iRODS — A Large-Scale Rule-Oriented Data Management System

Wayne Schroeder
Data Intensive Computing Environments,
San Diego Supercomputer Center,
University of California San Diego

schroede@sdsc.edu
http://diceresearch.org
http://www.irods.org
Topics

" Who We Are

" Our Software
  " Storage Resource Broker (SRB)
  " Integrated Rule Oriented Data management System (iRODS)

" How we use DBMS

" Informal Comparison of PostgreSQL and Oracle
"Team of about a dozen
  " Dr Reagan Moore, Dr Arcot Rajasekar, Dr Richard Marciano
  " Michael Wan, Wayne Schroeder, other software engineers
  " Software Engineering is Key; Must be Useful and Work Well

"Data Intensive Computing Environments (DICE)
  " 1997 DARPA
  " Series of awards NARA, NSF
  " National and International Uses
  " Customer Driven

"San Diego Supercomputer Center
  " NSF Funded, Series of initiatives
  " National Resource
  " Started 1985 under General Atomics at UCSD
  " 2000 as part of University of California San Diego
  " High Performance Computing
My Own Background

" Software Developer (BS CS 1976)
" SDSC at Start, 1985
   " Enthused to Support Science, etc
   " LLNL (Fusion Energy Center, NMFECC) before SDSC
" Entropia (startup) 2000-2002
" DICE 2002
   " SRB Installation/Testing, Java GUI Admin, etc
   " iRODS Co-Developer
      " Michael Wan, Arcot Rajasekar (Raja), myself
   " Catalog (DBMS) Interface (ICAT)
   " Administration
   " Installation/Testing
   " Authentication (password, GSI)
   " Etc
SRB Projects (Old Slide)

" Astronomy
  " National Virtual Observatory
" Data Grids
  " UK e-Science CCLRC
  " Teragrid
" Digital Libraries and Archives
  " National Archives and Records Administration
  " National Science Digital Library
  " Persistent Archive Testbed
" Ecological, Environmental, Oceanographic
  " ROADnet
  " Southern California Earthquake Center
  " SIO Digital Libraries
" Molecular Sciences
  " Synchrotron Data Repository
  " Alliance for Cellular Signaling
" Neuro Sciences
  " Biomedical Information Research Network
" Physics and Chemistry
  " BaBar
" Many others

Over 650 Tera Bytes in 106 million files
## Sampling of Funded Projects

<table>
<thead>
<tr>
<th>Project</th>
<th>Start Year - End Year</th>
<th>Funding Agency(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Massive Data Analysis System (MDAS)</td>
<td>1995-1997</td>
<td>DARPA</td>
</tr>
<tr>
<td>Distributed Object Computation Testbed</td>
<td>1996-1999</td>
<td>DOD, USPTO</td>
</tr>
<tr>
<td>National Partnership for Advanced Computational Infrastructure</td>
<td>1997-2004</td>
<td>NSF</td>
</tr>
<tr>
<td>Information Power Grid</td>
<td>1998-2004</td>
<td>NASA</td>
</tr>
<tr>
<td>Data Visualization Corridor</td>
<td>1998-2001</td>
<td>DOE, ASCI</td>
</tr>
<tr>
<td>Persistent Archive Research</td>
<td>1999-</td>
<td>NARA</td>
</tr>
<tr>
<td>(20 + more, see SRB Web site)</td>
<td>2000 -</td>
<td>Various</td>
</tr>
</tbody>
</table>
Extremely Successful

" Storage Resource Broker (SRB) manages 2 PBs of data in internationally shared collections

" Data collections for NSF, NARA, NASA, DOE, DOD, NIH, LC, NHPRC, IMLS: APAC, UK e-Science, IN2P3, WUNgrid

<table>
<thead>
<tr>
<th>Domain</th>
<th>Infrastructure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Astronomy</td>
<td>Data grid</td>
</tr>
<tr>
<td>Bio-informatics</td>
<td>Digital library</td>
</tr>
<tr>
<td>Earth Sciences</td>
<td>Data grid</td>
</tr>
<tr>
<td>Ecology</td>
<td>Collection</td>
</tr>
<tr>
<td>Education</td>
<td>Persistent archive</td>
</tr>
<tr>
<td>Engineering</td>
<td>Digital library</td>
</tr>
<tr>
<td>Environmental science</td>
<td>Data grid</td>
</tr>
<tr>
<td>High energy physics</td>
<td>Data grid</td>
</tr>
<tr>
<td>Humanities</td>
<td>Data Grid</td>
</tr>
<tr>
<td>Medical community</td>
<td>Digital library</td>
</tr>
<tr>
<td>Oceanography</td>
<td>Real time sensor data, persistent archive</td>
</tr>
<tr>
<td>Seismology</td>
<td>Digital library, real-time sensor data</td>
</tr>
</tbody>
</table>

" Goal has been generic infrastructure for distributed data
<table>
<thead>
<tr>
<th>Project</th>
<th>Date</th>
<th>5/17/02</th>
<th>GBs of data stored</th>
<th>1000's of files</th>
<th>GBs of data stored</th>
<th>1000's of files</th>
<th>Users with ACLs</th>
<th>GBs of data stored</th>
<th>1000's of files</th>
<th>Users with ACLs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Grid</td>
<td>6/30/04</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NSF / NVO</td>
<td></td>
<td></td>
<td>17,800</td>
<td>5,139</td>
<td>51,380</td>
<td>8,690</td>
<td>80</td>
<td>88,216</td>
<td>14,550</td>
<td>100</td>
</tr>
<tr>
<td>NSF / NPACI</td>
<td></td>
<td></td>
<td>1,972</td>
<td>1,083</td>
<td>17,578</td>
<td>4,694</td>
<td>380</td>
<td>39,697</td>
<td>7,590</td>
<td>380</td>
</tr>
<tr>
<td>Hayden</td>
<td></td>
<td></td>
<td>6,800</td>
<td>41</td>
<td>7,201</td>
<td>113</td>
<td>178</td>
<td>8,013</td>
<td>161</td>
<td>227</td>
</tr>
<tr>
<td>Pzone</td>
<td></td>
<td></td>
<td>438</td>
<td>31</td>
<td>812</td>
<td>47</td>
<td>49</td>
<td>28,799</td>
<td>17,640</td>
<td>68</td>
</tr>
<tr>
<td>NSF / LDAS-SALK</td>
<td></td>
<td></td>
<td>239</td>
<td>1</td>
<td>4,562</td>
<td>16</td>
<td>66</td>
<td>207,018</td>
<td>169</td>
<td>67</td>
</tr>
<tr>
<td>NSF / SLAC-JCSG</td>
<td></td>
<td></td>
<td>514</td>
<td>77</td>
<td>4,317</td>
<td>563</td>
<td>47</td>
<td>23,854</td>
<td>2,493</td>
<td>55</td>
</tr>
<tr>
<td>NSF / TeraGrid</td>
<td></td>
<td></td>
<td>80,354</td>
<td>685</td>
<td>2,962</td>
<td>2,962</td>
<td>148</td>
<td>282,536</td>
<td>7,257</td>
<td>3,267</td>
</tr>
<tr>
<td>NIH / BIRN</td>
<td></td>
<td></td>
<td>5,416</td>
<td>3,366</td>
<td>148</td>
<td>148</td>
<td></td>
<td>20,400</td>
<td>40,747</td>
<td>445</td>
</tr>
<tr>
<td>NCAR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LCA</td>
<td></td>
<td></td>
<td>3,787</td>
<td>77</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Digital Library</td>
<td>11/29/07</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NSF / LTER</td>
<td></td>
<td>158</td>
<td>3</td>
<td></td>
<td>233</td>
<td>6</td>
<td>35</td>
<td>260</td>
<td>42</td>
<td>36</td>
</tr>
<tr>
<td>NSF / Portal</td>
<td></td>
<td>33</td>
<td>5</td>
<td></td>
<td>1,745</td>
<td>48</td>
<td>384</td>
<td>2,620</td>
<td>53</td>
<td>460</td>
</tr>
<tr>
<td>NIH / AfCS</td>
<td></td>
<td>27</td>
<td>4</td>
<td></td>
<td>462</td>
<td>49</td>
<td>21</td>
<td>733</td>
<td>94</td>
<td>21</td>
</tr>
<tr>
<td>NSF / SIO Explorer</td>
<td></td>
<td>19</td>
<td>1</td>
<td></td>
<td>1,734</td>
<td>601</td>
<td>27</td>
<td>2,750</td>
<td>1,202</td>
<td>27</td>
</tr>
<tr>
<td>NSF / SSEC</td>
<td></td>
<td>15,246</td>
<td>1,737</td>
<td>52</td>
<td>168,931</td>
<td>3,545</td>
<td>73</td>
<td>18,934</td>
<td>2,338</td>
<td>5</td>
</tr>
<tr>
<td>LLNL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CHRON</td>
<td></td>
<td></td>
<td>12,863</td>
<td>6,443</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Persistent Archive</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NARA</td>
<td></td>
<td>7</td>
<td>2</td>
<td></td>
<td>63</td>
<td>81</td>
<td>58</td>
<td>5,023</td>
<td>6,430</td>
<td>58</td>
</tr>
<tr>
<td>NSF / NSDL</td>
<td></td>
<td></td>
<td>2,785</td>
<td>20,054</td>
<td>119</td>
<td>84,984</td>
<td>136</td>
<td>7,499</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UCSD Libraries</td>
<td></td>
<td>127</td>
<td>202</td>
<td></td>
<td>29</td>
<td></td>
<td></td>
<td>5,205</td>
<td>1,328</td>
<td>29</td>
</tr>
<tr>
<td>NHPRC / PAT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2,576</td>
<td>966</td>
</tr>
<tr>
<td>RoadNet</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3,557</td>
<td>1,569</td>
</tr>
<tr>
<td>UCTV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7,140</td>
<td>2</td>
</tr>
<tr>
<td>LOC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6,644</td>
<td>192</td>
</tr>
<tr>
<td>Earth Sci</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6,136</td>
<td>652</td>
</tr>
<tr>
<td>TOTAL</td>
<td></td>
<td>28 TB</td>
<td>6 mil</td>
<td></td>
<td>194 TB</td>
<td>40 mil</td>
<td>4,635</td>
<td>1,023 TB</td>
<td>200 mil</td>
<td>5,539</td>
</tr>
</tbody>
</table>
iRODS Tutorials - 2008

January 31, SDSC
April 8 - ISGC, Taipei
May 13 - China, National Academy of Science
May 27-30 - UK eScience, Edinburgh
June 5 - OGF23, Barcelona
July 7-11 - SAA, SDSC
August 4-8 - SAA, SDSC
August 25 - SAA, San Francisco
iRODS Development

" NSF - SDCI grant Adaptive Middleware for Community Shared Collections
  " iRODS development, SRB maintenance

" NARA - Transcontinental Persistent Archive Prototype
  " Trusted repository assessment criteria

" NSF - Ocean Research Interactive Observatory Network (ORION)
  " Real-time sensor data stream management

" NSF - Temporal Dynamics of Learning Center data grid
  " Management of IRB approval
iRODS Development

2005: Planning, Some Initial Development
2006, December: iRODS .5 Released
2007, June: iRODS .9 Released
2008, January: iRODS 1.0 Released
Soon: iRODS 1.1
iRODS/SRB Flavors

" Data grids
  " Share data - organize distributed data as a collection

" Digital libraries
  " Publish data - support browsing and discovery

" Persistent archives
  " Preserve data - manage technology evolution

" Real-time sensor systems
  " Federate sensor data - integrate across sensor streams

" Workflow systems
  " Analyze data - integrate client- & server-side workflows
Using a Data Grid

User asks for data from the data grid
- The data is found and returned
- Where & how details are hidden
Using a Data Grid - Details

- User asks for data
- Data request goes to iRODS Server
- Server looks up information in DB catalog
- Catalog tells which iRODS server has data
- 1st server asks 2nd for data
- The 2nd iRODS server applies rules
Data Grid State Information

State Information in DBMS

" Files (DataObjects)
" Directories (Collections)
" Users
" Resources, etc

For Each File DBMS information includes:

" Location: Host and Directory
" Other System Metadata
" User-defined Metadata
" Replica, etc
Data Grid Capabilities

" Logical file name space
  " Directory hierarchy / soft links
  " Versions / backups / replicas
  " Aggregation / containers
  " Descriptive metadata
  " Digital entities

" Physically Distributed on Network

" Authentication and authorization
  " GSI, challenge-response, Shibboleth
  " ACLs, audit trails
  " Checksums, synchronization
  " Logical user name space
  " Aggregation / groups
Generic Infrastructure

"Data grids manage data distributed across multiple types of storage systems"
  "File systems, tape archives, object ring buffers"

"Data grids manage collection attributes"
  "Provenance, descriptive, system metadata"

"Data grids manage technology evolution"
  "At the point in time when new technology is available, both the old and new systems can be integrated"
Tension between Common and Unique Components

" Synergism - common infrastructure
  " Distributed data
    " Sources, users, performance, reliability, analysis
  " Technology management
    " Incorporate new technology

" Unique components - extensibility
  " Information management
    " Semantics, formats, services
  " Management policies
    " Integrity, authenticity, availability, authorization
Storage Resource Broker
A Data Grid Solution

" Collaborative client-server system that federates distributed heterogeneous resources using uniform interfaces and metadata

" Provides a simple tool to integrate data and metadata handling attribute-based access

" Blends browsing and searching

" Developed at SDSC
  - Operational for 11+ years;
  - Under continual development since 1997;
IRODS - the Next Generation of Data Grid Technology
"Rule-based
  " Rules Engine at core
  " Our own implementation (Raja)

"Rules invoke microservices and/or rules

"Complete rewrite, but based on experience with SRB

"Client/Server, Server-Server

"Open Source (BSD) (SRB is available to edu and gov sites)
Data Grids

" SRB - Storage Resource Broker
  " Persistent naming of distributed data
  " Management of data stored in multiple types of storage systems
  " Organization of data as a shared collection with descriptive metadata, access controls, audit trails

" iRODS - integrated Rule-Oriented Data System
  " Rules control execution of remote micro-services
  " Manage persistent state information
  " Validate assertions about collection
  " Automate execution of management policies
Currently seven clients

- iRODS rich web client
- Unix shell commands
  - iRODS/clients/icommands/bin
- FUSE user level file system
  - iRODS/clients/fuse/bin/irodsFs fmount
- Jargon Java I/O class library
  - iRODS/java/jargon
- PHP web browser and PHP client library
  - http://irods.sdsc.edu
- C library calls
- Parrot user level file system
  - Douglas Thain, Notre Dame University
iCommands

~/irods/clients/icommands/bin

" icd  " iget  " iqdel
" ichmod  " iput  " iqmod
" icp  " ireg  " iqstat
" ils  " irepl  " irexecmd
" imkdir  " irsync  " irule
" imv  " itrim  " iuserinfo
" ipwd  " ilsresc  " isysmeta
" irm  " irsync  " imeta
" ienv  " irmtrash  " iquest
" ierror  " irmtrash  " imiscsvrinfo
" ichksum  " irmtrash  " iadmin
" iinit  " irepl  " iadmin
" iexit  " irsync  " iadmin
iroadsssetup: Installation

" Linux, Mac, Mac/Intel, Solaris, AIX, 32/64 bit
" Prompt User
" Download, Configure, Build, Install, Run
  " PostgreSQL
  " ODBC (Unix or PostgreSQL)
" Configure, Build, Install, Run iRODS
" Install ICAT Database
" Bring Up System
" Basic Tests, Optional Advanced Tests
Testing

" iCommand test suite from IN2P3, France
  " Thomas Kachelhoffer, Jean-Yves Nief
" ICAT test suite all 204 SQL Forms
" Layers of Scripts
  " Tinderbox
  " installation (rewritten by Dave Nadeau)
  " irodsctl test the above two test suites
" NMI Build & Test Facility, U of Wisc
iRODS Development Status

" Production release is version 1.0
  " January 24, 2008
" Version 1.1 Soon
" International collaborations
  " SHAMAN - University of Liverpool
    " Sustaining Heritage Access through Multivalent ArchiviNg
  " UK e-Science data grid
  " IN2P3 in Lyon, France
  " DSpace policy management
iRODS Data Grid Capabilities

" Logical Name Space
" Logical Storage Space
  " Dynamic resource creation
  " Standard operations
  " Heterogeneous storage systems
  " Trash
  " Collective operations / storage groups

" Data transport
  " Parallel I/O
  " Small file transport
  " Message engine
  " Containers / tar files / HDF5
  " Aggregation of I/O commands - remote procedures
iRODS Data Grid Capabilities

" Remote procedures
  " Atomic / deferred / periodic
  " Procedure execution / chaining
  " Structured information

" Structured information
  " Metadata catalog interactions / 204 SQL forms
  " Information transmission
  " Template parsing
  " Memory structures
  " Report generation / audit trail parsing
"SRB CATALOG (MCAT)

" Oracle, DB2, Sybase, PostgreSQL, Informix, or MySQL4 (primarily Oracle and PostgreSQL)

" Binary Large Objects

" DB2, Oracle, Illustra

" Oracle in Production

" SDSC and Elsewhere

" PostgreSQL for Testing/Demos
iRODS DBMS

" Catalog (ICAT)
  " PostgreSQL or Oracle (primarily PostgreSQL)
  " MySQL Planned

" PostgreSQL In Production (soon)
" PostgreSQL for Test/Demo
iRODS ICAT

" Interface to RDBMS
  iRODS State Information
" Simplified Schema (Raja)
" Bind Variables for Performance/Security
" Three levels:
  API - High Level calls (~45)
  Mid-level/Helpers
  PostgreSQL/ODBC or Oracle/OCI
" Called by
  " MicroServices/Rules, Server Code, Client/Server calls
" GeneralQuery, GeneralAdmin, SimpleQuery
" iadmin interface for Administration
Ingestion Rate as Collection Size Increases

- X-axis: Cumulative Number of Files in iRODS - in 1000s
- Y-axis: Time in Seconds - per 1000 file ingestion

The graph shows a trend where the ingestion rate increases as the collection size grows, with the time taken to ingest files increasing as the number of files in iRODS increases.
PostgreSQL Advantages

" Freely Downloaded/Installed for:
  " Testing, SRB/iRODS
  " Integrated Installation
    " SRB Demos/Tutorials
    " SRB in a Box (Shipboard Environmental Science)
    " iRODS Demos/Tutorials/Production Use

" Faster
  " i-cmd/ICAT test suite >2x Oracle
  " Same Host, Small DB

" Open Source
" psql vs sqlplus
iRODS WebSite-Wiki

http://irods.sdsc.edu

Descriptions of the technology

Publications / presentations

Download

Performance tests

Tinderbox system (continual build/test)

ters-chat page
Planned Development

- GSI support (1)
- Time-limited sessions via a one-way hash authentication
- Python Client library
- GUI Browser (AJAX in development)
- Driver for HPSS (in development)
- Driver for SAM-QFS
- Porting to additional versions of Unix/Linux
- Porting to Windows
- Support for MySQL as the metadata catalog
- API support packages based on existing mounted collection driver
- MCAT to ICAT migration tools (2)
- Extensible Metadata including Databases Access Interface (6)
- Zones/Federation (4)
- Auditing - mechanisms to record and track iRODS metadata changes
For More Information

Wayne Schroeder
San Diego Supercomputer Center
schroede@sdsc.edu

http://diceresearch.org
http://www.irods.org