Policy Based Data Management

Reagan W. Moore
Arcot Rajasekar
Mike Wan
Wayne Schroeder
Mike Conway
Jason Coposky

{moore,sekar,mwan, schroeder}@diceresearch.org
michael_conway@unc.edu
http://irods.diceresearch.org
Policy-based Data Environments

- **Purpose** - reason a collection is assembled
- **Properties** - attributes needed to ensure the purpose
- **Policies** - controls for enforcing desired properties, mapped to computer actionable rules
- **Procedures** - functions that implement the policies, mapped to computer actionable workflows
- **Persistent state information** - results of applying the procedures, mapped to system metadata
- **Assessment criteria** - validation that state information conforms to the desired purpose, mapped to periodically executed policies
Policy-based Data Sharing

Consensus on Policies and Procedures controlling the shared data
Applications

• Data grids – PB-size distributed collections
  – Astronomy – NOAO, CyberSKA, LSST
  – High Energy Physics – BaBar, KEK
  – Earth Systems – NASA (MODIS data set)
  – Australian Research Collaboration Service

• Institutional repositories
  – Carolina Digital Repository

• Libraries
  – Texas Digital Libraries
  – Seismology - Southern California Earthquake Center

• Archives
  – Ocean Observatories Initiative
Overview of iRODS Architecture

User w/Client
Can Search, Access, Add and Manage Data & Metadata

iRODS Middleware

iRODS Data Server
Disk, Tape, etc.

iRODS Rule Engine
Track Policies

iRODS Metadata Catalog
Track information

Access distributed data with Web-based Browser or iRODS GUI or Command Line clients.
Data Virtualization

- **Access Interface**
  - Map from the actions requested by the client to multiple policy enforcement points.

- **Policy Enforcement Points**
  - Map from policy to standard micro-services.

- **Standard Micro-services**
  - Map from micro-services to standard Posix I/O operations.

- **Standard I/O Operations**
  - Map standard I/O operations to the protocol supported by the storage system.
iRODS Components

• Clients – currently 48
  – Browsers / Digital library / File system / Grid tools / I/O libraries / Portal / Unix tools / Web services / Workflows

• Policy enforcement points – currently 71
  – Manage policies controlling actions, pre-action policies, and post-action policies

• Distributed rule engine
  – Control rule execution
  – Manage deferred and periodic policies
Highly Controlled Environment

• All accesses are authenticated
  – GSI / Kerberos / Challenge-response / Shibboleth

• All operations are authorized
  – ACLs on files, storage
  – Constraints on each rule

• Local rule base controls interactions with local storage
  – Local rules are enforced first
iRODS Extensible Infrastructure

- **Clients** – specific to discipline and life cycle state
- **Policies** – specific to discipline
- **Procedures** – specific to discipline
- **Remaining infrastructure is generic**
  - Network transport
  - Authentication / Authorization
  - Distributed storage access
  - Remote execution
  - Metadata management
  - Message passing
  - Rule engine
Capabilities

- Replication
- Registration of files into the data grid
- Synchronization of remote directory
- Managed file transport (iDrop)
- Automated metadata extraction
- Queries on metadata, tags
- Server-side workflows (loop over result sets)
- Parallel I/O streams & RBUDP transport
Policies

- Retention, disposition, distribution, arrangement
- Authenticity, provenance, description
- Integrity, replication, synchronization
- Deletion, trash cans, versioning
- Archiving, staging, caching
- Authentication, authorization, redaction
- Access, approval, IRB, audit trails, report generation
- Assessment criteria, validation
- Derived data product generation, format parsing
- Federation of independent data grids
Open Source Software

• **Community driven software development**
  – Focus on features required by user communities
  – Focus on bug-free software
  – Focus on highly reliable software
  – Focus on highly extensible software
  – Approximately 3-4 software releases per year

• **Distributed under a BSD license**
  – International collaborations on software development
  – IN2P3 (France), SHAMAN (UK), ARCS (Australia), Academia Sinica (Taiwan)
iRODS - Open Source Software

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”