CS 595

Hot Topics in Distributed Systems: Data-Intensive Computing

http://www.cs.iit.edu/~iraicu/teaching/CS595-F10/index.html

Dr. Ioan Raicu
Fall 2010
Monday/Wednesday, 1:15PM - 3:10PM

DETAILED COURSE TOPICS:

- **Distributed Systems**
- **Supercomputing**
- **Grid Computing**
- **Cloud Computing**
- **Many-core Computing**
- **Data Intensive Computing**
- **Storage Systems**
- **Distributed and Parallel File Systems**
- **Parallel I/O**
- **Local Resource Management**
- **Scientific Computing and Applications**
- **Parallel Programming Systems and Models**
- **MapReduce**
- **Case Studies: Sphere/Sector, Hadoop/ HDFS, Parrot/Chirp, Swift/Falkon**
- **Data-Intensive Computing with GPUs**
- **Data-Intensive Computing with Databases**
- **Open Research Questions**

Dr. Ioan Raicu is an assistant professor in CS at IIT. He was a NSF/CRA Computation Innovation Fellow at Northwestern University in 2009 - 2010, and obtained his Ph.D. in Computer Science from University of Chicago in 2009. He is a 3-year award winner of the GSRP Fellowship from NASA Ames Research Center. His research work and interests are in the general area of distributed systems. His work has focused on defining and exploring both the theory and practical aspects of realizing Many-Task Computing across a wide range of large-scale distributed systems. He is particularly interested in efficient task dispatch and execution systems, resource provisioning, data management, scheduling, and performance evaluations in distributed systems. His work has been funded by the NASA, DOE, NSF, and CRA. Ioan's research interests include resource management in large scale distributed systems with a focus on many-task computing, data intensive computing, cloud computing, grid computing, and many-core computing. He is a member of the ACM and IEEE.

This course is a tour through various research topics in distributed data-intensive computing, covering topics in cluster computing, grid computing, supercomputing, and cloud computing. We will explore solutions and learn design principles for building large network-based computational systems to support data intensive computing. This course is geared for junior/senior level undergraduates and graduate students in computer science.

The support for Data Intensive Computing is critical to advancing modern science as storage systems have experienced an increasing gap between its capacity and its bandwidth by more than 10-fold over the last decade. There is an emerging need for advanced techniques to manipulate, visualize and interpret large datasets. Building large scale distributed systems that support data-intensive computing involves challenges at multiple levels, from the network (e.g., transport, routing) to the algorithmic (e.g., data distribution, resource management) and even the social (e.g., incentives). Our readings and discussions will help us identify research problems and understand methods and general approaches to design, implement, and evaluate distributed systems to support data intensive computing. Topics include resource management (e.g. discovery, allocation, compute models, data models, data locality, virtualization, monitoring, provenance), programming models, application models, and system characterization. Our discussions will often be grounded in the context of deployed distributed systems, such as the TeraGrid, Amazon EC2 and S3, various top supercomputers (e.g. IBM BlueGene/P, Sun Constellation, Cray XT5), and various software/ programming platforms (e.g. Google's MapReduce, Hadoop, Dryad, Sphere/Sector, Swift/Falkon, and Parrot/Chirp). The course involves lectures, outside invited speakers, discussions of research papers, and a major project (including both a written report and an oral presentation).