Roadmap for Applying Hadoop Distributed File System in Scientific Grid Computing

Haifeng PI

Open Science Grid, UCSD, US

Original work on integrating Hadoop distributed file system (HDFS) into a complete grid storage solution has shown the potential of HDFS as a mainstream technology for the general scientific grid computing. We evaluate the requirements of grid storage technology, scalability and reliability of the overall storage system and its major components, computing architecture of storage and computing elements, and operation and maintenance procedure. According to the trends of CPU- and data-intensive scientific grid computing and networking technology for the next 5-10 years, the roadmap of HDFS can well fits into the continuous challenging environment. The testing and release infrastructure for HDFS-based storage element is developed at Open Science Grid (OSG) to meet the future needs. Recent applications of HDFS in CMS Tier-2 sites, in the Scale Testing for the Experiment Programme '09 (STEP09), and in the network bandwidth challenge at Super Computing 2C09, are presented.