The Course

- The challenges of distributed systems imposed by the data intensive applications
- State-of-the-art solutions proposed to overcome these challenges

- The topics include:
  - Introduction to Distributed Computing
  - Data Intensive Science & Applications
  - Computational and Data Grids
  - Distributed and Mass Storage Systems
  - Global/Parallel/Shared File Systems
  - Remote I/O and Data Staging
  - Distributed Data Management and Scheduling
  - Complex Workflow Management
  - Distributed and Remote Visualization

Course Material

- There is no required textbook.
- We have a reading list mostly consisting of scientific papers published in this area:
  [http://www.cct.lsu.edu/~kosar/csc7700/reading_list.html](http://www.cct.lsu.edu/~kosar/csc7700/reading_list.html)
- 2-4 papers for each topic
- Will be discussing 2 papers each class
- Form discussion groups (consisting of three students)
- Read and discuss papers before coming to class

Paper Presentations

- Each student will present 2-3 papers from the reading list
- 25 minutes each + 5 minutes Q&A
- Select 3 papers from the reading list (papers 9-48)
  - Each from a different topic
  - Send me your selections by this Friday
  - Send also three alternative papers you would consider
- Before each presentations, meet me 2 times
  - One week before the presentation: the outline of your talk
  - Two days before the presentation: your slides

Research Projects

- Team projects (team of 2)
- Example Projects Topics:
  - Effects of Staging vs Remote I/O
  - Comparison of Distributed/Parallel File Systems
  - Development of Data Scheduling Algorithms
  - Application of Grids to a Specific Science Problem
- Predefined projects will be available online by this weekend
- You can propose your own projects, come and see me
- Projects may include programming, but not extensively
- Most projects will include experimentation and analysis
Research Projects (cont)

- You will need to submit:
  - Research plan
    - Scope of the project, timeline, deliverables
  - Monthly progress reports (2)
    - Progress so far, changes in the original research plan
    - Final project report (technical paper)
    - Project presentation (30 min)
- **Best two papers** will be selected, to be:
  - extended and submitted to an Int. Workshop/Conference
  - presented at the CS/CCT seminar

Grading

- The end-of-semester grades will be composed of:
  - Active Contribution: 10%
  - Paper Presentations: 30%
    - 3 Presentations: 10% Each
  - Project: 60%
    - Research Plan: 10%
    - 2 Progress Reports: 10% each
    - Final Project Report: 15%
    - Project Presentation: 15%

Teaching Philosophy

- **Goal:**
  - For instructor: teaching the material
  - For student: learning and applying the material in real life
- Grades are of second degree importance
- Try to understand the material
- Be critical, do not need to always agree

Roadmap

- Course Syllabus
- Administrative Details
- **Introduction to the Course Material**
  - The Data Deluge

**The Large Hadron Collider (LHC)**

![Diagram of the Large Hadron Collider](image)

Major Science Driver:
Large Hadron Collider (LHC) @ CERN

- 27 km Tunnel in Switzerland & France
- Search for:
  - Origin of Mass
  - New fundamental forces
  - Supersymmetry
  - Other new particles
  - 2007 – ?
CMS: “Compact” Muon Solenoid

Inconsequential humans

LHC: Beyond Moore’s Law

Estimated CPU Capacity at CERN

Intel CPU (2 GHz) = 0.1 K SI95

LHC CPU Requirements

Moore’s Law (2000)

1 SI95 = 40 MIPS

LHC: Petascale Global Science

- Complexity: Millions of individual detector channels
- Scale: PetaFlops (CPU), 100s of Petabytes (Data)
- Distribution: Global distribution of people & resources

CMS Example - 2007
5000+ Physicists
250+ Institutes
60+ Countries

BaBar/D0 Example - 2004
100+ Physicists
35+ Countries

Science Driver: Protein Folding

- BLAST: decode genetic information and map the genomes of humans, and other species.
- Uses comparative genomics: compares unknown genetic sequences (~billions) to known genomes in search of similarities.
- Current dataset:
  - Several Petabytes
- Future:
  - Exponential Growth: SCARY!

Biology Applications: Tomography

Tomography:
- Derivation of 3D structure from a series of 2D electron microscopic projection images,
- Reconstruction and detailed structural analysis
  - complex structures like synapses
  - large structures like dendritic spines.
- Acquisition and generation of huge amounts of data
- Large amount of state-of-the-art image processing required to segment structures from extraneous background.

Dendrite structure to be rendered by Tomography
Astronomy

• Mapping of Universe, detection of new galaxies and stars...

Current Datasets

<table>
<thead>
<tr>
<th>Project</th>
<th>Data Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>MACHO</td>
<td>3 TB</td>
</tr>
<tr>
<td>2MASS</td>
<td>12 TB</td>
</tr>
<tr>
<td>SDSS</td>
<td>40 TB</td>
</tr>
</tbody>
</table>

Future Productions

<table>
<thead>
<tr>
<th>Project</th>
<th>Data Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>WFCAM</td>
<td>20 TB/year</td>
</tr>
<tr>
<td>VISTA</td>
<td>100 TB/year</td>
</tr>
<tr>
<td>LSST</td>
<td>1000 TB/year</td>
</tr>
</tbody>
</table>

---

Sloan Digital Sky Survey (SDSS)
Using Virtual Data in GriPhyN

Sloan Data

Galaxy cluster size distribution

---

The LIGO Project

Laser Interferometer Gravitational-wave Observatory:

• A $300M+ project funded by the National Science Foundation
• Two Observatories in the US looking for gravitational waves emitted by astrophysical sources.
• Detection of Gravitational Waves will be a milestone discovery, based on predictions of Einstein’s General Relativity.
• Observations of gravitational waves will open a new window in Astronomy for studying black holes, neutron stars...
• "Initial LIGO" (today): test technology and ability to analyze data, few predicted sources.
  – Instruments will be ready ~next year
  – Three data-taking Science Runs in 2010-2014
  – Data analysis in progress.
• "Advanced LIGO" (~2009+): better technology, many predicted detections/year.
An Educational Technology Project:
WCER Educational Video Processing

- Build histories of student learning for use in education research and instruction relying on video data.
- Analyze and share large amount of video.
- 1 hour DV video is ~13 GB
  - A typical educational research video uses 3 cameras => 39 GB for 1 hour
- Current data set:
  - > 500 Terabytes
- Future:
  - Several Petabytes

Other Application Areas

- Hurricane Track Prediction
- Storm Surge Modeling
- Coastal Erosion Modeling
- Numerical Relativity
- Petroleum Engineering
- Computational Fluid Dynamics (CFD)

How to access and process distributed data?

Any Questions?

Reading Assignment

- Read papers [1] [2] for background information and motivation.
- Read papers [3] [4] and discuss them in your reading group. These papers will be discussed in class this Thursday.

- Don’t forget to send me the list of 3 (+3) papers you would like to present by this Friday.