CS 595:
Hot Topics in Distributed Systems: Data-Intensive Computing

Syllabus

Ioan Raicu
Computer Science Department
Illinois Institute of Technology

CS 595
Hot Topics in Distributed Systems: Data-Intensive Computing
August 23rd, 2010

Introductions

- Ioan Raicu

  - More information at:
    - http://www.cs.iit.edu/~iraicu/
- Everyone else
  - Background?
  - What do you want to get out of this course?

Course Overview

- Data Intensive Computing is critical to advancing modern science
  - Applies to cluster computing, grid computing, supercomputing, and cloud computing
- Increasing gap between compute capacity and storage bandwidth
- Need for advanced techniques to manipulate, visualize and interpret large datasets
- Building large-scale distributed systems is hard
  - network (e.g., transport, routing)
  - algorithmic (e.g., data distribution, resource management)
  - social (e.g., incentives)

Course Overview (cont)

- Understand methods and approaches to:
  - Design, implement, and evaluate distributed systems
- Topics include:
  - Resource management (e.g. discovery, allocation, compute models, data models, data locality, virtualization, monitoring, provenance), programming models, application models, and system characterization
- Course involves:
  - Lectures, outside invited speakers, discussions of research papers, homeworks, and a major project

Prerequisites

- Topics
  - Programming (C, C++, or Java)
  - Networking
  - Operating systems
  - Distributed systems

Course Topics

- Distributed Systems: Clusters, Supercomputers, Grids and Clouds
- Data Intensive Computing Overview
- Local Resource Management Systems
- Storage Systems
- Shared, Distributed and Parallel File Systems
- Parallel I/O
- Scientific Computing and Applications
- Parallel Programming Systems and Models

Course Topics (cont)

- MapReduce & Hadoop
- Sphere/Sector
- Parrot and Chirp
- Swift/Falkon
- Data-Intensive Computing with GPUs
- Data-Intensive Computing with Databases
- Many-core Computing Era and New Challenges
- Open Research Questions in Data-Intensive Computing

Computer Usage

- falkon.cs.iit.edu
  - Request account from iraicu@cs.iit.edu
  - Intel Xeon, 16-cores @ 2.33GHz, 48GB RAM, 7TB RAID5 disk, 1Gb/s network
    - Linux Suse 11.2 x64

Computer Usage (cont)

- PADS Cluster at University of Chicago (1K cores x64)
- IBM BlueGene/P at Argonne National Laboratory (160K PPC)
- SiCortex at Argonne National Laboratory (5832 MIPS)
- ANL/UC TG Cluster at Argonne National Laboratory (~200 IA32)
- TeraGrid (150K of all variety of CPUs)
- Sun Constellation at TACC (62K x64)
- Magellan at Argonne National Laboratory (10K x64 cloud)
- Amazon EC2

Research Papers Reading and Discussion

- 1~2 papers per lecture
- Each paper must be summarized in writing
- Serve as background to the lecture
- Serve as basis for discussion
  - Each paper will have a student discussion leader

Homeworks

- Up to 5 assignments
- Will give hand-on experience with some specific technology or theoretical concept
- Generally will have 1 week to complete
- Must be completed individually

Projects

- Major quarter long project
  - Topic of choice of the student
  - Can work in groups
  - May require the following things:
    - Reading research papers
    - Using open source software
    - Implementation of a real/simulated system
    - Analysis of theoretical work
    - Performance evaluation of theoretical/real systems
    - Written report(s)
    - Oral presentation(s)

Project Ideas

- Distributed file systems
- Data aware scheduling algorithms
- Distributed operating systems
- Distributed job management systems
- Parallel programming languages
- Distributed workflow systems
- Distributed monitoring systems

Project Ideas (cont)

- Scientific computing with GPUs
- Scientific computing with MapReduce
- Distributed caching strategies
- Distributed cache eviction policies
- Distributed hash tables
- Virtualization impact for data-intensive computing

Useful Software for your Projects

- **Operating systems:** Linux, Windows
- **Scripting:** BASH
- **Source control:** SVN
- **Programming languages:** Java, C/C++
- **Job submission systems:** GRAM, PBS, Condor, Cobalt, SGE, Falkon
- **Programming models:** MapReduce (Hadoop), MPI (MPICH), Multi-Threading (PThreads), Workflows (Swift, Pegasus/DAGMan, Nimrod, Taverna, BPEL)
- **File systems:** FUSE

Useful Software for your Projects (cont)

- **Parallel file systems:** GPFS, PVFS, Lustre
- **Distributed file systems:** GPS, HDFS
- **Data services:** GridFTP
- **Grid middleware:** Globus
- **Cloud middleware:** Nimbus, Eucalyptus, OpenNebula
- **Distributed hash tables:** Chord, Tapestry
- **Simulation environments:** GridSim, SimGrid, OptorSim, GangSim, Bricks
- **Virtualization:** Sun Virtual Box, XEN, VMWare

Grading

- **Participation in paper discussions**: 15%
- **Homeworks**: 20%
- **Mid-quarter oral presentation**: 5%
- **Final oral presentation**: 10%
- **Project / Report**: 50%

Course Outcomes

- Understand the importance of data-intensive computing
- Understand the difference between cluster computing, grid computing, supercomputing, and cloud computing
- Understand how to build large scale distributed systems
- Understand applications that require data-intensive computing
- Understand trends in many-core computing and challenges that will come with them
- Build distributed systems
- Be familiar with multiple programming models
- Read and understand a research paper
- Make a formal presentation on a technical topic
- Write up a formal report (and a research paper) on the project

Miscellaneous

- Required texts
  - None
  - Readings will be from online material

Questions

- Write me:
  - iraicu@cs.iit.edu
- Skype me:
  - ioan.raicu
- Call me:
  - 1-312-567-5704
- Mailing list
  - cs595-f10@mailer.iit.edu