Hybrid K-Means Clustering

Rahul Chittimalla
Sr. no: 06-02-01-10-51-17-1-14585
rahulc@iisc.ac.in

Sulbewar Vishwakiran
Sr. no: 06-02-01-10-51-17-1-14511
sulbewark@iisc.ac.in

1 Problem statement

Clustering is a basic search method for hidden patterns that may exists in the data. It is a process of grouping data objects into disjointed clusters.K-means algorithm is one of the most famous unsupervised clustering algorithms. By dividing a cluster of data objects into k sub-clusters as shown in Fig 1. It represents all the data objects by the mean values or centroids of their respective sub-clusters. This algorithm has the advantages of fast convergence and ease of implementation, but it has poor performance in some applications with large datasets.

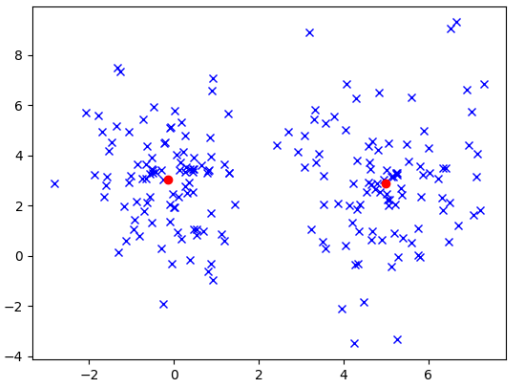

Figure 1: k-means cluster with $k = 2$

Nowadays, most computers are equipped with GPU's with plenty powerful Single Instruction Multiple Data(SIMD) Processors. The rapid advance in GPU's performance, coupled with recent improvements in its programming made it possible to parallelize K-means on computer systems.

2 Related work

Many recent works related to K-means based on the GPUs and using the parallel implementations using **MPI** have shown a great improvement in K-means performance. In [1] author explained a method of implementing K-means on Commodity GPU with CUDA. Another implementation in [2] using MPI, authored showed an improved performance and relatively stable and portable and it performs with low overhead time on large volumes of datasets. With increasing technology multiple GPUs have became common and in [3] author developed a method to scale clustering method using the multiple GPUs.

3 Proposed approach

Improving Technology has lead to the utilization of computing resources by developing different hybrid methods which use the GPU as well as CPU which further improves the performance. This work presents a hybrid approach that collaboratively combines the GPUs and CPUs available in a computer and applies it to the problem. Proper orchestration and decomposition of workload in such a heterogeneous system is an issue which is to be addressed in an efficient way.

Here we use an on-demand strategy whereby the computing devices request a new piece of work to do when idle. Hybrid approach thus takes advantage of the whole computing power available in modern computers and further re-

duces the processing time.

4 Baseline

In [1] author has implemented the K-means using the GPU and CUDA. We compare the approach presented in that paper & the sequential implementation with our proposed approach.

Table 1: Tentiative Project plan

Sl.No.	**Milestone/Target**
1	Implementing Sequential K-Means
2	Parallel implementation of K-Means on GPU using CUDA
3	Implementing the K-Means on CPU-GPU/hybrid method
4	Analysis of results

References

[1] B. Hong-tao, H. Li-li, O. Dan-tong, L. Zhanshan and L. He,K-Means on Commodity GPUs with CUDA,*WRI World Congress on Computer Science and Information Engineering*, Los Angeles, CA, 2009, pp. 651-655.

[2] J. Zhang, G. Wu, X. Hu, S. Li and S. Hao, A Parallel K-Means Clustering Algorithm with MPI,*Fourth International Symposium on Parallel Architectures, Algorithms and Programming*, Tianjin, 2011, pp. 60-64.

[3] M. K. Wasif and P. J. Narayanan,Scalable clustering using multiple GPUs,*18th International Conference on High Performance Computing*, Bangalore, 2011, pp. 1-10.