COMS3008 - Parallel Computing

Seale Rapolai (1098005)
Mahlekenyane Ts'eole (1118248)
Meriam Elabor (1076589)
Nkosikhona Sibisi (915702)

October 20, 2017

1 Introduction

This report focuses on the common problem of clustering data. It involves classifiying different data items that share similar characteristics within the same data set, into groups. By doing this, various clusters of data items that share similar characteristics can be identified. One of the problems that arises in trying to cluster data includes the amount of time required to classify each data item into a certain group (cluster). This report will focus on solving the performance problem related to clustering.

To solve the clustering problem introduced, an algorithm that classifies and groups the data based on their characteristics should be developed. Since the process of clustering is quite time consuming in itself, a proposed solution to this is to parallelise the entire process (i.e the process of clustering the input data). Parallelising this process will allow us to divide the amount of work involved in clustering the data amongst different processing items and hence cluster multiple data items concurrently.

There were different algorithms to consider in order to solve the problem of clustering. Due to the fact that exclusive clustering is a requirement in our solution to the problem, we will focus on the *K-Means* algorithm, instead of counterparts like *C-Means* and *Hierarchial* clustering algorithms. The *K-Means* works by picking k initial centroids then through iterating will produce k exclusive clusters. It is apparent that when the size of the dataset becomes increasingly large, the k-means and its clustering process will become increasingly slow. This will lead us to computing sections of the algorithm simultaneously through parallelisation techniques.

2 Solution technicalities

There are various clustering algorithms that could be used to cluster set of data into various clusters. This project will focus on using as well as parallelising the k-means clustering algorithm. In order to solve the above mentioned problem, the k-means cluster algorithm will be parallelised. In order to do this, various factors had to be taken into concideration, such as task dependencies, interactions between various tasks and mapping of tasks to processes. From this we will be able to determine the various factors such as the degree of concurrency.

2.1 Task Dependencies and Interactions

2.2 Task Decomposition

2.3 Process Mapping

2.4 Parallel Algorithm

How the code was parallelised

2.5 Limitation

3 Results analysis

3.1 Perfomance Analysis

3.2 Accurancy

4 Conclusion