Quiz 3 - Results

Attempt 1 of 1

Written Oct 16, 2025 8:47 PM - Oct 16, 2025 8:54 PM

Attempt Score **23 / 25 - A-**

Overall Grade (Highest Attempt) **23 / 25 - A-**

Question 1 1 / 1 point

The style of parallelism supported on GPUs is best described as:

- ○ MISD - Multiple Instruction Single Data
- ○ Data parallelism
- ○ SISD - Single Instruction Single Data
- ✓ ◉ SIMT - Single Instruction Multiple Thread

Question 2 2 / 2 points

Shared memory in CUDA is accessible to:

- ○ All threads associated with a single kernel
- ✓ ◉

All threads in a single block

Both the host and GPU

Question 3 2 / 2 points

Which of the following correctly describes the relationship between Warps, thread blocks, and CUDA cores? Select all correct answers.

- ✓ A warp is the smallest unit of threads that the GPU can schedule for execution.
- ✓ A warp is divided into a number of thread blocks, and each thread block executes on a single CUDA core
- ✓ A thread block is assigned to a warp, and each thread in the warp is executed on a separate CUDA core
- ✓ Thread blocks are divided into warps (typical of size 32) to efficiently utilize the GPU's parallel processing capabilities.
- ✓ All threads in a warp execute the same instruction at the same time

Question 4 2 / 2 points

Which of the following correctly describes a GPU kernel

- ✓ All thread blocks involved in the same computation use the same kernel
- A kernel may contain a mix of host and GPU code
- A kernel is part of the GPU's internal micro-operating system, allowing it to act as an independent host

Question 5 1 / 1 point

Functions annotated with the *__global__* qualifier may be executed on the host or

the device

- ○ True
- ✓ ◉ False

Question 6 1 / 1 point

Which of the following descriptions best captures the characteristics of GPU global memory system compared to CPU memory?

- ✓ ◉ longer latency, higher memory bandwidth
- ○ shorter latency, higher memory bandwidth

Question 7 1 / 1 point

High-performance computing with GPUs is often called heterogeneous computing because:

- ○ There are very many models of GPUs to choose from when constructing a system
- ○ GPUs typically run a variety of kernels over the course of a computation
- ✓ ◉ Computing with GPUs typically also involves CPUs, so there are two different kinds of hardware with different strengths

Question 8 2 / 2 points

In CUDA's memory hierarchy, local thread memory is on-chip memory.

- ○ True
- ✓ ◉ False

Question 9 3 / 3 points

If 5 blocks are assigned to an SM, and each block has 256 threads, how many Warps are in an SM?

- ○ 24
- ○ 32
- ○ 80
- ✔ ◉ 40

Question 10 0 / 2 points

Consider a GPU whose SM can take up to 2048 threads, and the maximum number of blocks that can be executed simultaneously on this SM is 32. For Matrix Multiplication using multiple blocks, should you use 4x4, 8X8, 16X16, or 32X32 block configurations for this GPU to maximize the SM occupancy? Please note that the nxn block terminology refers to having n threads in each of the 2D dimensions. For example, an "8x8 block configuration" means each block consists of 64 threads (8 threads in each dimension, the x and y dimensions)

- ○ 4x4
- ✖ ◉ 16x16
- ➡ ○ 8x8
- ○ 32x32

Question 11 2 / 2 points

In your code, cudaMemcpy() is about 2x slower than in your friend's code. What could be a likely reason?

- ○ In your code the source or destination of cudaMemcpy() in the device memory is not pinned.

- [x] In your code the source or destination of cudaMemcpy() in the host memory is not pinned.
- [ ] In your friend's code the source or destination of cudaMemcpy() in the device memory is not pinned.
- [ ] In your friend's code the source or destination of cudaMemcpy() in the host memory is not pinned.

Question 12 2 / 2 points

Which of the following are true when specifying virtual and real architectures during CUDA compilation?

- [x] To get best performance real architecture should be chosen as low as possible
- [x] To get best performance real architecture should be chosen as high as possible
- [x] Choosing virtual arch as low as possible maximizes portability of the PTX code
- [x] Choosing virtual arch as high as possible maximizes portability of the PTX code

Question 13 2 / 2 points

What is the primary benefit of memory coalescing in GPU computing?

- [ ] It enhances the GPU's ability to execute more threads simultaneously.
- [ ] It decreases the power consumption of the GPU during intensive computational tasks.

✓ It reduces the number of memory accesses to global memory by combining several memory accesses into one.

It increases the computational power of the GPU by adding more cores.

Question 14 2 / 2 points

Consider the following data structures designed for use in GPU kernels. Identify which of these data structures are properly aligned for efficient memory access on a GPU:

```
struct Data1 {
    float4 a; // Aligned to 16 bytes
    float2 b; // Aligned to 8 bytes
    float c;  // Aligned to 4 bytes
};
```

```
struct Data2 {
    float a;    // Aligned to 4 bytes
    char b;     // Aligned to 1 byte
    float2 c;   // Aligned to 8 bytes
};
```

✓ Only Data1 data structure is properly aligned.

Only Data2 data structure is properly aligned.

- Both Data1 and Data2 are properly aligned.
- Neither Data1 nor Data2 are properly aligned.

Done