Policy Based Data management

Integrated Rule Oriented Data System – iRODS

Reagan W. Moore (DICE-UNC)

Arcot Rajasekar (DICE-UNC)

http://irods.diceresearch.org

D·I·C·E

i·R·O·D·S

THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

UCSD

NSF

THE
NATIONAL
ARCHIVES
ARCHIVES.GOV

renci

iRODS

- □ **Integrated Rule Oriented Data System**
  - DICE group – Reagan Moore
  - Concepts – Arcot Rajasekar
  - Architect – Mike Wan
  - Security / metadata / production – Wayne Schroeder
  - Rule engine – Hao Xu
  - User interface (Java) – Mike Conway
  - Applications – Antoine de Torcy
  - Administration – Sheau-Yen Chen

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

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 executable workflows**

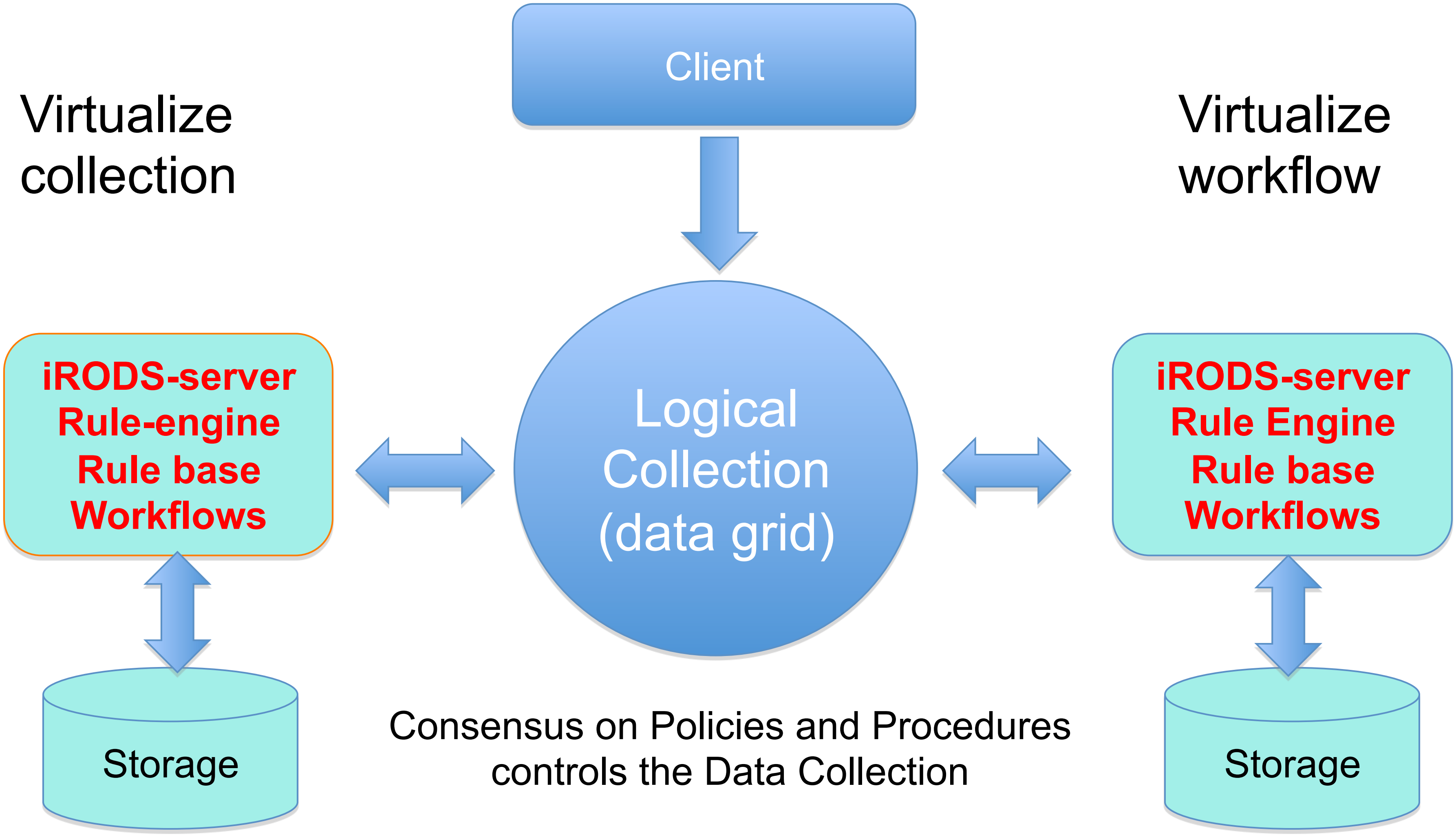
□ ***Persistent state information***

- Results of applying the ***procedures***
- **mapped to system metadata**

□ ***Property verification***

- Validation that ***state information*** conforms to the desired ***purpose***
- **mapped to periodically executed policies**

Policy-based Data Management



Virtualize collection

Client

Virtualize workflow

iRODS-server
Rule-engine
Rule base
Workflows

Logical Collection (data grid)

iRODS-server
Rule Engine
Rule base
Workflows

Storage

Storage

Consensus on Policies and Procedures controls the Data Collection

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

Building Community Resources

- ❑ **Digital libraries use collections to define context**
  - Provenance information
  - Descriptive information
  - Administrative information
- ❑ **Policy-based data management use procedures to encapsulate domain knowledge**
  - Workflows for generation of data
  - Workflows for administration of data
  - Workflows for enforcement of management policies
  - Workflows for verifying collection properties

Computer Actionable Knowledge

❑ Data	objects	bits
❑ Information	names	metadata
❑ Knowledge	relationships between names	procedures
❑ Wisdom	relationships between relationships	policy points

❑ Data	bits	Posix I/O
❑ Information	metadata	Relational database
❑ Knowledge	procedures	Workflows
❑ Wisdom	policy points	Rule engine

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

Sharing Domain Knowledge

- ❑ **Reproducible science**
  - Register workflows
  - Automate provenance management
- ❑ **Collaboration environments**
  - Share data
  - Share workflows
- ❑ **Reference collections**
  - Build community resources of shared data and workflows

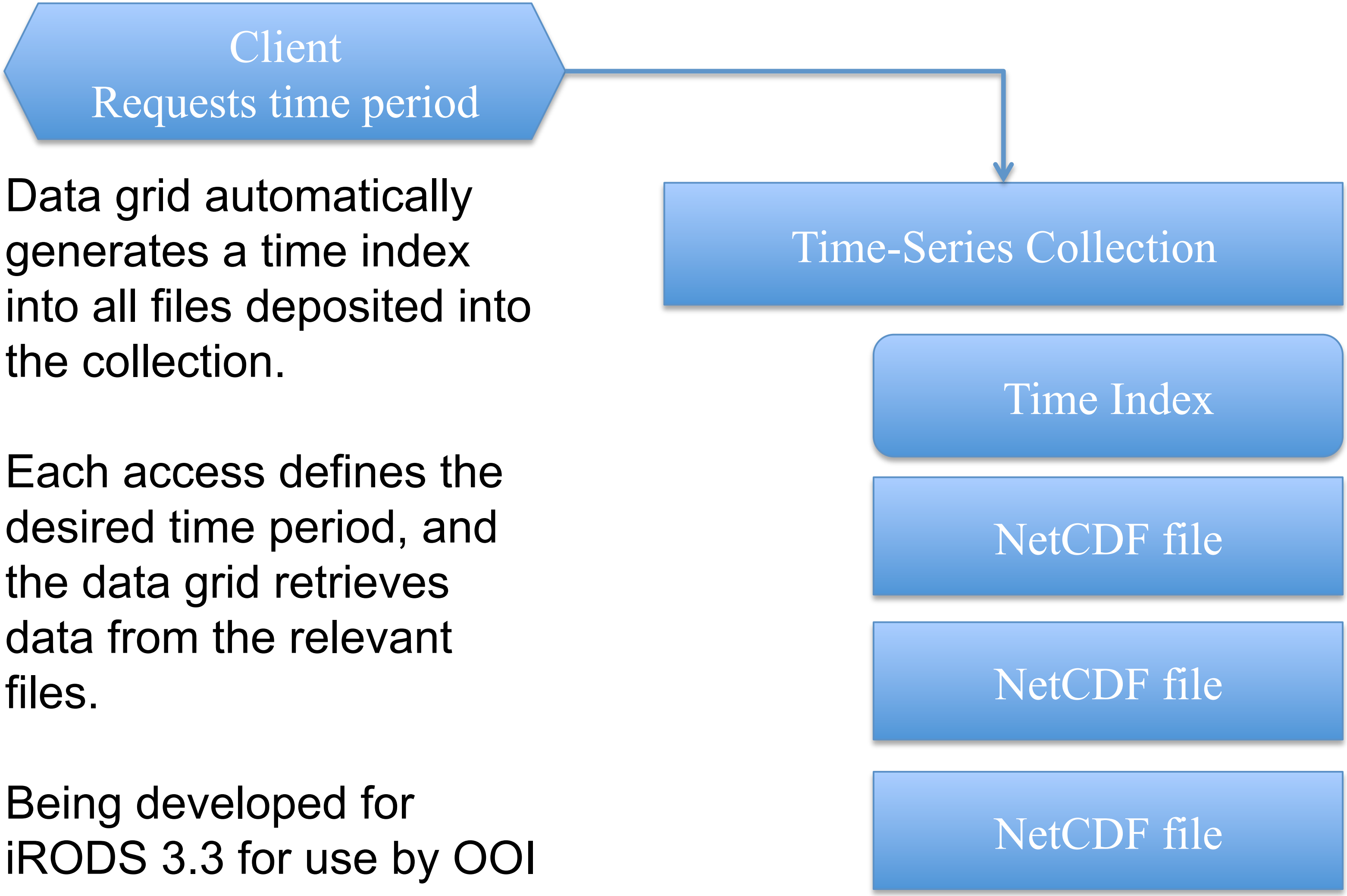
New Development

❑ **Active objects**
- Soft link - Micro-service structured object
  - Registers a remote object into the shared collection
  - Clicking on the object invokes the required protocol for retrieving the object
  - Can cache a local copy
- Can create soft links to
  - Web sites
  - FTP sites
  - Z39.50
  - SRB data grid
  - iRODS data grid

New Development

- ❑ **Active Collections**
  - Mounted collection
    - Can register a remote directory into the collection
    - Can then view contents, list files, retrieve files
  - Tar collection
    - Can view contents of a tar file
  - Time-series collection
    - Can request data stream for arbitrary time interval
  - Workflow collection
    - Can automate capture of workflow provenance

Automating Time Series Data Access



Client
Requests time period

Time-Series Collection

Time Index

NetCDF file

NetCDF file

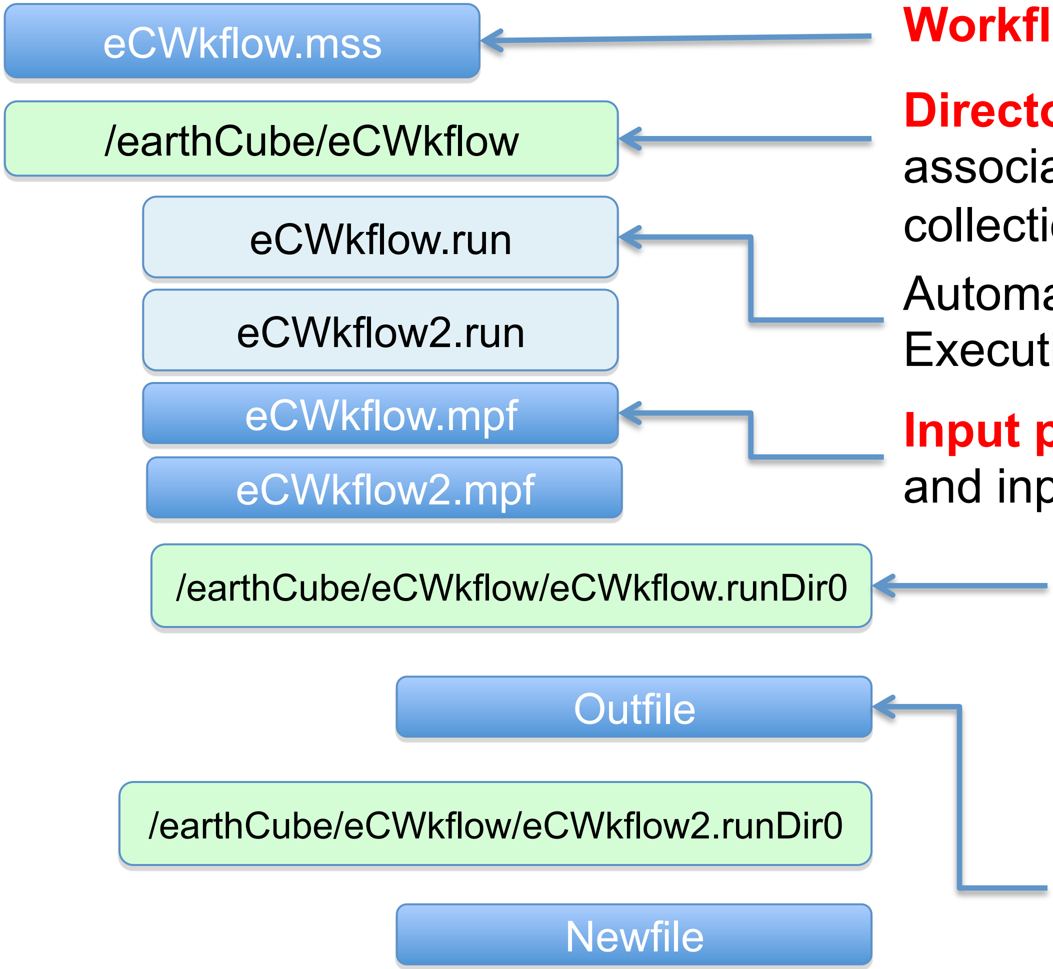
NetCDF file

Data grid automatically generates a time index into all files deposited into the collection.

Each access defines the desired time period, and the data grid retrieves data from the relevant files.

Being developed for iRODS 3.3 for use by OOI

Capturing Workflow Provenance



eCWkflow.mss

Workflow file

/earthCube/eCWkflow

Directory holding all input and output files associated with workflow file (mounted collection that is linked to the workflow file)

eCWkflow.run

eCWkflow2.run

Automatically generated run file for Executing each input file

eCWkflow.mpf

eCWkflow2.mpf

Input parameter file, lists parameters and input and output file names

/earthCube/eCWkflow/eCWkflow.runDir0

Directory holding all output files generated from invocation of eCWkflow.run, the version number is incremented for each execution

Outfile

Output files created for eCWKflow.mpf

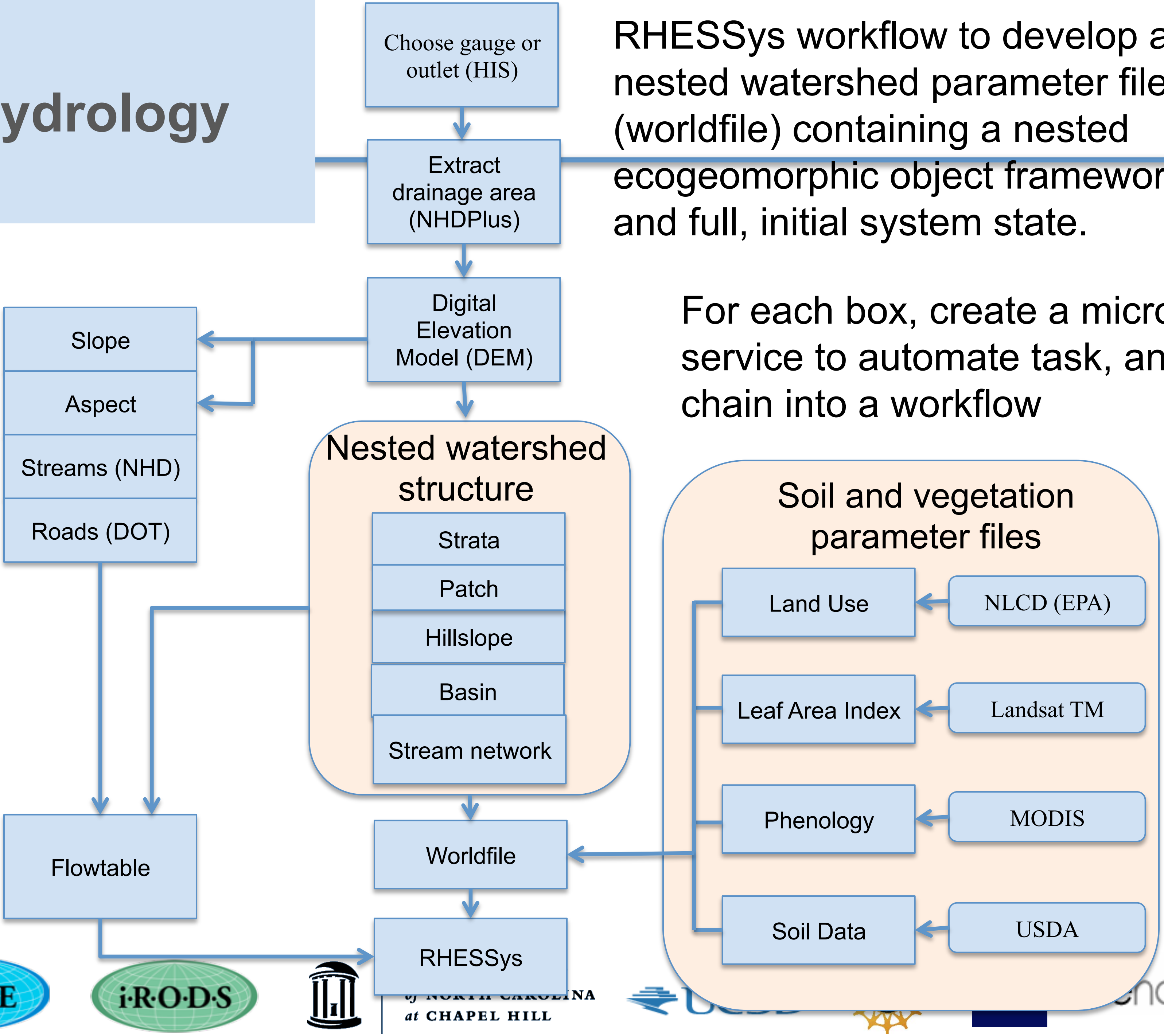
/earthCube/eCWkflow/eCWkflow2.runDir0

Newfile

Eco-Hydrology

RHESSys workflow to develop a nested watershed parameter file (worldfile) containing a nested ecogeomorphic object framework, and full, initial system state.

For each box, create a micro-service to automate task, and chain into a workflow



iRODS - Open Source Software

- http://irods.diceresearch.org
  - Distributed under BSD license
- **Current version is iRODS 3.2**
  - Typically have three releases per year
- **Scale of capabilities:**
  - 338 system attributes (users, files, collections, resources, rules)
  - 272 basic functions (micro-services)
  - 80 policy enforcement points
  - 22 basic storage operations (POSIX I/O plus staging)
  - 10 storage system drivers
  - More than 50 clients
- **Downloads**
  - 39 countries
  - 62 US academic institutions

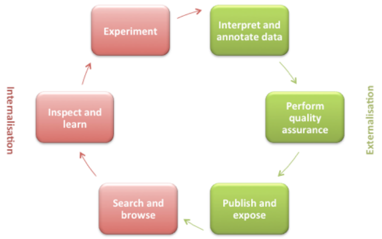
Examples of “National” Infrastructure

□ Data Grids (data sharing)

- National Optical Astronomy Observatory
- Ocean Observatories Initiative
- The iPlant Collaborative
- Babar High Energy Physics
- Broad Institute genomics data grid
- WellCome Trust Sanger Institute genomics data grid

□ Digital Libraries (data publication)

- French National Library
- Texas Digital Library
- UNC-CH SILS LifeTime Library



□ Repositories / Archives (data preservation)

- NASA Center for Climate Simulation
- Carolina Digital Repository

D·I·C·E

i·R·O·D·S

THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

UCSD

NSF

THE
NATIONAL
ARCHIVES
ARCHIVES.GOV

renci

Publications

❑ Rajasekar, R., M. Wan, R. Moore, W. Schroeder, S.-Y. Chen, L. Gilbert, C.-Y. Hou, C. Lee, R. Marciano, P. Tooby, A. de Torcy, B. Zhu, “iRODS Primer: Integrated Rule-Oriented Data System”, Morgan & Claypool, 2010.

❑ Ward, R., M. Wan, W. Schroeder, A. Rajasekar, A. de Torcy, T. Russell, H. Xu, R. Moore, “The integrated Rule-Oriented Data System (iRODS 3.0) Micro-service Workbook”, DICE Foundation, November 2011, ISBN: 9781466469129, Amazon.com

iRODS - Open Source Software

Reagan W. Moore

rwmoore@renci.org

http://irods.diceresearch.org

NSF OCI-0940841 “DataNet Federation Consortium”
NSF OCI-1032732 “Improvement of iRODS for Multi-Disciplinary Applications”
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 renci