15-440: Distributed Systems
Rubrics of Project 3

School of Computer Science
Carnegie Mellon University, Qatar
Fall 2014

DNA Dataset Generator (Points: 5%)

- Generator takes at least the number of clusters and number of points per cluster as parameters (5%)
  NOTE: Simply producing random strands will be awarded 1% only

Sequential Versions (Points: 12%)

- K-Means for 2D data points (6%)
- K-Means for DNA strands (6%)

MPI versions (Points: 60%)

- K-Means using MPI for 2D data points (30%)
- K-Means using MPI for DNA strands (30%)

Details (these are same for DNA and 2D data points versions):

- ✓ The code compiles and runs (1%)
- ✓ Initialize MPI and get the size and rank (1%)
- ✓ The master loads data from a file and sends workers their shares (3%)
- ✓ The workers receive data from the master (3%)
- ✓ The workers and the master apply K-means to their portions of data (adopting a good mechanism to compute new centroids) (6%)
- ✓ The workers send their intermediate data (either the summations and the count, or the count and average) to the master (5%)
- ✓ The master receives intermediate data from workers (3%)
- ✓ The master aggregates data and calculates the new means (3%)
- ✓ The master loops over for a new round (3%)
- ✓ After done, cleanly abort the program using the right MPI functions (1%)
- ✓ Compute the runtime correctly (1%)

Write-up (Points: 20%)

- Three scalability studies (plots and analysis) (15%)
- Experience in applying MPI to the K-Means algorithm (trade-offs of sequential vs. parallel) (2%)
- Thoughts on the applicability of K-Means to MPI (2%)
- Paper structure, level of writing, and language (1%)

Code Style (Points 3%)

- Method Comments, Block comments, Readability, Dead code, Code Design (3%)
  NOTE: Well-document functions that re-compute 2D and DNA centroids