CS6023 Assignment 3 - Report

E Santhosh Kumar (CS16B107)

March 27, 2019

Q1

The table shows the average execution time (averaged over 20 iterations) of the kernel for different configurations.

Blocks per Grid	Threads per Block	Average Execution Time (ms)
977	1024	0.2524
1954	512	0.2465
3907	256	0.2422
7813	128	0.2404
15625	64	0.3010
31250	32	0.5132

The best configuration was observed to have 128 threads per block, 7813 blocks, and average exectution time of 0.2404 ms.

Q2

The table shows the best observed configuration and average execution time (averaged over 20 iterations) for different values of N.

N	Blocks per Grid	Threads per Block	Average Execution Time (ms)
1	7813	128	0.6482
2	7813	128	0.6008
3	7813	128	0.3741
4	7813	128	0.3979
5	7813	128	0.3948

As N increases, the number of collisions while accessing the global memory bins decreases. This generally outweighs the increase in computation required as N increases. Hence as N increases, execution time decreases.

Q3

The table shows the best observed configuration and average execution time (averaged over 20 iterations) for different values of N.

N	Blocks per Grid	Threads per Block	Average Execution Time (ms)
1	7813	128	0.4279
2	7813	128	0.3578
3	3907	256	0.3995

Q4

The following table was observed to be the frequencies of word lengths in the text_8_1M.txt file.

Word Length	Frequency
1	65,381
2	161,672
3	189,224
4	200,067
5	128,392
6	84,545
7	64,099
8	106,620
9	0
10	0

From the above, the four most common word lengths are 2,3,4 and 5.

Heuristic Used: The hot-spot bins that are cached on shared memory correspond to n-count-grams that have all word lengths in {2,3,4,5}.

The following table shows the best observed configuration and average execution time (averaged over 20 iterations) for different values of N.

N	Blocks per Grid	Threads per Block	Average Execution Time (ms)
4	1954	512	0.4334
5	977	1024	0.6131

Q5

From the word length frequencies given in Q4, the three most common word lengths are 2,3,4.

Heuristic Used: The hot-spot bins that are cached on shared memory correspond to N-count-grams that have all word lengths in {2,3,4}. Each group of 8 threads have the same set of hot-spot bins as their local copy. We use 256 threads per block. Thus the amount of shared memory used per block is given by

$$\textit{shared memory used per block} = (256/8) * (3^N) * sizeof(int)$$

For N=5, this evaluates to 31,104 Bytes and can be accommodated on the shared memory (NVIDIA Tesla K40 GPU allows a maximum of 49,152 Bytes of Shared Memory per Block).

The table shows the configuration and average execution time (averaged over 20 iterations) for different values of N.

N	Blocks per Grid	Threads per Block	Average Execution Time (ms)
1	3907	256	0.7150
2	3907	256	1.0947
3	3907	256	0.8994
4	3907	256	2.2302
5	3907	256	15.0386

Q6

Heuristic Used: The hot-spot bins that are cached on shared memory correspond to n-count-grams that have all word lengths in {2,3,4,5}. Each thread maintains 4^N local bins and counts the occurences of N-count-grams over 1000 (obtained empirically) n-size windows. Each block contains 32 threads only, so that shared memory size per block isn't exceeded for N=4.

The table shows the configuration and average execution time (averaged over 20 iterations) for different values of N.

N	Blocks per Grid	Threads per Block	Average Execution Time (ms)
1	32	32	1.4630
2	32	32	1.9518
3	32	32	2.7688
4	32	32	7.9388