IRODS: the Integrated Rule-Oriented Data-Management System

Wayne Schroeder, Paul Tooby
Data Intensive Cyber Environments Team (DICE)

DICE Center, University of North Carolina at Chapel Hill;
Institute for Neural Computation (INC), University of California San Diego

irods.org, dice.unc.edu, diceresearch.org


THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL


Who Are We?

Computer Scientists and Software Engineers

- Started in 1997
- Grew out of High Performance Computing
  - Broadened support to Digital Libraries/Preservation
- Doing applied research
  - Digital Preservation and Data-Grids
- Develop and distribute Integrated Rule-Oriented Data System (iRODS)
  - Open Source; PCs to High-Performance Computing

What Problems Are We Solving?

Researchers may have millions of computer files

- Keep them safely stored and replicated (remotely)
- Distribute them across the network; remote access
- Automate handling; rules, work-flows
- Keep track of what they are (meta-data)
- Be able to find the right ones quickly (queries)
- Share them, in a controlled manner (authentication, access control, audit trails)
- Preserve them; change storage transparently

Data Management Applications

- Data grids
  - Build shareable collection
- Processing pipelines
  - Apply standard procedures on files
- Digital Libraries
  - Publish data collections
- Preservation Environments
  - Ensure properties of a collection
  - Migrate collections to new technology
- Federation
  - Build collection that spans multiple data grids

What Does iRODS Do? (1 of 3)

- Remote High-Performance Data Access
  - get/put, read/write
  - Encapsulate small message in request to send
  - Parallel threads for large transfers with TCP/IP
  - Reliable Blast UDP
- Unified View Of Disparate Data
  - Organizes distributed data into a shareable collection
  - Separates physical from logical (logical name-space)
  - Keeps track of names and locations of files
- Storage System Independent
  - Unix/Windows File Systems
  - HPSS (Archival Storage)
  - Universal mass storage system interface

What Does iRODS Do? (2 of 3)

- Replication/Backup
  - Physical copies
- Metadata (RDBMS)
  - System and user-defined
  - PostgreSQL, Oracle, MySQL
  - Queries/Information Discovery
- Access Control
  - users/groups
  - Secure Passwords, Grid Security Infrastructure (GSI), Kerberos, Shibboleth soon

What Does iRODS Do? (3 of 3)

- Policies / computer actionable rules
  - Highly configurable
  - Enforce management policies
  - Automate administrative functions
  - Validate assessment criteria
- Workflows
  - Rules/Micro-services (immediate, delayed, periodic)
- Management of Large Collections
  - irsync - synchronize remote directory into the data grid
  - Audit trails
  - Descriptive metadata - extensible schema

Sharing Data in iRODS Data System

Scientist A
Adds data to Shared Collection

Scientist B
Accesses and analyzes shared Data

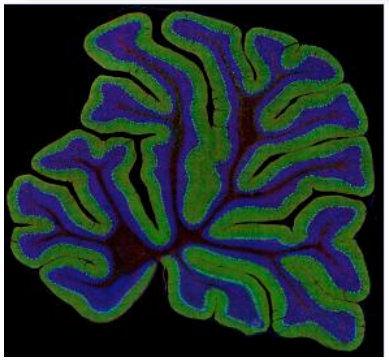
iRODS Data System

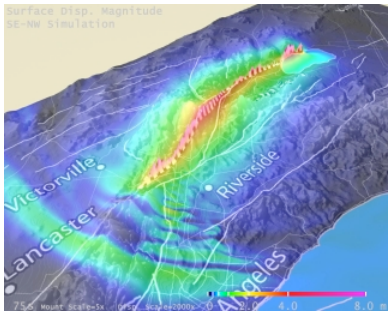
- ***Brain*** **Data Server, CA**
- ***Audio*** **Data Server, NJ**
- ***Video*** **Data Server, TN**
- **iRODS Metadata Catalog**

Scientists can use iRODS as a "data grid" to share multiple types of data, near and far. iRODS Rules also enforce and audit human subjects access restrictions.

D·I·C·E Data Intensive Cyber Environments · i·R·O·D·S · THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL · UCSD · NSF · THE NATIONAL ARCHIVES ARCHIVES.GOV

DICE Technologies Helping UCSD Projects





- **The National Center for Microscopy and Imaging Research** (NCMIR) is using DICE SRB and testing iRODS in the Cell Centered Database project.
- DICE iRODS helps computational seismologists from the **Southern California Earthquake Center** (SCEC) manage large-scale earthquake simulation data at **SDSC** and other TeraGrid sites.
- **UCSD Libraries Digital Asset Management System** (DAMS) using DICE technologies, including SRB.
- DICE iRODS helps **Ocean Observatories Initiative (OOI)** with Scripps and Calit2 manage large-scale diverse ocean data, including real-time streaming data.
- And others including **CineGrid**, **TDLC**, etc.

D·I·C·E Data Intensive Cyber Environments

i·R·O·D·S

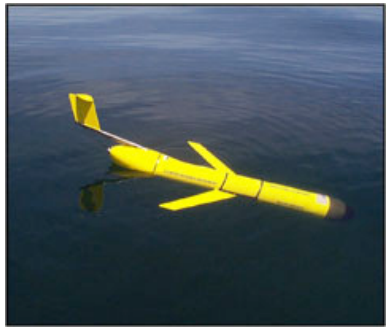
THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

UCSD

NSF

THE NATIONAL ARCHIVES ARCHIVES.GOV

Connecting Data Collections for New Science



- **"Federating" isolated "silos" of data enables new collaborations**
  - OOI ocean data flows in iRODS data grid to NOAA National Climatic Data Center (NCDC)
  - NCDC climate data is accessed through data grid for CUAHSI hydrology research on floods
  - CUAHSI hydrology data connects to Odom Institute for social science research on human impacts and response to floods
  - OOI climate data discovered and flows to iPlant Consortium for designing drought-resistant plants for climate change adaptation

D·I·C·E Data Intensive Cyber Environments

i·R·O·D·S

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

UCSD

NSF

THE NATIONAL ARCHIVES ARCHIVES.GOV

Growing Use of iRODS Data System

- ***Astronomy:*** NOAO, NVO, Observatoire de Strasbourg, France; CADAC, etc.
- ***Geo:*** NOAA NCDC; OOI; SCEC, etc.
- ***HPC:*** TeraGrid sites, SDSC, TACC, NICS, etc. NASA NCCS
- ***Bio:*** TDLC, NICMIR, iPlant, etc.
- ***Preservation:*** NARA TPAP, French National Library; Texas Digital Library; Fedora Commons; Dspace, etc.
- ***Workflow:*** Kepler, Taverna, etc.
- ***International:*** EU SHAMAN; Australian ARCS; UK e-Science; KEK (Japan); Academica Sinica (Taiwan); CC-IN2P3 HEP, France; etc.
- ***Industry:*** IBM, Oracle/Sun, Atos Origin, Microsoft, DataDirect

DICE Team

- Data Intensive Cyber Environments Center
  - University of North Carolina at Chapel Hill (UNC)
    - UNC School of Information and Library Science (SILS)
    - Renaissance Computing Institute (RENCI)
      - Reagan Moore (Professor)
      - Arcot Rajasekar (Professor)
      - Antoine de Torcy, Chien-Yi Hou, Mike Conway
  - UC San Diego
    - Institute for Neural Computation (INC)
      - Mike Wan
      - Wayne Schroeder
      - Sheau-Yen Chen, Bing Zhu, Paul Tooby
  - iRODS development is supported by
    - NSF OCI-0848296 "NARA Transcontinental Persistent Archives Prototype" (2008-2012)
    - NSF SDCI 0721400 "Data Grids for Community Driven Applications" (2007-2010)