**Blue Waters Petascale Semester Curriculum v1.0**

**Unit 7: CUDA**

**Lesson 7: Parallel Reduce in CUDA**

**Instructor Guide**

*Developed by Sanish Rai for the Shodor Education Foundation, Inc.*



*Except where otherwise noted, this work by The Shodor Education Foundation, Inc. is licensed under CC BY-SA 4.0. To view a copy of this license, visit*[*https://creativecommons.org/licenses/by-sa/4.0*](https://creativecommons.org/licenses/by-sa/4.0)

*Browse and search the full curriculum at*[*http://shodor.org/petascale/materials/semester-curriculum*](http://shodor.org/petascale/materials/semester-curriculum)

*We welcome your improvements! You can submit your proposed changes to this material and the rest of the curriculum in our GitHub repository at*[*https://github.com/shodor-education/petascale-semester-curriculum*](https://github.com/shodor-education/petascale-semester-curriculum)

*We want to hear from you! Please let us know your experiences using this material by sending email to* [*petascale@shodor.org*](mailto:petascale@shodor.org)

1. Exercise 1: Review the given source code (reduction\_example.cu) for parallel reduction using sum operator for 2^20 elements

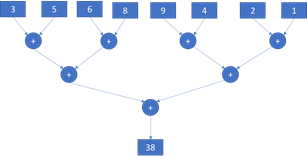
To run the program:

nvcc <filename>

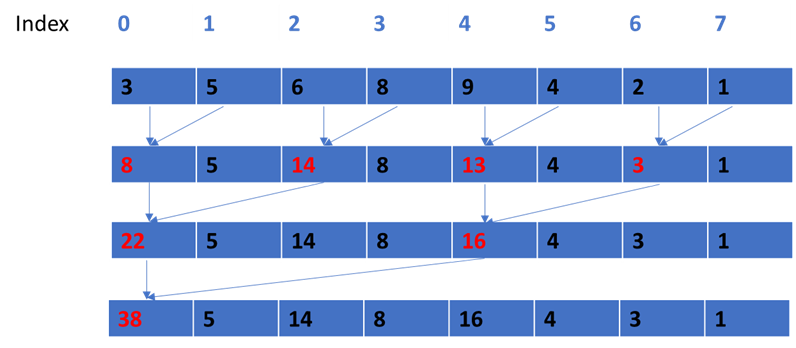
a //(in windows)

or ./a //(in linux)

1. Exercise 2: Write a kernel that performs reduction as shown in figure using shared memory for 2^20 elements.



Here is how it can be designed



Elements in red represent the reduced values.

Note: Since each block will compute a single result, you will need to add all the block results at the end to get the final reduced result.

Instructor note: The solution is provided in reduction\_exercise.cu

The for loop changes the stride by double in every iteration so that it reduces elements at stride 1 in the first iteration, at stride 2 in second iteration, at stride 4 in third iteration and so on

We can use this for loop to do so: for( int i=1; i < blockDim.x; i \*= 2)

We can use this if (t\_id % (2\*i) == 0) to check that we are adding the correct index

Since we use blocks, after first kernel computation is completed, each of the block stores the result in the output array. So we need to add them all to get final result.

1. Exercise 3: Write a program for parallel reduction using max operation to find the max value among 2^20 elements
   1. First define host variables and a serial method to find max
   2. Define device variables and initialize them
   3. Create a kernel to find max similar to exercise 1 using global memory
   4. Create a kernel to find max similar to exercise 1 using shared memory

**Common Pitfalls for Students and Instructors**

* Mistakes in proper allocation of memory in CPU and GPU
* Mistakes in writing kernel and using atomic functions
* Students might be confused with using proper number of threads and blocks
* Students might make mistake in copying data from CPU to GPU or vice versa
* Students should be careful in writing loops for proper addition with the list element
* Students should synchronize threads for proper updates of computation