CSE 510: Data Management in Cloud Computing Environments

Course Logistics

WELCOME TO CSE 510

Course Information

- **Instructor**: Iman Elghandour
  Assistant Professor at Alexandria University
- **Email**: ielghand@alexu.edu.eg
- **Office Hours**: Monday 12:00 PM – 1:00 PM or by appointment.

- **Lectures**:
  Monday 1:00 PM – 3:45 PM
  Room XXXX

Course Webpage

- http://www.alexeng.edu.eg/~ielghand/teaching/cs713/
- https://piazza.com/alexu.edu.eg/fall2012/cs713/home (updates + announcements)

Class Organization

- 3 papers every class
- For every paper:
  - 30 min presentation
  - 30 min discussion
- Read papers before coming to class
- Submit reviews
- Present papers

Grading

- Presentation (20%)
- Paper reviews (15%)
- Participation (15%)
- Semester-long Project (50%)
-

Presentations

- 2 -3 presentations per semester.
- Presentations are to be sent to the instructor prior to class by Sunday midnight.
- Describe any required background, the motivation for the work, and a brief description of the solution.
- Bring up interesting points that can open discussion with other students.

Paper Reviews

- Write one page review for papers we are presenting in class.
- Reviews are due on Sunday at midnight (the day before class).
- The review of each paper should be divided into three parts:
  - **Summary:** One paragraph (10 sentences maximum) to summarize the problem addressed in the paper and the proposed solution (contribution of the paper).
  - **Strong Points**: 2-3 strong points about this paper.
  - **Weak Points:** 2-3 weak points about this paper.
- *No need to submit paper reviews if you are presenting a paper the same week*.

Class Participation

- You need to read papers that we will discuss in class before coming to class.
- You are also expected to participate in discussion about the presented papers.

Useful Links

- **Reading Papers**

  http://www.cs.columbia.edu/~hgs/netbib/efficientReading.pdf

  http://blizzard.cs.uwaterloo.ca/keshav/home/Papers/data/07/paper-reading.pdf
- **Writing Reviews**

  http://www.cs.utexas.edu/users/mckinley/notes/reviewing-smith.pdf
- **Presenting a Paper**

  http://pages.cs.wisc.edu/~markhill/conference-talk.html

  http://infolab.stanford.edu/~widom/conference-talks.html
- **Writing Technical Papers**

  http://infolab.stanford.edu/~widom/paper-writing.html

  http://www.cs.uky.edu/~raphael/writing.html

Project

- It is worth 50% of your grade.
- Projects can be conducted individually, or in teams of two students.
- **Timeline**
  - **October 19, 2012**: Project proposal is due.
  - **January xx, 2012**: Project presentation in class.
  - **January xx, 2012**: Project report is due.

Project Proposal

One page document that includes the following:

- Title of the project.
- Names of the team members.
- Motivation for the problem.
- Description of the problem.
- Description of how the members of the team will attempt to solve this problem.
- Resources/platforms needed to solve the problem.

Project Deliverables

- A project report:
  - Written as a research paper.
  - Includes: motivation of the work, details of the solutions, implementation, and evaluation experiments.
  - Page limit is 10 pages.
  - Format: ACM SIG Proceeding Templates.
- A 30 minute presentation of the work.
- Source code and any required scripts required to run your experiments.
- A demo of your solution if applicable.

Project Ideas

- You are encouraged to come up with project ideas.
- Projects can be extensions to papers in the reading list.
- Projects can be about non-database problems that can be scaled into and exported into the cloud computing environment.
- Two types of projects:
  - **Research Project**
  - **Experiments and Analysis**

Introduction

Definitions

- **Database**: A large and persistent collection of (more-or-less similar) pieces of information organized in a way that facilitates efficient retrieval and modification.
- **DBMS**: A program (or set of programs) that manages details related to storage and access for a database.

Application of Databases

- There is lots of data (mass storage)
- Data is formatted
- Requirements:
  - persistence and reliability
  - efficient and concurrent access
- Issues:
  - many files with different structure
  - shared files or replicated data
  - need to exchange data (translation programs)

Database Management System

- Data Model
  all data stored in a well defined way
- Access control
  only authorized people get to see/modify it
- Concurrency control
  multiple concurrent applications access data
- Database recovery
  nothing gets accidentally lost
- Database maintenance

Transactions

An application-specified atomic and durable unit of work.

Properties of transactions ensured by the DBMS:

- Atomic: a transaction occurs entirely, or not at all
- Consistency: each transaction preserves the consistency of the database
- Isolated: concurrent transactions do not interfere with each other
- Durable: once completed, a transaction's changes are permanent

However!

- New emerging applications that are complex.
- Huge data.
- Examples of emerging applications:
  - Data stream management systems (continuous data).
  - Spatio-temporal applications (space and time change).
  - Scientific data management.
  - Large-scale data analytics.
  - Online Analytical Processing (OLAP).

Parallel and Distributed Databases

Why?

- Large data
- Nature of the application
- Improve the query performance (faster response time)

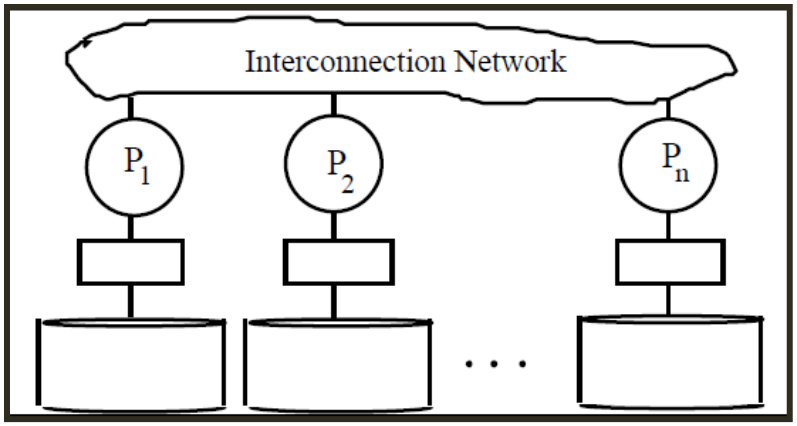
Overview

- **Parallel Databases**
  - Machines are physically close to each other, e.g., same server room.
  - Machines connects with dedicated high-speed LANs and switches.
  - Communication cost can be ignored.

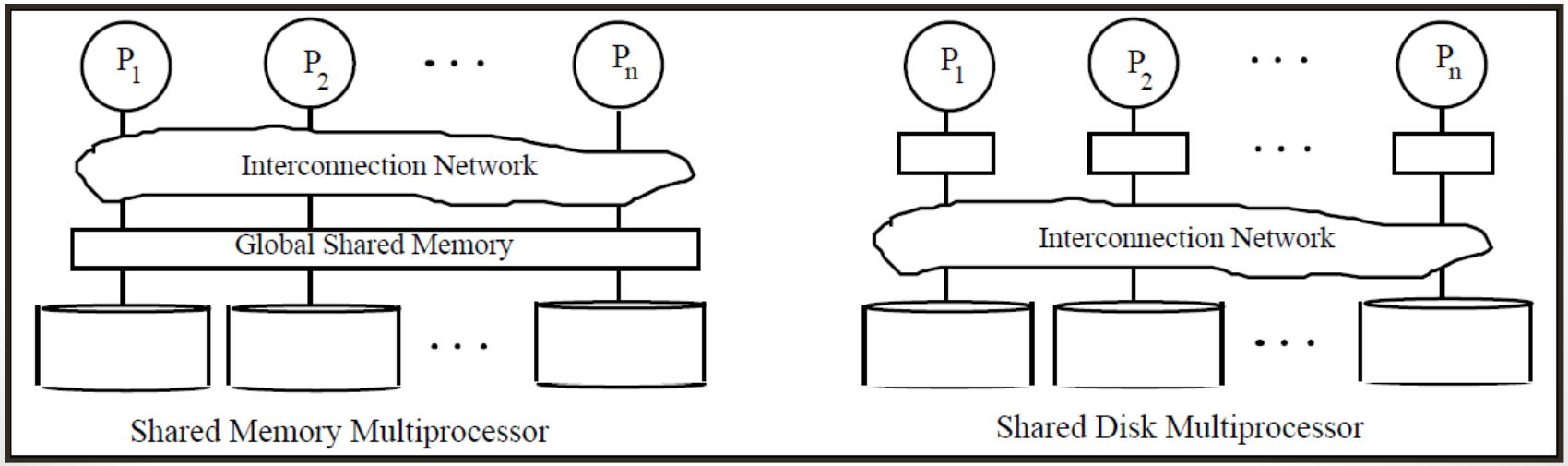
    shared-memory, shared-disk, or shared-nothing architecture
- **Distributed Databases**
  - Machines are not be physically located at the same place.
  - Can be connected using public network, e.g., Internet.
  - Communication cost and problems cannot be ignored.

    shared-nothing architecture

Systems with Multiprocessors



Shared-nothing design



Cloud Computing

- Key: hardware and/or software that is provides as services over the network (typically Internet).
- Common Categories:
  - IaaS: Infrastructure as a Service (e.g. Amazon EC2).
  - PaaS: Platform as a Service (e.g WaveMaker, AppEngine).
  - SaaS: Software as a Service (e.g. Google documents).
  - DaaS: Data as a Service (MySQL).

Topics

- **Background**
- **Distributed and Parallel Processing**
- **Data Analysis Frameworks**
- **Pricing in the Cloud**
- **Efficient Data Analysis**
- **Debugging Analysis Workloads- Performance Measuring**
- **Consistent Data Storage**
- **Cloud Databases**