Exam : GP-GPU

Instructions

All exercises are independent.

1 General understanding

Answers should be short.

Question 1.
`i=blockIdx.x*blockDim.x*8 + threadIdx.x`

Question 2.
`1024`

Question 3.
```
dim3 DimGrid(ceil(n/16.0), ceil(m/16.0), 1); dim3 DimBlock(16, 16, 1);
```

Question 4.
Control divergence is the fact that threads in a warp do not execute the same code. We have to count the threads that hit outside the image : 49

Question 5.
The answer depends on the hardware used. One element was missing in the question. Assuming that the streaming multiprocessor has 32K registers and 64KB of shared memory, the maximum number of blocks that will run simultaneously is 2.

Question 6.
The maximum throughput is $10^9/2000 = 5.10^5$ operation per seconds.

Question 7.
Each warp is going to access 256 bytes of consecutive data. This is a fully coalesced access so the maximal achievable bandwidth is 240 GB/s.

2 CUDA Basics

Question 8.

1. 10
2. 32
3. 10240
4. Yes, there is control divergence, which is caused by the if statement in line 4. Since the total number of threads in the grid will be 10240, larger than the size of arrays, warp 313 will have divergence : the first 16 threads in the warp will take the true path and the other 16 threads will take the false path.
5. No, there will be no divergence. The total number of warps needed is $20000/32 = 625$. The number of wraps created is 640. The last 15 warps will be inactive and will not create divergence. For the warp 625, all threads will be used.

Question 9.

1. 1

2. 11
3. There is control divergence if `stride < 32`.
4. There will be 5 for loop iterates with control divergence.
5. 1
6. 1

3 Shared memory

Question 10.
- `(Width-1)/TILE_WIDTH +1`
- `Row<Width&&Col<Width`
- `Row<Width&&Col<Width`
- `TILE_WIDTH`