Assignment 5: GPU basics

Preliminary

To use CUDA, add `module load cuda` in your `.bashrc` or `.bash_profile` on mamba.

To submit a CUDA job, use -lnodes=1:ppn=7:gpus=1

Processing on the GPU can be asynchronous, which can lead to time measurement of 0 seconds. Use `cudaDeviceSynchronize()` to ensure previously submitted tasks have completed.

1 Polynomial expansion (60 pts)

(Code for polynomial expansion on the CPU is given.)
Question: Write a simple CUDA code that allocates and fill an array on the CPU and transfer it to the GPU. (Take array size as a parameter)
Question: Compute the polynomial expansion of each element of the array on the GPU. (Take block size and degree of the polynomial as a parameter.)
Question: Bring the results back on the CPU and confirm the GPU code is correct.

2 Measurements (40 pts)

Question: Measure PCI-express latency. (That is the time for an array of size 1.)
Question: Measure PCI-express Bandwidth. (The initial memory copy for different size of the array.)
Question: Measure GPU memory bandwidth. (Exclude memory copies and use a low degree polynomial.)
Question: Measure GPU flops rating. (Exclude memory copies and use a high degree polynomial.)