Harnessing Grid Resources to Enable the Dynamic Analysis of Large Astronomy Datasets

Ioan Raicu
Computer Science Department
The University of Chicago
iraicu@cs.uchicago.edu

Ian Foster
Math and Computer Science Div., Argonne National Lab
Computer Science Dept., The University of Chicago
foster@mesa.anl.gov

Alex Szalay
Dept. of Physics and Astronomy
The Johns Hopkins University
szalay@jhu.edu

Abstract

Grid computing has emerged as an important new field focusing on large-scale resource sharing and high-performance computation. The astronomy community has an abundance of imaging datasets at its disposal which are essentially the "raw data" for the astronomical community. However, these astronomy datasets are generally terabytes in size and contain hundreds of millions of objects separated into billions of files - factors that make many analyses impractical to perform on small computers. The key question we answer is: "How can we leverage Grid resources to make the analysis of large astronomy datasets a reality for the astronomy community?" Our answer is: "AstroPortal," a grid resource tailored for the astronomy community. To address this question, we have developed a collection of Web Services-based systems that use grid computing to federate large computing and storage resources for dynamic analysis of large datasets. Building on the Globus Toolkit, we have built a prototype consisting of several systems (AstroPortal, DYRE, Dynamic Resource pool Loggers, 3DcacheGrid, etc.) Dynamic, Distributed Data cache for Grid applications, and CompuStore - Computational Scheduler to enable a "stacking" analysis, the analysis of multiple regions of the sky, a function that can help both identify variable sources and detect faint objects. We have deployed AstroPortal and the related systems on the TerGrid distributed infrastructure and applied the stacking function to the Sloan Digital Sky Survey (SDSS), EDR5, which comprises more than 250 billion objects dispersed over 1.5 million files. A tool of 4.5 terabytes of compressed data, with promising results. AstroPortal gives the astronomy community a new tool to advance their research and to open new doors to opportunities never before possible on such a large scale.

Optimizations

3DcacheGrid

DYRE

Architecture Overview

Stacking Results

Related Material, Further Reading, and Acknowledgements


For more information, please visit: http://people.cs.uchicago.edu/~irmaci/ and http://people.cs.uchicago.edu/~irmaci/research/AstroPortal.

This work was funded by the NASA Grant (from June 2005 - September 2006), the work is currently being funded by the NASA/DOE GRID Proposal October 2006 - September 2007.