SMs in the GPU?

A. 16,256 KB
B. 32,128 KB
C. 16,128 KB
D. 32,256 KB

Answer: C

Solution:
Shared memory per SM = 16 KB
Number of SMs in the GPU = No. of TPC × SMs per TPC
$=2\times4 = 8$
Number of SPs in the GPU=128
Number of SPs per SM=Number of SPs/Number of SMs
$=128/8=16$
Total Shared Memory=Shared memory per SM × Number of SMs
$=16\ KB\times8=128\ KB$

Question 6:
In GPGPU,how many warps are selected by the SM warp scheduler in each operation cycle.

A. 10
B. 32
C. 1
D. 8

Answer: C

Solution:
In each operation cycle, the SM warp scheduler selects one of the 24 warp

Question 7:
Match column 1 with the correct option in column 2

a) Global memory	i) Shared only within a single SM
b) Shared memory	ii) Shared across all SMs
c) Local memory	iii) Private to a wrap
d) Constant memory	iv) Private to a thread
	v) Shared across group of SMs but not among all SMs

A. a-ii,b-i,c-iv,d-iii
B. a-ii,b-i,c-iv,d-ii
C. a-ii,b-ii,c-iv,d-v
D. a-ii,b-i,c-v,d-iii

Answer: B

Solution:
The memory shared among all SMs is Global Memory and Constant Memory.Shared memory is shared among all the threads within a single SM while local memory is private to a thread.

Question 8:
PTX instructions have format: opcode.type d, a, b, c; where d is the

A. Destination
B. Domain
C. Data
D. Double

Answer: A

Question 9:
GPGPU programming uses _____ execution model.

A. SIMD
B. SIMT
C. SPMD
D. SPMT

Answer: B

Question 10:
Amdahl's law is an expression used to find the maximum expected improvement to an overall system when only part of the system is improved. It is often used in parallel computing to predict the theoretical maximum speedup. Amdahl's law for overall speedup is given by-

$$Overall\ Speedup = \frac{Old\ execution\ time}{New\ execution\ time} = \frac{1}{((1 - Fraction_{enhanced}) + (\frac{Fraction_{enhanced}}{Speed\ Up_{enhanced}}))}$$

A program has two parts: one that is parallelizable and one that is not. Suppose 60% of the program can be parallelized, while the remaining 40% is inherently sequential. If the program is run on a system with 8 processing cores, calculate the maximum speedup achievable according to Amdahl's Law.

A. 2.208
B. 2.105
C. 2.575
D. 3.820

Answer: B
Solution:
$Fraction_{enhanced}$d=0.6
Speed $Up_{enhanced}$=8
Putting the given values in the formula.We get,
S=1/((1-0.6)+(0.6/8))=2.105

*********END**********