AstroPortal: A Science Gateway for Large-scale Astronomy Data Analysis

Ioan Raicu
Distributed Systems Laboratory
Computer Science Department
University of Chicago

Joint work with:
Ian Foster: Univ. of Chicago, CS & Argonne National Laboratory, MCS
Alex Szalay: Dept. of Physics and Astronomy, Johns Hopkins University
Gabriela Turcu: Univ. of Chicago, CS

TeraGrid Conference 2006
June 13th, 2006
Introduction

• Science Portals: gateway to Grid resources
• Potential Applications Characteristics
  – Large data sets
  – Large number of users
  – Easy (but not necessarily trivial) parallelization
• Applicable fields:
  – Astronomy
  – Medicine
  – Others
Astronomy Field

• Astronomy datasets (i.e. SDSS) are the crown-jewels
  – SDSS DR4
    • 1.3M images
      – 300M+ objects
      – 3TB compressed images (2MB x 1.3M)
      – 8TB raw images (6.1MB x 1.3M)
    • 100K worldwide potential users (100s of big users)

• Applications:
  – Stacking
  – Montage
AstroPortal: Stacking Service
Results

User ID: iranl
Password: ******
Stacking Description: stacking_description.txt
Stacking Size: 20
AstroPortal Web Service Location: http://ig-so.login.ui.beragrid.org/50001/wasB/services/AstroPortal/core/WS/APEFactoryService

RESULT:

Size: 43 KB
Dimensions: 100x100 pixels
Download result: stacked_result.fits

Time to complete Stacking: 5.164 seconds
Number of physical resources utilized: 16
Number of Stacks completed successfully: 18
Number of Star Objects not found in the SDSS dataset: 1
List of Star Objects [ra, dec, band] not found:
  • [194.989060213455, 13.9018344168167, r]

Number of Data Objects not found in the data cache: 1
List of Data Objects [ra, dec, band] filename [x_coord y_coord] not found:
  • [(194.959705877549, 2.93855950426612, r),
    /disks/scratch/gps1/iranl/sdc.gz/das.sds.org/DR4/dataimaging/752/400/coor64pC-0000752-x-6-0245.fits.gz [0 x 0]]

To start a new stacking, go back to the main Stacking Service.
Open Research Questions

• Site level
  – advanced reservations
  – resource allocation
  – resource de-allocation

• Data management
  – Data location and replication
  – Data caching hierarchies

• Resource management
  – Distributed resource management between various sites
Questions?

- More information: http://people.cs.uchicago.edu/~iraicu/research/
- Related materials and further readings:
  - Ioan Raicu, Ian Foster, Alex Szalay. “Harnessing Grid Resources to Enable the Dynamic Analysis of Large Astronomy Datasets”, under review at SuperComputing 2006.
  - Ioan Raicu, Ian Foster, Alex Szalay, Gabriela Turcu, Catalin Dumitrescu. “Enabling Large-scale Astronomy Data Analysis with the AstroPortal,” under preparation for the HPC Analytics Challenge at SC06.