Page: 1
Date: / /

Assignment NO: 1

- TITLE : a] Implement parallel Reduction using Min, Max, Sum, average operations.
  b] write a CUDA Program that given an N-element vector find :-
  i] the maximum element in vector
  ii] the minimum element in vector
  iii] the arithmatic mean of the vector
  iv] the standard deviation of the values in the vector.

- Objective : i] To understand parallel reduction operations
  ii] TO understand vector operations.

- Outcome :
  i] understand the parallel reduction operations and vector operations.

- Software and Hardware requirements : • A system with 4GB RAM, 500 GB HDD, and GPU functionalities
  • Google colab.

- Date of completion :

- Theory :

CUDA :-

a] CUDA (Compute Unified Device Architecture) is a parallel computing platform and application programming interface model created by NVIDIA.

b] It allows software Developers and software engineers to use CUDA enable graphics Processing unit for general Purpose Processing an approach termed GPU.

c] The CUDA platform is also a software layer that gives direct access to the GPU's virtual instructions set and parallel computional elements for the execution of compute kernels.

d] The CUDA platform is design to work with Programming languages such as C, C++, & Fortan.

e] This accessibility makes it easier for the specialist in Parallel Programming to use GPU resources in contact to prior API's like Direct 3D and open GL.

f] which required advanced skills in graphics Programming Also CUDA supports Programming frameworks.

g] Such as open ACC & open CL when it was first introduced by Nvidia the name CUDA was an acronym for compute unified Device Architecture byt Nivdia subsequently dropped the use of the acronym.

- Max operation : This method returns the larger element a and b compare function can be omitted. If there is no compare function used in max() the elements compared with operator by default compare function is used to determine which one of the object is larger when the object a & b are non-numeric type.

  Syntax :- Max(object-type a, object-type b, compare-function).

- Min operation : This method returns the smaller element a & b compare function can be omitted. If there is no compare function used min() then elements are compared with operator by default.

- Arithmatic Mean operation : This basic arithmatic operations are addition, substraction, multiplication & division. The arithmatic mean of a set of data is found by taking the sum of the data and then dividing the sum by the total number of values in the set A mean is commonly referred to as a Average.

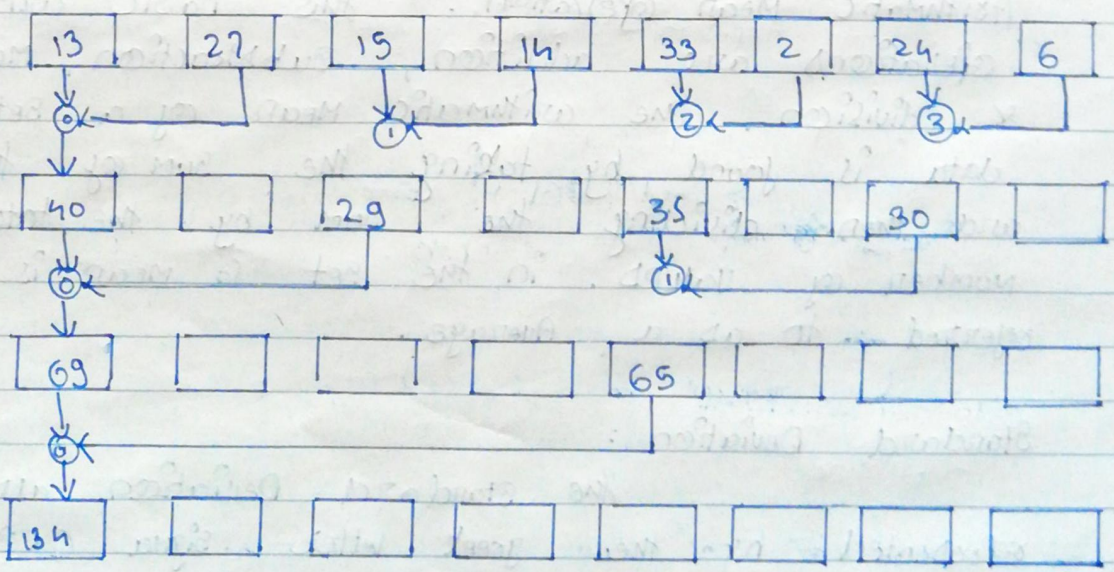
- Standard Deviation :

  The standard Deviation also represented by the greek letter sigma (σ) or the latin is a measure that is used to quantify the amount of variation or dispersion

of a set of data values. Standard Deviation is a number used to tell how how measurements for a group are spread out from the average (mean) or expected value.

- Parallel Reduction :

Implementing Parallel Reduction in CUDA. Reduction operations are those that reduce a collection of values to a single value. Operation which are associative and commutative can be reduction operation. Some of them are addition, Manipulation. Multiplication bitwise AND/OR/NOT logical AND/OR/NOT finding Maximum/Minimum amongst a given set of number sequence computation complexity can be O(log n)



- Test cases :

Functions	Input Size	Sequential Time	Parallel Time	Efficiency
Average	n = 250 average = 245	0.132	0.128	1.15
SUM	n = 250 SUM: 61411	0.01	0.18	0.05
MIN	n = 150 Min : 11	0.134	0.114	0.9
Max	n = 250 Max = 996	0.136	0.118	1.05
Standard Deviation	n = 250 std = 261.928	0.133	0.175	0.76

$$\text{Efficiency} = \frac{\text{wcSA}}{\text{wcPA}}$$

- Conclusion : Hence we successfully implemented parallel Reduction operation using Min, Max, Sum, average operations and perform the operations on N-element vector.