B649: Topics in Systems: Cloud Computing for Data Intensive Sciences

Pre-requisites:
B649 Cloud Computing online course is a programming intensive course. It has similar requirements to the CS graduate level residential version. Students are expected to have weekly (or biweekly) programming homework. General programming experience with Windows or Linux using Java (2-3 years) and scripts is required. A background in parallel and cluster computing is a plus, although not necessary.

- **Introductory Video and Course Homepage**
  http://cloudmooc.appspot.com/preview

Course description:
New computing paradigms are emerging from data-driven discovery and scalable data processing. These include virtual machine based utility computing environments such as Amazon AWS and Microsoft Azure. Furthermore, a set of new MapReduce programming models coming from Information retrieval field has been shown to be effective for scientific applications. This class covers many of the basic concepts and solutions required for data analytics at massive levels. Data science techniques include cloud computing, parallel algorithms, storage and high level language support to develop proficiency with a complex ecosystem of tools; these techniques are applicable to many disciplines that can make use of data mining and optimization.

Data Science topics:

- Data intensive sciences and the data center model
- Clouds with infrastructure, platform and software as a service
- Virtualization technologies and tools
- Introduction to FutureGrid as an experimental testbed
- Parallel programming using MapReduce vs. MPI
- MapReduce and data parallel applications using Hadoop
- Iterative MapReduce and data mining algorithms using Twister (expectation maximization, clustering, multi-dimensional scaling, latent Dirichlet allocation, Bayes networks)
- MapReduce on Multicore/GPU (CUDA)
- NoSQL databases (Google BigTable and Hadoop HBase) and parallel query processing
- High-level language (Hive and Pig)
- Amazon EC2 and Microsoft Azure and their applications

In a final project, students will build data intensive science applications using a list of software technologies including Openstack and FutureGrid, Hadoop, Twister, many data mining algorithms, CUDA, HBase, Hive and Pig.