Problem set 6, Part 1

TDT4200, Fall 2016

Deadline: 09.11.2016 at 20.00 Contact course staff if you cannot meet the deadline.

Evaluation: Pass/Fail

Delivery: Use It's Learning. Deliver exactly one file:

- *yourusername_ps6.pdf*, with answers to the theory questions

General notes: All problem sets are to be done **INDIVIDUALLY**.

Problem 1

For each of the memories, describe a situation/application/memory access pattern where it can be beneficial to use it (on the GPU) and explain why it is beneficial to use the given memory type over others.

a) Texture memory

b) Shared memory

c) Constant memory

Problem 2

Which of the following code snipets will (if executed on a GPU) cause branch divergence? Explain your answers.

a)
```
if(blockIdx.x > 16){
     foo();
}
else{
     bar();
}
```

b)
```
if(threadIdx.x > 16){
     foo();
}
else{
     bar();
}
```

c)
```
for(int i = 0; i < threadIdx.x; i++){
     foo();
}
```

Problem 3

For embarassingly parallel applications (like the Mandelbrot calculation) it is possible to divide the work between both the CPU and GPU. If one of the devices is faster than the other, it should be given more of the work. To minimize execution time, the devices should finish at the same time.

Consider an application whose input and output both have size n. The execution time, on both CPU and GPU only depends upon the input/output size, on the GPU it is nR_{gpu} and on the CPU nR_{cpu}. We can divide the work between the CPU and GPU in any way we want, the ratio of work (which will be between 0 and 1) assigned to the GPU is b. This means the input/output size on the GPU is bn, and the execution time bnR_{gpu}.

The bandwidth between the CPU and GPU memories is B. Transfering an input of size m from CPU to GPU memory therefore takes mB time.

a) Find an expression for the execution time of the part of the work assigned to the CPU.

b) Find and expression for the execution time of the part of the work assigned to the GPU. Include data transfer times.

c) Find an expression for b which will divide work evenly.

d) Will it ever be beneficial to not use the GPU at all?

Problem 4

a) What are coaleced memory accesses?

b) What are shared memory bank conflicts?