



Managing Petabytes of data with iRODS at CC-IN2P3

Jean-Yves Nief



Talk overview



- What is CC-IN2P3 ?
- Who is using iRODS ?
- iRODS administration:
 - Hardware setup.
- iRODS interaction with other services:
 - Mass Storage System, backup system, Fedora Commons etc...
 - iRODS clients usage.
- Architecture examples with collaborating sites.
- Rules examples.
- Prospects.

CC-IN2P3 activities



- Federate computing needs of the french scientific community in:
 - Nuclear and particle physics.
 - Astrophysics and astroparticles.
- Computing services to international collaborations:
 - CERN (LHC), Fermilab, SLAC,
- Opened now to biology, Arts & Humanities.

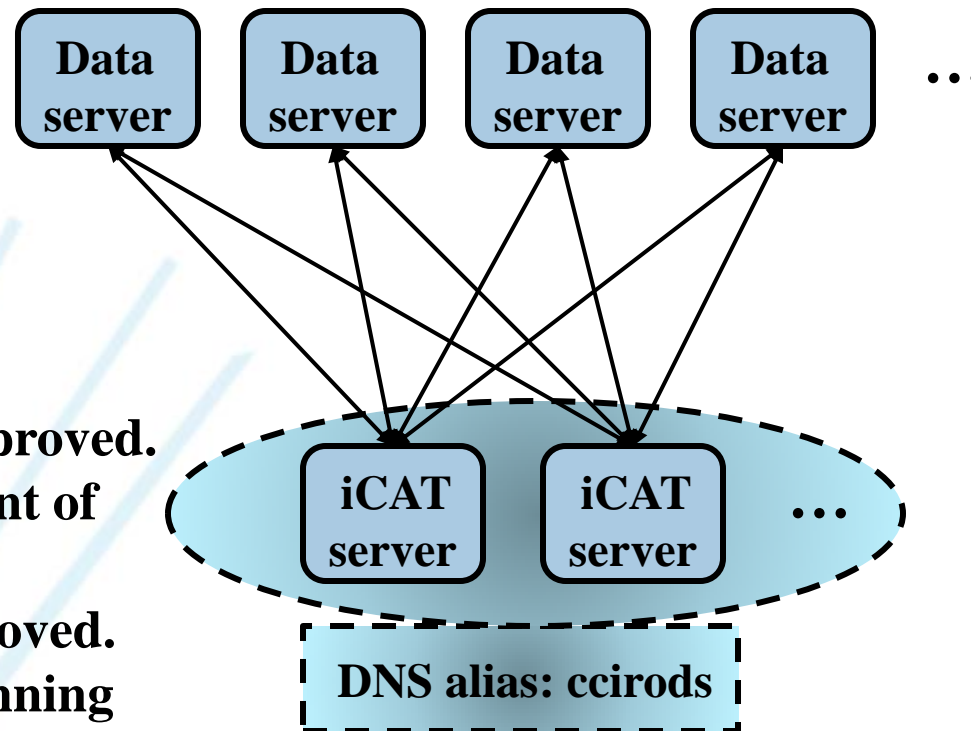


iRODS setup @ CC-IN2P3



- Being used since the beginning in 2006.
- In production since early 2008.
- 15 servers:
 - 2 iCAT servers (metacatalog): Linux SL4, Linux SL5
 - 13 data servers (737 TB): Sun Thor x454 with Solaris 10, DELL R510, R720xd with Linux SL5.
- Metacatalog on a dedicated Oracle 11g cluster (2 servers).
- Monitoring and restart of the services fully automated (*crontab* + *Nagios*).
- Automatic weekly reindexing of the iCAT databases.
- Accounting: daily report on our web site.

iRODS setup @ CC-IN2P3



DNS alias:

- load balanced.
- redundancy improved.
- avoid single point of failure.
- scalability improved.
- all instances running on the same iCAT servers.

iRODS monitoring: Nagios



Nagios®

General

- Home
- Documentation

Current Status

- Tactical Overview
- Map
- Hosts
- Services
- Host Groups
 - Summary
 - Grid
- Service Groups
 - Summary
 - Grid
- Problems
 - Services (Unhandled)
 - Hosts (Unhandled)
 - Network Outages

Reports

- Availability
- Trends
- Alerts
 - History
 - Summary
 - Histogram
- Notifications
- Event Log

System

- Comments
- Downtime
- Process Info

Quick Search:

Current Network Status

Last Updated: Tue Feb 28 11:09:09 CET 2012
Updated every 90 seconds
Nagios® Core™ 3.2.3 - www.nagios.org
Logged in as centre

[View Service Status Detail For All Host Groups](#)
[View Host Status Detail For This Host Group](#)
[View Status Overview For This Host Group](#)
[View Status Summary For This Host Group](#)
[View Status Grid For This Host Group](#)

Host Status Totals

Up	Down	Unreachable	Pending
2	0	0	0

All Problems

All Types

0	2
---	---

Service Status Totals

Ok	Warning	Unknown	Critical	Pending
38	0	0	0	0

All Problems

All Types

0	38
---	----

Service Status Details For Host Group 'irods-hostgroup'

Host ↑	Service ↑	Status ↑	Last Check ↑	Duration ↑	Attempt ↑	Status Information
ccirods02	Checking ADONIS port 5589	OK	2012-02-28 10:43:47	20d 18h 25m 22s	1/3	Server connexion test to ccirods02 on port 5589 successful!
	Checking AMS port 5592	OK	2012-02-28 10:48:47	20d 19h 20m 22s	1/3	Server connexion test to ccirods02 on port 5592 successful!
	Checking Auger port 5512	OK	2012-02-28 11:04:31	20d 19h 34m 38s	1/3	Server connexion test to ccirods02 on port 5512 successful!
	Checking BAO port 5545	OK	2012-02-28 11:08:50	19d 20h 0m 19s	1/3	Server connexion test to ccirods02 on port 5545 successful!
	Checking BIOEMERGENCE port 5531	OK	2012-02-28 10:43:50	20d 19h 25m 19s	1/3	Server connexion test to ccirods02 on port 5531 successful!
	Checking BaBar port 5571	OK	2012-02-28 10:43:47	20d 19h 25m 22s	1/3	Server connexion test to ccirods02 on port 5571 successful!
	Checking CODALEMA port 5547	OK	2012-02-28 10:43:47	20d 19h 25m 22s	1/3	Server connexion test to ccirods02 on port 5547 successful!
	Checking DCHOOZ port 5585	OK	2012-02-28 10:43:47	20d 19h 25m 22s	1/3	Server connexion test to ccirods02 on port 5585 successful!
	Checking EROS port 5540	OK	2012-02-28 10:43:38	20d 18h 25m 31s	1/3	Server connexion test to ccirods02 on port 5540 successful!
	Checking Fazio port 5537	OK	2012-02-28 10:44:41	19d 21h 54m 28s	1/3	Server connexion test to ccirods02 on port 5537 successful!
	Checking GENERAL port 5570	OK	2012-02-28 11:04:16	20d 19h 34m 53s	1/3	Server connexion test to ccirods02 on port 5570 successful!
	Checking GRILLE RHONE ALPES port 5588	OK	2012-02-28 11:01:59	20d 19h 37m 10s	1/3	Server connexion test to ccirods02 on port 5588 successful!
	Checking ILC port 5542	OK	2012-02-28 10:49:39	20d 16h 19m 30s	1/3	Server connexion test to ccirods02 on port 5542 successful!

iRODS interaction with other services



- **Mass storage system: HPSS.**
 - Using compound resources.
 - Interfaced using the universal MSS driver (RFIO protocol used).
 - Staging requests ordered by tapes using tape requests scheduling.
- **Backup system: TSM.**
 - Used for projects who do not have the possibility to replicate precious data on other sites.
- **Fedora Commons:**
 - Storage backend based on iRODS using FUSE.
 - Rules to register iRODS files into Fedora.
- **External databases:**
 - Rules using RDA.

iRODS servers migration



- Almost 200 TBs of disk space decommissioned in 2012.
- Moved the data without stopping production.
- Easy for file system resources (even for data movement to remote sites).
- More tricky with group resources (archive resource is MSS).



iRODS clients



- Clients:
 - Access from batch jobs, virtually from anywhere.
- Authentication: password or X509 certificates.
- iCommands: most popular.
 - From any platform: Windows, Mac OSX, Linux (RH, CentOS, Debian...), Solaris 10.
- Java APIs: interaction with iRODS within workflows.
- C APIs: direct access to files (open, read, write) to do « random access ».
- FUSE for legacy web sites and Fedora Commons.
- Windows explorer and iDrop.

Who is using iRODS ?



- **High energy and nuclear physics:**
 - BaBar: data management of the entire data set between SLAC and CC-IN2P3: total foreseen 2PBs.
 - dChooz: neutrino experiment (France, USA, Japan etc...): 600 TBs.
- **Astroparticle and astrophysics:**
 - AMS: cosmic ray experiment on the International Space Station (1 PB).
 - TREND, BAOradio: radioastronomy (170 TBs).
- **Biology and biomedical applications:** phylogenetics, neuroscience, cardiology (50 TBs).
- **Arts and Humanities:** Adonis (74 TBs).



iRODS use cases



- Data sharing and transfers for wide spread communities.
- Online data access from any kind of front-end app (web, home grown clients, batch farm...) allowing data policies to be run on the data underneath.
- Data archival.

**Not intended for massive I/O ops from a batch farm!
(not a parallel file system)**

**➔ Still lots of access from the batch farm
(potentially more than 1k clients).**

Who is using iRODS ?



iRods disk space usage & files per experiment at IN2P3 Computing Center



Area maintained by Thomas Kachelhoffer

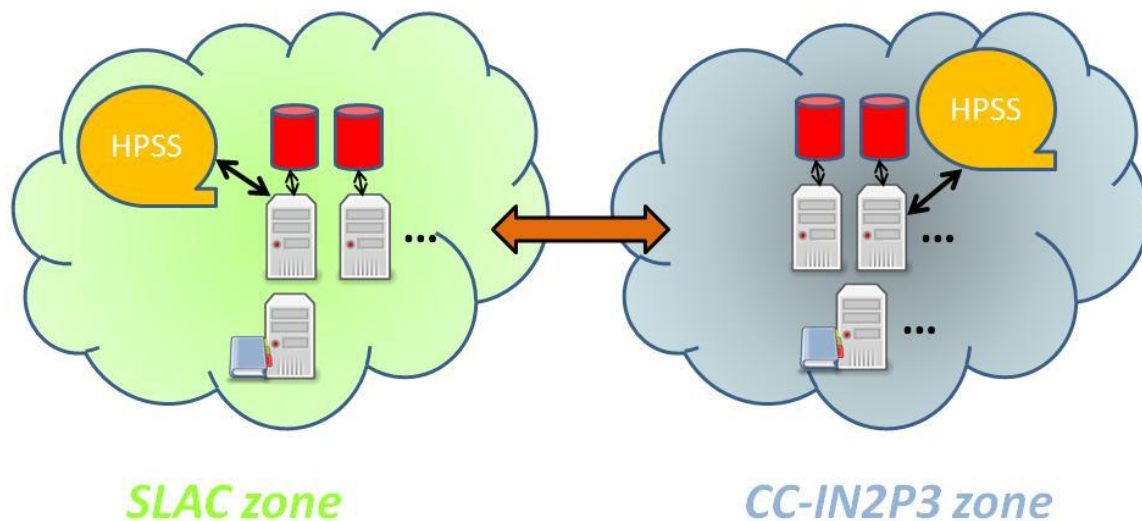
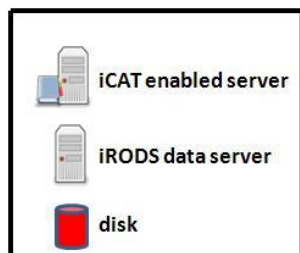
Description:

6 381 TB are used at this time. These values were collected the **2013-02-27** at **22:27:01**. By clicking on the instance name below, you will find the values corresponding to the selected instance and their evolutions.

List of iRods instances:

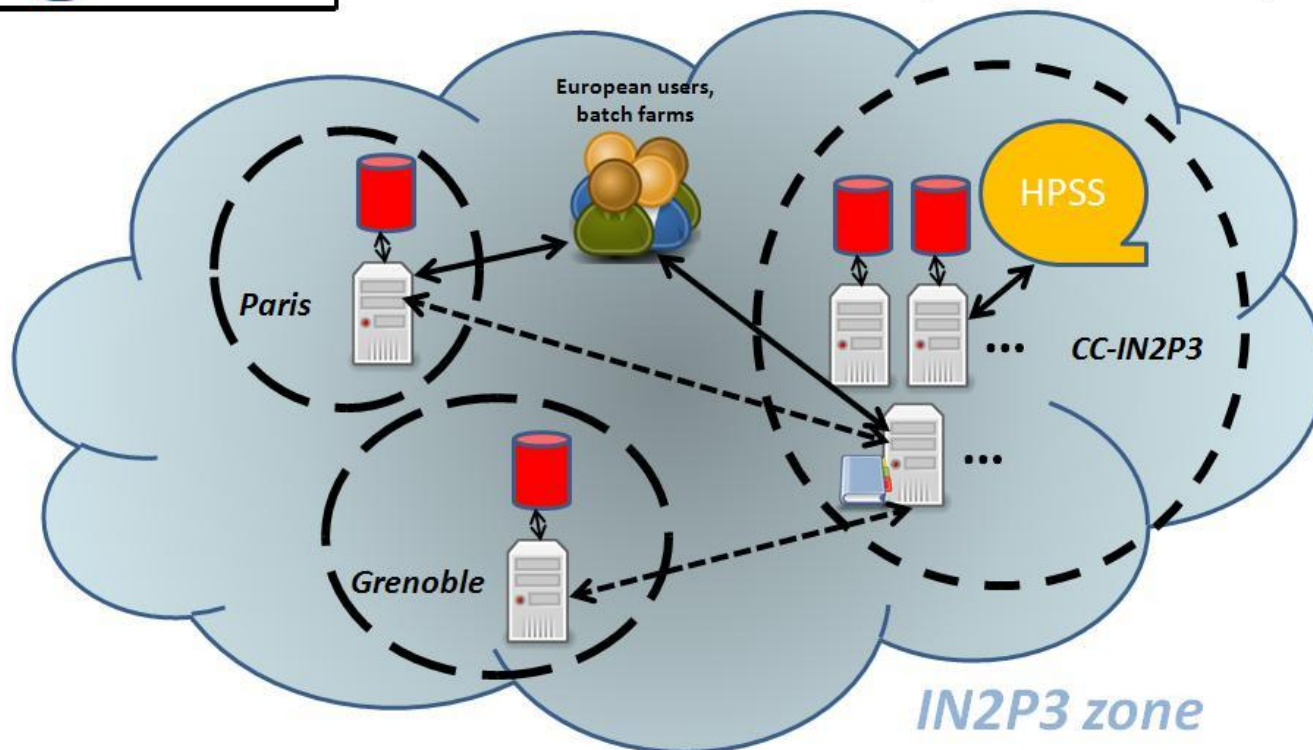
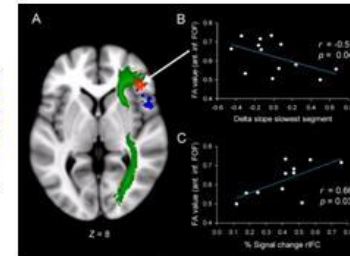
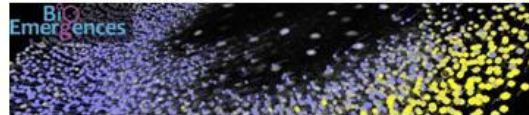
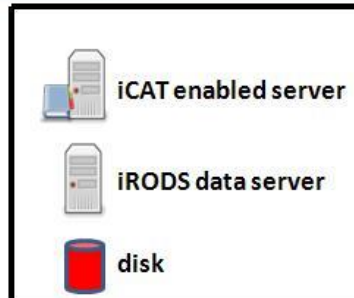
adonis	75 183 GB	6 925 190 files
ams	1 054 779 GB	644 051 files
auger	1 GB	1 195 files
babar	1 550 638 GB	1 592 947 files
bao	148 373 GB	2 341 203 files
bioemergence	13 807 GB	3 837 665 files
codalema	1 953 GB	520 447 files
dchooz	661 673 GB	877 974 files
edelweiss	30 744 GB	9 725 files
fazia	3 352 GB	8 470 files
general	21 959 GB	227 399 files
imxgam	2 141 GB	62 483 files
indra	15 964 GB	72 548 files
ipm	1 179 GB	1 114 832 files
lsst	1 GB	2 files
qcd	2 057 208 GB	812 651 files
test	1 920 GB	302 036 files
tidra	23 535 GB	8 164 832 files
tidra-neuro	9 153 GB	1 170 730 files
trend	46 754 GB	1 316 601 files
virgo	814 076 GB	581 165 files
	6 534 395 GB	30 584 146 files

Architecture example: BaBar



- archival in Lyon of the entire BaBar data set (total of 2 PBs).
- automatic transfer from tape to tape: 3 TBs/day (no limitation).
- automatic recovery of faulty transfers.
- ability for a SLAC admin to recover files directly from the CC-IN2P3 zone if data lost at SLAC.

Architecture example: embryogenesis and neuroscience





Rules examples (I)



- Delayed replication to the MSS:
 - Data on disk cache replication into MSS asynchronously (1h later) using a delayExec rule.
 - Recovery mechanism: retries until success, delay between each retries is doubled at each round.
 - Automatic purge of the cache for the oldest files.
 - Automatic file bundle before migration to MSS.
- ACL management:
 - Rules needed for fine granularity access rights management.
 - Eg:
 - 3 groups of users (**admins**, **experts**, **users**).
 - ACLs on /<zone-name>*/rawdata => **admins** : r/w, **experts** + **users** : r
 - ACLs on all others subcollections => **admins** + **experts** : r/w, **users** : r

Rules examples (II)



- Fedora Commons:
 - Tar balls content stored in iRODS are automatically registered into Fedora Commons.
 - 1. Automatic untar of the files + checksum on the iRODS side:
msiTarFileExtract.
 - 2. Automatic registration in Fedora-commons (delayed rule):
msiExecCmd of a java application.
- Automatic metadata extraction from DICOM files (neuroscience...):
 - A given predefined list of metadata is extracted from the files using DCMTK (thanks to Yonny Cardenas), then user metadata are created for each file.

SRB to iRODS migration



- SRB almost completed, still 2 projects to migrate.
 - ➔ Finish at the end of the first semester of this year!
- Migration to iRODS already made for BioEmergence (embryogenesis) in 2010:
 - Data workflow was using Jargon: transparent.
 - Migration from Scommands to icommands was needed.
 - 2 hours of downtime to complete the migration (scripts were needed).
- Migration headache:
 - SRB is deeply embedded in data management workflows and projects can't live without SRB.
 - ➔ Main issue: migration should be as « transparent » as possible in order to keep up with the data activity.



To-do list



- Connection control (CCMS):
 - Very useful: some servers could be under heavy stress (one iCAT needed to be rebooted a couple of times!).
 - Connections can come from anywhere especially batch farms on the data grid.
 - Servers can be overwhelmed (network, disk activity for hundreds of connection in //).
 - Causes clients to exit with an error → not good.
 - Improved version of CCMS (connection control) is needed.
- Connection pooling on the Oracle side.



Prospects



- 6.3 PBs in iRODS as of Feb 2013 (should be at least **8 PBs** at the end of this year).
- Future projects:
 - LSST (astro): summer data challenge including NCSA (Illinois) + CC-IN2P3: iRODS will be used for data distribution between the two sites (100 TBs).
 - Replication of archival data from an other data centre (Cines, France).
 - Private companies (data encryption needed).



Acknowledgement



- Thanks to:
 - Pascal Calvat.
 - Yonny Cardenas.
 - Rachid Lemrani.
 - Thomas Kachelhoffer.
 - Pierre-Yves Jallud.