Modeling and Predicting Resource Management for Parallel Implementations of K-means

Author: Janki Bhimani
Advisor: Prof. Ningfang Mi
Electrical and Computer Engineering Dept.
Northeastern University
Boston, MA
Email: bhimani@ece.neu.edu

Advisor: Prof. Miriam Leeser
Electrical and Computer Engineering Dept.
Northeastern University
Boston, MA
Email: mel@coe.neu.edu

Acceleration using HPC is required for many time consuming applications including those involving big data and unsupervised learning. Currently, researchers perform experiments trying different hardware configurations to find the best performance point. This is time consuming, and moreover usage of more resources does not guarantee performance improvement. Hence, it would be advantageous if we could predict the optimal resources required for the best performance.

To address this problem, we have developed a performance-modeling framework that predicts resources. The input to our model is application and platform details and the output gives the amount of resources appropriate to use to obtain the optimal performance. Our analytical model combines statistical Markov modeling with machine learning techniques and queuing theory to predict resource requirements. It predicts the total time an algorithm spends in performing calculations with a thread level Markov model. It learns the hardware behavior during training, such that no instrumentation of hardware details like operational cycles and stalls is required during prediction. Our model also predicts the total time spent in network communications through a simulation queuing model. The queuing model provides an abstraction of real world systems in terms of job arrival rate, waiting time and service rate, which helps us to estimate system communication time. We developed different models to capture various patterns of data communication such as: uplink, downlink, broadcast and one-to-one.

We evaluate effectiveness of our modeling technique by using it to predict performance of the K-means clustering algorithm that is implemented on a distributed memory platform. K-means is one of the most popular clustering algorithms, which aims to partition \( n \) observations into \( K \) clusters. Each observation is assigned to a cluster with the nearest mean and based on the updated assignment, new means are computed. This process is repeated iteratively until the desired quality of clusters is achieved. Performance of K-means depends mainly on three parameters: the size of the input dataset, the number of clusters to be formed and the convergence criteria to decide the number of iterations. Additionally, in a parallel implementation, the total time to run such an algorithm further depends on the amount of parallel resources. To qualify the accuracy of our prediction technique, we rank resources and compare the predicted and actual ranking lists. Our experimental results show that the correlation between actual and predicted ranking is in the range 0.82 – 0.99 that indicates good accuracy.
There are four major features of our technique which makes it unique: 1) it is hardware independent; 2) it is capable of predicting resource utilization statistics for calculation and communication separately; 3) it is independent of algorithm related parameters; 4) different from most analytical models, it can be trained on datasets that need not be similar to the actual data. In the future, we plan to evaluate our model on other applications. We also plan to expand its scope to predict the performance of computing platforms with accelerators such as GPUs.