COMS3008 - Parallel Computing

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

October 17, 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 these problems include the amount of time required to classify each data item into a certain group (cluster). This report will focus on solving the perfomrance problem related to clustering.

Since the process of clustering has been know to be quite time consumming, a proposed solutions 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 to cluster the data amonsts diffent processing items and hence cluster multiple data items concurrently.

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 parallelising the k-means clustering algorithm. In order to solve the above mentioned problem of parallelising the clusterning of data items, it was necesary to

first identify the various dependencies that exist within the clustering algorithm.

To do this, a serial version of the k-means clustering algothim was designed and in such a way so as to minimise the dependencies within the algorithm. This

3 Results analysis

4 Conclusion