CS240A Project

- **Project proposal**
  - Teams of 2+ students.
  - Challenging parallel application or system
  - State-of-art: Conference-quality papers as a leveraging point
    - ACM/IEEE/SIAM/USENIX conferences (SuperComputing, SIGIR, SIGMOD, OSDI, WWW, VLDB, WSDM, KDD, etc)
  - Success metrics
    - Understanding performance, tuning, scaling, etc.
    - Quantify why it is a challenging problem, and why you get a great performance

Some Project Ideas

- **Examples**
  - Parallel applications
    - Data mining. Ranking (parallel algorithms).
    - Duplicate detection
    - Secure search (encrypted data, slow to process)
    - Search engines/graph search etc.
    - Matrix multiplication for similarity/recommendation
  - Systems
    - Speedup Mapreduce I/O. Incremental computing
    - Integrate MPI with Mapreduce
  - Parallel storage systems.

Timeline

- Feb 1- 4: preliminary proposal
- Select paper(s) for reviewing.
- Feb 21 week (probably delay 1 week): Project progress/technology presentation
- Final/exam week
  Final project presentation. Short report.

Sample Example: Deduplication in Versioned Dataset and Cloud Archival

- Build a parallel cloud backup system with deduplication support with various constraints
  - Manage each file as a set of chunks
  - Given n files on a storage, find other files that have the identical chunks
  - Compare a set of new chunks with an existing set of chunks

References:

W. Zhang, et. al *Low-Cost Data Deduplication for Virtual Machine Backup in Cloud Storage* . USENIX HotStorage'2013. Slides.

Parallel Secure Inverted Index Search

- Develop a prototype inverted index search code on a multi-threaded multicore shared memory system
  - Input : inverted index for a set of words
  - Output: find results matching a query
- **Reference: Dynamic Searchable Encryption in Very-Large Databases: Data Structures and Implementation**
  David Cash, Joseph Jaeger, Stanislaw Jarecki, Charanjit Jutla, Hugo Krawczyk, Marcel Rosu, Michael Steiner
  NDSS 2014. [eprint]

Fast Regression Tree Computation

Fast Boosted Regression Trees for Classification and Ranking.

- Compute a score for a document vector using a set of decision trees (for example, add partial scores from many trees).
- Develop cache-aware computation. Possible parallel extension.
- Claudio Lucchese, Franco Maria Nardini, Salvatore Orlando, Raffaele Perego, Nicola Tonellotto, and Rossano Venturini. Quickscorer: A fast algorithm to rank documents with additive ensembles of regression trees. In SIGIR, pages 73–82, 2015

Fast Similarity Comparison

- Similarity computation.
  - Two items are similar if their vector multiplication value > threadshold.
  - N- item similarity is a matrix multiplication problem.
- Data:
  - Wikipedia, Twitter. Yahoo music
  - Application: Recommendation system uses item similarity computation.
- PL2AP: Fast Parallel Cosine Similarity Search.

David C. Anastasiu, George Karypis. 2015 Supercomputing Workshop.

http://glaros.dtc.umn.edu/gkhome/node/1178.

TREC data

- **TREC Disk 4/Disk 5**
  http://www.nist.gov/tac/data/data_desc.html

  News articles, congressional records

  Test queries/evaluation from
  http://trec.nist.gov/data/t13_robust.html

  http://trec.nist.gov/data/qa/t8_qadata.html for TREC-Q&A queries/evaluation
- **AQUAINT data (news articles)**
  - TREC-2005, Robust track queries/evaluation

  http://trec.nist.gov/data/t14_robust.html
- **TREC enterprise 2008 data**
  - Text data/queries/evaluation
    http://trec.nist.gov/data/t17_enterprise.html

Other Datasets

- **Wikipedia data sets**
  - http://www.search-engines-book.com/collections/
  - 121K documents, 715MB.
  - More from http://en.wikipedia.org/wiki/Wikipedia:Database_download
  - http://dumps.wikimedia.org/ for old dumps

- **Edmunds Car review & Tripadvisor Hotel review**
  - from Tripadvisor (~259,000 reviews) and Edmunds (~42,230 reviews).
  - http://archive.ics.uci.edu/ml/datasets/OpinRank+Review+Dataset.

Processed Datasets

- **Microsoft Web page ranking data** (LETOR)
  - Feature vectors extracted from query-url pairs along with relevance judgment training data (10K, 30K).
- **Yelp competition**
  - http://www.yelp.com/dataset_challenge/
  - 11,537 businesses, 8,282 checkin sets, 43,873 users, 229,907 reviews