RISE:2015
RESEARCH, INNOVATION AND SCHOLARSHIP EXPO

Category: Engineering and Technology
Degree Level: Graduate
Abstract ID# 892

Northeastern University

Predicting Parallel Processing of Machine Learning Algorithms

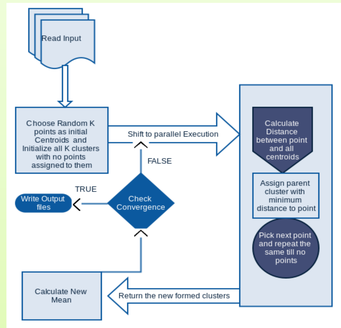
Janki Bhimani Jennifer Dy Miriam Leeser Ningfang Mi

1. Overview

- Develop new techniques and approaches to speed up machine learning algorithms for large-scale data processing.
- Objectives
  - ✧ Explore **parallel implementations** of machine learning algorithms on computer clusters w/ and w/o a variety of accelerators
  - ✧ Develop new **capacity planning models** to give a data analyst quick information about the execution time of a machine learning algorithm
- Driven applications: the classic k-means clustering algorithm parallelized using *MATLAB Parallel Computing Toolbox*.

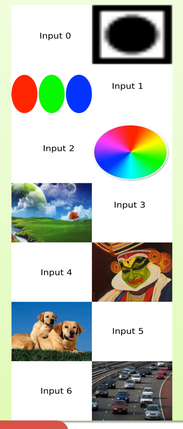
2. Parallelization

- Three general classes of unsupervised learning methods:
  - ✧ K-means clustering
  - ✧ Mixture models
  - ✧ Latent Dirichlet Allocation (LDA)
- K-means clustering
  - ✧ Parallelize the "assignment" in each iteration
  - ✧ MATLAB Parallel Computing Toolbox
    - *parfor, spmd*
  - ✧ CUDA or MPI programming

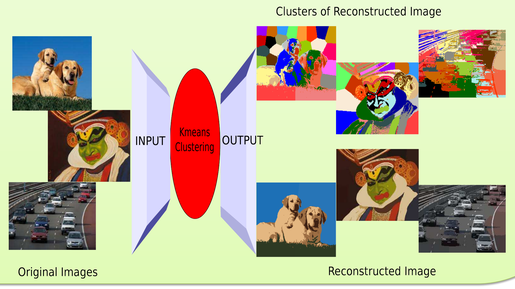


3. Case Study

- Parallel k-means clustering for image processing

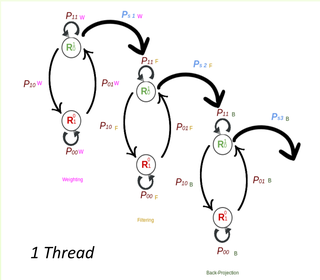


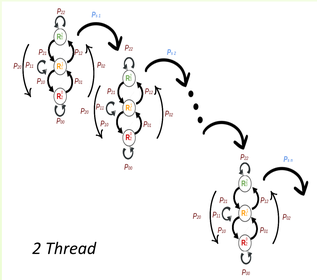
Input	Dimension
Input0	15x15 pixels
Input6	275x183 pixels
Input4	260x194 pixels
Input5	251x201 pixels
Input3	320x200 pixels
Input2	300x300 pixels
Input1	640x320 pixels



4. Performance Prediction

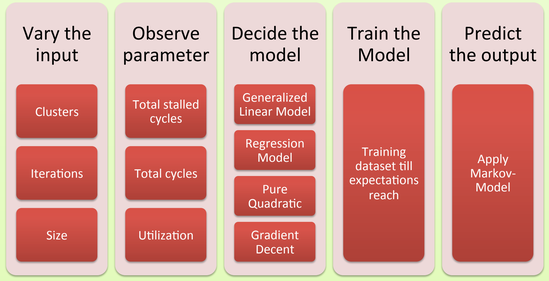
- *FiM* – a Fine-grained Markov model for predicting execution times
  - ✧ state of each thread: *active* (R_0^1) or *passive* (R_1^0)
  - ✧ partition the processing of an application into multiple stages

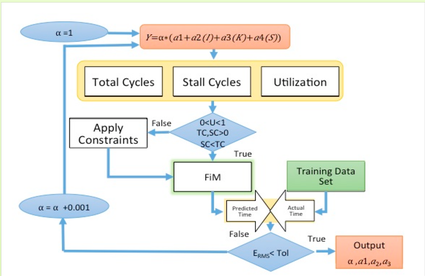




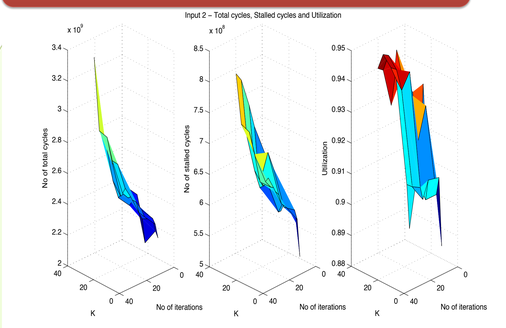
- *Transition probabilities*: switch between two states (P_{10}) or self-loop in the same state (P_{11})
- *Completion probabilities*: transfer from one stage to the other one (e.g., P_{s1} and P_{s2})
- Instrumentation tools (*perf stat*) to obtain required data
- *Execution* time per stage: $T_i = 1/(P_{si} \cdot F)$

- *Learning trend to* refine models to be independent on hardware parameters
  - ✧ develop a new regression model with a multiplicative tuning factor

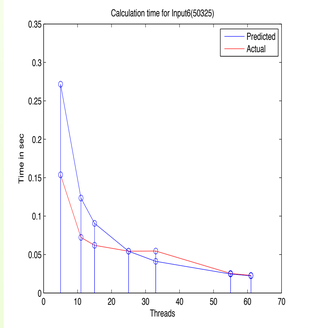


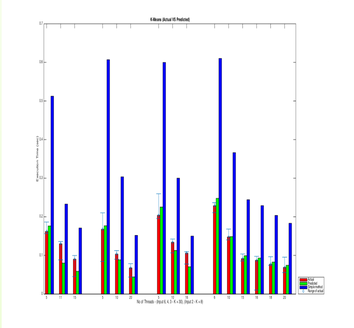


5. Preliminary Results

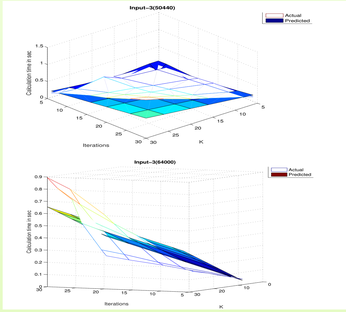


(a) Learning trend of total cycles, stalled cycles and utilizations





(c) Actual v.s. Predicted results



(b) Predicated results using the modified linear regression model

Summary of this project

- ❖ Capture the characteristics of heterogeneous hardware architectures.
- ❖ Predict the behavior of an application running on an array of computing platforms.
- ❖ Configure the right computing platform and choose a good way (MATLAB Parallel Computing Toolbox, CUDA or MPI) to parallelize unsupervised learning algorithms.

- Referances:

- ✧ T. M. A. Xi E. Chen, "A first-order fine-grained multithreaded throughput model," in High Performance Computer Architecture, 2009. HPCA 2009. IEEE 15th International Symposium on. IEEE, 2009, pp. 329–340.
- ✧ D. E. C. Rafael H. Saavedra-Barrera, "An analytical solution for a markov chain modeling multithreaded," 1991.

MathWorks AFOSR http://nucsrl.coe.neu.edu