



GPU Teaching Kit
Accelerated Computing



Module 7.1 – Parallel Computation Patterns (Histogram)

Histogramming

Objective

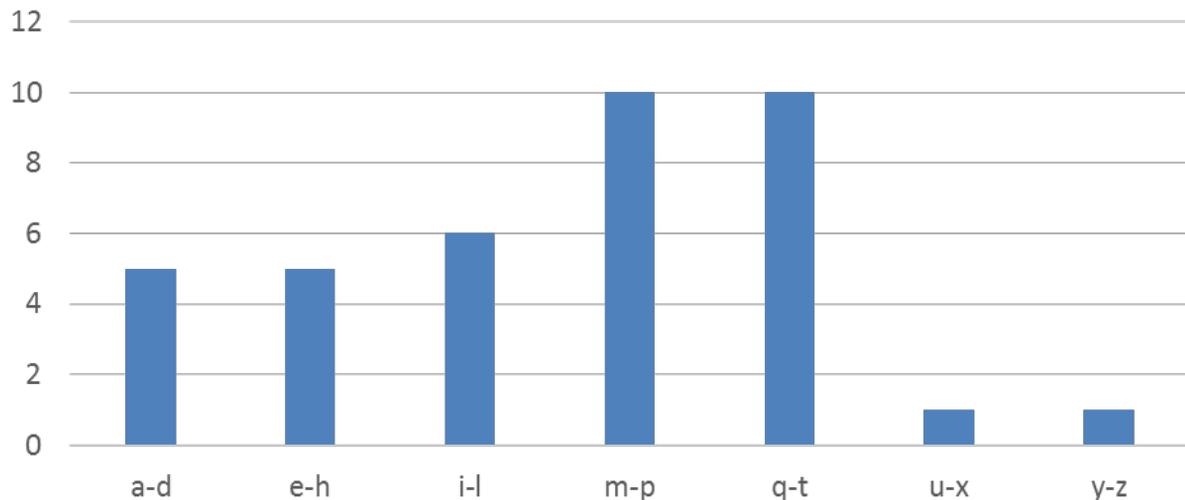
- To learn the parallel histogram computation pattern
 - An important, useful computation
 - Very different from all the patterns we have covered so far in terms of output behavior of each thread
 - A good starting point for understanding output interference in parallel computation

Histogram

- A method for extracting notable features and patterns from large data sets
 - Feature extraction for object recognition in images
 - Fraud detection in credit card transactions
 - Correlating heavenly object movements in astrophysics
 - ...
- Basic histograms - for each element in the data set, use the value to identify a “bin counter” to increment

A Text Histogram Example

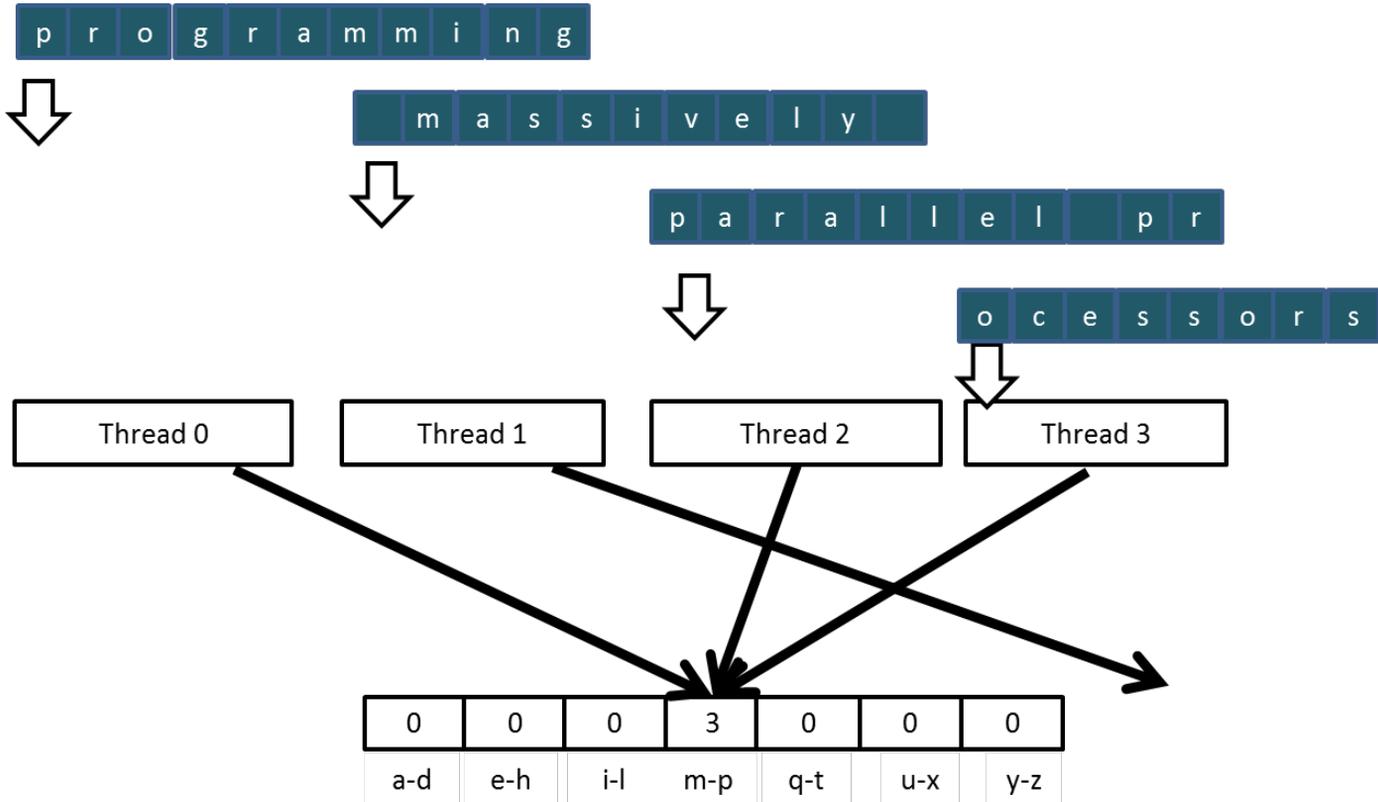
- Define the bins as four-letter sections of the alphabet: a-d, e-h, i-l, n-p, ...
- For each character in an input string, increment the appropriate bin counter.
- In the phrase “Programming Massively Parallel Processors” the output histogram is shown below:



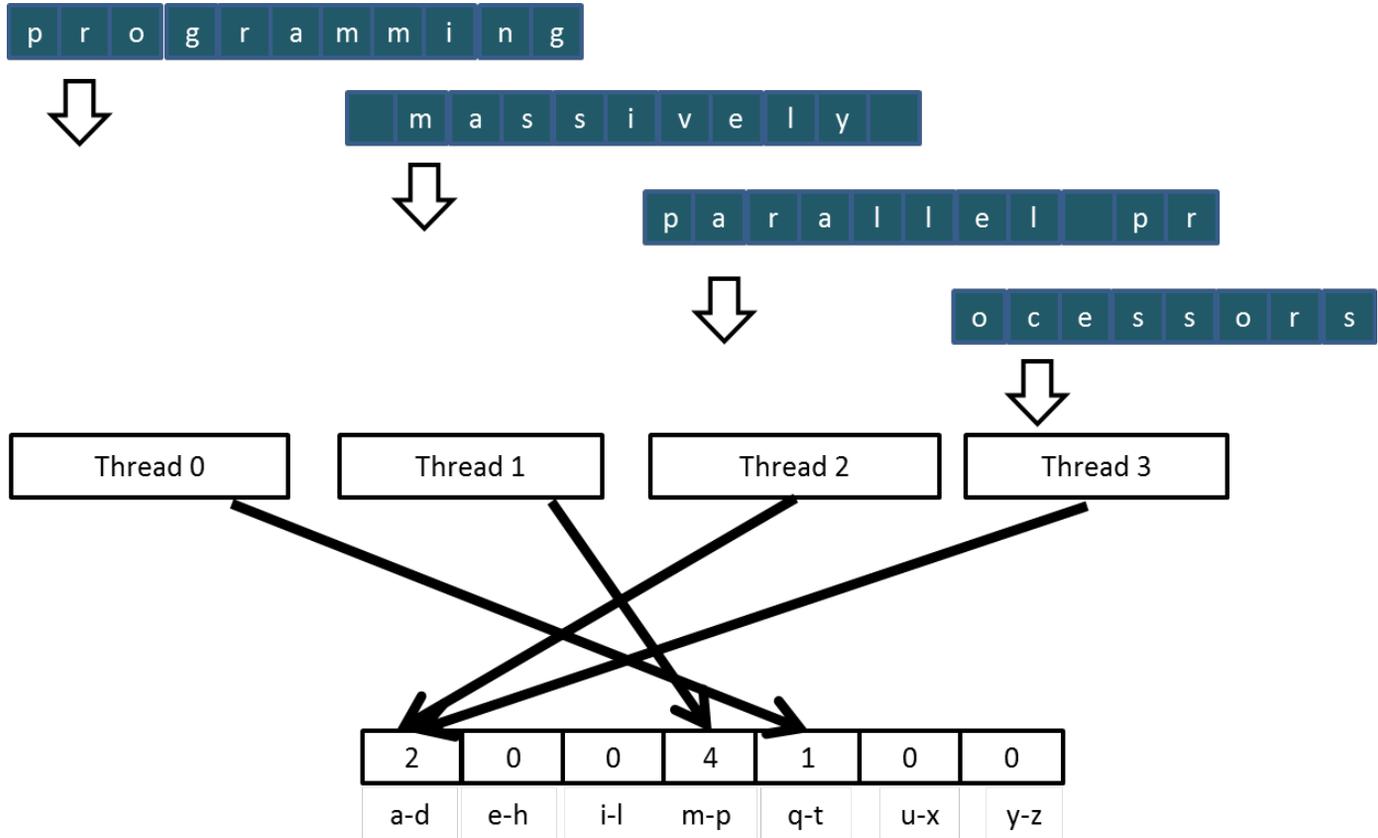
A simple parallel histogram algorithm

- Partition the input into sections
- Have each thread to take a section of the input
- Each thread iterates through its section.
- For each letter, increment the appropriate bin counter

Sectioned Partitioning (Iteration #1)



Sectioned Partitioning (Iteration #2)



Input Partitioning Affects Memory Access Efficiency

- Sectioned partitioning results in poor memory access efficiency
 - Adjacent threads do not access adjacent memory locations
 - Accesses are not coalesced
 - DRAM bandwidth is poorly utilized



Input Partitioning Affects Memory Access Efficiency

- Sectioned partitioning results in poor memory access efficiency
 - Adjacent threads do not access adjacent memory locations
 - Accesses are not coalesced
 - DRAM bandwidth is poorly utilized

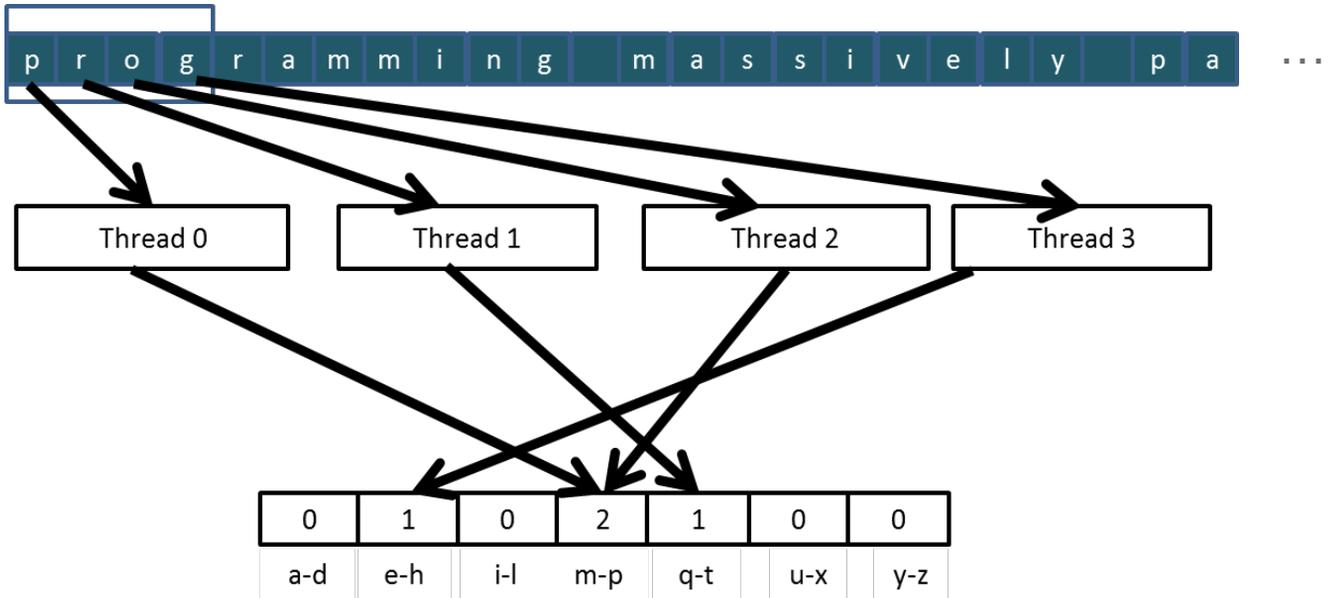


- Change to interleaved partitioning
 - All threads process a contiguous section of elements
 - They all move to the next section and repeat
 - The memory accesses are coalesced

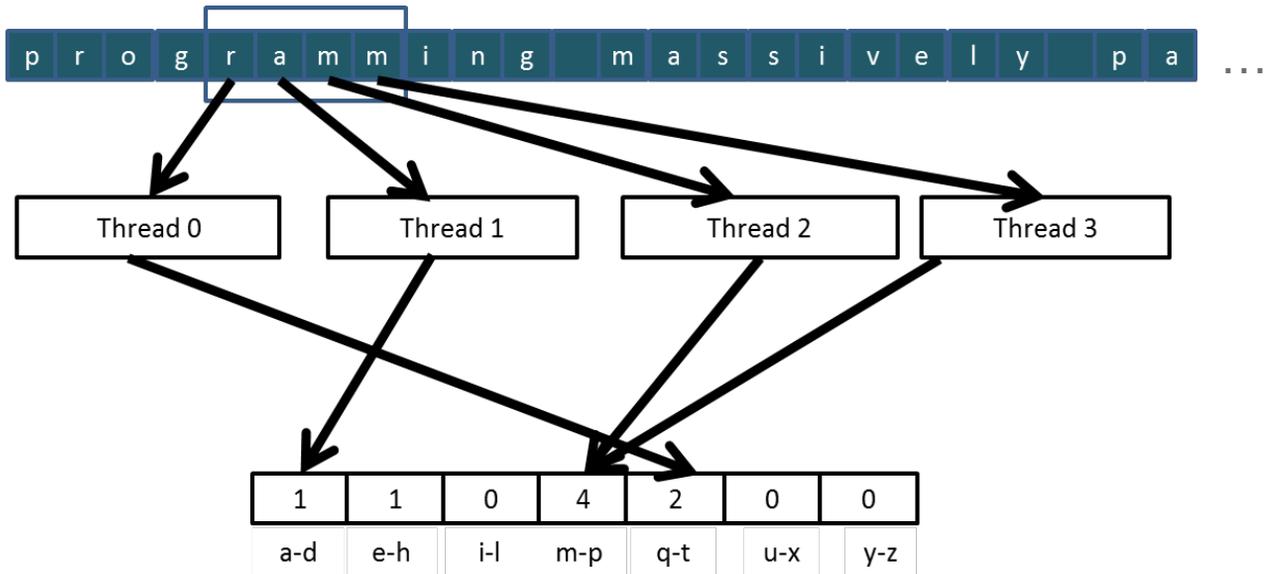


Interleaved Partitioning of Input

- For coalescing and better memory access performance



Interleaved Partitioning (Iteration 2)





GPU Teaching Kit

Accelerated Computing



The GPU Teaching Kit is licensed by NVIDIA and the University of Illinois under the [Creative Commons Attribution-NonCommercial 4.0 International License](https://creativecommons.org/licenses/by-nc/4.0/).