**ECEN 5593 – Advanced Computer Architecture**

**CUDA Vector Reduction**

**Name**: Adarsh Anand Kulkarni

1. Execution time and memory transfer time for different data sizes:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Input Size** | **Blocksize** | **GPU Execution Time (ms)** | **Memory Transfer Time (ms)** | **CPU Time (to add partial sums) (ms)** | **Overall Execution Time (Memory + GPU Execution) (ms)** |
| 1000 | 32 | 0.038000 | 0.047000 | 0.000000 | 0.085000 |
| 10000 | 32 | 0.041000 | 0.058000 | 0.001000 | 0.099000 |
| 100000 | 32 | 0.138000 | 0.208000 | 0.005000 | 0.346000 |
| 1000000 | 32 | 0.651000 | 1.567000 | 0.051000 | 2.219000 |
| 2000000 | 32 | 1.101000 | 3.012000 | 0.097000 | 4.113000 |
| 1000 | 64 | 0.038000 | 0.039000 | 0.000000 | 0.077000 |
| 10000 | 64 | 0.040000 | 0.056000 | 0.001000 | 0.097000 |
| 100000 | 64 | 0.117000 | 0.203000 | 0.003000 | 0.321000 |
| 1000000 | 64 | 0.447000 | 1.530000 | 0.024000 | 1.977000 |
| 2000000 | 64 | 0.692000 | 2.928000 | 0.050000 | 3.620000 |
| 1000 | 128 | 0.037000 | 0.039000 | 0.000000 | 0.077000 |
| 10000 | 128 | 0.039000 | 0.057000 | 0.000000 | 0.096000 |
| 100000 | 128 | 0.111000 | 0.204000 | 0.002000 | 0.316000 |
| 1000000 | 128 | 0.368000 | 1.496000 | 0.012000 | 1.865000 |
| 2000000 | 128 | 0.513000 | 2.821000 | 0.025000 | 3.334000 |
| 1000 | 256 | 0.038000 | 0.040000 | 0.000000 | 0.078000 |
| 10000 | 256 | 0.039000 | 0.056000 | 0.000000 | 0.095000 |
| 100000 | 256 | 0.114000 | 0.196000 | 0.001000 | 0.310000 |
| 1000000 | 256 | 0.384000 | 1.448000 | 0.006000 | 1.833000 |
| 2000000 | 256 | 0.545000 | 2.792000 | 0.012000 | 3.337000 |

1. Graph the result using Excel or any plotting tool, clearly showing the overall execution time versus Input Size and Block Size.
2. Results using AtomicAdd :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Input Size** | **Blocksize** | **Previous total execution (CPU+GPU) (ms)** | **Total execution (atomic support in GPU) (ms)** | **Speedup (%)** |
| 1000 | 32 | 0.085 | 0.077000 | 9.411765 |
| 10000 | 32 | 0.1 | 0.099000 | 1 |
| 100000 | 32 | 0.351 | 0.346000 | 1.424501 |
| 1000000 | 32 | 2.269 | 2.272000 | -0.13222 |
| 2000000 | 32 | 4.21 | 4.159000 | 1.211401 |
| 1000 | 64 | 0.077 | 0.077000 | 0 |
| 10000 | 64 | 0.097 | 0.097000 | 0 |
| 100000 | 64 | 0.323 | 0.327000 | -1.23839 |
| 1000000 | 64 | 2.001 | 1.994000 | 0.349825 |
| 2000000 | 64 | 3.67 | 3.625000 | 1.226158 |
| 1000 | 128 | 0.076 | 0.077000 | -1.31579 |
| 10000 | 128 | 0.096 | 0.095000 | 1.041667 |
| 100000 | 128 | 0.317 | 0.306000 | 3.470032 |
| 1000000 | 128 | 1.876 | 1.850000 | 1.385928 |
| 2000000 | 128 | 3.359 | 3.346000 | 0.38702 |
| 1000 | 256 | 0.078 | 0.078000 | 0 |
| 10000 | 256 | 0.095 | 0.095000 | 0 |
| 100000 | 256 | 0.311 | 0.307000 | 1.286174 |
| 1000000 | 256 | 1.838 | 1.836000 | 0.108814 |
| 2000000 | 256 | 3.349 | 3.306000 | 1.283965 |

1. **Results for performing the runtime check in the Kernel :**

|  |  |  |  |
| --- | --- | --- | --- |
| **Input Size** | **GPU Overall Execution Time (blocksize=32)**  **with the if statement present** | **GPU Overall Execution Time (blocksize=32)**  **without the if statement present** | **Percentage different in performance (% Speedup)** |
| 1024 | 0.077000 | 0.078000 | -0.1 |
| 4096 | 0.085000 | 0.084000 | 0.1 |
| 16384 | 0.114000 | 0.122000 | -0.8 |
| 262144 | 0.843000 | 0.822000 | 2.1 |
| 1048576 | 2.391000 | 2.384000 | 0.7 |