15-440/640 Distributed Systems

Lab 4 Report: MPI: Clustering and DNA

Team:

Abhishek Bhowmik (abhowmi1),

Neil Rajesh Dhruva (ndhruva)

Date:

12/05/2014

**Contents**

**Problem Definition ………………………………………………………………….. 3**

**Solution Overview …………………………………………………………………… 4**

**Framework ……………………………………………………………………………… 6**

**Code Organization …………………………………………………………………… 11**

**Conclusion ………………………………………………………………………………. 13**

**Problem Definition**

Clustering is a task of grouping objects that are more similar to each other as compared to other objects in the dataset. Hence, the objects with the closest ‘distance’ to each other are put in one group. The goal of this project is derived from the idea of grouping similar objects using a distance metric in a parallel computation setting. The two main objectives of the project include:

1. Understand and implement a parallel computation algorithm for clustering in large datasets using the OpenMPI framework.
2. To analyze the performance of a parallel algorithm, and compare it with a sequential implementation of the same, for varying dataset sizes and varying degrees of parallelism.

**Solution Overview**

For this project, we have created parallel algorithms for clustering points in 2D space using the K-means clustering technique, and generalized it for use in the DNA clustering application. In addition, we have provided implementations for generating random datasets of user-defined sizes for both these problems.

For solving the clustering problem using K-means, we use the Euclidean distance between data points to find the data center closest to a given point. The point is then a ‘member’ of the cluster centered at that data center. In an iterative manner, all points closest to a given cluster center are identified, and the cluster center is recalculated as an average of these points. This calculation is parallelized by distributing the dataset equally among all participating nodes, calculating the membership of points locally on those nodes, and recalculating cluster centers using OpenMPI for communication between the nodes.

This idea is then extended to calculate cluster centers in a dataset containing DNA strands (randomly generated using a script provided with the code). The closest center to each strand is calculated using the maximum frequency technique for each DNA base at a given position in the DNA strand. Using OpenMPI for communication, K-means is parallelized for better performance across several nodes.

**Clustering Overview**

--flowchart

https://www.gliffy.com/go/html5/launch?app=1b5094b0-6042-11e2-bcfd-0800200c9a66