CA Lab Project

Ishaan R Dharamdas
Department of Computer Engineering
National Institute of Technology Karnataka
Surathkal, Mangalore 575025
Email: ishaanrd6@gmail.com

M R Ashrit
Department of Computer Engineering
National Institute of Technology Karnataka
Surathkal, Mangalore 575025
Email: ashuabhi260898@gmail.com

I. Problem Statement

Accelerating K-Means on the Graphics Processor via CUDA

A. Description of K-Means clustering algorithm

K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the K-means clustering algorithm are:

The centroids of the K clusters, which can be used to label new data

Labels for the training data (each data point is assigned to a single cluster)

II. Project Execution Plan

Implement the K-means clustering algorithm in serial and parallel to determine the performance difference between the two methods.
Implementation will be done using the NVIDIA CUDA Architecture in C++. The input will be taken from a file and will consist of:

1. A set X of n data points $x_i \epsilon R^d$, i = 1, . . . , n.

2. Number of clusters k which belongs to the set of positive integers and is less than n.

A cluster C_j which is a subset of X, j = 1, . . . , k with a centroid $c_j \epsilon R^d$ is composed of all points in X for which c_j is the nearest centroid using euclidean distance.

The Output file will consist of k lines indicating the co-oridnates of the cluster center.

III. Project Timeline

31/10/18 - Finish implementation of the serial code for K-means clustering
09/11/18 - Finish implementation of the parallel code for K-means clustering
15/11/18 - Complete comparing the results and submit the project and relevant files

IV. Work Distribution

Both the members will contribute to the project equally and share all the work involved like gathering resources, input data sets for testing, and the algorithms and their implementation.

V. Progress Report

Studied and got a basic idea about what the K-means clustering algorithm is and where it is used. Skimmed over the serial code for K-means clustering. Decided the programming language and framework to be used for implementation of the algorithm.

References

[1] M. Zechner and M. Granitzer, *Accelerating K-Means on the Graphics Processor via CUDA*,, 2009 First International Conference on Intensive Applications and Services, Valencia, 2009, pp. 7-15.