Federated Observational and Simulation Data in the NASA Center for Climate Simulation

Data Management System Project

Glenn Tamkin, Scott Sinno, Roger Gill, David Fladung,
Dave Ripley, Savannah Strong & John Schnase

NASA Center for Climate Simulation (NCCS)
NASA Goddard Space Flight Center
NCCS Mission

• **Traditional**
  
  • Enable scientists to increase their understanding of the Earth and the universe by providing state-of-the-art high performance computing, storage, network, and application solutions
  
  • Provide large-scale compute engines, analytics, data sharing, and high-end computing services

• **Future**
  
  • Develop a data services capability to better support the climate research communities and prepare the way for technology advances
New Challenges

• **Finding** observational and model data for use in climate and weather studies
• **Accessing** the geographically distributed data
• **Managing** the massive digital holdings, which are measured in petabytes and hundreds of millions of files
• **Maintaining** the data, which must often be preserved for decades
• **Supporting** data sharing, data publication, and data stewardship
NCCS Test Bed

- iRODS abstracts physical location of data
- iRODS assists with archive management
Preliminary Tests – Ingest/Registration

- iRODS rules and microservices allow data to be stored in configurable collections based on data policies
- Replication to backup storage resources also supported
Preliminary Tests - Search

• iRODS rules and microservices can be used to assign metadata
• iRODS provides advanced search capabilities over the metadata
Preliminary Tests – Observational Data

- Developed an iRODS data grid that publishes Moderate Resolution Imaging Spectroradiometer (MODIS) observational data
  - 54 million registered files, 630 TB of data, and over 300 million defined metadata values
- Developed an iRODS data grid that focuses on a small-scale, multi-product, application-specific data service
  - The Invasive Species Data Service (ISDS) manages a collection of MODIS data products for ecological forecasting applications
Preliminary Tests – Simulation Data

- Developed an iRODS data grid that manages Modern Era Retrospective-Analysis for Research and Applications (MERRA) simulation data
  - 360 files, 47 GB of data, and 4000 metadata values

- Developed an iRODS data grid that publishes Year of Tropical Convection (YOTC) data sets
  - 134,000 files, 12 TB of data, and 400,000 metadata values
Preliminary Tests – Federation

- Tested and evaluated iRODS data federation
  - Federated the YOTC and MODIS grids to simulate the union of observational and simulation data
- Explored the integrated management of observational and simulation data
  - Implemented an interface that enables comingling of remote and local observational and simulation data for advanced scientific study
Preliminary Results

- iRODS is a promising technology for exposing services for data management, publication, and analysis
- The iRODS catalog (ICAT) demonstrated adequate scaling for data registration
  - Optimization desired for searching huge datasets
- Good collaboration with the iRODS development team
- NCCS has made the decision to operationalize iRODS
New Goals

- **IPCC / AR5**
  - Provide the data management services and analytical tools necessary to support the publication requirements of the Intergovernmental Panel on Climate Change (IPCC).

- **Observation/Simulation Data Integration**
  - Bring the climate modeling and observational communities together to work toward the goal of integrating model outputs and observational data.

- **Next Generation HEC Requirements for Modeling and Assimilation**
  - Contribute emerging technologies to address computing requirements for Earth system modeling that will increase significantly in the coming years.
Questions