PARALLEL K-MEANS CLUSTERING WITH MPI

Author: Meghana N Prasanna

Course: CSE 633 Parallel Algorithms

Instructor : Russ Miller

University at Buffalo The State University of New York

CONTENT:

Introduction to Clustering

K-Means Algorithm

Parallel Approach

Output Analysis

Graphs

References

CLUSTERING

- A cluster refers to a collection of data points aggregated together because of certain similarities.
- Clustering refers to the process of automatically grouping together data points with similar characteristics and assigning them to “clusters.”
- Group similar data points together and discover underlying patterns.
- Dividing the data into clusters can be on the basis of centroids, distributions, densities, etc

K-MEANS CLUSTERING

- Notion of similarity is derived by the closeness of a data point to the centroid of the clusters.
- The no. of clusters required at the end have to be mentioned beforehand, which makes it important to have prior knowledge of the dataset.
- K-means algorithm has three major advantages covering simple implementation, efficient when handling a large data sets and a solid theoretical foundation based on the greedy optimization of Voronoi partition

NOTE: In mathematics, a Voronoi diagram is a partition of a plane into regions close to each of a given set of objects..

K-MEANS ALGORITHM

1. Specify the desired number of clusters K
2. Randomly assign each data point to a cluster
3. Compute cluster centroids
4. Re-assign each point to the closest cluster centroid
5. Re-compute cluster centroids
6. Repeat steps 4 and 5 until no improvements are possible

Complexity: O(input*K*iterations*dimensions)

EXAMPLE

- U = {1,6,10,18,3,14} , k=2
- Assume cluster centers to be c1 = 1, c2 = 6
- Cluster c1: {1,3}

  Cluster c2: {6,10,18,14}
- Update centre c1 = avg {1,3} = 2

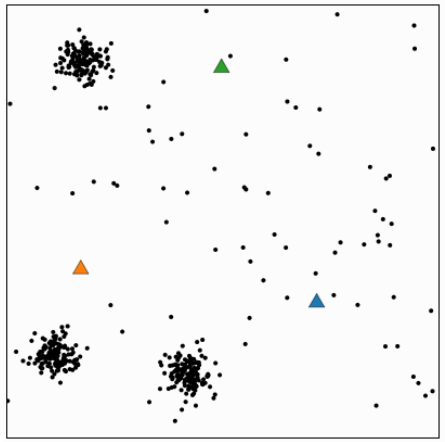
  Update centre c2 = avg {6,10,18,14} = 12
- Updated cluster c1: {1,3,6}

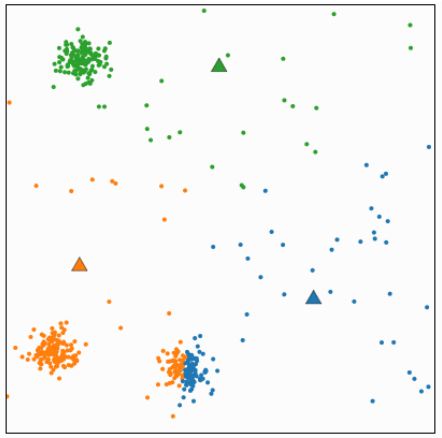
  Updated cluster c2: {10,18,14}
- Update centre c1 = avg {1,3,6} = 3.333

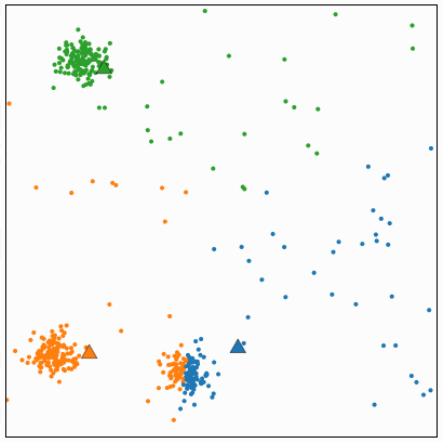
  Update centre c2 = avg {10,18,14} = 14
- Updated cluster c1: {1,3,6}

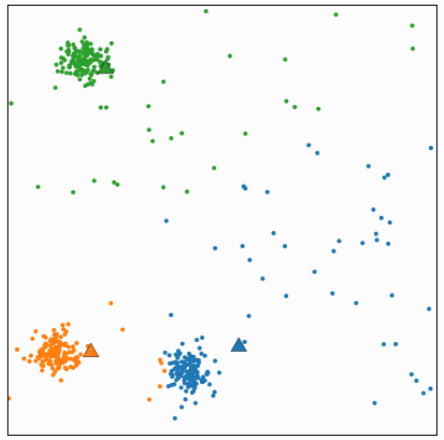
  Updated cluster c2: {10,18,14}
- No change in cluster (convergence)
- Stop

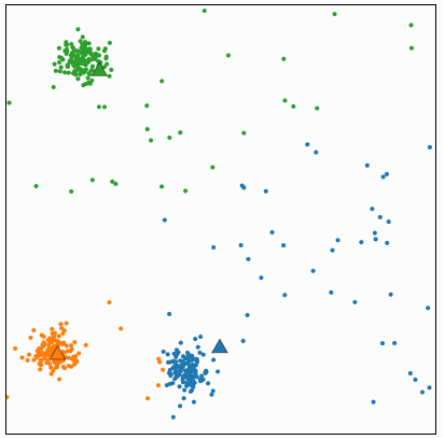
VISUALIZATION

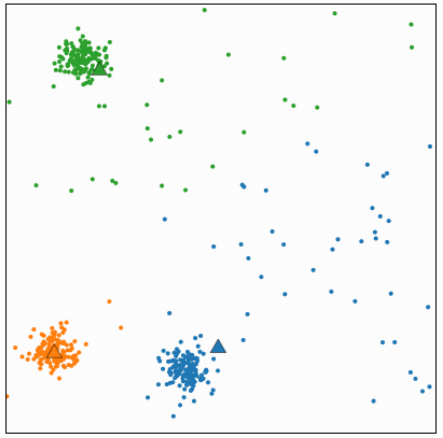












PARALLEL APPROACH

1.Divide data among each processors equally

2. The processor with rank 0 initializes k random centroids and broadcasts it to all other processors.

3. In each processor,

    a) calculate distance of a point from each centroid and divide into K clusters

    b) locally calculate the sum of each cluster and returns the sum and length of each cluster to the processor with rank 0

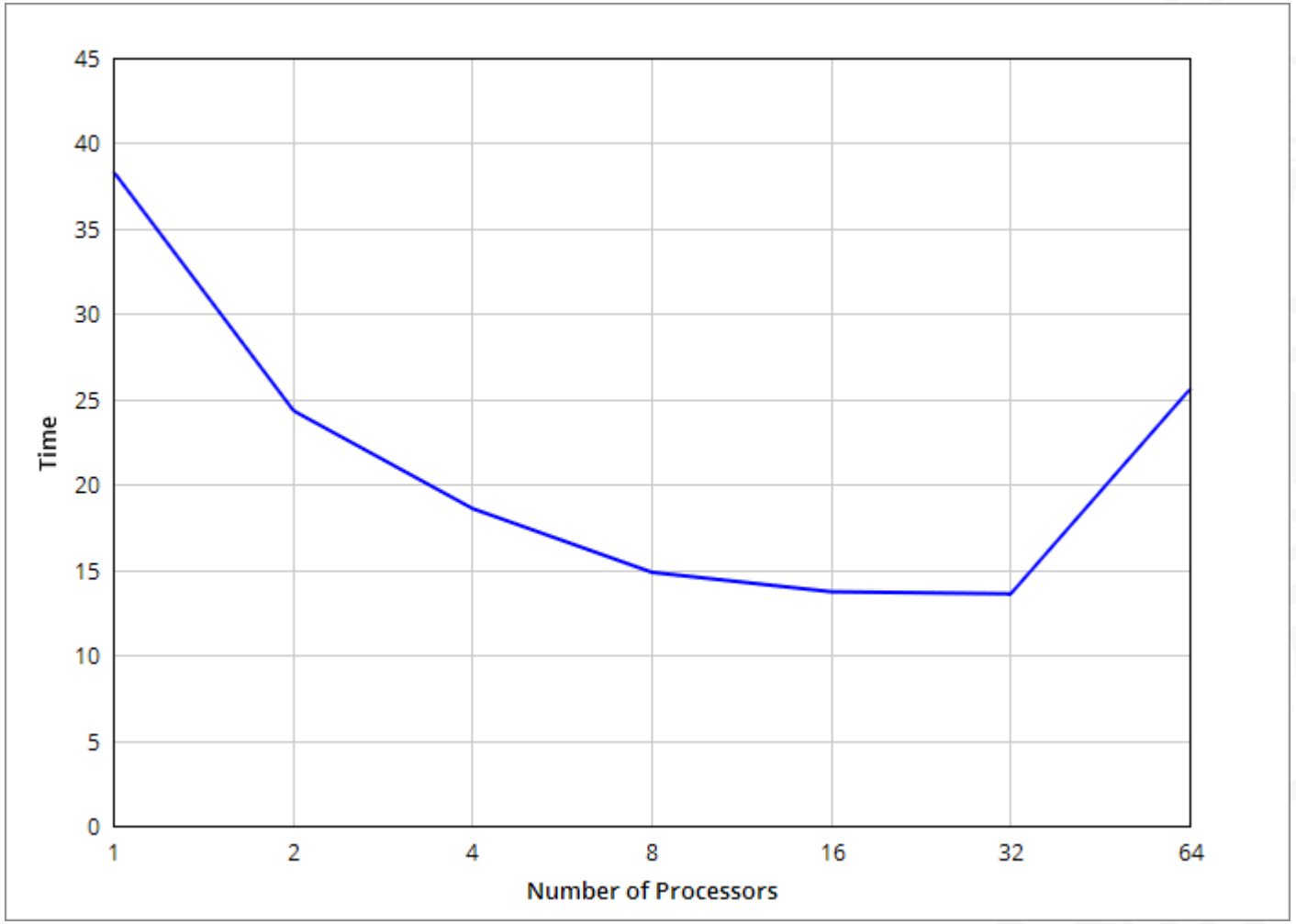
    c) processor with rank 0, receives the sum and length of the clusters and calculate the new clusters centroids, and broadcast it to all the processors

4. Repeat step 3 for n iterations

Output Analysis 1: Increasing number of processors

Data Points	1 Processor	2 Processors	4 Processors	8 Processors	16 Processors	32 Processors	48 Processors	64 Processors
2,000	0.04448549	0.04001685	0.033114235	0.035625418	0.044011593	0.047050953	1.366001209	1.646006505
20,000	0.286171277	0.216140787	0.125030597	0.099642396	0.089014967	0.10159413	1.489562472	1.681152662
200,000	2.81931746	1.854880174	1.308184942	1.036558032	0.910197198	0.892894506	3.073674162	3.129479885
2,000,000	38.29015589	24.32618809	18.5834624	14.84745844	13.70954235	13.59111623	25.21348433	25.6014063
20,000,000	514.4130695	404.2367609	377.4870376	357.0828284	371.6022041	408.4592911	771.4760255	784.7449865

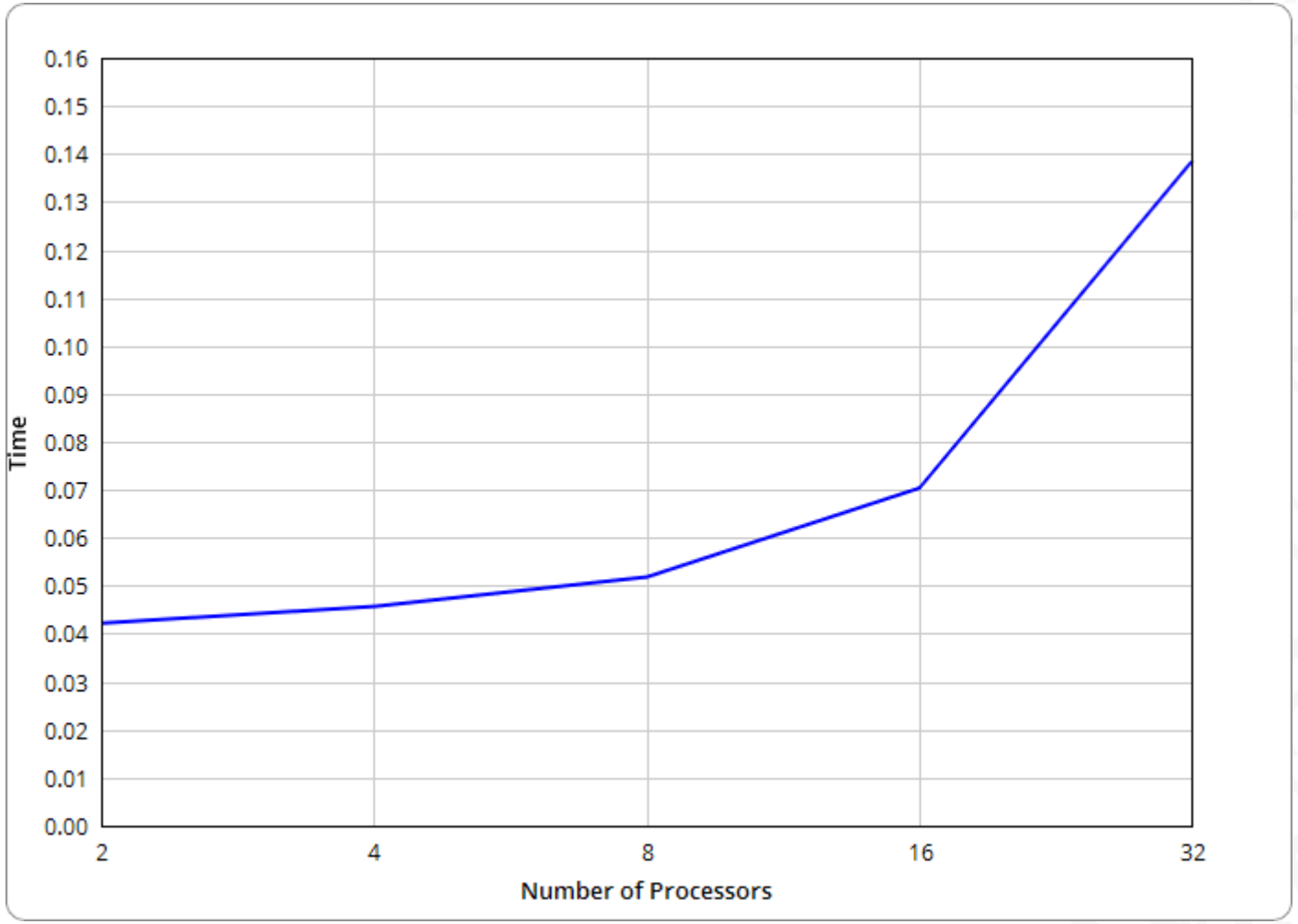
Amdahl's Graph (for 2,000,000 data points)



Output Analysis 2: Double Datapoints - Processors

2000 - 2	4000 - 4	8000 - 8	16000 – 16	32000 - 32
0.04221034	0.045693398	0.051833153	0.07037425	0.138515949

Gustafson's Graph



REFERENCES

- Algorithms Sequential & Parallel: A Unified Approach (Dr. Russ Miller, Dr. Laurence Boxer)
- https://ubccr.freshdesk.com/support/solutions/articles/13000026245-tutorials-and-training-documents (Dr. Matthew Jones)
- A Parallel K-Means Clustering Algorithm with MPI (Jing Zhang, Gongqing Wu, Xuegang Hu, Shiying Li, Shuilong Hao)
- Parallel K-Means Algorithm for Shared Memory Multiprocessors by Tayfun Kucukyilmaz ,Computer Engineering Department, University of Turkish Aeronautical Association, TR06800, Ankara, Turkey
- J. Bhimani, M. Leeser and N. Mi, "Accelerating K-Means clustering with parallel implementations and GPU computing," 2015 IEEE High Performance Extreme Computing Conference (HPEC), Waltham, MA, USA, 2015, pp. 1-6, doi: 10.1109/HPEC.2015.7322467.