iRODS User Group Meeting

Hosted by

Dr. Hermann Lederer

Rechenzentrum Garching der Max-Planck-Gesellschaft

Max-Planck-Institut fuer Plasmaphysik

Garching, Germany

Feb 28 - March 1, 2013

DICE Team

- Wayne Schroeder (architect)
- Mike Wan (retired)
- Arcot Rajasekar (innovation)
- Hao Xu (rule engine)
- Mike Conway (Jargon)
- Antoine de Torcy (micro-services)
- Sheau-Yen Chen (administration)
- Reagan Moore (projects)

DFC DataNet FEDERATION CONSORTIUM

NSF

Usage Statistics

- Based on 6-months of downloads of version 3.1
  - Accessed by projects in 39 countries
  - Accessed by 62 academic institutions in the US
  - Accessed by 16 companies
  - Accessed by 10 US federal agencies
- Total of 858 downloads
  - National data grids
  - Digital libraries
  - Archives
  - Institutional repositories
  - Collaboration environments

DFC DataNet FEDERATION CONSORTIUM

NSF

Supporting Projects

- NSF SDCI – development of new features, support for user communities (Sept 2013)
- NSF EarthCube – interoperability with other systems (2013)
- NSF DataBridge – socio-metric analysis of use of data and application of processes (2015)
- NSF DataNet Federation Consortium – application to research domains (2016)

DFC DataNet FEDERATION CONSORTIUM

NSF

Future Directions

- Keep iRODS focused on generic middleware
  - Extensible
  - Policy-based
  - Pluggable
- Extend into new domains
  - Policy-based storage controllers (DDN)
  - Policy-based networks (Future Internet Architecture)
- Research Data Alliance Practical Policy working group
  - Standard policy sets
  - Identification of best practices (administration, verification)
  - Invite all to post policies http://centosext1.irods.renci.org/irodswiki/index.php/Practical_Policy

DFC DataNet FEDERATION CONSORTIUM

National Science Foundation Cooperative Agreement: OCI-0940841

Policy-Based Data Management

- ***Purpose*** - reason a collection is assembled
- ***Properties*** - attributes needed to ensure the **purpose**
- ***Policies*** - enforce and maintain collection **properties**
- ***Procedures*** - functions that implement the **policies**
- ***Persistent state information*** - results of applying **procedures**
- ***Property assessment criteria*** – validation that **state information** conforms to the desired **purpose**
- ***Federation*** - controlled sharing of **logical name spaces**

- These are the necessary elements for collaborative research

DFC DataNet FEDERATION CONSORTIUM

NSF

Cyberinfrastructure Evolution

- Managing data — File systems
  - Share files
- Managing information — Digital libraries
  - Share metadata context
- Managing knowledge — Policy-based systems
  - Share analysis workflows

- Expect cyberinfrastructure to extend into the Internet
  - Future Internet Architecture and policy-based networks

DFC DataNet FEDERATION CONSORTIUM

NSF

DFC Vision for Prototype

- Build national data cyberinfrastructure
  - Federation of existing data management systems
- Enable collaborative research
  - Sharing of workflows, information, and data
- Support reproducible science
  - Encapsulate knowledge in shared workflows
- Enable student participation in research
  - Policy-controlled access to “live” data

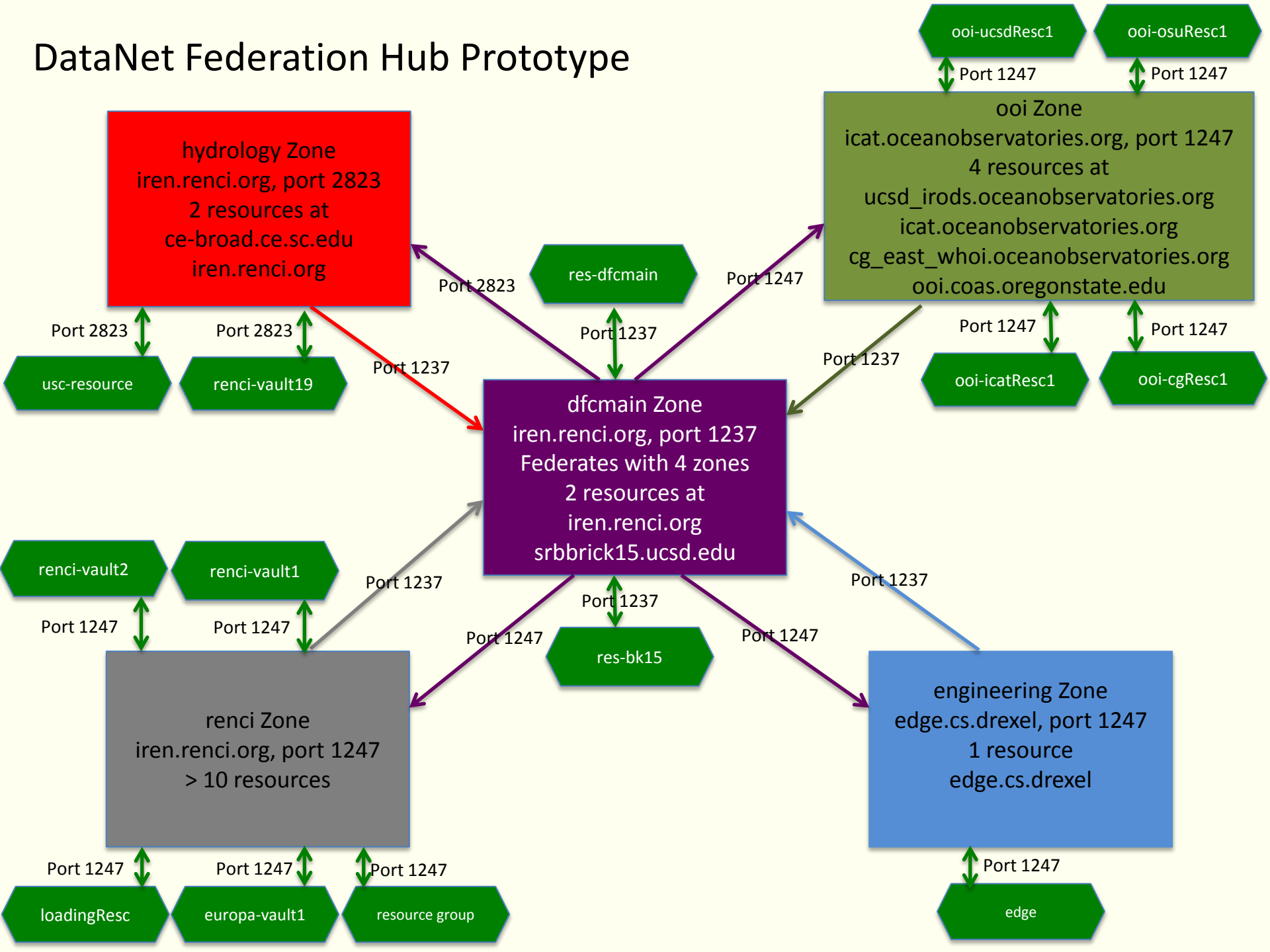
Approach: Data Driven Science

- (Cyberinfrastructure) Prototype designed to integrate domain infrastructure
  - Oceanography
    - Archiving of real-time sensor data streams,
    - Manipulation of NetCDF files,
    - Access through OpenDAP protocol
  - Engineering
    - Engineering digital library mediawiki integration
    - Automation of format transformations
  - Hydrology
    - Automation of hydrology workflows
    - Testbed for NSF EarthCube interoperability demonstrations
- (Researcher support) Survey of domain researchers for collaboration requirements
  - Continuing assessment
- (Broader Impact) Development of policy kits that simplify use by new domains
  - Cognitive Science, Plant Biology, Social Science domains
  - Out year activity

DFC DataNet FEDERATION CONSORTIUM

NSF

DataNet Federation Hub Prototype



Community Based Collection Life Cycle
(NSF Data Management Plans)

Each collection life cycle stage re-purposes the original collection

Project Collection	Data Grid	Data Processing Pipeline	Digital Library	Reference Collection	Federation
Private	Shared	Analyzed	Published	Preserved	Sustained
Local Policy	Distribution Policy	Service Policy	Description Policy	Representation Policy	Re-purposing Policy

Stages correspond to addition of new policies for successively broader communities
Virtualize the stages of the collection life cycle through policy evolution

DFC DataNet FEDERATION CONSORTIUM

NSF

Reproducible Science – Knowledge Management

- Capturing domain knowledge in workflows
  - Registration and sharing of workflows
  - Support for acquisition of data from remote resources
  - Support for transformation of input data
  - Automated management of workflow input and output files
  - Support for re-execution of workflows

Agenda

Thursday Morning February 28

9:00 AM	Welcome	Hermann Lederer, RG-MPG
9:15 AM	Introduction	Reagan Moore, UNC-CH
9:30 AM	iRODS version 3.2 features and bug fixes	Wayne Schroeder, UCSD
10:00 AM	iDrop and Jargon status	Mike Conway, UNC-CH
10:30 AM	Break	
11:00 AM	Composable resources	Terrell Russell, Jason Coposky, RENCI
11:30 AM	New PAM/LDAP authentication	Wayne Schroeder, UCSD

DFC DataNet FEDERATION CONSORTIUM

NSF

Agenda

Thursday afternoon February 28

1:30 PM	Administrative interface	Terrell Russell, RENCI
2:00 PM	Managing Petabytes of data with iRODS	Jean-Yves Nief, CC-IN2P3 at CC-IN2P3
2:30 PM	IDS – the INCF Dataspace	Raphael Ritz, Sina Khaknezhad, Sean Hill, INCF, Chris Smith, Distributed Bio
3:00 PM	Break	
3:30 PM	iRODS at ZIH and LSDMA project introduction	Richard Grunzke, ZIH
4:00 PM	Cloud S3 Storage Using iRODS - A Possible Access Paradigm, Solving Security Problems with iRODS	Alan Hall, NCDC
4:30 PM	Pragmatic approaches for enabling data driven collaborations for plant sciences and beyond	Andy Lenards, Edwin Skidmore, iPC
5:00 PM	E-iRODS consortium	Charles Schmitt, RENCI

DFC DataNet FEDERATION CONSORTIUM

NSF

Agenda

Friday morning March 1

9:00 AM	New Zealand Best Grid	Vladimir Menci, UC
9:30 AM	HIVE reserved vocabularies	Mike Conway, UNC-CH
10:00 AM	Using E-iRODS in the Management of Human Genomic Data for Research and Clinical Use	Charles Schmitt, Chris Bizon, Phil Owen, Joshua Sailsbery, Jason Reilly, Xiaoshu Wang, Erik Scott, Michael Shoffner, Nassib Nassar, Kirk Wilhelmsen, RENCI
10:30 AM	Break	
11:00 AM	AmpliStor - Unbreakable Distributed Storage	Steve Rietveld, Amplidata
11:30 AM	Embedding iRODS and integration with Object Storage	James Coomer, DDN

DFC DataNet FEDERATION CONSORTIUM

NSF

Agenda

Friday afternoon March 1

1:30 PM	Practical Use of iRODS in a Cross-Disciplinary Federation	Peter Wittenburg, Willem Elbers, Daan Broeder, John Kennedy, Beatriz Sanchez Bribian, MPI RZG
2:00 PM	CGHub: Kick-starting the Worldwide Genome Web	Christopher Wilks, Mark Diekhans, David Haussler, UCSC, Dan Maltbie, Annai Systems
2:30 PM	Distributed Storage System and Storage Federation at ASGC	Eric Yen, ASGC
3:00 PM	Break	
3:30 PM	PyRODS and Cheshire	Jerome Fuselier, John Harrison, SHAMAN
4:00 PM	Plans for next version	Wayne Schroeder, UCSD
4:30 PM	Solicitation of new features	Reagan Moore, UNC-CH
5:00 PM	Emerging businesses	Wayne Schroeder, UCSD, Reagan Moore, UNC-CH,

DFC DataNet FEDERATION CONSORTIUM

NSF

Usage Information

- Would like to generate a map identifying locations where IRODS is being used, similar to
- http://www.moosefs.org/who-is-using-moosefs.html

- Please provide information on street address, data sizes, number of files, zone name

DFC DataNet FEDERATION CONSORTIUM

NSF