



**Yahoo! Academic Relations & The Center
for Autonomic Computing at Rutgers
University Present:**

Grid Computing at Yahoo!

Viraj Bhat
Grid Engineer, Yahoo!



Date: Thursday, October 1st, 2009

Time: 2:30 p.m.

Location: Room 538, CoRE Bldg.

Busch Campus, Piscataway, NJ

The Yahoo! Grid Initiative is an effort underway to build a massive computing environment that supports and is augmented by an open-source software framework. The computing environment includes more than 30,000 nodes. The software framework is based on the Hadoop project from the Apache Software Foundation, an open-source implementation of the map/reduce programming model and a distributed file system that places the data close to the computations. When combined, the computing environment and software framework enable distributed, parallel processing of huge amounts of data. The Grid is used for a variety of research and development projects and for a growing number of production processes from across Yahoo!, including key components of search, advertising, data pipelines and user-facing properties.

The focus of this talk is on Apache Hadoop and its related sub-projects such as Pig and Zookeeper, which form the building blocks of the Grid infrastructure. Hadoop is a framework for running applications on large clusters built of commodity hardware. The Hadoop framework transparently provides applications both reliability and data motion. Hadoop implements a computational paradigm named Map/Reduce, where the application is divided into many small fragments of work, each of which may be executed or re-executed on any node in the cluster. In addition, it provides a distributed file system (HDFS) that stores data on the compute nodes, providing very high aggregate bandwidth across the cluster. Both Map/Reduce and the distributed file system are designed so that node failures are automatically handled by the framework. Additional projects that augment Hadoop include Pig, a high-level data-flow language and execution framework and Zookeeper, a high-performance coordination service for distributed applications.

Viraj Bhat graduated with a Ph.D. degree (May 2008) in the Department of Electrical and Computer Engineering under the guidance of Professor Manish Parashar. Currently he is employed with Yahoo! as a Grid engineer, where he works on building, porting and parallelizing several data-intensive applications. He recently received the Yahoo! award for evangelizing Grid technologies, profiling and optimizing Map Reduce applications.

