GPU(CUDA) HISTOGRAM AND ATOMICS

Mounika Reddy Edula & Rahul Yamasani

MAY 12, 2017

UNIVERSITY OF COLORADO BOULDER

Mounika Reddy Edula & Rahul Yamasani
ECEN 5593 - HPCA
HW: GPU (CUDA) Vector Reduction

Histogram and Atomics

Due to the insufficient time, available because of the project. The last checkpoint is a teamwork of Mounika and Rahul. Sorry for the inconvenience.

Histogram

The data randomly generated is divided among the threads and histograms are generated for individual data and merged at the CPU like vector reduction.
Time – Memory transfer between CPU and GPU
Time - Histogram for partial data at GPU
Time – Memory transfer of partial histograms between GPU and CPU
Time – Total time for histogram on CPU

The data is moved to the shared memory to increase the performance. The partial histogram is moved back to the global memory.
After completion of all threads then we merge the histogram which is transferred to the host.
The partial histogram is converted to GPU at CPU.

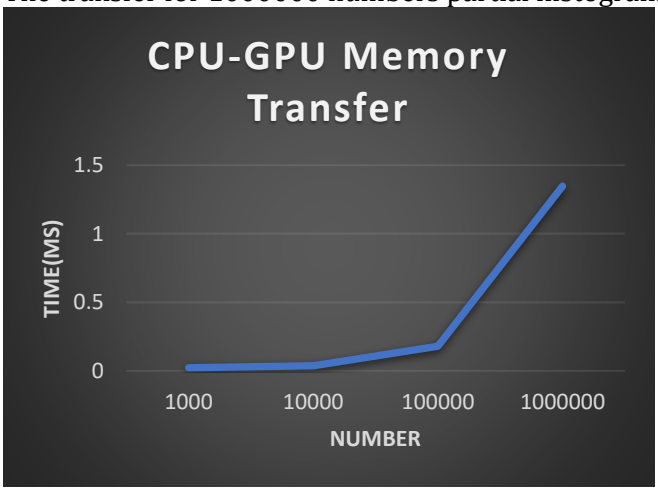
Execution times using the event timers (T=32, H=64)

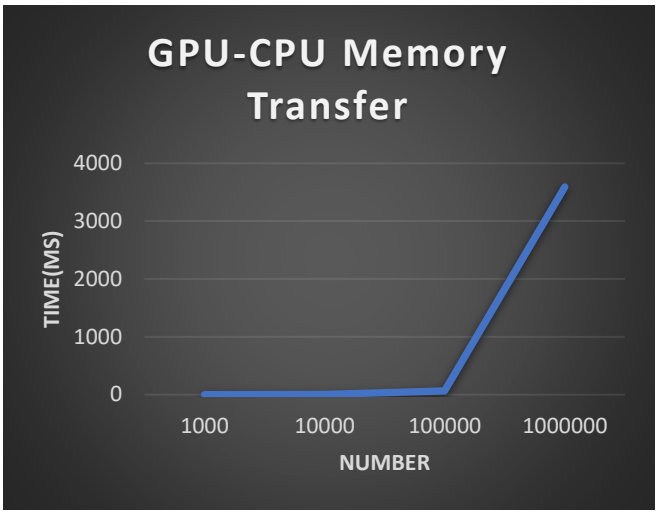
Array Size (N)	CPU Execution	GPU Execution + CPU clean up : histogram kernel	GPU Speedup(%)
1000	0.003	0.035	-10.667
10000	0.025	0.036	0.44
100000	0.254	0.039	0.846
1000000	2.518	0.043	98.22

This analysis describes that for a 1000 array size usage of GPU decreased the performance. The decrease is due to time wasted in the memory transfer. Higher the array size the performance improvement will be improved.

Memory Transfer Analysis

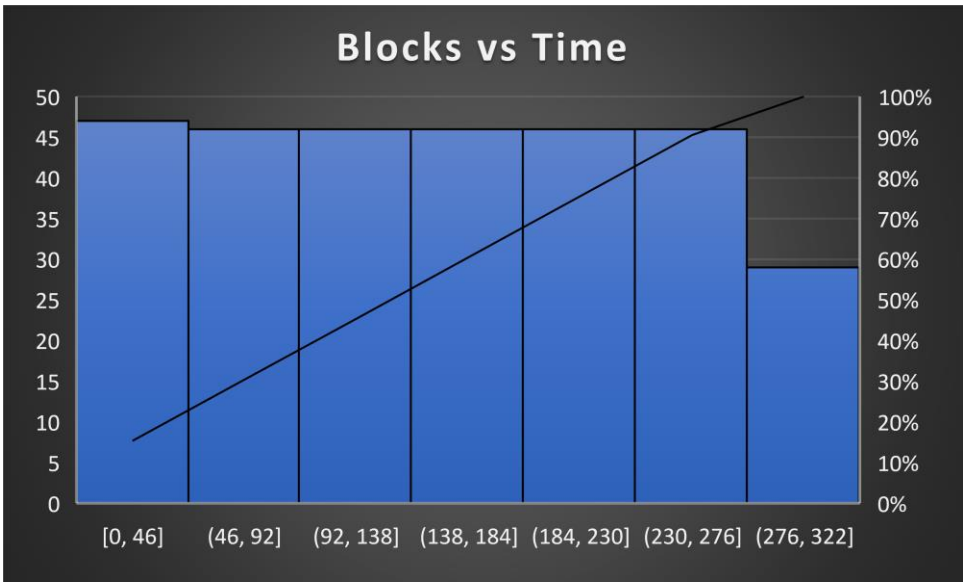
Transfer of the numbers generated between CPU and GPU will be less when compared to GPU-CPU transfer for individual histogram. The individual histograms are merged into one at the end.
The transfer for 1000000 numbers partial histograms is greater than 3000(ms).





Clock for performance evaluation

Start a clock for every block and stop it after the execution of block. The time values are stored in an array which will be transferred to the host.



Histogram Atomic Kernel

Array Size (N)	GPU Execution + CPU clean up : histogram_kernel	GPU Execution + CPU clean up : histogram_atomic_kernel	Speedup
1000	0.035	0.03	14.2
10000	0.036	0.03	16.67
100000	0.039	0.034	12.82
1000000	0.043	0.038	11.62

Clock for performance evaluation

Start a clock for every block and stop it after the execution of block. The time values are stored in an array which will be transferred to the host.