THE PAST, PRESENT AND FUTURE OF IRODS AT TACC

Chris Jordan

TACC Data Management and Collections Group

IRODS UGM, June 10 2020

TEXAS ADVANCED COMPUTING CENTER

- Organized Research Unit of the University of Texas at Austin
- Nowadays known for big NSF systems: Frontera and Stampede (1 & 2) – Among the fastest and most-utilized high performance compute resources globally
- Diverse compute, visualization and data resources serving wide array of research needs, at all scales and independent of geography
- Corral storage resource, and iRODS as an access mechanism, have supported persistent research data collections for over a decade as of 2020
- Also support ~100PB of high-performance cluster storage, >100PB of Tape Archive

WHY IRODS AT TACC? WHY SO LONG?

- In a past life, your presenter worked at SDSC in UC San Diego, with the iRODS group
- Moved to TACC with the charge to improve provision for research data management in Texas
- IRODS was a natural fit for bringing on some new collections and testing data workflows
  - Arctos natural history collection images with web access
  - Archeological artifact media collections (highly structured)
  - Early experiments with archive integration
- Over the last decade+, application modes and collection patterns have evolved

CORRAL AND IRODS

- Corral commissioned in 2009 as a ~1PB Lustre resource using private donation
  - Major advantages: no explicit lifespan for data, freedom to allocate based on research needs/partnerships, explicit tie to data services for access
  - iRODS one of the first services offered on Corral
- Corral 2 deployed 2013: 4PB GPFS with offsite replication
- Corral 3 in 2016 expands to 12PB capacity, retains replication
  - At each stage, iRODS was deployed and collections in iRODS persisted across the hardware change

IRODS DEPLOYMENTS

- TACC Supports 3 separate iRODS deployments on Corral
  - General Purpose – available to any UT System researcher, TACC authentication support, integrates web publication mechanisms: ~1PB total usage
  - CyVerse – Resource Server in the CyVerse grid – Replicated data from primary Arizona resource servers. Data available for local TACC usage: ~2PB
  - Galaxy - New iRODS zone for testing of Galaxy with iRODS as the data store. Long-term storage, supporting retrieval of data for computational usage and storage of results.

GENERAL PURPOSE IRODS: USE CASES

- iRODS now primarily works best with 2 major use cases:
  - Large, complex, long-term collections - ~100TB on average, with specialized metadata needs, or common workflows. Instrument integration for research data generation a major scenario
    - Mid-American Geospatial Information Center (MAGIC)
    - University of Texas CT Facility – General CT Scanning facility, long-term data archive
  - Web-focused collections, iRODS used to manage data for web publication
    - Arctos Institutions: MVZ/UC Berkeley, U Alaska Museum of the North, Museum of Southwest Biology/UNM
- Also involved in various digital preservation and publishing efforts (DPN, Chronopolis, TDL)

WEB ACCESS TO IRODS COLLECTIONS

- Simple setup with top-level /corralZ/web/ directory mapped to a web root
- Allows projects to manage both private and public collections within iRODS
- Makes clear when data is published to the web via data move or copy
- Used by MAGIC, Arctos institutions, several smaller projects for web publication of research data

CYVERSE INTEGRATION

- iRODS at TACC provides only a resource server – catalog services provided at ASU
- Separate, dedicated server utilizing Corral storage
- Lots of customization at the level of the rule engine – tools for verifying checksums, dealing with NetCDF files, etc
  - See Tony Edgin's presentation for info on CyVerse use of iRODS
- Historically, mostly just a replica site for data, but opportunity exists for computation at TACC to take advantage of data at TACC

GALAXY INTEGRATION

- "Galaxy is an open source, web-based platform for data-intensive biomedical research"
- "Galaxy", or the Galaxy community, has over 3.5PB of data on Corral
  - Currently just stored in POSIX file system and accessed by Galaxy software via NFS mounts
- Project begun in April 2020 to test use of iRODS for the "static" file storage, with data retrieved to "scratch" storage for computational workflows
- Once interoperability is established, plans to make use of Audit tools, storage hierarchies/policies, and metadata capabilities

A WHOLE GRAVEYARD OF TECH DEMOS

- Diverse needs over long periods of time means you try many, many things
  - Periodic experiments with replication to tape archives
  - Testing of iRODS with a DDN WOS demo unit
  - Re-implementing i-Commands on the REST API using BASH + Curl
  - A homebrew audit trail mechanism using the rule engine
  - The original WebDAV server, subsequent WebDAV implementations, various FUSE implementations, and so on

FUTURE OF IRODS AND RESEARCH DATA MANAGEMENT

- Interfaces, Interfaces, Interfaces – Users still most interested in easy access to data, and metadata. Still few commonalities in the research community
- Securing data, managing access, other compliance issues
- PAM + Multi-Factor Authentication
- Lots of interest in S3/Object store interfaces generally – need to understand why S3 is being asked for and how to provide for it
- More dynamic deployment scenarios (Containers, Orchestration, etc)

CONTACTS, Q&A

- https://www.tacc.utexas.edu – General Information
- Email: data@tacc.utexas.edu for general data-related queries
- Email Chris: ctjordan@tacc.utexas.edu