Big Data Computer Systems: Course Overview

Kai Shen

General Course Information

- Course Web page
  - http://www.cs.rochester.edu/~kshen/csc296-fall2013/
- Course email address
  - cs576@cs.rochester.edu
- Text and references
  - No official textbook, will use online resources and papers
- Account in computer science research network
  - If you don't have one, sign up for one
  - UG account is not enough

Interaction, Please!

- This isn't an established course; I am learning along with you.
- I strongly encourage discussions and interactions.
- Extra credits for strong participation.

Assignments and Final Project

- Three programming assignments
  - First on data collection
  - The other two on parallel data processing/analysis
- A survey on a big data subject of your choice
- A final course project on a topic of your choice
  - Proposal
  - Presentation
  - Demo

Pre-requisite and Programming Requirement

- No formal prerequisite
- Desire good programming skills
  - Comparable to CSC252
  - Need to know Java
  - Abilities of learning new programming tools/techniques – e.g., scripting language, threads, network sockets, …
  - NO need to know C/C++

What is Big Data?

- The recognition that data is at the center of our digital world and that there are big challenges in collecting, storing, processing, analyzing, and making use of such data.
- To me, "big" may refer to very large data volume, but not necessarily so.

A Computer Systems Course

- Big data is a broad concept that covers many aspects of computer science.
- We focus on the computer systems aspect---for instance,
  - How various parts of a big data computer system (hardware, system software, and applications) are put together?
  - What are the appropriate approaches to realize high performance, scalability, reliability, and security in practical big data computer systems?
- Probably not the right course if you are hoping to learn about algorithmic design and theoretical/mathematical foundations for machine learning and data mining.

Kinds of Data

- Web data and web data accesses
- Emails, online chats, tweets, …
- Telephone data
- Public databases – GeneBank, …
- Private datasets – medical records, stock trades, credit card transactions, …
- Sensor-ed data – camera surveillance, wearable sensors, seismic data around an earth fault line or volcano, …
- Byproduct of computer systems operations – power signal, CPU events, …
- … …

Data is Valuable

- Google and Facebook build their businesses on mining user data (web searches, social network interactions) for advertising purposes.
- Hedge fund companies analyze financial records, (real-time) transactions, or web/social media for opportunities of profitable trades.
- Health, medical data is processed for enhanced health care and treatment.
- Highway is monitored for traffic analysis and control.
- What else?

Data Centers

- Containing racks of machines and storage
- Size of warehouses
- Sometimes built next to rivers because
  - Cheap energy from nearby dams
  - Good corporate image for using renewable energy

"Big" Data in "Small" Systems

- Data sources are sometimes in remote areas ⇨ Systems are necessarily small due to deployment and power constraints
- But data collection and processing is still the center of these systems

Photos are from videos in work by Wolff et al. 2012

Collection of Big Data

- Collection/acquisition of big data is challenging
  - Difficult to get access to valuable data
  - It stresses the computer system's ability to acquire a lot of data efficiently
  - It is also difficult to collect useful information from a sea of irrelevant data
  - Data collection should not negatively impact the target system's operations
- An example – Web data collection
  - How to crawl the web efficiently, on the right topic, without affecting the normal uses of the web?

Processing and Analysis of Big Data

- Processing large datasets is time consuming
  - Parallel data processing is necessary
  - But parallel data processing is challenging
- Mapreduce
  - Parallel data processing with easy programming and automatic support of data movement, load balancing, and fault-tolerance
  - Originated in web data processing (counting words); suitable for easily parallelizable workloads
  - But limited semantics
- Threaded and networked data processing in parallel

Data Representation and Organization

- Relational databases and SQL
  - Hard to scale for big data, no mainframe is big enough for big data
- Key-value, nosql stores
  - Bigtable
- Specialized indexes
  - Inverted indexes for web search
  - Multi-dimensional data organization

Storage and I/O

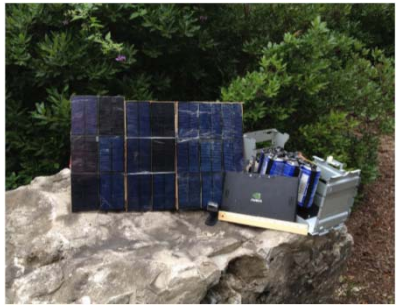
- Storage and I/O are critical for big data performance and reliability
- Hardware: disks, Flash, SSD, nonvolatile memory
- Parallelism: RAID, parallel data storage and file system
- Data durability and consistency

Energy

- Energy efficiency in data centers
  - A huge financial and environmental issue
- Data center construction from low-power computers [Anderson et al. 2009]
  - Think of a stack of tablets
  - Low joules per unit of work compared to conventional data center
- Data centers on renewable energy
  - Hydro-power, wind, solar, …

Sustainability

- Minimize environmental harm in field data collection
  - Use renewable energy; no batteries



Data Privacy and Protection

- Mis-uses of big data is a big concern
  - E.g., information of a person's online activities may reveal every aspect of the person's life
- Systems provide clear guidelines on data privacy and protection
  - E.g., sensitive clinical information is not propagated to datasets used for medical research
  - Computer systems are equipped with proper mechanisms to ensure data privacy and protection
- A user needs to understand the ways that the big data world operates on