The Virtual Climate Data Server (vCDS):
An iRODS-Based Data Management Software Appliance
Supporting Climate Data Services and
Virtualization-as-a-Service
in the NASA Center for Climate Simulation

John L. Schnase¹, Glenn S. Tamkin²,³, W. David Ripley, III²,³,
Savannah Strong²,³, Roger Gill²,⁴, and Daniel Q. Duffy²¹

Office of Computational and Information Science and Technology³
NASA Center for Climate Simulation (NCCS)³
Computer Science Corporation (CSC),
Innovim, LLC
NASA Goddard Space Flight Center
Greenbelt, MD 20771
The Data Management System Project

Part 1 – Background (5 Minutes)
• vCDS Concept and Rationale
• vCDS 0.9 Anatomy / vCDS 0.9 Products

Part 2 – Where We Are Now (10 Minutes)
• NetCDF/IPCC Toolkit
• Administrative Extensions
• Repetitive Provisioning
• Operational Deployment
• Amazon Cloud vCDS-IPCC-ESG-v0.9

Part 3 – Wrap Up (5 Minutes +)
• Next Steps
• Discussion
Concept and Rationale

Scenario
A customer approaches the NCCS with a new dataset they want us to manage ...

Q. What technology is needed to quickly meet that customer’s requirement under the follow constraints:
  • The solution should be: simple, fast, and cheap;
  • provide core capabilities to get started, but extendable to accommodate future needs;
  • be flexible, with the ability to use, optimize, and change deployment configurations in response to resource availability;
  • allow the new dataset to be integrated into an existing data collection; and
  • come with a help desk and user support?

Definitions

Customer – an individual scientist, a lab, project, or mission.

Dataset – may be products generated by a GCM, may be observational data or a subset thereof, reanalysis data, or specialized products of value to an individual scientist or lab.

Manage – may refer to short-term file storage, long-term archival preservation; data may be used online by a person or application.

Examples
Abound:
  • IPCC AR5 data for ESG.
  • MODIS Atmospheres data for CMIP5.
  • MERRA downscaled meteorological and environmental data.
  • AgMIP, CERES, SMOS, Laboratory for Atmospheres, the Snowfake project ...

The DMS Project has been looking at iRODS data grid software as a potential solution ...
Managed Collections: (1) Publication Datasets

- GISS Collection – Served to ESG in Amazon
  - Ingest IPCC AR5 data into CDS 0.9
  - Operationally harden CDS 0.9 to CDS 1.0 (TRL 9)
  - Expose collection to ESG publication system
  - Develop Collection Administrator’s interface

*A production system in Nebula that mirrors the NCCS’s current capabilities...*

Develop requirements, implement Collection Administration policies and mechanisms, and specify OAIS Policy Metadata – all relatively easy with Publication Datasets.

<table>
<thead>
<tr>
<th>Published Collections</th>
<th>Estimated Current Size (TB)</th>
<th>Estimated Final Size (TB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GISS IPCC</td>
<td>3.5</td>
<td>60</td>
</tr>
<tr>
<td>MERRA</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>CERES</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>AgMIP</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GMAO IPCC</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>4.1</strong></td>
<td><strong>120.6</strong></td>
</tr>
</tbody>
</table>

Current estimates as of 9/1/2011. Total estimated size of IPCC AR5 200 TB.
Our approach has been to build a core suite of general purpose scientific kits – such as NetCDF, HDF, and GeoTIFF – that sit in the vertical stack above iRODS and below application-specific climate kits such as IPCC, MERRA, and SMOS...

**CDS 0.9 Products**

1. **IPCC / NetCDF Module**
   - iRODS microservices, rules, configuration settings, and software utilities required to implement canonical CRUD operations for IPCC/NetCDF system kernel...
   - **Administrative Extensions**
     - iRODS Postgres extensions and utilities to log system-level object provenance and provide QA for OAIS metadata compliance (plus associated Rich Web Browser GUI extensions)...
   - **Repetitive Provisioning**
     - RPM scripts to build software stacks for the SLES 11 SP1 (IaaS), iRODS AE (PaaS), and CDS/IPCC (SaaS) virtual images...
   - **Deployment and Distribution**
     - Product library, documentation, and SLA infrastructure for distribution, deployment, and help desk support...

---

The DMS Project Team - NASA Goddard Space Flight Center

The Data Management System Project
## vCDS V0.9 Products

<table>
<thead>
<tr>
<th>Module</th>
<th>Description</th>
<th>Images</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>IPCC / NetCDF Module</strong></td>
<td>iRODS microservices, rules, configuration settings, and software utilities required to implement canonical CRUD operations for IPCC/NetCDF system kernel</td>
<td><img src="image1.png" alt="Diagram" /></td>
</tr>
<tr>
<td><strong>Administrative Extensions</strong></td>
<td>iRODS Postgres extensions and utilities to log system-level object provenance and provide QA for OAIS metadata compliance (plus associated Rich Web Browser GUI extensions)</td>
<td><img src="image2.png" alt="Diagram" /></td>
</tr>
<tr>
<td><strong>Repetitive Provisioning</strong></td>
<td>RPM script to build software stacks for the SLES 11 SP1 (IaaS), iRODS AE (PaaS), and CDS/IPCC (SaaS) virtual images</td>
<td><img src="image3.png" alt="Diagram" /></td>
</tr>
<tr>
<td><strong>Deployment and Distribution</strong></td>
<td>Product library, documentation, and SLA infrastructure for distribution, deployment, and help desk support</td>
<td><img src="image4.png" alt="Diagram" /></td>
</tr>
</tbody>
</table>

- Microservice Code
- Microservice Utilities
- IPCC / NetCDF Rules
- Configuration File
- OAIS Object Views
- Object Action Logging
- PHP Browser Extensions
- Automatic Installation
- VaaS Architecture
- Tech Transfer Plan
- Tech Transfer Team
- UNC RENCE Partnership
IPCC NetCDF Metadata

Right now application-independent metadata appears in several places...

A. Self-describing file header
B. Filesystem path
C. File attributes
D. File name

Metadata ownership / rules come from several places...

• Producer / CMIP5 DRS
• Producer + Admin / Policy
• Operating system / Operating System
• Producer / CMIP5 DRS

CDS 0.9 Products

1. IPCC / NetCDF Module
   iRODS microservices, rules, configuration settings, and software utilities required to implement canonical CRUD operations for IPCC/NetCDF system kernel...

   • Administrative Extensions
     iRODS Postgres extensions and utilities to log system-level object provenance and provide QA for OAIS metadata compliance (plus associated Rich Web Browser GUI extensions)...

   • Repetitive Provisioning
     RPM scripts to build software stacks for the SLES 11 SP1 (IaaS), iRODS AE (PaaS), and CDS/IPCC (SaaS) virtual images...

   • Deployment and Distribution
     Product library, documentation, and SLA infrastructure for distribution, deployment, and help desk support...

```bash
> pwd
/portal/GISS/AR5/piControl/E2-R_piControl_r1i1p1

> ls -al
-rw-r--r--  1 giss  admin  20828964  Jul 18  08:05  cSoil_Lmon_GISS-E2-R_piControl_r1i1p1_398101-400511.nc
-rw-r--r--  1 giss  admin  20828964  Jul 18  08:05  cSoil_Lmon_GISS-E2-R_piControl_r1i1p1_398101-400512.nc
```
Open Archival Information System (OAIS)

An OAIS is an archive, consisting of an organization of people and systems that has the responsibility to preserve information and make it available for a designated community ...

The reference model addresses a full range of archival information preservation functions including:

- ingest, data management, access, and dissemination;
- the migration of digital information to new media and forms;
- the data models used to represent the information, the role of software in information preservation, and the exchange of digital information among archives.

And it identifies both internal and external interfaces to the archive functions;

- it identifies a number of high-level services at these interfaces;
- it provides various illustrative examples and some ‘best practice' recommendations;
- and it defines a minimal set of responsibilities for an archive to be called an OAIS.
Administrative Extensions

CDS 0.9 Products

1. IPCC / NetCDF Module
   iRODS microservices, rules, configuration settings, and software utilities required to implement canonical CRUD operations for IPCC/NetCDF system kernel ...

   • Administrative Extensions
     iRODS Postgres extensions and utilities to log system-level object provenance and provide QA for OAIS metadata compliance (plus associated Rich Web Browser GUI extensions) ...

   • Repetitive Provisioning
     RPM scripts to build software stacks for the SLES 11 ...

---

Major Functions

Basic system-level capabilities to log object provenance and provide OAIS package views of object metadata ...

- iRODS Postgres extensions
- iRODS microservice extensions
- Rich Web Browser extensions

Administrative Extensions

To appear in iRODS 3.0 release ...
First, about our development environment ...

The DMS Project has worked in a virtualized environment – including MacTops with VMware Fusion and a VMware vSphere dev/test server farm.

This environment has influenced the way we are thinking about building, distributing, and deploying CDS components ...

Rationale: Why create an RPM?

- Automate installation
  Goal is to be able to conveniently install iRODS and our vCDS software stack in different environments and on different platforms ...

- Reduce installation errors and eliminate the user interface
  Installing iRODS “out of the box” is cumbersome, and manual installation leads to errors and unstable systems ...

The Data Management System Project
Operational Deployment - Amazon Cloud vCDS-IPCC-ESG-v0.9

- Index Node (i.e., Gateway)
- Data Node
- Identity Provider Node (IDP)

Amazon Public IP

- iRODS FUSE Server Interface
- IPCC/OAIS Kit (Rules & Microservices)
- NetCDF Kit
- iRODS 2.5 AE
- S3fs Client Interface
- SLES 11 SP 1

Climate Data Server V0.9
Amazon EBS Instance 22 GB

Earth System Grid P2P
Amazon EBS Instance 6 GB

“Read Only” FUSE File System

Non-Federated ESG Node (http)

Amazon S3 Attached

Copy

IPCC AR5 Subset (Next)
NCCS Data Portal

IPCC AR5 Subset (Next)
IPCC AR5 Canonical Test Data (Now)
Amazon S3 Storage

IPCC AR5
Subset (Next)
... then move to other primary and derived products ...

**Managed Collections: (1) Publication Datasets**

- **MERRA Collection**
  - Develop iRODS MERRA kit
  - Develop iRODS HDFS, Swift, S3 drivers
  - Create testbed HDFS, Swift, S3 repositories
  - Expose collection to ESG publication system

- **CERES, AgMIP, GMAO Collections**
  Replicate/refine the Managed Collections processes as needed to accommodate customer interest, response, and needs – and budget, time, and political constraints ...

- **NASA / RENCI jointly developed products:**
  1) CDS V1.0 Enterprise Edition
  2) NetCDF, HDF science kits
  3) IPCC, MERRA climate kits
  4) HDFS, Swift, S3 drivers
  5) iDROP Collection Administrators GUI

<table>
<thead>
<tr>
<th>Published Collections</th>
<th>Estimated Current Size (TB)</th>
<th>Estimated Final Size (TB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓ GISS IPCC</td>
<td>3.5</td>
<td>60</td>
</tr>
<tr>
<td>MERRA</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>CERES</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>AgMIP</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GMAO IPCC</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>Total</td>
<td>4.1</td>
<td>120.6</td>
</tr>
</tbody>
</table>

Current estimates as of 9/1/2011. Total estimated size of IPCC AR5 200 TB.
Managed Collections: (2) Research Datasets

- Operational Research
  - Transition from “Archive” to “Managed Collections”
  - Approach will be stepwise, incremental, logical
  - Need established technology, a process, and an expert team
  - And a conceptual model for how this is done...

Develop requirements, implement user policies and mechanisms, and specify OAIS Policy Metadata and Discovered Metadata – this is where we add layers to the kernel.

<table>
<thead>
<tr>
<th>Research Collections</th>
<th>Estimated Current Size (TB)</th>
<th>Estimated Final Size (TB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td>???.?</td>
<td>???.?</td>
</tr>
<tr>
<td>Person</td>
<td>???.?</td>
<td>???.?</td>
</tr>
<tr>
<td>Project</td>
<td>???.?</td>
<td>???.?</td>
</tr>
<tr>
<td>Project</td>
<td>???.?</td>
<td>???.?</td>
</tr>
<tr>
<td>Lab</td>
<td>???.?</td>
<td>???.?</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>???.?</td>
<td>???.?</td>
</tr>
</tbody>
</table>

... then to the challenging task of active research collections.

FY13 and beyond ...
Discussion

The NCCS Data Management System