CIS 700/005
Networking Meets Databases

Boon Thau Loo
Spring 2007
Lecture 1

Note: Some slides are courtesy of Jan Helmer and Scott Shenker

Overview

- Introductions
- Course logistics
- Overview of the course
- Review of database concepts

Introductions

- Assistant Professor since Jan ’07
- Areas of interests: databases and networking
- Ph.D. from UC Berkeley, Dec ’06
- Industry collaborations:
  - Intel Research Labs at Berkeley (’05/’06)
  - Visiting researcher at Microsoft Research (Aug-Dec ’06)
- Current areas of focus:
  - Distributed query processing and optimizations,
  - Internet-scale query processing,
  - Extensible data-centric network architectures
  - Large-scale DB-IR infrastructures

You?

- Name, year, background, interest, advisor(s), etc.

Contact Information

- Office: 605 Levine Hall
- Office hours: Wednesday 3-4pm or by appt
- Email: boonloo@cis.upenn.edu
- Home page: http://www.cis.upenn.edu/~boonloo

Course Logistics

- Enroll for credit:
  - Paper summaries: 30%
  - Class participation: 20%
  - Class presentation: 10%
  - Project: 40%
- Welcome to audit the class (officially or unofficially)
- Mix of lectures + paper discussions
- 1-1.5 papers per class
Paper summaries and presentations

- Summaries due noon of the day of class
  - Main contributions
  - Limitations
  - Propose improvements
- Pick your favorite paper to present
  - Week in advance notice
  - Work with me on presentations

Class Project

- Two options:
  - Research project (individual or groups of 2)
  - Survey paper (individual)
  - End-result: 6 pages workshop paper + in-class presentation
- A good project can become the basis for:
  - Publication (VLDB, SIGMOD, NSDI, SIGCOMM, IMC, etc)
  - Masters/Ph.D. thesis
  - WPE-II critical review
- My role is to work with you:
  - Suggest/develop ideas, regular meetings, provide tools
    - http://www.cis.upenn.edu/~boonloo/cis700-sp07/ideas/ideas.html

Primary Goal of the Course

- Explore topics at the intersection of database and networking
- Study how database concepts can influence the way networks are designed and built
- Equivalent courses in other universities:
  - Networked Information Systems, Distributed Data Management, P2P networks, sensor databases, etc
- Main difference:
  - Stronger networking flavor
  - Aim for impact in the networking community

For the Rest of Today

- From DB to NetDB: Emerging Synergies
- Topic Highlights
- Review Database Concepts

Database Systems Then

Slides courtesy of CS186: Fall 2008, UC Berkeley
Napster, Gnutella, KaZaA, etc.

Points to key social issues driving adoption of decentralized technology:

- Fairly trivial technology
- IP, SMTP, HTTP, SNMP log formats, firewall logs, etc.
- DoS attacks cross administrative boundaries
- Network forensics and accountability (various papers)

Knowledge plane (SIGCOMM ’03)
Information plane (OSDI ’04, OSDI ’06)

Who Needs Internet-Scale Querying?

Example 1: Filenames

- Simple ubiquitous schemas:
  - Filenames, Sizes, ID3 tags
  - Early P2P filesharing apps
  - Napster, Gnutella, KaZaA, etc.
  - Fairly trivial technology
  - But...
    - Points to key social issues driving adoption of decentralized systems
    - Provide real workloads to validate more complex designs for the future
- PIER + Gnutella infrastructure (VLDB ’04)

Example 2: Network Traces

- Schemas are mostly standardized:
  - IP, SMTP, HTTP, SNMP log formats, firewall logs, etc.
- Network administrators are looking for patterns within their site AND with other sites:
  - DoS attacks cross administrative boundaries
  - Tracking sources of viruses/worms
  - Timeliness is very helpful
- Hot topic today:
  - PlanetLab (distributed research test bed) is mostly filled with people monitoring the network status
  - Network forensics and accountability (various papers)
Future: Data-centric Internet

- "Clean-slate" Internet Designs:
  - NSF FIND and GENI programs
- Traditional Internet:
  - Host-centric protocols
  - Get data from here to there
  - Protocols defined in terms of IP addresses
- Web-based Internet:
  - DNS, URLs
  - Data-centricity in a host-centric world
  - Key-enabler: search engines

Routing Protocols as Dataflows

- P2 declarative networking system: distributed query processor for the core network architecture
- Network routing as database execution plans:
  - Query processors are dataflow engines.
  - So are routers (e.g., CLICK modular router toolkit).

Networking meets Databases

- Great research opportunities
- Potentially high-impact area:
  - New classes of data-intensive applications
  - Cross-disciplinary (networks, algorithms, databases, security, languages)
- Emerging publication venues:
  - VLDB Information Infrastructure track
  - NetDB (co-located with NSDI/ICDE),
  - MineNet (co-located with SIGCOMM)
  - Others (IPTPS, WORLDS, etc)
Topic Highlights: Building blocks

- Content-addressable networks:
  - Chord (SIGCOMM ’01), all others (survey paper)
- Internet-scale query processing:
  - PIER (VLDB ’03), SDMIS (SIGCOMM ’04)
  - Internet-scale publish-subscribe
  - SIENA (SIGCOMM ’03), ONYX (VLDB ’04)
- Cost-based network optimizations:
  - CORONA (NSDI ’06), XPORT (SIGMOD ’06)

Topic Highlights: Applications

- Data-centric network architectures:
  - 3 (SIGCOMM ’03), ROFL (SIGCOMM ’06)
  - Declarative networking (SIGCOMM ’05, SOSP ’05)
  - Revisit active networking research
- Network monitoring and security
  - Query languages and systems for monitoring
  - Network/database forensics and accountability
  - Uncertainty, heterogeneity and data reduction

Topic Highlights: Rest of semester

- Stream processing:
  - Continuous and adaptive query processing
  - Systems: TelegraphCQ, Aurora, HiFi, Borealis
- Other possible topics:
  - Sensor databases. E.g. TinyDB
  - Data management in delay-tolerant networks
  - Large-scale web infrastructures
  - Your suggestions and interests matter!

For the Rest of Today

- Emerging synergies between networking and databases
- Topic Highlights
- Review key database concepts

What is a DBMS?

- A Database Management System is a software system designed to provide data management services
- Examples of DBMS
  - Oracle, DB2 (IBM), SQL Server (Microsoft),
  - PostgreSQL, MySQL,…

Relational Data Model

- Collection of relations (or tables)
  - Each relation has a list of attributes (or columns)
  - Each relation contains a set of tuples (or rows)
  - Duplicates not allowed

<table>
<thead>
<tr>
<th>Student</th>
<th>Course</th>
<th>Grade</th>
<th>Enroll</th>
</tr>
</thead>
<tbody>
<tr>
<td>462</td>
<td>Unix</td>
<td>5</td>
<td>140</td>
</tr>
<tr>
<td>123</td>
<td>C++</td>
<td>3</td>
<td>122</td>
</tr>
<tr>
<td>500</td>
<td>SQL</td>
<td>4</td>
<td>697</td>
</tr>
<tr>
<td>456</td>
<td>Reln</td>
<td>2</td>
<td>426</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Student</th>
<th>Course</th>
<th>Grade</th>
<th>Enroll</th>
</tr>
</thead>
<tbody>
<tr>
<td>462</td>
<td>Unix</td>
<td>5</td>
<td>140</td>
</tr>
<tr>
<td>123</td>
<td>C++</td>
<td>3</td>
<td>122</td>
</tr>
<tr>
<td>500</td>
<td>SQL</td>
<td>4</td>
<td>697</td>
</tr>
<tr>
<td>456</td>
<td>Reln</td>
<td>2</td>
<td>426</td>
</tr>
</tbody>
</table>
Relational Algebra

- Core set of operators:
  - Selection, projection, cross-product, union, difference and renaming
  - Set operations
- Additional derived operator:
  - Join, etc

SQL Example

- Names of students in CIS 700 with GPA 4.0

```sql
SELECT Student.name
FROM Student, Enroll
WHERE Enroll.CID = CIS700
AND Enroll.SID = Student.SID AND Student.GPA = 4.0;
```

Life of a Query

Advantages of DBMS

- Data independence: applications insulated from how data is structured and stored
- Optimizability
- Reduced application development time
- Efficient data accesses
- Data integrity and security
- Concurrency control and crash recovery
- So why not use them always?
  - Expensive/complicated to set up & maintain
  - This cost & complexity must be offset by need
  - General-purpose, not suited for special-purpose tasks (e.g. text search)

Next lecture

- Thursday
  - Review of distributed and parallel databases
    - Loosely based on Ramakrishnan’s textbook, R*, Gamma
  - No readings due
- Due by Thursday noon:
  - Introduction email (year, background, research interest, advisor, audit/enroll)