iRODS User Group

integrated Rule Oriented Data System

Reagan Moore

{moore, sekar, mwan, schroeder, bzhu, ptooby, antoine, sheauc}@diceresearch.org

{chienyi, marciano, michael_conway}@email.unc.edu

Wireless

SSID: UNC-1

WEP Key: 2003acce55

Agenda - Wednesday

- **Session I** (9:00- 10:30)
  - **Introduction to iRODS** (30 min) Moore
  - **iRODS Version 2.3** (30 min) Schroeder
  - **Intro on micro-services** (30 min) Moore
- Break (30 min)
- **Session II (**11:00-12:30)
  - **Intro to policies** (30 min) Moore
  - **Policy session,** how to build a set of policies for your collection (1 hour) Rajasekar
- Lunch **(**12:30 – 1:30)
- **Session III (**1:30- 3:00)
  - **Micro-service session,** how to write a micro-service (1 hour) Wan
  - **Advanced iCommands** (30 min) Wan
- Break (30 min)
- **Session IV** (3:30-5:00)
  - **iCat interactions** (1 hour) Schroeder / Rajasekar
  - **Questions** (30 min)

Agenda - Thursday

- **Session V** (9:00-10:30)
  - **User application sessions, how communities have applied iRODS**
    - ***High Availability iRODS System (HAIRS)*** Yutaka Kawai (KEK, Japan), Adil Hasan (University of Liverpool) (teleconference)
    - ***iRODS at CC-IN2P3*** Jean-Yves Nief, Pascal Calvat, Yonny Cardenas, Pierre-Yves Jallud, Thomas Kachelhoffer (CC-IN2P3, Lyon, France)
    - ***Using iRODS to Preserve and Publish a Dataverse Archive***, Mason Chua (Odum Institute, UNC), Antoine de Torcy (DICE Center, UNC), Jewel H. Ward (SILS, UNC), Jonathan Crabtree (Odum Institute, UNC)
    - ***Distributed Data Sharing with PetaShare for Collaborative Research***, PetaShare Team @LSU (poster)
    - ***University of North Carolina Information Technology Services***, William Schultz (poster)
- Break (30 Min)
- **Session VI** (11:00-12:30)
    - ***The ARCS Data Fabric***, Shunde Zhang, Florian Goessmann, Pauline Mak (poster)
    - ***A Service-Oriented Interface to the iRODS Data Grid***, Nicola Venuti, Francesco Locunto, Michael Conway, Leesa Brieger
    - ***iExplore for iRODS Distributed Data Management***, Bing Zhu (DICE group, UCSD)
    - ***A GridFTP Interface for iRODS***, Shunde Zhang
- Lunch (12:30-1:30)

Agenda - Thursday (Cont)

- **Session VII** (1:30-3:00)
  - **Clients for iRODS**
    - ***The Development of Digital Archives Management Tools for iRODS,*** Tsung-Tai Yeh, Hsin-Wen Wei, Shin-Hao Liu (Academia Sinica, Taiwan), Pei-Chi Huang (Tsing Hua University, Taiwan), Tsan-sheng Hsu (Academia Sinica, Taiwan), Yen-Chiu Chen (Tsing Hua University, Taiwan)
    - ***Building a Trusted Distributed Archival Preservation Service with iRODS***, Jewel H. Ward, Terrell G. Russell, and Alexandra Chassanoff (poster)
    - ***Conceptualizing Policy-Driven Repository Interoperability (PoDRI) Using iRODS and Fedora***, David Pcolar, Daniel W. Davis, Bing Zhu, Alexandra Chassanoff, Chien-Yi Hou, Richard Marciano
    - ***Community-Driven Development of Preservation Services***, Richard Marciano
- Break (30 min)
- **Session VIII** (3:30-5:00)
  - ***Enhancing iRODS Integration: Jargon and an Evolving iRODS Service Model*** Mike Conway (DICE Center, UNC)
  - Questions on user porting of clients

Agenda - Friday

- **Session IX** (9:00-10:30)
  - **Prioritization of tasks** (1 1/2 hour) Moore
- Break (30 min)
- **Session X** (11:00-12:30)
  - **Question and Answers** (1 1/2 hours) Moore
- Lunch (12:30 – 1:30)
- **Session XI** (1:30 – 3:00)
  - **Integration session,** how to integrate your favorite workflow/ client with iRODS (60 min) Conway
  - **Data Intensive Cyberinfrastructure Foundation session,** coordinating development across interested communities. (30 minutes) Tooby

Goal - iRODS User Group Meeting

- Present most recent developments
  - Within the DICE group
  - By iRODS collaborators
- Gain feedback:
  - Use experience
  - Desired features
  - Production environments
  - Production policies
- Prioritize
  - New development
  - New clients

Development Team

- iRODS development and application support
  - Sheau-Yen Chen - Data Grid Administration
  - Mike Conway - Java (Jargon)
  - Chien-Yi Hou - Preservation Micro-services
  - Richard Marciano - Preservation Development Lead
  - Reagan Moore - PI
  - Arcot Rajasekar - iRODS Development Lead
  - Wayne Schroeder - iRODS Product Mgr., Developer
  - Paul Tooby - Documentation, Foundation
  - Antoine de Torcy - Preservation Micro-services
  - Mike Wan - iRODS Chief Architect
  - Bing Zhu - Fedora, Windows
- Graduate Students
  - Christine Cheng - metadata
  - Rahul Deshmukh - MakeFlow / NetCDF
  - William Miao - protocol documentation
  - Russell Terrell - user interface
  - Jewel Ward - policy set comparison
  - Hao Xu - rule engine

Goal - Generic Infrastructure

- Manage all stages of the data life cycle
  - Data organization
  - Data processing pipelines
  - Collection creation
  - Data sharing
  - Data publication
  - Data preservation
- Create reference collection against which future information and knowledge is compared
  - Each stage uses similar storage, arrangement, description, and access mechanisms

Preservation is a Stage in the Data Life Cycle

Each data 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 a broader community

Virtualize the stages of the data life cycle through policy evolution

Interoperability across data life cycle representations

Policy-based Data Management

- *Purpose* - reason a collection is assembled
- *Properties* - attributes needed to ensure the **purpose**
- *Policies* - control for ensuring maintenance of **properties**
- *Procedures* - functions that implement the **policies**
- *State information* - results of applying the **procedures**
- *Assessment criteria* - validation that **state information** conforms to the desired **purpose**
- *Federation* - controlled sharing of **logical name spaces**

These are the necessary elements for data life cycle management

iRODS - Policy-based Data Management

- Turn policies into computer actionable rules
- Compose rules by chaining standard operations
  - Standard operations (micro-services) executed at the remote storage location
- Manage state information as attributes on namespaces:
  - Files / collections /users / resources / rules
- Validate assessment criteria
  - Queries on state information, parsing of audit trails
- Automate administrative functions
  - Minimize labor costs

Policy-based Preservation - Authenticity

- *Purpose* - Maintain authenticity of records
- *Properties* - Define template for required representation information
- *Policies* - Extract and register representation information for each file on ingestion
- *Procedures* - Parse record / XML file to extract metadata
- *State information* - Register representation information into metadata catalog
- *Assessment criteria* - Compare registered metadata with template defining required values

A preservation environment should automate each of these steps

DICE Data Intensive Cyber Environments · iRODS · The University of North Carolina at Chapel Hill · UCSD · NSF · The National Archives archives.gov

13

Assessment Criteria

- NARA Electronic Records Archive capabilities list
  - 853 defined capabilities
  - Mapped to 174 computer actionable rules
  - Mapped to 212 state information attibutes
- RLG/NARA Trusted Repository Audit Checklist
  - Mapped to 105 computer actionable rules
  - Included 66 rules specific to preservation
- ISO Mission Operations Information Management System repository audit checklist
  - 106 policies for operation and control
  - Mapped to 52 computer actionable rules

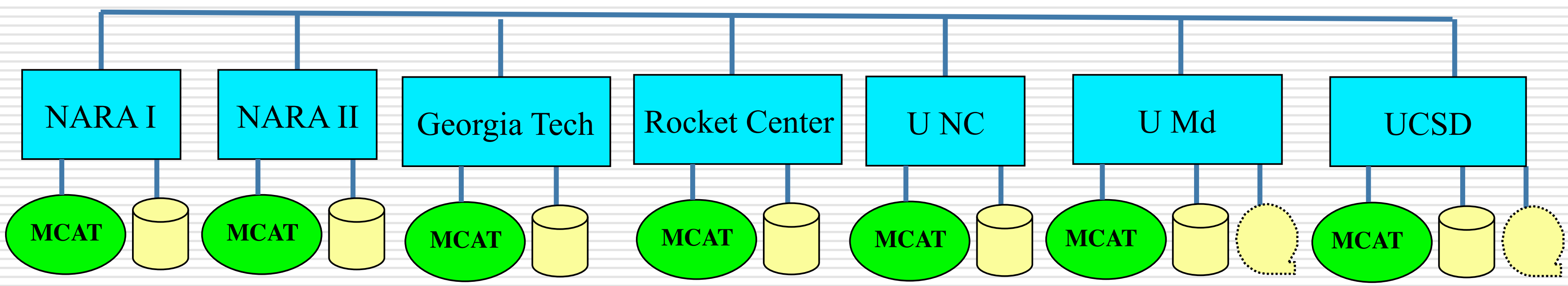
Examples of Assessment Criteria

- Specify
  - a template that governs the representation information required for a specific record series
  - content of a Submission Information Package (SIP)
  - content of an Archival Information Package (AIP)
  - number of replicas
- Verify
  - compliance of SIP with specification
  - compliance of AIP with specification
  - compliance with required replica number
  - integrity of the replicas

iRODS User Communities

- NARA Transcontinental Persistent Archive Prototype
  - Develop policies to automate preservation of selected digital holdings
- National Optical Astronomy Observatory
  - Accession images from a telescope in Chile
- Carolina Digital Repository
  - Preserve institutional collections

Federation of Seven Independent Data Grids

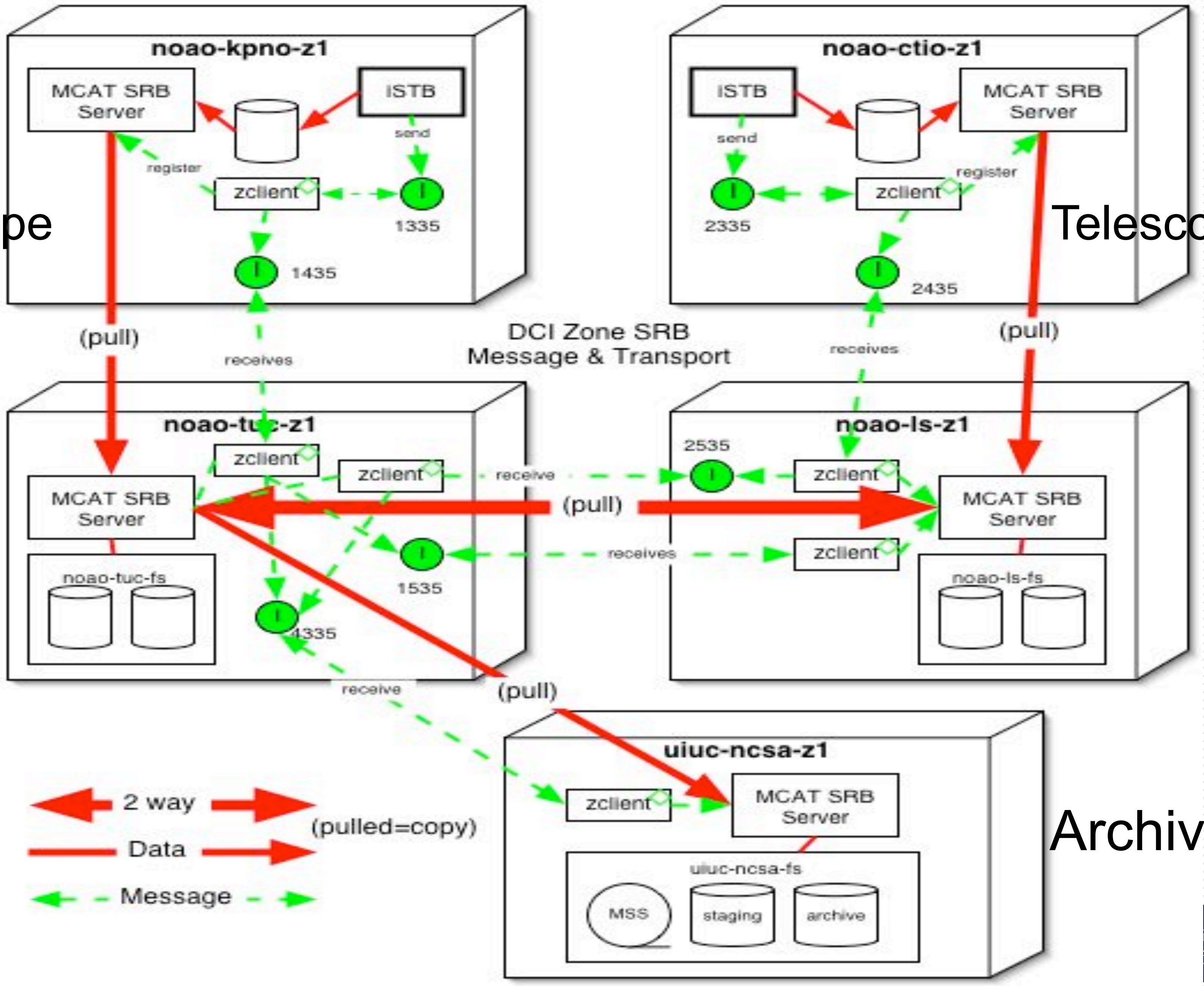


Extensible Environment, can federate with additional research and education sites. Each data grid can use different vendor products.

Policy to coalesce authentic records from independent data grids. Choose whether write to central archive, or use soft links.

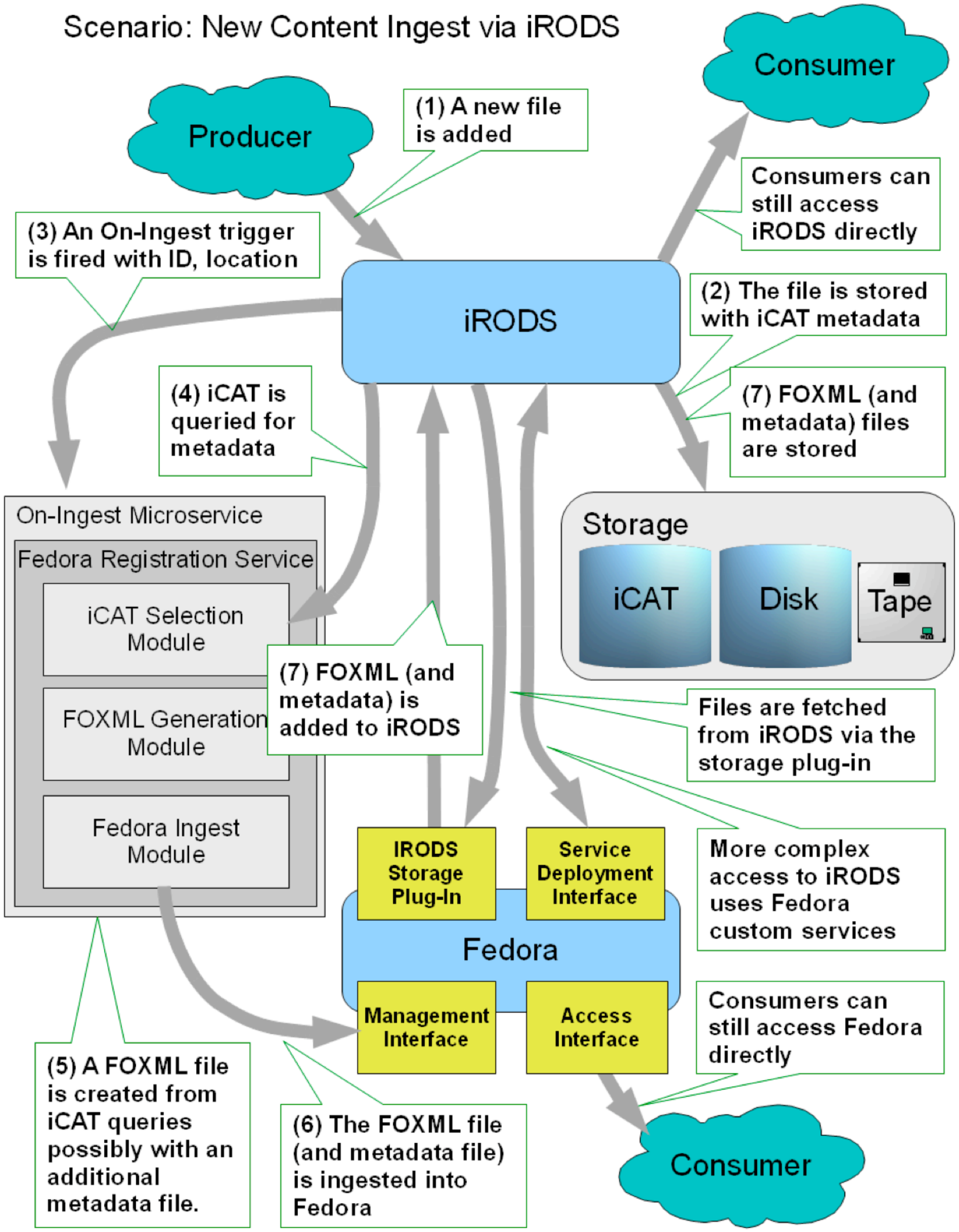
NOAO SRB Zone Architecture

Telescope



Telescope

Archive



Carolina Digital Repository

Architecture:

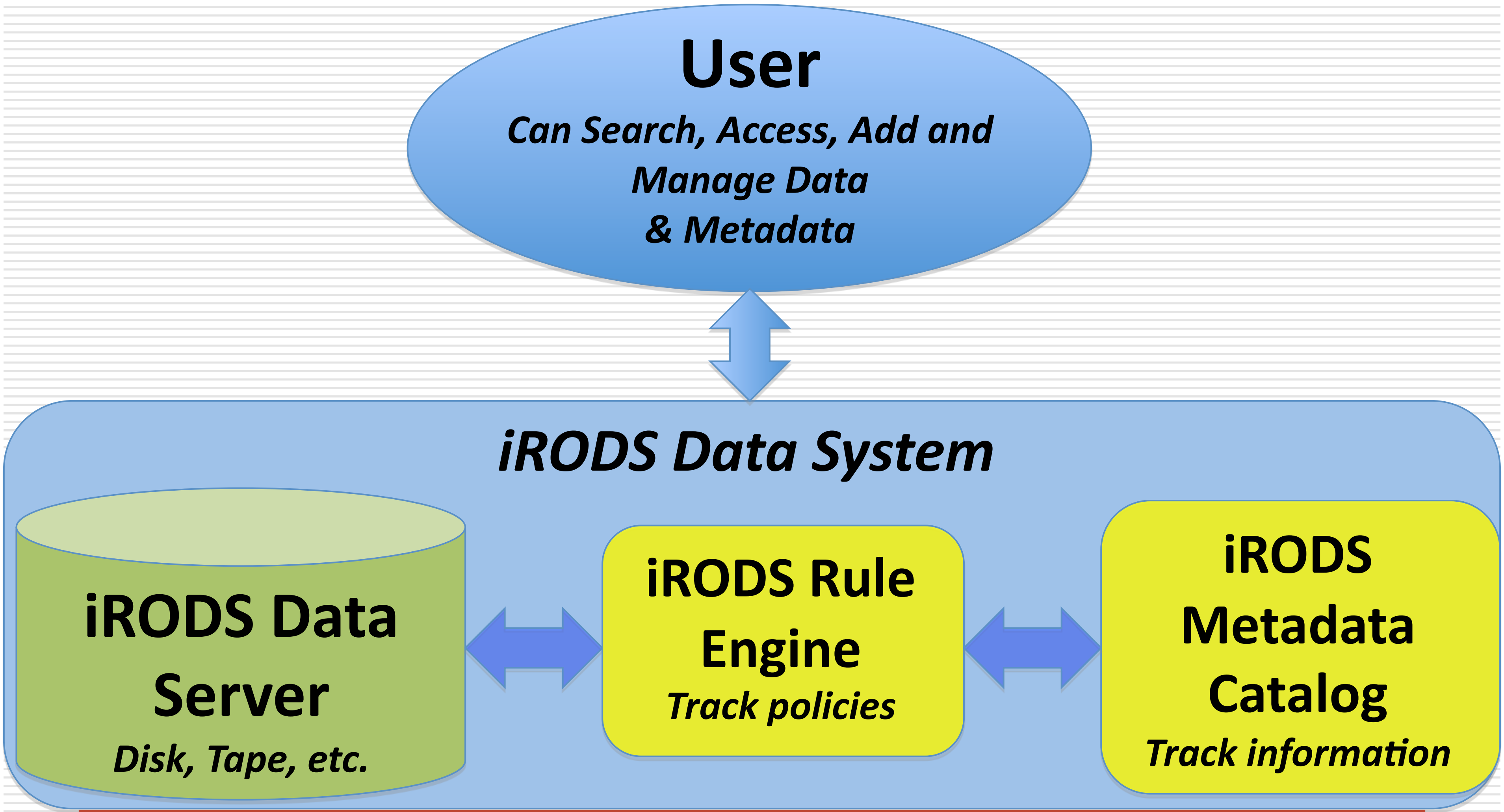
- Web interface
- Fedora digital library middleware
- iRODS data grid

Supports:

- Registration of file into iRODS
- Generation of FOXM
- Registration into Fedor
- Query through Fedor
- Synchronization of catalogs

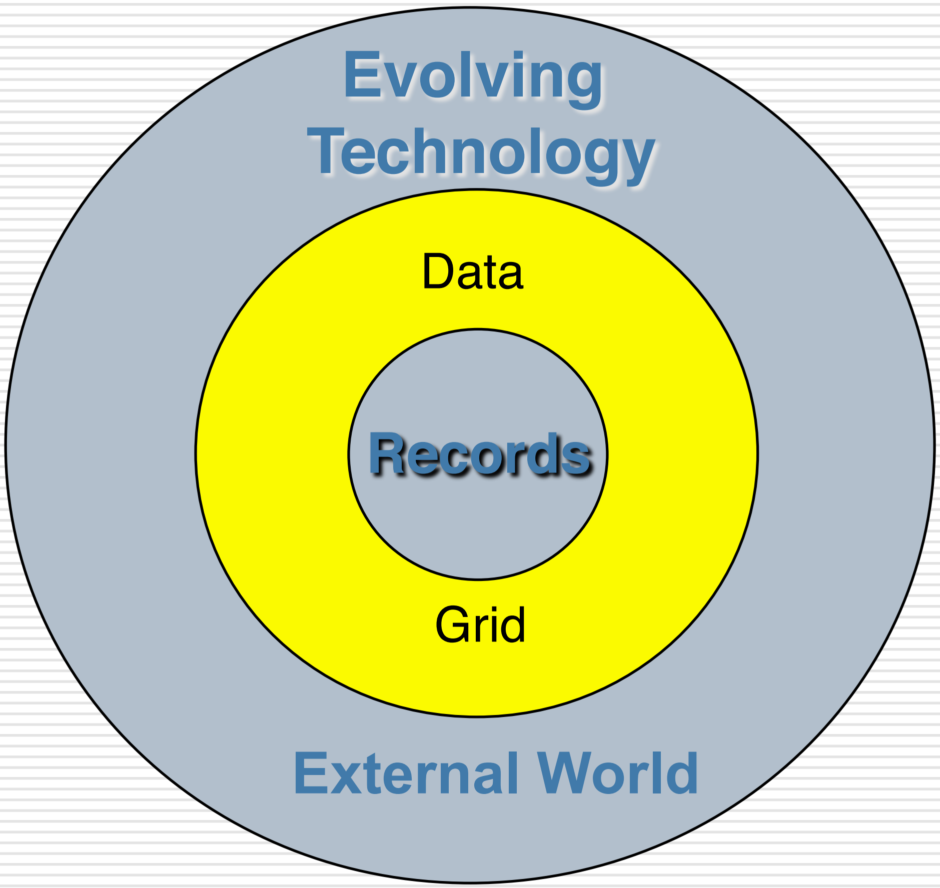
From Conceptualizing Policy-Driven Repository Interoperability (PoDRI) Using iRODS and Fedora (Pcolar, Davis, Zhu, Chassanoff, Hou, Marciano)

Overview of iRODS Architecture



*Access data with Web-based Browser or iRODS GUI or Command Line clients.

Infrastructure Independence



Data grid middleware insulates records from changes in the external world.

Data grid maps from procedures to new operating systems, protocols, and clients

Data grid provides interoperability mechanisms between old and new technologies

Migration of Micro-services

Access Interface

Map from actions requested by the access method to a standard set of Micro-services.

Standard Micro-services

Data Grid

Standard Operations

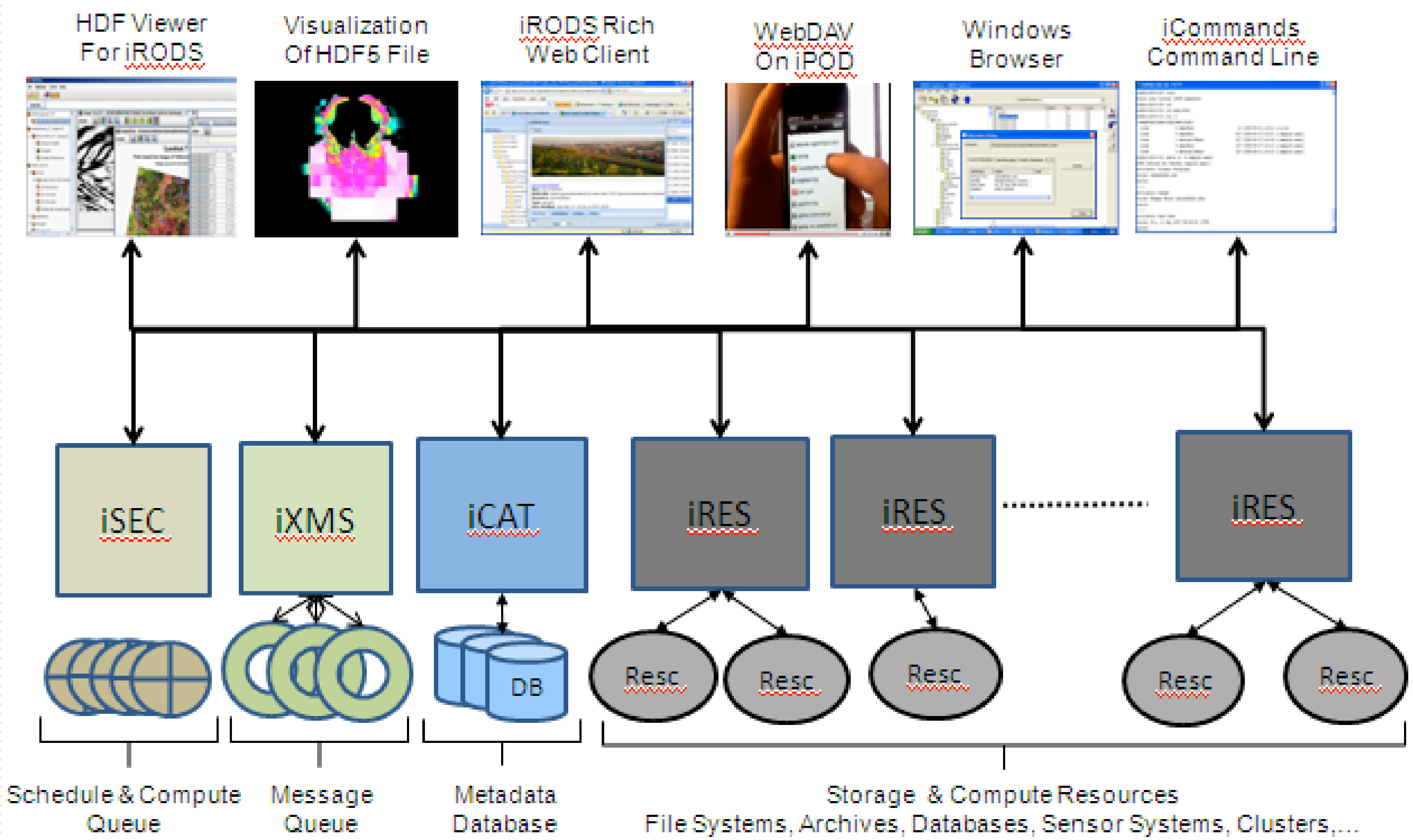
Map the standard Micro-services to standard operations.

Storage Protocol

Storage System

Map the operations to protocol supported by the operating system.

iRODS - Distributed Operating System



Future Development

- Development of simple preservation environment interfaces
  - Template based presentation as in Islandora
- Preservation management features
  - Format parsing routines
  - Representation metadata
- Automated creation of assessment policies
  - Given a template, create rule to validate use
- Development of standard preservation policy sets
  - Starter policy kits for communities

Research Coordination

- iRODS Development
  - NSF SDCI - supports development of core iRODS Data Grid infrastructure
- iRODS Applications
  - NSF NARA - supports application of Data Grids to preservation environments
  - NSF OOI - future integration of Data Grids with real-time sensor data streams and grid computing
  - NSF TDLC - production TDLC Data Grid and extension to remaining five Science of Learning Centers
  - NSF SCEC - current production environment
  - NSF Teragrid - production environment
- iRODS collaborations
  - Exchange of open-source technology with projects in the UK, France, Taiwan, Australia, Japan, US

Funding

- First generation Data Grid - Storage Resource Broker (SRB)
  - DARPA Massive Data Analysis System (1996)
  - DARPA/USPTO Distributed Object Computation Testbed (1998)
  - NARA Persistent Archive (1999)
  - Application driven development (2000-2005)
- Second generation Data Grid - iRODS
  - NSF ITR 0427196, “Constraint-based Knowledge Systems for Grids, Digital Libraries, and Persistent Archives” (2004)
  - NARA supplement to NSF SCI 0438741, “Cyberinfrastructure; From Vision to Reality” - “Transcontinental Persistent Archive Prototype” (TPAP) (2005)
  - NSF SDCI 0721400, "SDCI Data Improvement: Data Grids for Community Driven Applications” (2007)
  - NARA/NSF OCI 0848296, “NARA Transcontinental Persistent Archive Prototype” (2008)

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

Proposals Submitted

- NSF DataNet
  - Explore creation of national infrastructure linking federal repositories and NSF research initiatives
  - $20 million, 10 institutions, 6 science and engineering consortia, 5 years
- NSF SDCI
  - Continue development of iRODS
  - $3 million, 3 years
- DOE data management at extreme scale
  - Integrate with Open Science Grid, Earth Systems Grid
  - $1.3 million, 3 years
- NARA Transcontinental Persistent Archive Prototype
  - Build preservation policies
  - $2.7 million, 3 years

Data Grid Development Costs

- Storage Resource Broker middleware
  - 300,000 lines of code
  - Six year development / ten year deployment
  - 10-15 professional software engineers
- Total cost ~ $15,000,000
  - $17 / line for design, development, testing, documentation, bug fixes
  - $14 / line for interoperability (clients)
  - $12 / line for application use support
  - $7 / line for management / administration
  - Total cost ~ $50 / line
- Development and application funded by:
  - NSF / NARA / DARPA / DoE / NASA / NIH / IMLS / NHPRC / LoC / DoD
  - More than 20 funded projects to sustain development
  - International collaborations on use, development, bug fixes, support

Foundation

- Data Intensive Cyber Environments Foundation
  - Nonprofit open source software development
  - Promotes use of iRODS technology
  - Supports standards efforts
  - Coordinates international development efforts
    - IN2P3 - quota and monitoring system
    - King's College London - Shibboleth
    - Australian Research Collaboration Services - WebDAV
    - Academia Sinica - SRM interface

iRODS is a "coordinated NSF/OCI-Nat'l Archives research activity" under the auspices of the President's NITRD Program and is identified as among the priorities underlying the President's 2009 Budget Supplement in the area of Human and Computer Interaction Information Management technology research.

Reagan W. Moore

rwmoore@renci.org

http://irods.diceresearch.org

NSF OCI-0848296 "NARA Transcontinental Persistent Archives Prototype"

NSF SDCI-0721400 "Data Grids for Community Driven Applications"