# Advanced Computer Architecture

# GPU (CUDA) Histogram and Atomics

## CUDA Code

Please see the attached .cu files. When compiling the code, the Makefile and structure assumes the code is located with the NVIDIA C sample directory (/NVIDIA/C/src/). The program is run using the following command:

```
./histogram -length <Desired length of random input sequence> -a
```

- Length: enter in the desired size of the input array from which to make a histogram from
- a: adding the -a flag will turn the atomicAdd function on

The default bin size is 64.

## Timing Results

The graph below summarizes the timing results for both the non-atomic and atomic GPU operations. As can be seen in the charts below, there is a slowdown when using the GPU for the smaller operations (input vector size N <4000). However, as the vector grows in size, there is seen a dramatic increase in the speedup factor when using the GPU -- a speedup of 27.5 and 29.1 for non-atomic and atomic GPU execution times, respectively.

All charts are included in the attached excel file.

| Array Size | CPU Execution Time | No Atomics GPU | | | |
|---|---|---|---|---|---|
| | | GPU Execution Time | Memory Transfer Time | Total GPU Time | Speedup |
| 4000 | 0.008 | 0.046 | 0.00008 | 0.04608 | 0.173611 |
| 40000 | 0.054 | 0.07 | 0.0001 | 0.0701 | 0.770328 |
| 400000 | 0.522 | 0.075 | 0.00027 | 0.07527 | 6.935034 |
| 4000000 | 5.222 | 0.188 | 0.00203 | 0.19003 | 27.47987 |

**Fig. 1a: First half of table summarizing execution times for histogram program**

| CPU Execution Time | Atomic Operations GPU | | | | GPU Speedup |
|---|---|---|---|---|---|
| | GPU Execution Time | Memory Transfer Time | Total GPU Time | Speedup | |
| 0.007 | 0.041 | 0.00008 | 0.04108 | 0.170399 | 1.121714 |
| 0.054 | 0.064 | 0.0001 | 0.0641 | 0.842434 | 1.093604 |
| 0.522 | 0.068 | 0.00026 | 0.06826 | 7.647231 | 1.102696 |
| 5.222 | 0.177 | 0.00202 | 0.17902 | 29.16993 | 1.061502 |

**Fig. 1b: Second half of table summarizing execution times for histogram program**

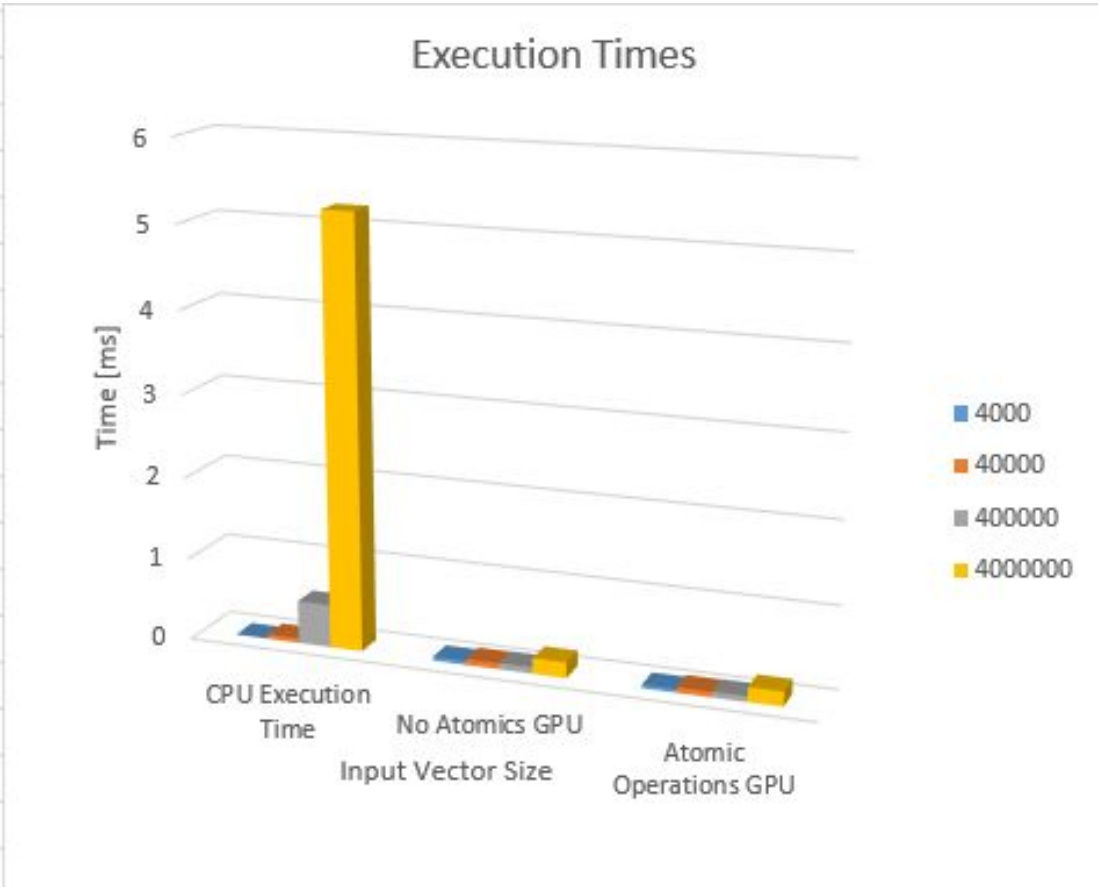


**Fig. 2 Total execution times**

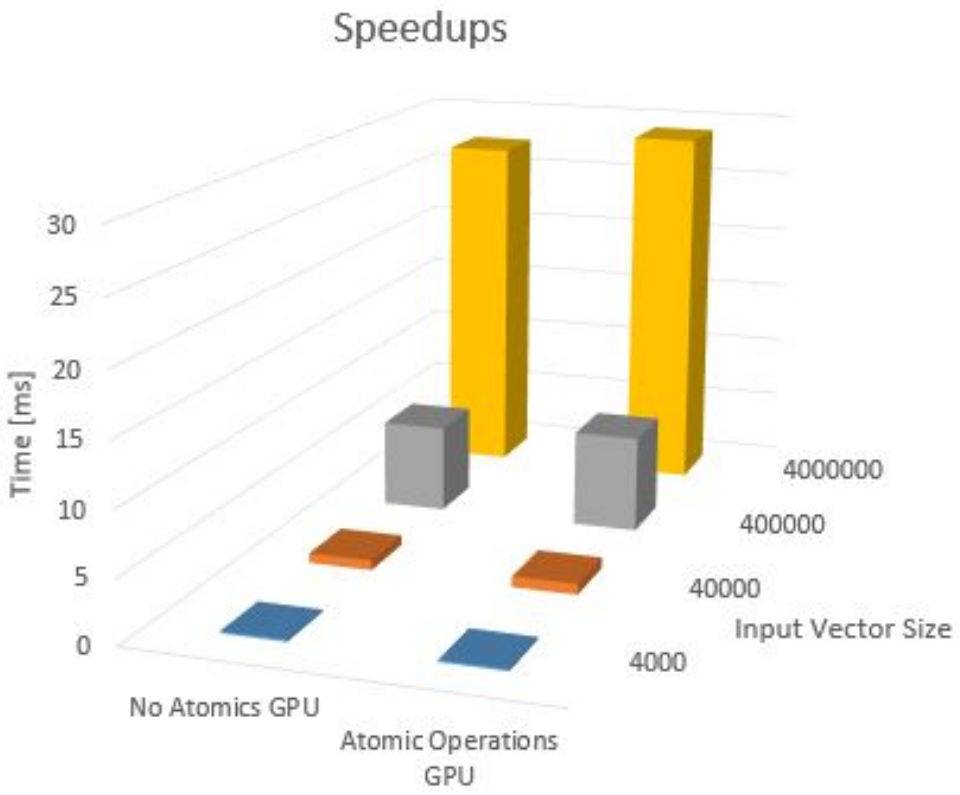


**Fig. 3 GPU Speedups for Input Size**

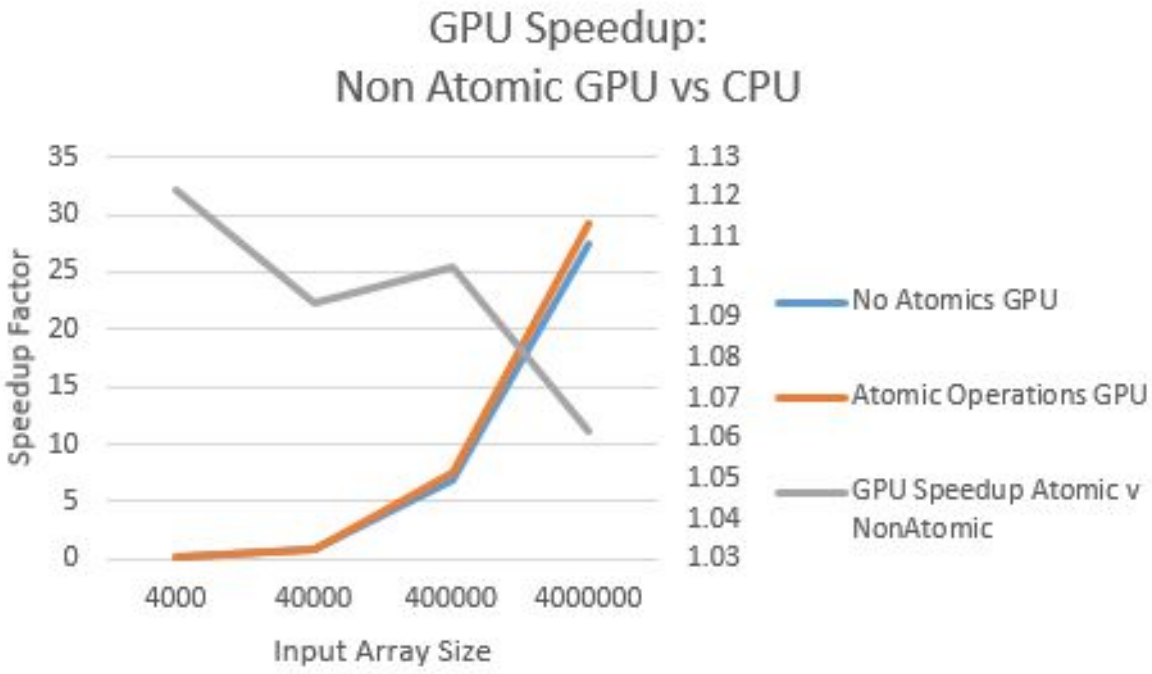


**Fig. 4 Speedup Comparisons**

# Conclusion

The largest speedup over the CPU is when doing atomic operations on the GPU with the largest input array size. The initial sizes of the input were too small for there to be any advantage to using the GPU for calculations when taking into account the time to transfer memory between the CPU and GPU. On the larger input sizes seen in this exercise, there is not a very noticeable difference between the GPU using atomic vs non-atomic operations. As the input size increased, however, there was a downward trend in the speedup seen in the atomic vs non-atomic operation, indicating that there was a time expense to using atomic operations as to be expected.