

CS6023: GPU Programming

Assignment 3

Abhishek Nair
EE16B060

Optimization Methods Used

- Instead of reading in and storing the words in an array, only the lengths of words are stored. This saves space and reduces the time taken to copy data onto the GPU.
- No functions are used in the host code. Even the removal of punctuation is carried out in the main function itself. This saves the time involved in context switching.
- 1-D block and grid sizes are chosen since they reflect the geometry of the problem well. The number of threads per block and blocks per grid are tuned to obtain maximum performance.
- Privatisation of histograms : Each block keeps a private copy of the array of counts. This reduces the latency involved in updating the counts and also reduces the latency involved with using atomic updates as fewer threads are competing to update values.