Assignment III: CUDA Basics II

Aritra Bhakat

November 10, 2023

GitHub repo

Exercise 1: Histogram and Atomics

1. **Describe all optimizations you tried regardless of whether you committed to them or abandoned them and whether they improved or hurt performance.**

   The first optimisation I tried (after implementing a naive version that only accesses global memory), was to keep the bins in shared memory, so each block had its local version. Then, I performed a reduction between the blocks, adding the bins together atomically. I let each thread within a block be responsible for adding a regions of the bins. This did not work unfortunately and increased the time initially. Using 1024 threads gave the best performance for the method, but it still did not outperform the naive method.

   To reduce contention in global memory I didn't perform the atomic add if a bin is empty. Even though it leads to branching it reduced the histogram kernel execution time.

2. **Which optimizations you chose in the end and why?**

3. **How many global memory reads are being performed by your kernel? Explain**

   Each thread reads its corresponding input data from global memory, which gives `inputLength` reads. In each block all the bins in global memory are also updated, so there are a further (up to) `gridSize = ceil(inputLength / TPB_HIST)` reads, where `TPB_HIST` is the amount of threads per block for the histogram kernel.

4. **How many atomic operations are being performed by your kernel? Explain**

5. **How much shared memory is used in your code? Explain**

6. **How would the value distribution of the input array affect the contention among threads? For instance, what contentions would you expect if every element in the array has the same value?**
7. **Plot a histogram generated by your code and specify your input length, thread block and grid.**
8. **For a input array of 1024 elements, profile with Nvidia Nsight and report Shared Memory Configuration Size and Achieved Occupancy. Did Nvsight report any potential performance issues?**

Exercise 2: Histogram and Atomics

1. **Describe the environment you used, what changes you made to the Makefile, and how you ran the simulation.**
2. **Describe your design of the GPU implementation of mover_PC() briefly.**
3. **Compare the output of both CPU and GPU implementation to guarantee that your GPU implementations produce correct answers.**
4. **Compare the execution time of your GPU implementation with its CPU version.**