Data Management using iRODS

Fundamentals of Data Management

September 2014

Albert Heyrovsky
Applications Developer, EPCC
a.heyrovsky@epcc.ed.ac.uk

Course outline

- Why talk about iRODS?
- What is iRODS?
- The main features of iRODS
- What can iRODS do?
- Who uses iRODS?

- After completing this lesson, you should:
  - Have an overview of iRODS
  - Know what iRODS can be used for

Why talk about iRODS?

- It is a data management system widely used by many organizations worldwide (including EPCC)
- It is open source software
- It is being actively developed and supported
- It is free

What is iRODS?

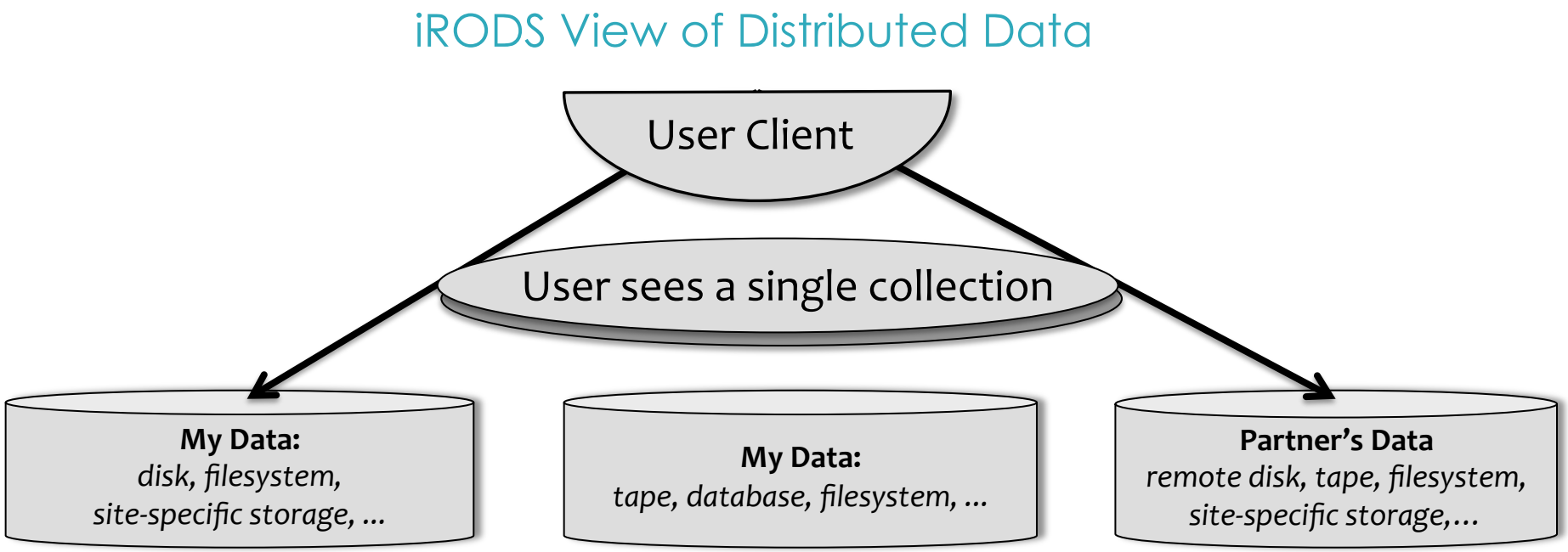
- iRODS stands for **I**ntegrated **R**ule-**O**riented **D**ata **S**ystem
- It is an open-source data grid middleware
- As per Wikipedia (http://en.wikipedia.org/wiki/Data_grid): "A data grid is an architecture or set of services that gives individuals or groups of users the ability to access, modify and transfer extremely large amounts of geographically distributed data for research purposes."
- It is developed and supported by the iRODS Consortium

The main features of iRODS

- Supports large numbers of users (1000s) and user groups in a single data grid
- Supports heterogeneous data storage resources, e.g.:
  - Unix File Systems
  - Amazon S3 buckets
  - DataDirect Networks (DDN) Web Object Scaler (WOS) appliances
  - High Performance Storage System (HPSS) data stores
  - And other storage resources, more are being developed
- Files stored in these heterogeneous storage resources are exposed to users in a single unified namespace

The main features of iRODS

- iRODS Unified Virtual Collection



- iRODS installs over heterogeneous data resources
- Access and manage distributed data as a single collection

The main features of iRODS

- Handles big data (petabytes)
- A high-performance network data transfer protocol
  - Parallel I/O for large files
  - Comparable to GridFTP
- A metadata catalogue named iCAT
  - Stores system metadata and user-defined metadata
  - Manages access control
  - Manages mappings between logical and physical name spaces
  - And some other services
- Easy backup and replication to multiple storage devices and locations

The main features of iRODS

- Security - Authentication
  - iRODS usernames / passwords
  - Supports Pluggable Authentication Modules (PAM)
    - can use an LDAP authentication server
  - Grid Security Infrastructure (GSI)
    - provides authentication using X.509 digital certificates
  - Kerberos
  - Shibboleth

The main features of iRODS

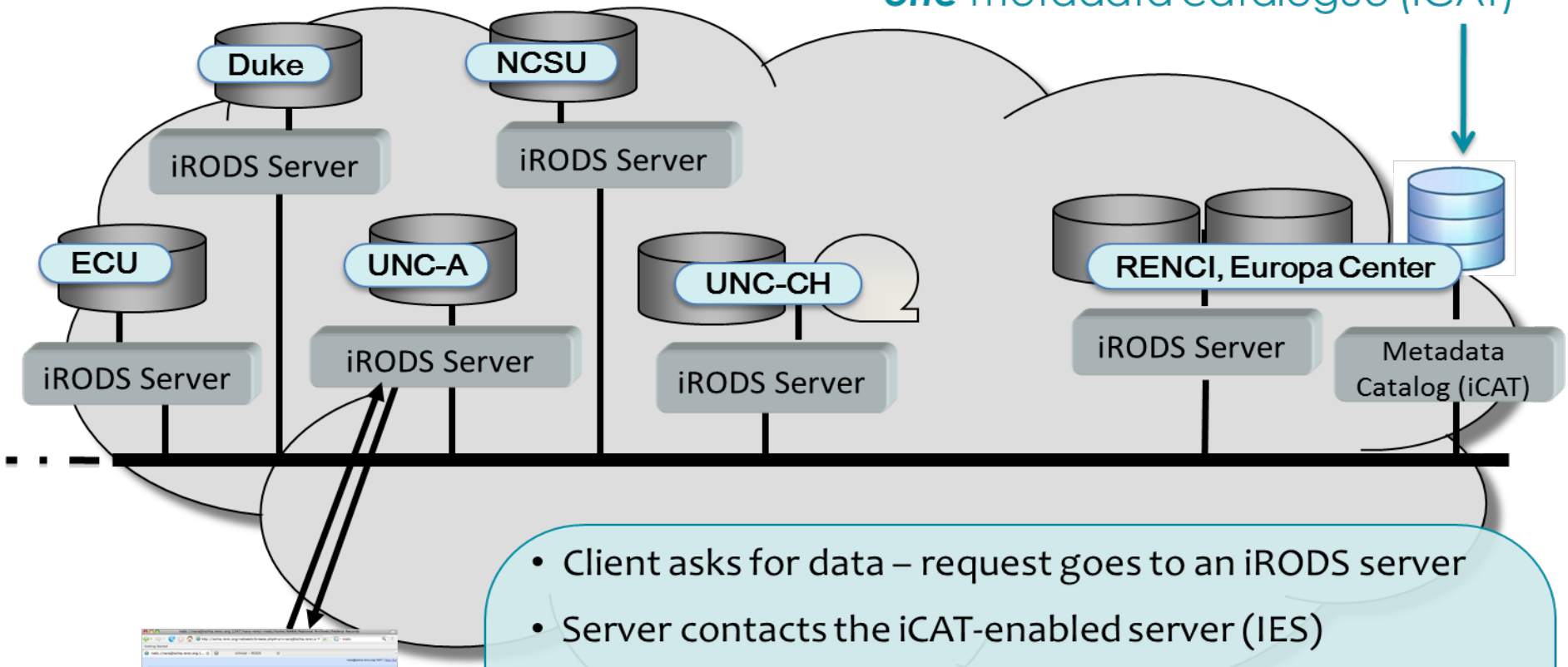
- A Rule Engine
  - Enables automation of data operations, e.g.
    - Validating file checksums, backing up files, archiving unused data, logging data operations, file access permissions, etc.
  - Implements / enforces data management policies, e.g.
    - Records retention and privacy protection policies
    - Audit trails to verify compliance with policies
  - Enables rule-based workflows
- Data grid federation
  - Independent data grids can be federated with one another to allow controlled access to remote grids operated by separate workgroups

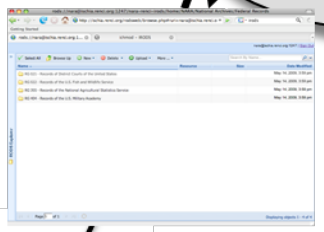
The main features of iRODS

- iRODS Client Applications and APIs
  - More than 50
  - Command line clients (e.g. iRODS i-Commands)
  - Web clients (e.g. iDrop Web)
  - iDrop Desktop – a desktop GUI client
  - PyRods – a Python client API to iRODS
  - Jargon – a Java client API to iRODS
  - Prods – a PHP client API to iRODS
  - Custom clients

A RENCI Data Grid

A complete data grid (**zone**) has ***one*** metadata catalogue (iCAT)





- Client asks for data – request goes to an iRODS server
- Server contacts the iCAT-enabled server (IES)
- Information (location, access rights, etc) is retrieved from the iCAT
- Server holding data is signaled to send data to authorized client

What can iRODS do?

- For Data Centre Managers it simplifies data grid management
- For Users it simplifies data discovery, data validation and data processing
- Data Preservation – Digital Archives
- Data Maintenance
- Data Sharing and Access
- Policy Enforcement
- Data Protection and Security
- Data Curation – Digital Libraries
- Automated Data Processing
- Distributed Data Management

Who uses iRODS?

Science and Engineering Domains, e.g.

Astrophysics	Auger supernova search
Atmospheric science	NASA Langley Atmospheric Sciences Center
Biology	Phylogenetics at CC IN2P3
Climate	NOAA National Climatic Data Center
Cognitive Science	Temporal Dynamics of Learning Center
Computer Science	GENI experimental network
Cosmic Ray	AMS experiment on the International Space Station
Dark Matter Physics	Edelweiss II
Earth Science	NASA Center for Climate Simulations
Ecology	CEED Caveat Emptor Ecological Data
Engineering	CIBER-U
High Energy Physics	BaBar / Stanford Linear Accelerator
Hydrology	Institute for the Environment, UNC-CH; Hydroshare
Genomics	Broad Institute, Wellcome Trust Sanger Institute, NGS
Medicine	Sick Kids Hospital
Neuroscience	International Neuroinformatics Coordinating Facility
Neutrino Physics	T2K and dChooz neutrino experiments
Oceanography	Ocean Observatories Initiative
Optical Astronomy	National Optical Astronomy Observatory
Particle Physics	Indra multi-detector collaboration at IN2P3
Plant genetics	the iPlant Collaborative

Who uses iRODS?

Science and Engineering Domains, e.g.

Quantum Chromodynamics	IN2P3
Radio Astronomy	Cyber Square Kilometer Array, TREND, BAOradio
Seismology	Southern California Earthquake Center
Social Science	Odum, TerraPop

Arts and Humanities Domains, e.g.

Digital Library	French National Library, Texas Digital Libraries
Indexing	Cheshire
Institutional repository	Carolina Digital Repository
Preservation	Adonis
Reference collections	SILS LifeTime Library

Commercial Users, e.g.

DOW Chemical

Beijing Genome Institute

and many others, e.g.

the cross-domain European Data Infrastructure (EUDAT) consortium

Summary

- iRODS is a data grid management system
- It is scalable
  - It can manage millions of files and millions of metadata annotations totalling petabytes of data
  - It can support thousands of users
- It is widely used by many organizations
- There are other data grid management systems with similar features, e.g.
  - DSpace
  - Fedora Commons

Acknowledgements

- Thanks to the iRODS Consortium for providing materials for this lecture.