MapReduce and Data Intensive Applications

Judy Qiu
Computer Science
Indiana University
150 South Woodlawn Avenue
Bloomington, IN 47405

xqiu@indiana.edu

Geoffrey Fox
Pervasive Technology Institute
Indiana University
2719 E. 10th Street
Bloomington IN 47408

gcf@indiana.edul

ABSTRACT

We are in the era of data deluge and future success in science depends on the ability to leverage and utilize large-scale data. This proposal follows up our successful first meetings in this series of "MapReduce application and environments" at TeraGrid 2011. Further we will use it to kick start an XSEDE forum. It aligns directly with several NSF goals including Cyberinfrastructure Framework for 21st Century Science and Engineering (CF21) and Core Techniques and Technologies for Advancing big Data Science & Engineering (BIGDATA). In particular, MapReduce based programming models and run-time systems such as the open-source Hadoop system have increasingly been adopted by researchers of HPC, Grid and Cloud community with data-intensive problems, in areas including bio-informatics, data mining and analytics, and text processing. While MapReduce run-time systems such as Hadoop are currently not supported across XSEDE systems (it is available on some systems including FutureGrid), there is increased demand for these environments by the science community. Figure 1 shows the statistics of projects on FutureGrid testbed, where Hadoop, MapReduce, and Twister (MapReduce variant) have been used extensively as a framework for experiments in scalable data processing.

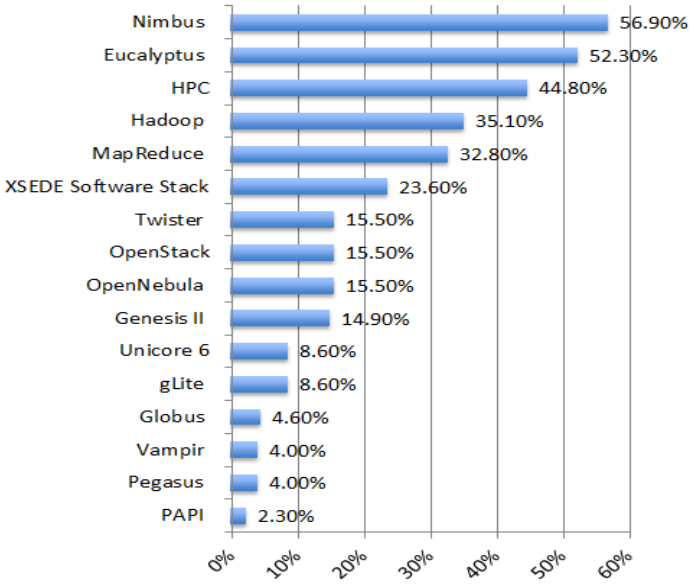

Figure 1 Applications and Technologies usage on FutureGrid in 2011 by ratio of projects

This BOF session will provide a forum for discussions with users on challenges and opportunities for the use of MapReduce as an interoperable framework on HPC, Grid and Cloud. It will be moderated by Judy Qiu who will start with a short overview of MapReduce and the applications for which it is suitable. These include pleasingly parallel applications and many loosely coupled data mining and data analysis problems where we will use genomics, information retrieval and particle physics as examples. We will discuss the interest of users, the possibility of using XSEDE and commercial clouds, and the type of training that would be useful. The BOF will assume only broad knowledge and will not need or discuss details of technologies like Hadoop, Dryad, and Twister except to discuss the key features that determine functionality. We will discuss some important issues of storage models used by MapReduce.

ACKNOWLEDGMENTS
Our thanks to the support by NSF Grant No.0910812 to Indiana University for "FutureGrid: An Experimental, High Performance Grid Test-bed.", NSF CAREER award, and NIH Grant number RC2HG005806-02