EECS 684:
Current Topics in Databases

Barzan Mozafari

mozafari [at] umich.edu

4769 CSE

About Me

- PhD from UCLA 2006-2011
- Postdoc at MIT CSAIL 2011-2013
- Assistant Professor in CSE 2013-now
- Research:

  Applying theory to build systems

  - large-scale data-intensive systems, database-as-a-service, and crowdsourcing
- Office: CSE 4769

  http://web.eecs.umich.edu/~mozafari/

Today's Class in a Glance

- Introduction
  - What is EECS 684?
  - Getting to know each other

Course Overview

- Regular class meetings:
  - TuTh 1:30-3 PM *(Class starts at 1:40 on Michigan time)*
- Classroom
  - 1003 EECS
- Office Hours:
  - By appointment only (4769 CSE)
- Course Website:

  http://web.eecs.umich.edu/~mozafari/fall2013/eecs584/
- Prerequisites:
  - EECS 484 or 485 or 584, or equivalent course, or instructor's permission

Course Syllabus

- Current Topics in Databases
  1. Predictability of performance
  2. Human-database interaction
  3. Distributed transaction processing
  4. Multi-core databases
  5. Large-scale machine learning

Course Objectives

- Learn advanced data management concepts
  - for understanding the state-of-the-art
  - for doing research in data management
  - for effective use of current technology

- Learn important research skills
  - Read & critically evaluate research papers
  - Present technical material (orally and in written form)
  - Conduct (small-scale) original research

Papers & Reviews

- You will read 1 paper per class
  - No official text; papers available on course website
  - Background / reference: Database Management Systems (3rd Edition) by Ramakrishnan & Gehrke
- Read paper and post a 300-400 word *summary* by 11:59 PM the Sunday *before* class
  - What problem is addressed? Why important?
  - 1-2 main technical contributions? Describe.
  - 1-2 weaknesses or open questions? Describe and discuss.
- Note: no reviews needed for paper presented before Jan 21, i.e. the *first submission due* *Jan 19*

Seminar Format

- Typical class meeting:
  - Student-prepared paper presentation (40 minutes)
  - Class discussion (30 minutes)
- Paper presentation goals:
  - Motivate the paper, provide background
  - Highlight key contributions
  - Explain important technical points
    - **Examples, examples, examples!**
- Use template on course website

Presentation Advice

- Good presentations come from lots of preparation
  - Give yourself plenty of time to read & understand paper
  - You may have to read cited & related papers to fully understand
  - Revise your slides several times
- Beware of presenting too much technical detail
  - Pick most important 1-2 technical items; summarize the rest
- Put yourself in authors' shoes
  - Why did they do what they did?
  - Are their decisions still good ones?

Discussion Format

- Discussion Format
  - Instructor will lead discussion
  - Be prepared with key questions from the paper
  - I'll ask specific students questions if needed
- Examples of good discussion questions:
  - Locking and optimistic concurrency control can both be used to provide transactional semantics. When would it be better to use locking? What would it take for a DBMS to implement both kinds of concurrency control?
- *All students* are expected to participate in discussion!

Course Project

- Major component of the course
- Small-scale original research project
- Opportunity to study a database-related problem in depth
- Work on a project individually
- Not in DB? Choose a project that fits your other interests
- Choose your project by early October

Grading

Paper Reviews 25%	*11:59 PM of the Sunday before* class Student graded (1-3)
Class Discussions 15%	May skip 2 lectures w/ legitimate reasons
Paper Presentations 20%	Graded 1-3 based on quality 2 mandatory presentations 1 optional (extra 5% credit)
Final Project 40%	Proposal, final presentation, and final paper

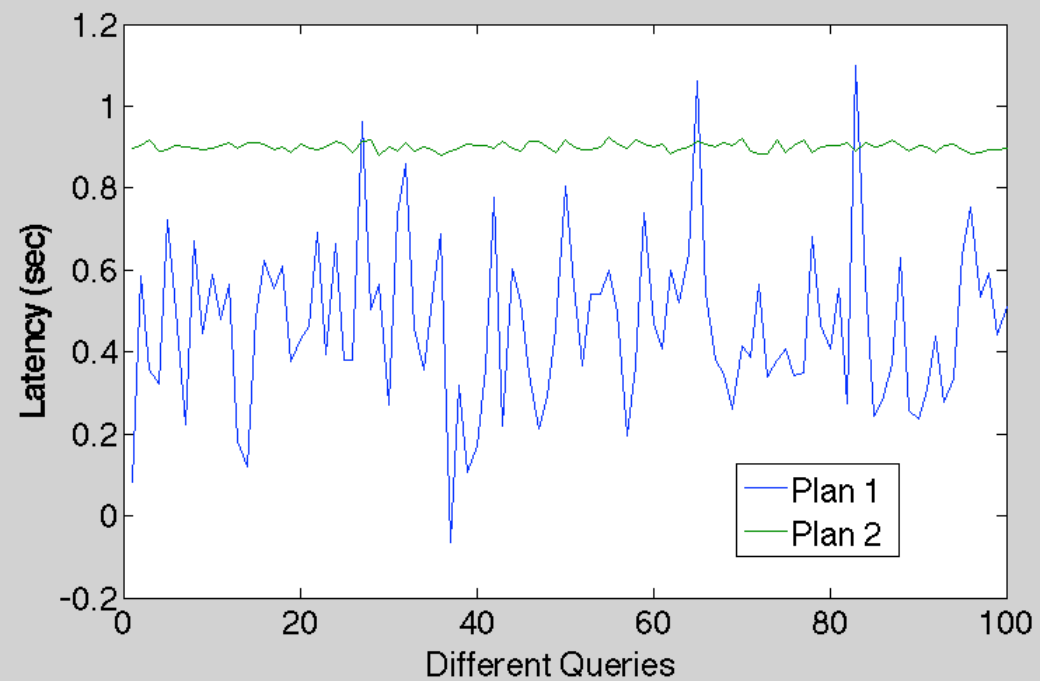
Presentation Logistics

- Decide on a date
- Decide on a paper (choose from reading list, or find your own paper!)
- Student presentations begin *Tuesday Jan. 21*

Course Overview

1. Predictability of performance
2. Human-database interaction
3. Distributed transaction processing
4. Multi-core databases
5. Large-scale machine learning

1. Predictability of Performance



Why do we care about predictability?

- Cloud Computing (Database-as-a-Service)

2. Human-Database Interaction

- Databases are a pain to use!
  - Operating a DB requires mad skills
  - Tuning a DB is black-art
  - Maintaining a DB is labor-intensive
  - DBs are not flexible
    - You need a rigid schema
    - You need to express your query in SQL
    - You need to provision resources in advance
- How can you make a DB easier to use?

3. Distributed Transactions

- What is a transaction?
- ACID properties:
  - Atomicity
  - Consistency
  - Isolation
  - Durability
- Why transactions need to be distributed?
- Trends:
  - NoSQL
  - NewSQL

4. Modern Hardware & Databases

- DBs were designed w/ 30 years, assuming:
  - Memory does not fit the entire data
  - Memory is volatile
  - Random reads slower than sequential seeks
  - There are few CPU cores
- Which one of these assumptions are still true?
- How to design DBs that can take advantage of modern hardware?

5. Machine Learning and Big Data

- Data is as useful as you make it!
  - Machine Learning (ML): a popular method for extracting value from raw data
    - Online Advertisement
    - Human-voice recognition
    - …
- ML techniques tend to be computationally intense!
- How to design data-intensive systems that can cater to ML workloads?
  - Map-Reduce/Hadoop, GraphLab, etc

Introduce Yourself

- Answer these questions
  - What is your name?
  - Where are you from?
  - What is your affiliation / program at UM?
  - What type of job are you interested in?