Introduction to Climate Data Services using iRODS

Data Management System Project

Savannah Strong Finch, Glenn Tamkin, Dave Ripley, Ed Luczak, Scott Sinno, Roger Gill, Deni Nadeau, John Schnase, Mark Mcinerney

NASA Center for Climate Simulation (NCCS)

NASA Goddard Space Flight Center
Topics

• NCCS Background
• Goals and Challenges
• Data Grid software (Overview)
• iRODS background
• Preliminary Tests with iRODS (Overview)
• Climate Data Server (Test Applications and Results)
• Current and Future Integration within NCCS (eCDS)
• Questions
NCCS Background

Current

- Provide state-of-the-art high performance computing, storage, network, and application solutions to enable scientists to increase their understanding of the Earth and the universe
- Provide large-scale compute engines, analytics, data sharing, and high-end computing services support
Goals

Develop a data services capability to better support the climate research communities and prepare the way for technology advances for:

- **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 High End Computing (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
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
Data Grid Software
Data grid “middleware” runs as an application in user space and provides a richer set of metadata descriptors and extended capabilities ...
Data Grid Software

... including federation and inter-collection discovery and access.
Data Grid Software

iRODS

Integrated Rule Oriented Data System
iRODS: integrated Rule-Oriented Data System

Background
- Open source data grid software system.
- Developed by the Data Intensive Cyber Environments (DICE) group, University of North Carolina.
- Historic roots in data grids, digital libraries, persistent archives, and real-time data systems R&D, and SRB.

Features
- Management of large collections
- Manages metadata
- Policies, Rules and Micro-services
- A unified view of disparate data
- Controlled access
- Easy back up and replication
- High-performance network data transfer
- Support for a wide range of physical storage

Major Concepts
- iRODS rules
  - Actions for policies
- iRODS microservices
  - Implementations of definitions of Actions
*With iRODS metadata providing the information necessary to perform these mappings....
iRODS-Based Climate Data Server

Core Components

- Application-specific microservices
- Application-specific metadata
- Application-specific rules
- Application-specific utilities
- Application-specific configurations

⇒ "Application-Specific Kit"

- A specific release of iRODS
- A specific operating system

⇒ "CDS Software Appliance"
iRODS ..

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

- iRODS rules and microservices allow data to be stored in configurable collections based on data policies
- Rich web client allows for “explorer” like view into collections of the registered data
- Can also perform command line interface “icommands”:
  Bash-4.1$ ils /merra_Zone/home/public/merra/1979: MERRA100.prod.assim.instM_3d.nc

*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
Climate Data Server (Test Applications)

- MODIS and ISDS
- Merra Monthly Means and Proxies for AR5 simulations
- vCDS in the Amazon Cloud
- ODAS workflow
Climate Data Server (Test Applications) – Observational Data

- Developed an iRODS data grid that published 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
Climate Data Server (Test Applications) – Analysis and Simulation Data

- Developed an iRODS data grid that manages Modern Era Retrospective-Analysis for Research and Applications (MERRA) Monthly means analysis data
  - 360 files, 47 GB of data, and 4000 metadata values
- Developed an iRODS data grid that published public GEOS-5 simulation data as a proxy for AR5 data sets
  - 134,000 files, 12 TB of data, and 400,000 metadata values
Climate Data Server (Test Application) – Federation

• Tested and evaluated iRODS data federation
  
  • Federated the GEOS-5 public data and MODIS grids to simulate the union of observational and simulation data

• the integrated management of observational and simulation data was explored
  
  • Implemented an interface that enables comingling of remote and local observational and simulation data for advanced scientific study
Climate Data Server (Test Application) – vCDS in the Amazon Cloud
Climate Data Server (Test Application) – Extending iRODS

- NetCDF kit knows how to read the file header based on file format
- IPCC/OAIS kit defines which metadata to store and how to store it

<table>
<thead>
<tr>
<th>Name</th>
<th>Value</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>var.VS.FillValue</td>
<td>1.0 e-15f</td>
<td></td>
</tr>
<tr>
<td>var.VS.min</td>
<td>-1.0 e15f</td>
<td></td>
</tr>
<tr>
<td>var.VS.max</td>
<td>1.0 e15f</td>
<td></td>
</tr>
<tr>
<td>var.VS.valid_range</td>
<td>-1.0 e15f, 1.0 e15f</td>
<td></td>
</tr>
<tr>
<td>var.VS.units</td>
<td>m s-1</td>
<td></td>
</tr>
<tr>
<td>var.VS.type</td>
<td>float</td>
<td></td>
</tr>
<tr>
<td>var.VS.standard_name</td>
<td>surface_Agrid_northward_velocity</td>
<td></td>
</tr>
<tr>
<td>var.VS.scale_factor</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>var.VS.missing_value</td>
<td>1.0 e15f</td>
<td></td>
</tr>
<tr>
<td>var.VS.long_name</td>
<td>surface_Agrid_northward_velocity</td>
<td></td>
</tr>
<tr>
<td>var.VS.missing_value</td>
<td>1.0 e15f</td>
<td></td>
</tr>
<tr>
<td>var.VS.dims</td>
<td>time, lat, lon</td>
<td></td>
</tr>
<tr>
<td>var.VS.coordinates</td>
<td>LON LAT</td>
<td></td>
</tr>
<tr>
<td>var.VS.xdim_offset</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>var.VS.FillValue</td>
<td>1.0 e15f</td>
<td></td>
</tr>
<tr>
<td>var.VS.min</td>
<td>-1.0 e15f</td>
<td></td>
</tr>
<tr>
<td>var.VS.max</td>
<td>1.0 e15f</td>
<td></td>
</tr>
<tr>
<td>var.VS.valid_range</td>
<td>-1.0 e15f, 1.0 e15f</td>
<td></td>
</tr>
<tr>
<td>var.VS.units</td>
<td>m s-1</td>
<td></td>
</tr>
<tr>
<td>var.VS.type</td>
<td>float</td>
<td></td>
</tr>
<tr>
<td>var.VS.standard_name</td>
<td>meanonal_velocity_of_surface_sea</td>
<td></td>
</tr>
</tbody>
</table>
Climate Data Server – Ocean Data Assimilation System (ODAS)

- Leveraged iRODS to monitor ODAS workflow status
- Developed a series of “Ocommands” that are wrappers around the iRODS “icommands”
- Ocommands were integrated into existing ODAS workflow scripts and perform functions such as:
  - Register data
  - Query the iRODS database for decision-making information
  - Maintain the status of the hierarchy of ODAS workflow artifacts as status changing events occur
  - Log relevant event metadata to the appropriate ODAS workflow artifact
  - Reprocess preparation
    - Remove relevant files and reset status in the hierarchy of ODAS workflow artifacts in preparation for reprocessing.
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

• Exercised enough diverse Test Cases to have confidence in performance leading to decision to be made to progress towards making iRODS-based Climate Data Services Operational
• Establish an Enterprise Climate Data Service (E-CDS) federated grid across the NCCS resources

• Starting with projects:
  • ODAS
  • ESGF
  • Archive
    ➢ Allows for operational capability of the ESGF to use the archive in the case that the data portal disks were unavailable

• Potential follow on projects MERRA2, NCA, UVCDAT, MODIS
The End.

Questions?
Moving Forward –
What does E-CDS mean to operations folks

- Account creation
- Config (firewall, security, etc.)
- e-CDS Dependency installation (unixODBC, postgres, perl, authd, etc.) and configuration
- Installation of E-CDS rpm (includes irods + extensions)
- Admin Support