

## Assignment No.3.

Title: Parallel Reduction.

Objective: To implement Min, Max, Sum and Average operations using Parallel Reduction.

Problem Statement: Implement Min, Max, Sum and Average operations using Parallel Reduction.

Theory:

Parallel reduction:

One common approach to this problem is parallel reduction. This can be applied for many problems, a min operation being just one of them. It works by using half the number of threads of the elements in the dataset. It involves dividing the dataset among multiple threads and performing operations such as minimum, maximum, sum, or average. In the case of a minimum operation, for example, each thread calculates the minimum value of its own element and another element, passing the result to the next round until only one element remains. It is important that thread utilization is optimised, ensuring that processing unit such as warp in CUDA, is efficiently utilized. To avoid branch divergence and maximise performance it's recommended to select elements for comparison in a way that minimizes the divergence within warps. Overall, parallel reduction is a powerful method for accelerating computations on large datasets in parallel computing environments.



To implement these operations using Parallel Reduction we need to follow these steps.

1. Divide the input data into equal-sized chunks among the available threads.
2. Calculate the local min, max, sum and Average for each thread on their respective chunks.
3. Combine the local min, max, sum and Average values across all threads using the Parallel Reduction Technique.
4. Finally, obtain the final min, max, sum and Average values by combining the results of all threads.

Min operation:

To perform a Min operation using Parallel Reduction, we can start by initializing a variable to a very large value (e.g.  $INT\_MAX$  for integer data). Each thread then calculates the local minimum for its chunk of data, and the global minimum is found by comparing the local minimum values from each thread and taking the smallest one.

Max operations:

Similar to the min operation, we can initialize a variable to a very small value (e.g.  $INT\_MIN$  for integer data) to perform a Max operation. Each thread calculates the local maximum for its chunk of data, and the global maximum is found by comparing the local maximum values from each thread and taking the largest one.



Sum operation:

To perform a Sum operation Using Parallel Reduction each thread calculates the local sum for its chunk of data, and the global sum is obtained by summing up the local sum values from each thread.

Average operation:

To perform an Average operation using Parallel Reduction, we first perform a Sum operation on the input data as describe above. Then, we divide the global sum by the total number of elements in the input data to obtain the global average.

Conclusion:

We have successfully implemented Min, Max, Sum and Average operations using Parallel Reduction.

*[Signature]*