| | |
|---|---|
| **Started on** | Thursday, September 29, 2022, 3:01 PM |
| **State** | Finished |
| **Completed on** | Thursday, September 29, 2022, 3:37 PM |
| **Time taken** | 35 mins 43 secs |

Question **1**

Complete

Points out of 1.00

For a tiled 1D convolution, the output tile width is 250 elements and mask width is 7 elements. If we assume that the input tile does not involve any ghost element, what would be the ratio of global memory reduction for generating the output tile by loading the input tile into the shared memory?

Select one:

- a. 250
- b. 250*7/256 ✓ ?
- c. 256*7/250
- d. None of the answers
- e. 7

Question **2**

Complete

Points out of 1.00

To perform an atomic add operation to add the value of an integer variable Partial to a global memory integer variable Total. Which one of the following statements should be used?

Select one:

- a. atomicAdd(Total, &Partial);
- b. atomicAdd(&Total, &Partial);
- c. atomicAdd(&Total, Partial);
- d. None of the answers
- e. atomicAdd(Total, 1);

Question **3**

Complete

Points out of 1.00

Assume that each atomic operation in a DRAM system has a total latency of 100ns. Assume that we privatize the global memory variable into shared memory variables in the kernel and the shared memory access latency is 1ns. All original global memory atomic operations are converted into shared memory atomic operation. For simplicity, assume that the additional global memory atomic operations for accumulating privatized variable into the global variable adds 10% to the total execution time. Assume that a kernel performs 5 floating-point operations per atomic operation. What is the maximal floating-point throughput of the kernel execution as limited by the throughput of the atomic operations?

Select one:

- a. None of the answers
- b. 0.45 GFLOPS
- c. 4.5 GFLOPS ✓
- d. 4500 GFLOPS
- e. 45 GFLOPS

All atomic done in shared, shared latency is 1ns, ns is pow-9 and giga is pow9 so...
1s = 5 GFLOPS
10% latency so 4.5 GFLOPS

Question **4**

Complete

Points out of 2.00

For tiling convolution shown in the lecture, we use Design 2. Assume that we load an entire input tile, including the halo elements into the shared memory when calculating an output tile. Further assume that the tiles are internal and thus do not involve any ghost elements. Each block thread loads 1 input element (i.e., the size of each thread block matches the size of an input tile). If the thread block size is 32x32 and mask size is 7x7, write down the output position for thread (1, 2) of block (40,30) during the calculation of an output tile. Your answer should be in the format of (x,y), e.g. (105,106), where there is no extra space in the expression.

Answer: 1281,962 ✗

Question **5**

Complete

Points out of 1.00

For a tiled 2D convolution, assume that we load an entire input tile, including the halo elements into the shared memory when calculating an output tile. Further assume that the tiles are internal and thus do not involve any ghost elements. If each output tile is a square with 12 elements on each side and the mask is a square with 5 elements on each side, which of the following best approximate the average number of times each input element will be accessed from the shared memory during the calculation of an output tile?

Select one:

- a. 4.9
- b. 256
- c. None of the answers
- d. 14 ✓
- e. 37

Question **6**

Complete

Points out of 1.00

For a tiled 2D convolution, if each output tile is a square with 12 elements on each side and the mask is a square with 5 elements on each side, how many elements are in each input tile?

Select one:

- a. 5*5 = 25
- b. (12+4) * (12+4) = 256
- c. 12*12 = 144
- d. (12+2) * (12+2) = 196
- e. None of the answers

Question **7**

Complete

Points out of 1.00

For a tiled 1D convolution, if the output tile width is 250 elements and mask width is 7 elements, what is the input tile width?

Select one:

- a. 250
- b. 256
- c. 7
- d. None of the answers
- e. 254

Question **8**

Complete

Points out of 1.00

Assume that each atomic operation in a DRAM system has a total latency of 100ns. Assume that a kernel performs 5 floating-point operations per atomic operation. What is the maximal floating-point throughput of the kernel execution as limited by the throughput of the atomic operations?

Select one:

- a. 0.0005 GFLOPS
- b. 5 GFLOPS
- c. 0.37 GFLOPS
- d. 0.05 GFLOPS ✓
- e. 500 GFLOPS

Question **9**

Complete

Points out of 1.00

In a tiled 2D convolution with 12x12 output tiles and 5x5 mask, how many warps in each thread block will have control divergence?

Select one:

- a. 16
- b. None of the answers
- c. 4
- d. 6
- e. 2