

iRODS at TACC: Secure Infrastructure for Open Science

Chris Jordan

What is TACC?

- Texas Advanced Computing Center
 - Cyberinfrastructure Resources for Open Science
 - University of Texas System
 - 9 Academic, 6 Health campuses
 - NSF and NIH researchers
 - Commercial and non-commercial partners
 - Very diverse community with contrasting needs

Data Intensive Computing

- Data Analysis and Statistics
- Data Management and Collections
 - Data management planning and cleanup
 - File and database-oriented collections
 - “Pure” data dissemination/data sharing
 - Long term storage and project partnerships
 - Web repositories and custom toolkits

The TACC Ecosystem

- Stampede – Top 10/Petaflop-class traditional cluster HPC system
- Stockyard– 20 Petabytes of combined disk storage for all data needs
- Ranch – 160 Petabytes of tape archive storage
- Maverick/Rustler/Rodeo – “Niche” systems for visualization, Hadoop, VMs, etc

Corral

- 5 Petabytes replicated DDN storage
- GPFS basic file system
- SAS and SATA disk tiers
- >100 Projects, 100s of users, >4PB of data stored

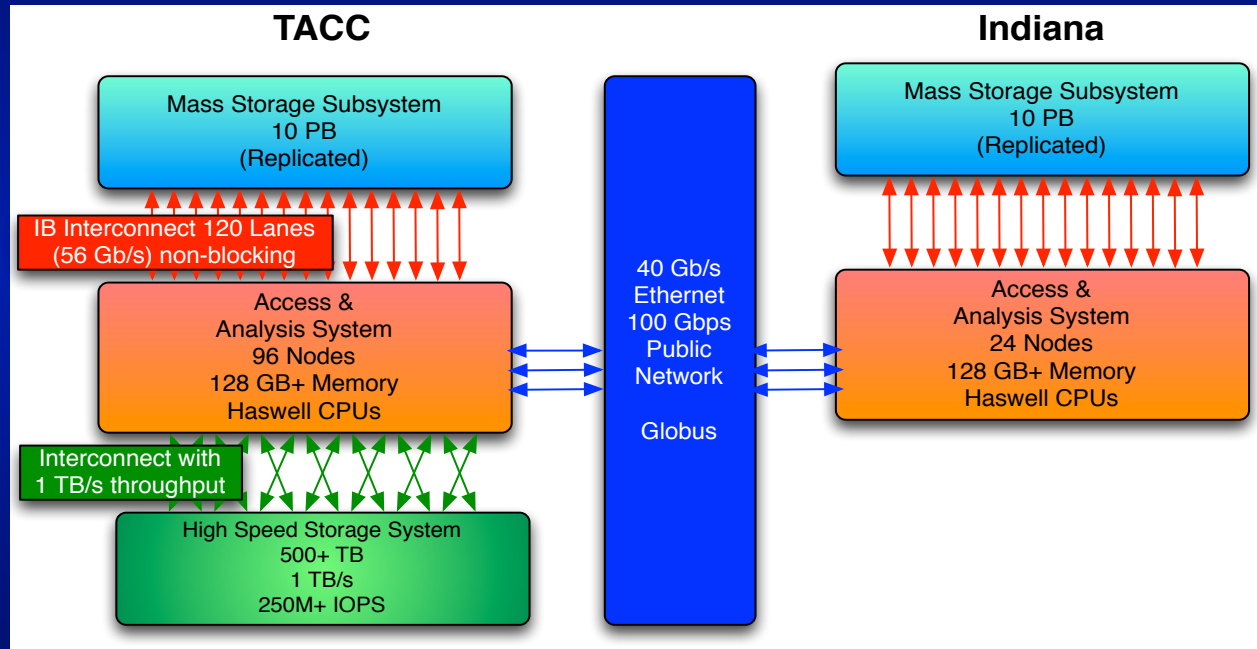
Corral iRODS

- iRODS 3.3.1, iDrop-Web, Davis
- Resources on replicated or unreplicated Corral file systems, Ranch tape archive
- Primarily used for data sharing, instrument facilities and other special projects

Corral Example: Institute of Classical Archeology

- Specialized metadata requirements
- Discipline-specific web interface
- Highly structured collection
- Automated metadata extraction on ingest, registration into PHP-based website
- Two-way metadata sync with website DB

Wrangler



Wrangler Service Model

- Need to dynamically provision a wide range of database, iRODS, web services
- Wrangler web portal provides simple user interface for requesting services
- Includes iRODS and iRODS policy/feature selection

Wrangler Dynamic Services

- Services can be dynamic or persistent
- Dynamic services run on compute nodes, for days or weeks
- Collaborative creation/deployment model
- Can deploy iRODS as dynamic service
 - Want to test iRODS database backends on the fastest storage system on the planet? OK

Wrangler Portal

The screenshot shows the TACC Wrangler Data Portal interface. At the top left is the TACC logo (Texas Advanced Computing Center) and the text 'WRANGLER DATA PORTAL'. To the right are social media icons for Facebook, Twitter, Google+, YouTube, and LinkedIn. Below the header is a navigation bar with a home icon, 'Documentation', and a user greeting 'Hello, Christopher'. The main content area is titled 'TG-Sta110019S : XSEDE SP TACC' and includes a 'Details' button. It is divided into three sections: 'Active and Pending Reservations' with 'Create Node Reservation' and 'Create Hadoop Reservation' buttons; 'Past Reservations' with a list of reservation IDs (e.g., dssd+TG-Sta110019S+211) and expandable icons; and 'DB Services' with a message 'Manage DB Services coming soon. Please submit a consulting ticket to create a persistent database service.' Below this is the 'iRODS Collections' section, showing 'test_nonpublic' and 'test_public' with expandable icons and a 'Create iRODS Collection' button.

TACC WRANGLER DATA PORTAL

Documentation Hello, Christopher

TG-Sta110019S : XSEDE SP TACC Details

Active and Pending Reservations

[Create Node Reservation](#) [Create Hadoop Reservation](#)

Past Reservations

- dssd+TG-Sta110019S+211
- dssd+TG-Sta110019S+208
- dssd+TG-Sta110019S+205
- dssd+TG-Sta110019S+202
- dssd+TG-Sta110019S+172
- dssd+TG-Sta110019S+166
- dssd+TG-Sta110019S+157
- hadoop+TG-Sta110019S+152
- hadoop+TG-Sta110019S+149
- hadoop+TG-Sta110019S+146
- dssd+TG-Sta110019S+143
- dssd+TG-Sta110019S+137
- dssd+TG-Sta110019S+119

DB Services

Manage DB Services coming soon.
Please submit a [consulting ticket](#) to create a persistent database service.

iRODS Collections

- test_nonpublic
- test_public

[Create iRODS Collection](#)

iRODS in Wrangler

- Primary Data Management mechanism
- Will support data publication w/ DOIs, audit trails, checksum/manifest verification, etc
- Web and WebDAV interfaces
- Secondary use as platform for experimentation (new rules development, resource hierarchies, website integration)

Use case: Research Instruments

- Genome sequencing, CT scanners, fMRI
- Telescopes, particle colliders, and bears ...
- Central challenge is data management and dissemination to customers
- iRODS used for metadata, direct output, access controls, short and long-term storage

HIPAA/FISMA Requirements

- Documentation, Policy, Documentation
- Secure replicated storage “at rest”
- Secure transmission/networking
- Effective access controls
- User education

Securing iRODS

- Best practices (mailing list, other big users)
- Difficult to mix “secure” and “less secure” installations (strict ACLs, web sharing, etc)
- De-identify or isolate secure data
- Need to look at infrastructure as a whole:
 - Network security is crucial
 - Firewall is required, VPN may be required

Securing iRODS 2

- Really need two-factor authentication
- Possible with PAM (?)
 - Looking at Toopher plus iRODS
- Need improved SSL support
- iRODS + SSL + PAM 2-factor + Attractive web interface = Killer app

Chris Jordan

ctjordan@tacc.utexas.edu

For more information:

www.tacc.utexas.edu

