A DATA CORRELATION-AWARE FRAMEWORK FOR SPARSE REGRESSION IN THE CLOUD

Jin Kyu Kim, Seunghak Lee, Eric Xing, Garth Gibson (Carnegie Mellon University)

MACHINE LEARNING AND THE CLOUD

- **Analytics on Big Data is increasingly about discovery, that is mechanized learning**
- **Machine learning techniques are mostly serial and often just mathematics**
- **Big Data drives machine learning to seek out scalable parallel algorithms and implementations**
- **For a while, this is best done case by case, implementing parallel ML algorithms using low level system abstractions. Now, they are seeking generalized frameworks**

GENOME-WIDE ASSOCIATION MAPPING

- **Goal is to find SNPs (genetic variations) which indicate/predict disease status or gene expression level**
- **As the cost of DNA sequencing decreases, large data sets become more readily available**
- **GWA analysis is popular for choosing safe therapy, estimating the risk of a disease, and making drugs for specific individuals**
- **One typical case: Alzheimer's disease analysis:**
  - **I.e. Four hundred patients provide DNA data. Each patient's information consists of 1 million SNPs and 20 thousand gene expressions**
  - **Computational demand of this regression solution is huge**

Need scalable algorithms and implementations

STATE OF ART PARALLEL LASSO REGRESSION

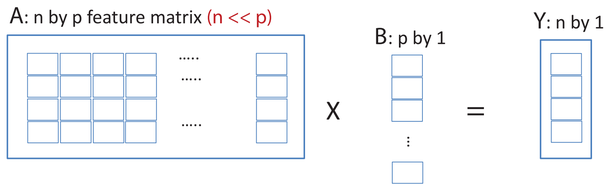
- **Shotgun is a recent parallelization of a Lasso Coordinate Descent Algorithm**
  - **It updates M parameters at a time, where M is the number of computing cores**
  - **Updated parameters are propagated immediately (synchronously)**
- **Potential disadvantages when applied to large computers**
  - **Each parameter update incurs a network transfer**
  - **Strict synchronization between two iterations limits the progress of algorithm to the slowest computing core**

INITIAL RESULTS WITH SMALL DATA

- **Correlation-aware execution prevents divergence as the number of cores increases**
- **Longer intervals between bulk updates does not hurt convergence rate and correctness while reducing the number of network transfers substantially**
- **Correlation-aware execution helps improve the convergence rate of a Shotgun approach**

TARGET PROBLEM: SPARSE REGRESSION

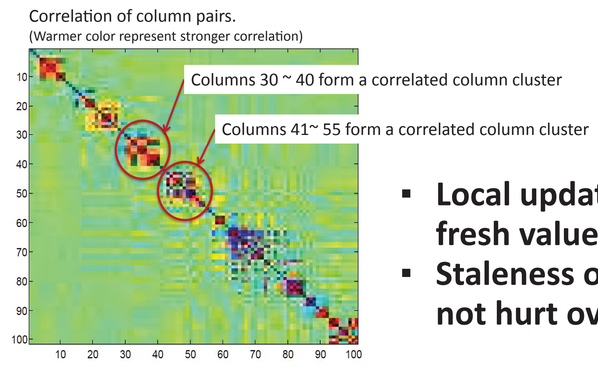
- **Goal is to find B "weight" vector for given A, Y**



The Lasso (least absolute shrinkage and selection operator) formulation of a least squares regression on underdetermined systems (n equations in p unknowns) prefers solutions with few non-zero unknowns, effectively selecting a small number of key dependencies (columns of A), e.g. Lasso is used for Genome-Wide Association (GWA) mapping.

KEY OBSERVATIONS ON BIO-DATA (SNPS)

Small number of columns (SNPs) form correlated column clusters where each cluster member is correlated to other members in the same cluster, but are independent of all others.

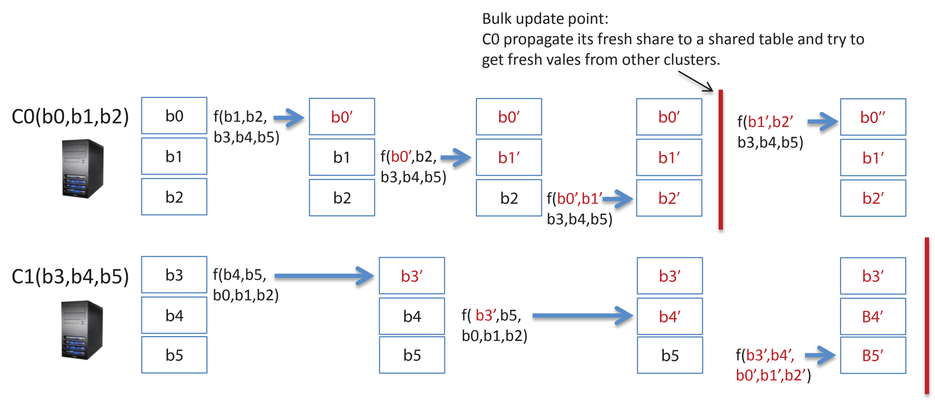


- **Local update in a cluster does not need fresh values from other clusters**
- **Staleness of non-local variables does not hurt overall correctness**

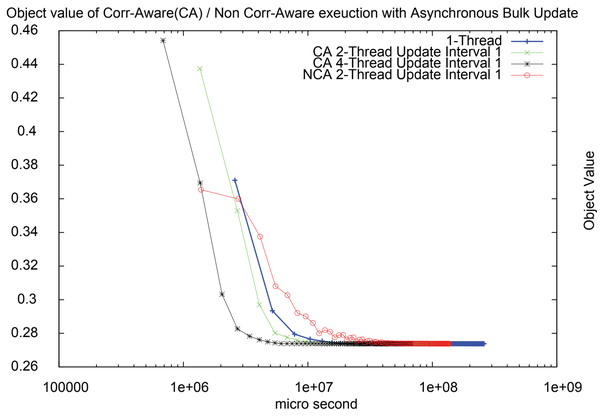
NEW IDEAS

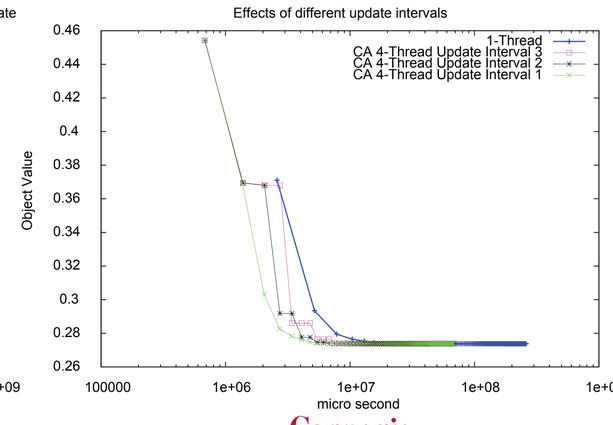
Correlation Aware Execution and Asynchronous Bulk Data Update

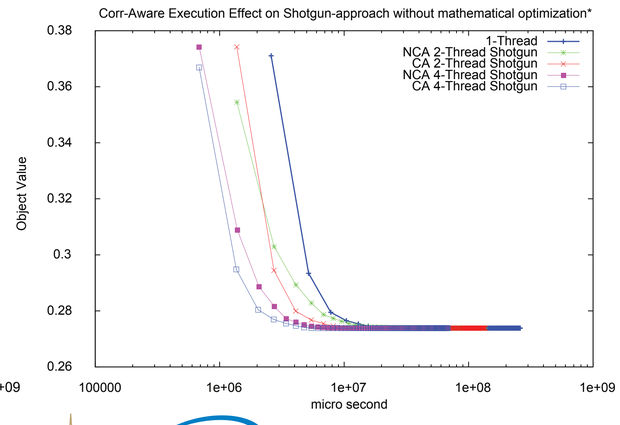
- **Within parameters in a correlation cluster: Serial execution and immediate synchronous data update**
- **Between uncorrelated parameters: Asynchronous execution and asynchronous bulk data update**



A: 463 patients with 1024 SNPs. Y: generated using ground truth B + noise.







Intel Science & Technology Center for Cloud Computing

Carnegie Mellon University · Georgia Tech · intel · PRINCETON UNIVERSITY · UC Berkeley