Project proposal – Mohammad Atif Faiz Afzal

HPC1

Date- 10/13/2014

In our research group, one of the projects is to discover trends in data derived from different quantum chemical methods. The work focuses on the analysis and comparison of results of various molecular properties derived from different flavors of Density Functional Theory (DFT). These patterns in data have significant implications for the utility of the employed approximations and the design of new quantum chemical techniques. There is more than million data available for each method that has been generated. Currently, linear regression models is being used to connect results from different methods. The code is currently written in python and the SciPy.stats library is being used for linear regression model. The model in the library only compare between two features at a time and is a serial code thus taking very long time (more than an hour) to get the fitting. As there are many flavors, large number of correlations are required and thus the overall time of computing is very high. I want to create a single linear regression model to fit all the features at once and simultaneously reduce the computation time. As part of the HPC project I would like to write the model in FORTRAN and then parallelize the code to decrease the computation time.

Second aspect of the project is to extract the data that do not fit in the close vicinity to the linear regression fit. This is currently done by calculating the distance of each data point from the linear fit and separating the data points that lie away from the linear fit. For a set of million data, the code takes about 2 days to run just for a pair of features. The CCR time constraint of 72 hours would limit the use of this code for more than two features. As part of the HPC project I would also like to write the code in FORTRAN and parallelize it so that I can reduce the computation time and thus can also use the code on more than two feature models.

Third aspect of the project is to study the data points that do not lie in the vicinity of the linear regression by using machine learning techniques. Clustering will be performed and then the data will be classified into special class of compounds and examine these special classes. As part of HPC project I would like to apply clustering algorithm such as K-means and use MapReduce to decrease the computation time.

Final Objectives

1. Parallelize the code for linear regression model using OpenMP/MPI
2. Write a parallel code to extract the data that does not fit regression model
3. Apply machine learning technique to find different clusters of the extracted data and also implement data parallelization.