Homework 3: Reduction in CUDA

|  |  |  |
| --- | --- | --- |
| Jasmine Bhanushali  Beverly Abadines | March 11,2018 | EECS 224 |

# Naïve Parallel Reduction

Bandwidth = **10GB/second**

N: 8388608

Timer: gettimeofday

Time to execute naive GPU reduction kernel: 0.003357 secs

Effective bandwidth: 10.00 GB/s

Time to execute naive CPU reduction: 0.116124 secs

## Bandwidth Calculation

Effective Bandwidth = (Size of input data / time to reduce the sum)

# Strided Access by Consecutive Threads

Time to execute strided index GPU reduction kernel: 0.002217 secs

Effective bandwidth: **15.14 GB/s**

Time to execute naive CPU reduction: 0.116105 secs

Strided is 1.514 times faster than Naïve Parallel Reduction.

# Sequential Access by Consecutive Threads

Time to execute sequential index GPU reduction kernel: 0.001767 secs

Effective bandwidth: **18.99 GB/s**

Time to execute naive CPU reduction: 0.09992 secs

# First Add Before Reduce

Time to execute first add GPU reduction kernel: 0.000937 secs

Effective bandwidth: **35.81 GB/s**

Time to execute naive CPU reduction: 0.104616 secs

# Unroll the Last Wrap

Time to execute unrolled GPU reduction kernel: 0.000621 secs

Effective bandwidth: **54.03 GB/s**

Time to execute naive CPU reduction: 0.101121 secs

# Algorithm Cascading

Time to execute multiple add GPU reduction kernel: 0.000396 secs

Effective bandwidth: **84.73 GB/s**

Time to execute naive CPU reduction: 0.116129 secs