Rule-Oriented Data Systems

A Grid-Based Cyberinfrastructure for Large Scale & Long-Term Data Management

Arcot (Raja) Rajasekar, Frank Vernon & Reagan Moore, UCSD

What is iRODS?

Data grid system—data virtualization
- A distributed file system, based on a client-server architecture
- Allows users to access files seamlessly across a distributed environment, based on their attributes rather than just their names or physical locations
- It replicates, syncs and archives data, connecting heterogeneous resources in a logical and abstracted manner

Distributed workflow system – policy virtualization
- Policy can be coded as functions (micro-services)
- Remote micro-services can be chained
- The chains (workflow) are interpreted at run-time
- The chains can be triggered on an event and condition (rules)
- They can also be recursive
- Micro-services communicate through parameters, shared contexts, and out-of-band message queues

Data Virtualization with iRODS

Logical name space
Location independent identifier; Persistent identifier
Collection owned data
Access controls; Audit trails;
Checksums; Descriptive metadata
Inter-realm authentication
Single sign-on system

Policy Virtualization with iRODS

MICRO-SERVICES: Functions with well-defined semantics; Transactional - recovery; Context of application; Message Queues
RULES: Triggered by events; Conditional execution of alternative rule declarations; System constructs: loops, recursion, branching
WORKFLOWS: Distributed Execution; Immediate, Deferred, Periodic

Automation of Management Tasks

Integrated Rule-Oriented Data System — iRODS
Express management policies as rules that control the execution of micro-services
- Micro-service is a standard operation performed on a remote storage system

Manage persistent state information that describes outcome of the micro-service
- Persistent Metadata catalog stores state information

Virtualize the management policies
- Logical name space for rules
- Logical name space for micro-services
- Logical name space for state information

iRODS System

Applications include
- Data grids for sharing data
- Digital libraries for publishing data
- Persistent archives for preserving data
- Real-time sensor data collections
- Large scale data analysis

Management Virtualization

Examples of management policies

Integrity
- Validation of checksums
- Synchronization of replicas
- Data distribution
- Data retention
- Access controls

Authenticity
- Chain of custody - audit trails
- Track required preservation metadata - templates
- Generation of Archival Information Packages

Rule-based Data Management

Associate rules with combinations of name spaces
- Rule set for a particular collection
- Rule set for a particular user group
- Rule set for a particular user group when accessing a particular collection
- Rule set for a particular storage system
- Rule set for a particular micro-service
- Generic rules based on SRB operations

Sample Rules

```
IngestInCollection(S) :- /* just store */  ingests(S), register(S).
IngestInCollection(S) :- /* store, check, backup & extract metadata */
  ingests(S), register(S),
  chkCond4(S),
  computeSerChkSum(S,C1),
  computeClntChkSum(S,C2),
  checkAndRegisterChkSum(C1,C2,S).
IngestInCollection(S) :- /* store, check, backup & extract metadata */
  ingests(S), register(S),
  chkCond4(S),
  computeSerChkSum(S,C1),
  computeClntChkSum(S,C2),
  checkAndRegisterChkSum(C1,C2,S).
```

Potential iRODS Applications

Virtual Organizations
Federation of Observatories
- ORION, Zen
- Shared Collections
  - NVO, NOAO, LSST – Astronomy, ROADNet, SEEK; UK eScience; Babar – INP3, KfK
- Persistent Archives
  - NASA – National Archives, SHAMAN