

CUDA Vector reduction

ECEN-5593



May 9, 2017

Rahul yamasani

rahul.yamasani@colorado.edu

**NOTE**: All the relevant files for this Checkpoint are listed in **Yamasani\_CheckP3\_Filters.rar**

1. Execution time and memory transfer time for vectorReduce.cu

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Input Size** | **Blocksize** | **GPU Execution Time** | **Memory Transfer Time** | **CPU Time (to add partial sums)** | **Overall Execution Time (Memory + GPU Execution)** |
| 1000 | 32 | 0.037 | 0.025 | 0 | 0.078 |
| 10000 | 32 | 0.041 | 0.042 | 0 | 0.101 |
| 100000 | 32 | 0.137 | 0.178 | 0.0051 | 0.356 |
| 1000000 | 32 | 0.673 | 1.42 | 0.049 | 2.245 |
| 2000000 | 32 | 1.1 | 2.717 | 0.0952 | 4.1032 |
| 1000 | 64 | 0.038 | 0.025 | 0 | 0.079 |
| 10000 | 64 | 0.041 | 0.041 | 0.001 | 0.1 |
| 100000 | 64 | 0.138 | 0.178 | 0.003 | 0.347 |
| 1000000 | 64 | 0.671 | 1.434 | 0.025 | 2.261 |
| 2000000 | 64 | 1.099 | 2.728 | 0.04 | 4.119 |
| 1000 | 128 | 0.037 | 0.026 | 0 | 0.079 |
| 10000 | 128 | 0.042 | 0.041 | 0 | 0.101 |
| 100000 | 128 | 0.135 | 0.177 | 0.001 | 0.344 |
| 1000000 | 128 | 0.672 | 1.425 | 0.0125 | 2.247 |
| 2000000 | 128 | 1.101 | 2.707 | 0.022 | 4.094 |
| 1000 | 256 | 0.044 | 0.025 | 0.001 | 0.085 |
| 10000 | 256 | 0.042 | 0.04 | 0.001 | 0.101 |
| 100000 | 256 | 0.155 | 0.177 | 0.001 | 0.361 |
| 1000000 | 256 | 0.679 | 1.429 | 0.0064 | 2.259 |
| 2000000 | 256 | 1.081 | 2.711 | 0.0111 | 4.078 |

1. Graphs for Overall Execution time vs Input size and Block Size

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Input Size | Block Size:32 | Block Size:64 | Block Size:128 | Block Size:256 |
| 1000 | 0.078 | 0.079 | 0.079 | 0.085 |
| 10000 | 0.101 | 0.1 | 0.101 | 0.101 |
| 100000 | 0.356 | 0.347 | 0.344 | 0.361 |
| 1000000 | 2.245 | 2.261 | 2.247 | 2.259 |
| 2000000 | 4.1032 | 4.119 | 4.094 | 4.078 |

1. Comparing the previous item with Atomic operations and including the SpeedUp

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Input Size** | **Blocksize** | **Previous total execution (CPU+GPU)** | **Total execution (atomic support in GPU)** | **Speedup** |
| 1000 | 32 | 0.078 | 0.078 | 1 |
| 10000 | 32 | 0.101 | 0.098 | 0.97029703 |
| 100000 | 32 | 0.356 | 0.355 | 0.997191011 |
| 1000000 | 32 | 2.245 | 2.249 | 1.001781737 |
| 2000000 | 32 | 4.1032 | 4.1125 | 1.002266524 |
| 1000 | 64 | 0.079 | 0.078 | 0.987341772 |
| 10000 | 64 | 0.1 | 0.12 | 1.2 |
| 100000 | 64 | 0.347 | 0.34 | 0.979827089 |
| 1000000 | 64 | 2.261 | 2.276 | 1.006634233 |
| 2000000 | 64 | 4.119 | 4.234 | 1.027919398 |
| 1000 | 128 | 0.079 | 0.078 | 0.987341772 |
| 10000 | 128 | 0.101 | 0.1 | 0.99009901 |
| 100000 | 128 | 0.344 | 0.346 | 1.005813953 |
| 1000000 | 128 | 2.247 | 2.234 | 0.994214508 |
| 2000000 | 128 | 4.094 | 4.123 | 1.007083537 |
| 1000 | 256 | 0.085 | 0.07934 | 0.933411765 |
| 10000 | 256 | 0.101 | 0.102 | 1.00990099 |
| 100000 | 256 | 0.361 | 0.354 | 0.980609418 |
| 1000000 | 256 | 2.259 | 2.267 | 1.00354139 |
| 2000000 | 256 | 4.078 | 4.234 | 1.038254046 |

1. Overhead with the border conditions

|  |  |  |  |
| --- | --- | --- | --- |
| **Input Size** | **GPU Overall Execution Time (blocksize=32)** | **GPU Overall Execution Time (blocksize=32)** | **Percentage different in performance [(A-B)/A\*100]** |
| **with the if statement present (A)** | **without the if statement present(B)** |
| 1024 | 0.081 | 0.079 | 1.092 |
| 4096 | 0.088 | 0.086 | 4.014 |
| 16384 | 0.117 | 0.114 | 0.8934 |
| 262144 | 0.818 | 0.853 | 4.1346 |
| 1048576 | 2.499 | 2.426 | 5.4194 |

**Conclusions:**

1. Vector reduction is performed with the given Input and Block sizes
2. It has been observed that the GPU execution increases with increase in input size and decreases with block size

**References:**

1. Professors video lectures
2. Code samples from the checkpoints/assignments