Data Intensive Collaboration
iRODS Overview
Policy-Driven Data Preservation

Reagan W. Moore

Arcot Rajasekar

Mike Wan

{moore,sekar,mwan}@diceresearch.org

http://irods.diceresearch.org


THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL
UCSD


Integrated Rule Oriented Data System

- Software to organize distributed data into a shared collection
  - Manage properties of the shared collection
- Developed by DICE group
  - University of North Carolina
    - 5 staff, 3 students
  - University of California, San Diego
    - 6 staff, 2 students
- Funded by

NSF OCI-0848296 "NARA Transcontinental Persistent Archives Prototype"
NSF SDCI-0721400 "Data Grids for Community Driven Applications"

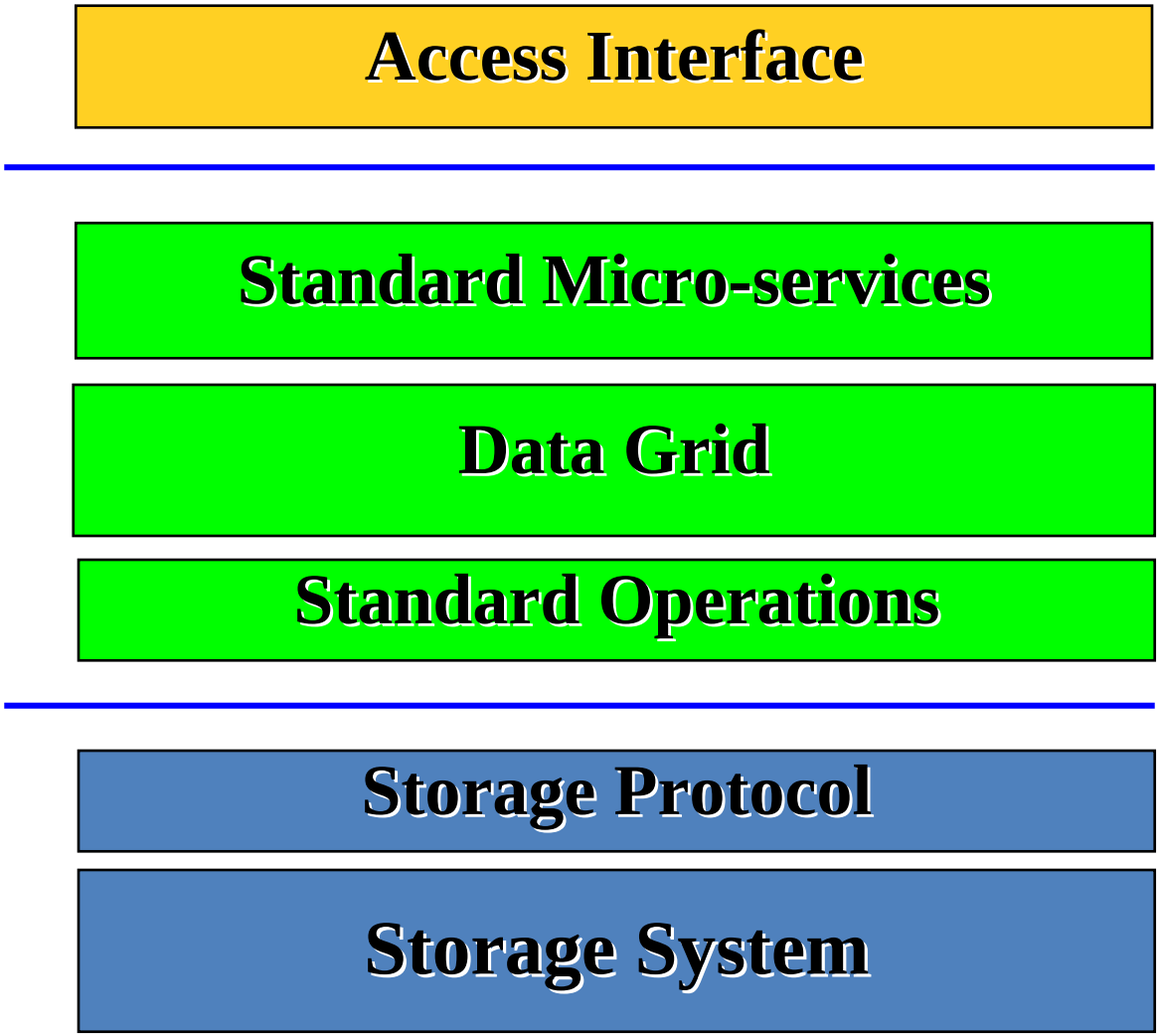
D·I·C·E i·R·O·D·S THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL UCSD NSF THE NATIONAL ARCHIVES ARCHIVES.GOV

Preservation Concept

- Maintain properties of records while underlying technology evolves
  - Records are permanent
  - Technology is ephemeral
- Implications
  - Active management of records
  - Multiple indirection mechanisms to protect records from dependencies upon technology
  - Validation of properties to verify assertions about records


THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL


Data Virtualization



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

The standard micro-services are mapped to standard operations.

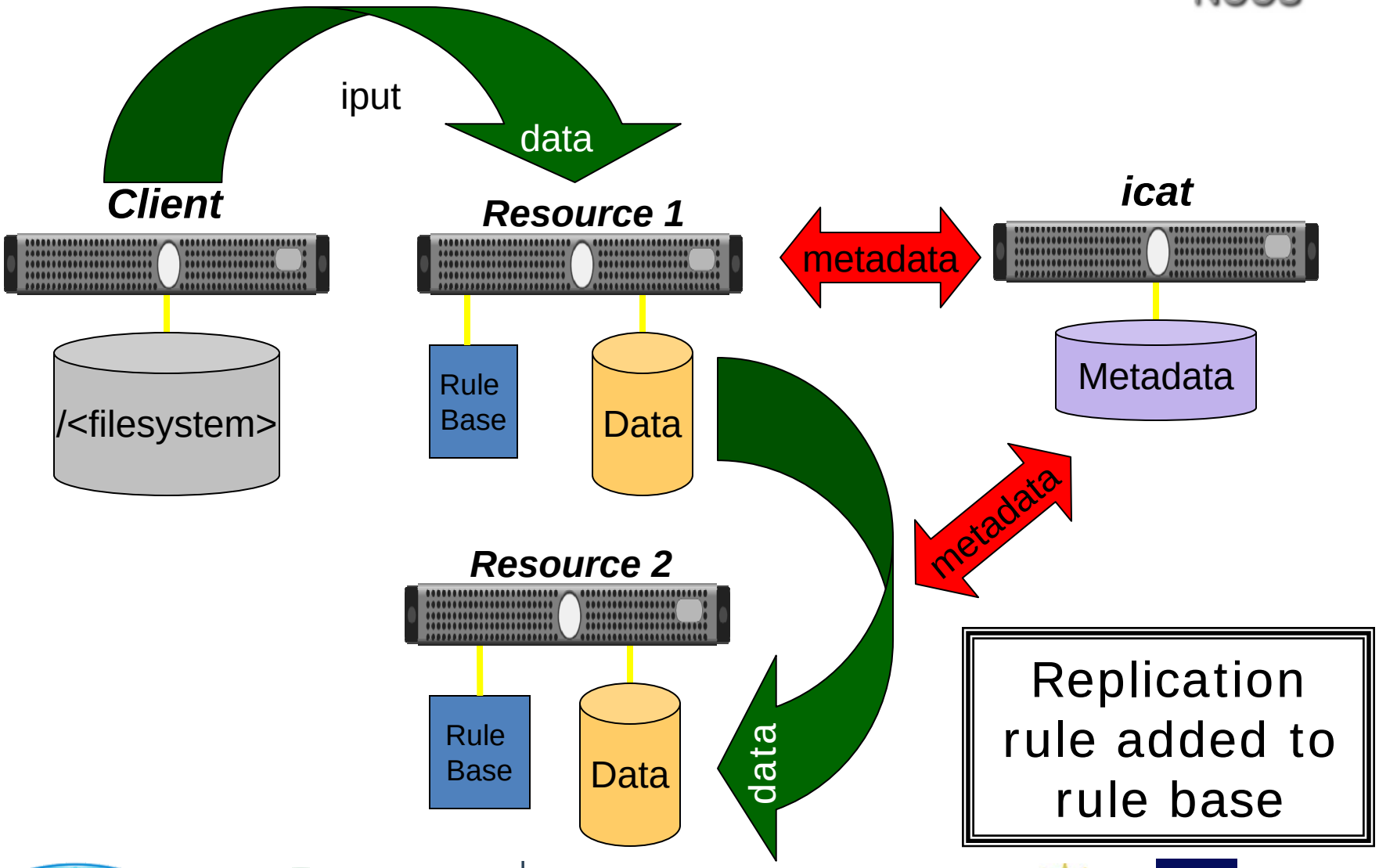
Operations are mapped to the protocol of the storage system

Policy-based Record Management

- Express policies as computer actionable rules
  - Define explicit locations in data management framework where policies will be enforced
- Express procedures as remotely executable micro-services
  - Create new preservation services by linking micro-services into a workflow
- Manage state information to track the application of preservation procedures
  - Maintain audit trails to track evolution of policies and procedures


THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL


iput With Replication

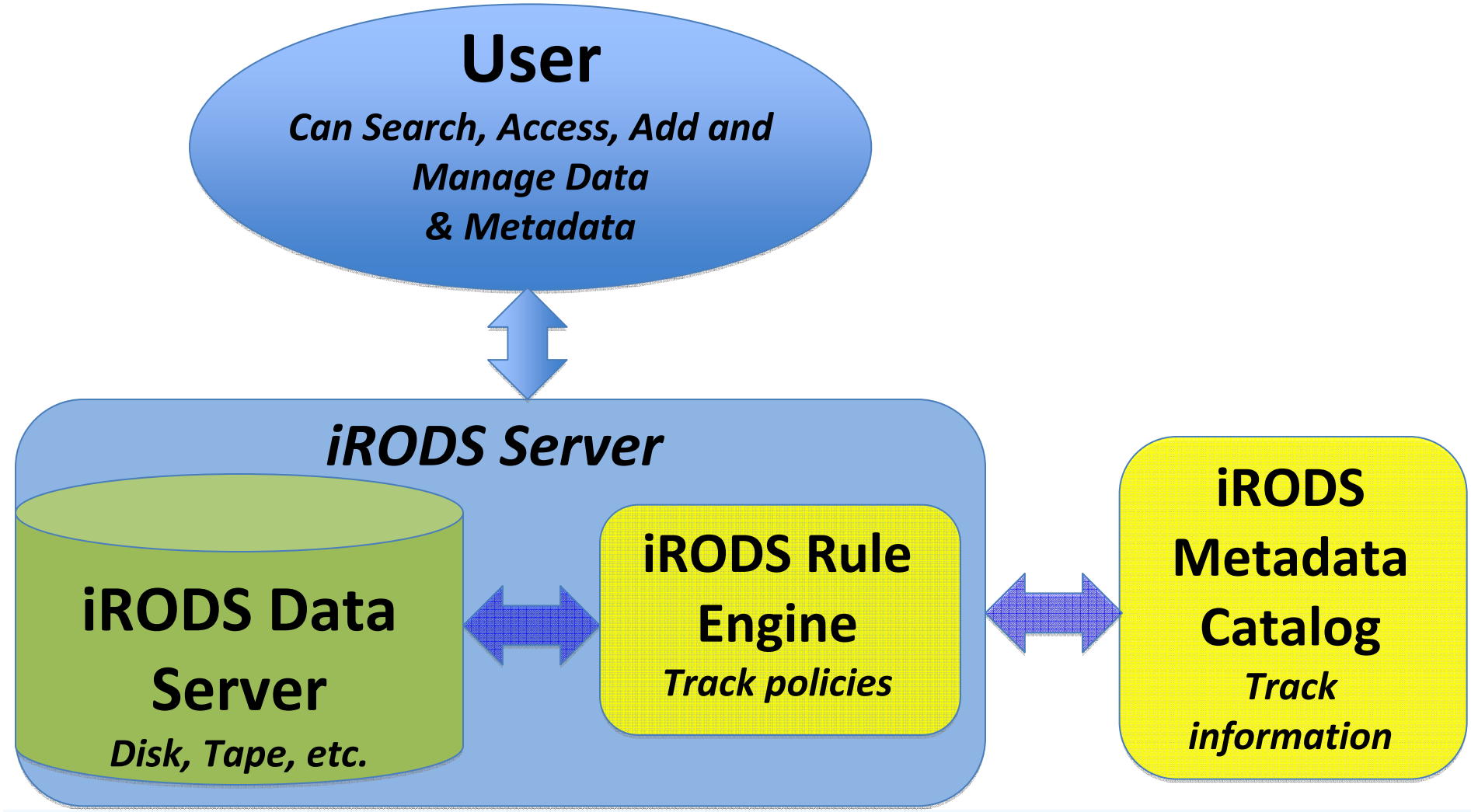


Policies

- Automation of preservation procedures
  - Transformative migration
  - Descriptive metadata extraction
  - Creation of archival form (AIP)
- Automation of administrative functions
  - Distribution, replication, retention, disposition
  - Report generation - usage, quotas, error tracking
- Periodic validation of assessment criteria
  - Trustworthiness, integrity, authenticity, chain of custody
  - Parsing of audit trails for compliance over time


THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL


Overview of iRODS Data System



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

Evolution of Policy Framework

- Compose policies by combining a set of rules
  - Store rules in a rule base
- Initially instrumented the locations where policy management was required by users of the Storage Resource Broker
  - About 30 locations
    - Create file, delete file
    - Create user, delete user,
    - Set number of parallel I/O streams,
    - Define vault path name
    - Select resource
- iRODS version 2.1
  - Provide pre-processing and post-processing hooks


THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL
UCSD

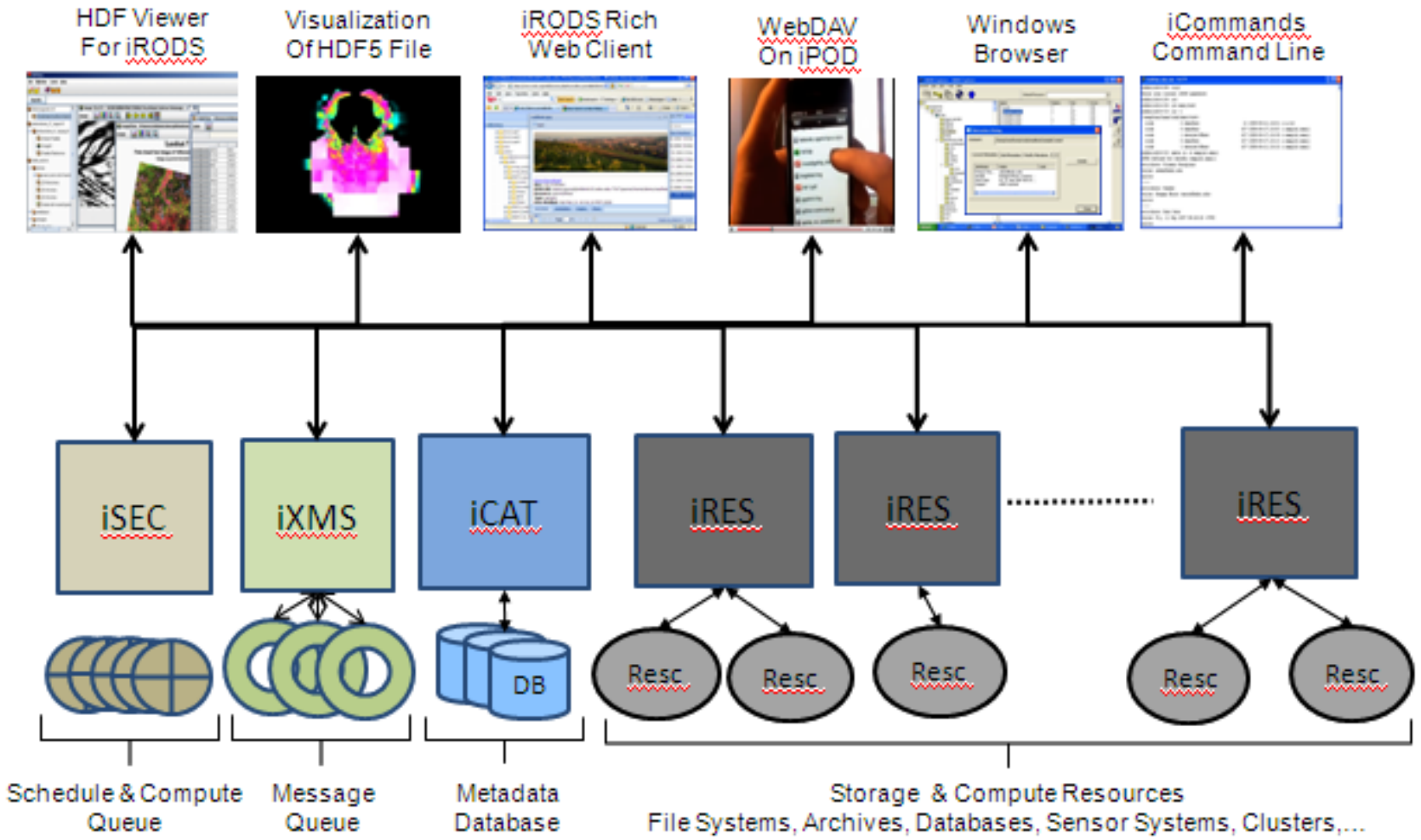

Management Framework

- Instrument data management infrastructure at all locations where policy should be enforced (65 hooks)
  - File create, open, read, write, delete
  - Collection create, delete
  - User create, modify, delete, group
  - Resource create, modify, delete, group
  - Metadata file modify, collection modify, descriptive
  - ACL modify
- Support pre-processing policy
  - Authorization, selection, redirection
- Support post-processing policy
  - Audit trails, redaction, derived product generation


THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL



iRODS Distributed Data Management



D·I·C·E i·R·O·D·S THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL UCSD NSF THE NATIONAL ARCHIVES ARCHIVES.GOV

Types of Rules

- Synchronous rules applied by framework at management hook locations
  - Stored in rule base; core.irb file
- Asynchronous rule that are queued for deferred or periodic execution
  - Batch system to manage queue
- Interactively executed rules defined by a user
  - Executed through irule command

iRODS Rules

- Server-side workflows

  Action | condition | workflow chain | recovery chain
- Condition - test on any attribute:
  - Collection, file name, storage system, file type, user group, elapsed time, IRB approval flag, descriptive metadata
- Workflow chain:
  - Micro-services / rules that are executed at the storage system
- Recovery chain:
  - Micro-services / rules that are used to recover from errors


THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL


Checksum Validation Rule

```
myChecksumRule{
    msiMakeQuery("DATA_NAME, COLL_NAME, DATA_CHECKSUM",*Condition,*Query);
    msiExecStrCondQuery(*Query,*B);
    assign(*A,0);
    forEachExec (*B)  {
            msiGetValByKey(*B,COLL_NAME,*C);
            msiGetValByKey(*B,DATA_NAME,*D);
            msiGetValByKey(*B,DATA_CHECKSUM,*E);
            msiDataObjChksum(*B,*Operation,*F);
            ifExec  (*E != *F)  {
                writeLine(stdout,file *C/*D has registered checksum *E and computed checksum *F);
            }
            else {
                assign(*A,*A + 1);
            }
    }
    ifExec(*A > 0)  {
          writeLine(stdout, have *A good files);
    }
}
```

*Condition can be COLL_NAME like '/ils161/home/moore/genealogy/%'


THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL


Highly Extensible

- Given a collection of records that are being managed with a set of policies, rules, and state information
- Define a new set of policies, new rules, and new state information
- Write a rule that automates the migration of a record collection from the old policies and procedures to the new policies and procedures

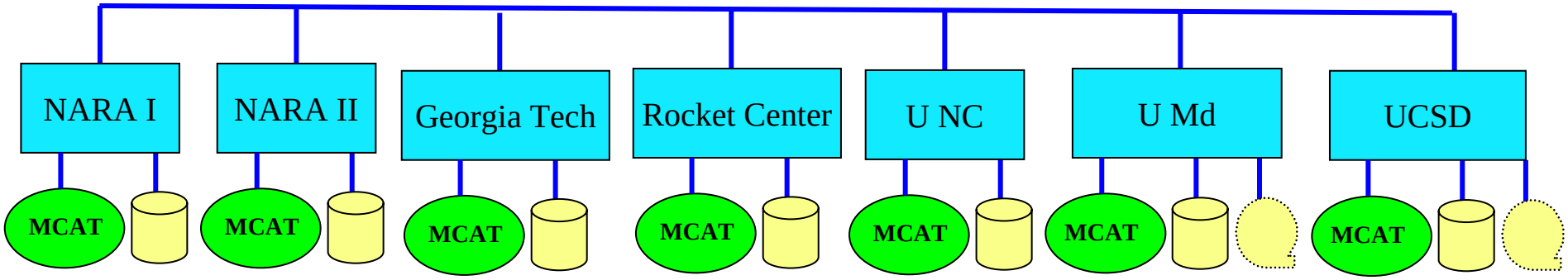
NARA Transcontinental Persistent Archive Prototype

- Use data grid technology to build a preservation environment
- Conduct research on preservation concepts
  - Infrastructure independence
  - Enforcement of preservation properties
  - Automation of administrative preservation processes
  - Validation of preservation assessment criteria
- Demonstrate preservation on selected NARA digital holdings
  - Integration of generic infrastructure with preservation technologies (Cheshire, MVD, JHOVE, Pronom, Fedora, Dspace)


THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL
UCSD
NSF
THE NATIONAL ARCHIVES ARCHIVES.GOV

National Archives and Records Administration
Transcontinental Persistent Archive Prototype

Federation of Seven
Independent Data Grids



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

D·I·C·E i·R·O·D·S THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL UCSD NSF THE NATIONAL ARCHIVES ARCHIVES.GOV

Collaborations

- IN2P3 (Lyon, France)
  - Quota system and monitoring system
- Australian Research Collaboration Service
  - External identity management, Web-DAV client
- Stanford Linear Accelerator
  - Port of metadata catalog to mySQL
- SHAMAN (University of Liverpool)
  - Integration with Cheshire and Multivalent parser


THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL
UCSD


iRODS Evaluations

- NASA Jet Propulsion Laboratory
  - iRODS selected for managing distribution of Planetary Data System records
- NASA National Center for Computational Sciences
  - iRODS chosen to manage archive of simulation output and serve as access data cache for distribution
- AVETEC appraisal for DoD HPC centers
  - iRODS now provides all required capabilities (added Kerberos authentication support)
- French National Library
  - iRODS rules to control ingestion, access, and audit functions
- Australian Research Collaboration Service
  - iRODS manages data distributed between academic institutions


THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL


iRODS is a "coordinated NSF/OCI-Nat'l Archives research activity" under the auspices of the President's NITRD Program

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"


THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL