CC-IN2P3 data repositories

Jean-Yves Nief

dapnia
cea
saclay
CENTRE NATIONAL
DE LA RECHERCHE
SCIENTIFIQUE

What is CC-IN2P3 ?

- Federate computing needs of the french community:
  – Particle and nuclear physics
  – Astroparticles and astrophysics.
  – Opened to biology, Arts & Humanities etc…

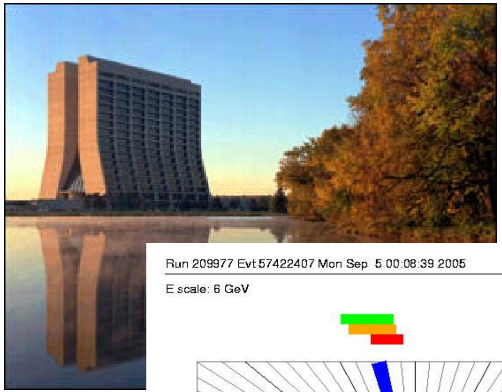
IN2P3

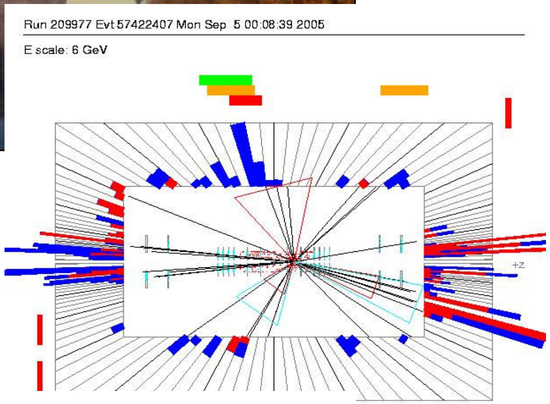
Institut National de Physique Nucléaire et de Physique des Particules

DSM cea dapnia

CCIN2P3

CC-IN2P3 goal





Physics experiments

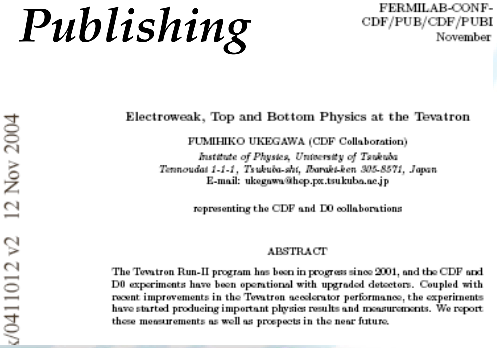
- *Particle & nuclear physics*
- *Astroparticles*

Data

101000 100111 0001001010001
00011101 100010001111 00010
101000 100111 0001001010001
00011101 100010001111 00010
101000 100111 110001 111010
0001001010001101000 100111
0001001010001 00011101 1110
100010001111 00010 101000 00 11
0001100111 0001001010001 00011101
100010001111 00010 101000 100111
0001001010001

Fundamental research

Publishing



FERMILAB-CONF-
CDF/PUB/CDF/PUBI
November

/0411012 v2 12 Nov 2004

Electroweak, Top and Bottom Physics at the Tevatron

FUMIHIKO UKEGAWA (CDF Collaboration)
Institute of Physics, University of Tsukuba
Tennoudai 1-1-1, Tsukuba-shi, Ibaraki-ken 305-8571, Japan
E-mail: ukegawa@hep.px.tsukuba.ac.jp

representing the CDF and D0 collaborations

ABSTRACT

The Tevatron Run-II program has been in progress since 2001, and the CDF and D0 experiments have been operational with upgraded detectors. Coupled with recent improvements in the Tevatron accelerator performance, the experiments have started producing important physics results and measurements. We report these measurements as well as prospects in the near future.

Data analysis

Computing needs at CC-IN2P3

- 5000 users, 80 groups:
  – Users can be also foreign collaborators.
- Access also through the grid (LCG/EGEE).
- Linked to other computing centres around the world.
- Around 10000 cores.
- Two batch farms (PCs):
  – Serial analysis.
  – Parallel analysis (MPI, PVM).
- CPU power doubles every year.

Storage needs at CC-IN2P3

- Multiple data storages:
  - Hardware:
    - Disks (3 PB).
    - Tapes (5 PB, max limit = 30 PB).
  - Software:
    - HPSS (mass storage system): up to 70 TB/day (read/write).
    - Parallel filesystem: GPFS.
    - Global filesystem: AFS.
    - « HEP home made filesystems » (dCache, xrootd).
    - Relational databases (Oracle, mySQL etc..).
    - First step towards virtualization.

Datasets stored @ CC-IN2P3

- Data coming from the experiments/projects:
  - Events from particle colliders.
  - Astrophysics events (supernovae, high energy cosmic rays, etc…).
  - Biology: embryogenesis (e.g: zebra fish).
  - Human related data: brain, heart MRI.
  - Arts & humanities: digital archives.
- Also simulations for these experiments/projects.

Science datasets: data integrity

- Scientific studies based on:
  - statistics like high energy physics.
  - unique events or unreproducible data (astro, biomedical applications).
- Data are more or less precious:
  - Must keep the data safe (backups, replication, data integrity check).
  - Should be available until the end of the experiment and above (migration to new storage media, data format migration).

Data integrity: examples

- Must be able to recover from disaster (broken tape, disks issues, software problem, human errors).
- HEP:
  - BaBar: duplication between SLAC, CC-IN2P3, Padova.
  - LHC: copies of the files on multiple sites.
- Astro, bio:
  - Data replication on the same site (e.g.: double copy on tape) or elsewhere.
  - Use of backup solutions (TSM): backup copy stored in other building.
  - Data integrity check (checksum).

Science datasets: data security

- Most of the groups don't have public data:
  – Users must belong to the group, virtual organization to have data access ➔ authentication (kerberos, user/pwd, certificate).
- Within the group, VO not all the data are available to everybody:
  – Access Control List (ACLs) on the data needed: private data, subgroups within the VO.
- Anonymization: medical records.

Data security: some examples

- LCG/EGEE:
  - Grid certificate.
  - ACLs not widely used (within a VO one can access all the VO data and remove them).
- Other HEP projects, astro, bio:
  - Hierarchy between data producers and others.
  - Can be more complex with groups within the VO.
- Medical records (brain fMRI, heart fMRI):
  - Research data should be anonymized when stored at CC-IN2P3 (non anonymized data outside the hospitals).
  - Must ensure that this policy is achieved.

Science datasets: data discovery

- Use of metadata.
- Can be simple (or too simple):
  - File catalog in a flat file.
- Usually using relational databases (Oracle, mySQL, PostGres) to do it:
  - Metadata organization can be complex and vary a lot from one project wrt an other.
  - Difficult to provide a standard framework: flexibility is needed.
- The relation between logical filename and physical filename must be provided:
  - Sometimes trivial: add a prefix to the logical filename to produce the file URL.
  - Or in a database.

Data storage and access: tools

- On disk:
  - GPFS: working space for high performance data access: not a permanent space.
  - dCache, Xrootd: HEP home-grown protocols for data access to experiment data.
  - Databases: mostly for metadata but some are using it for storing all their data (e.g. 100 TB for Opera in Oracle in the next 2-3 years).
- On tapes:
  - Mass storage System (HPSS): used by a lot of experiment as back end for the storage. Considered to be cheaper (?). Used as an online system with higher latency compared to disk access.
  - Backup system: TSM.

Data access with the outside world

- Can be simple with tools like scp, bbftp or AFS:
  – Provides limited capabilities: not enough.
- LHC data grid:
  – Have their own tools.
  – Heavy machinery, difficult to fit for other needs.
- SRB, iRODS:
  – Not simple data transfer.
  – Real data management tools at a global level (ie federating different data centres).

Data storage preservation @ CC-IN2P3

- What happen to the data after the end of data taking by the experiments ?
- Still kept here as long as needed, ie as long as collaborators are working on them (e.g: LEP experiments stop in 2000, still analysis in 2003), then discard them.
- What about astroparticle data ?
  - Keep them as long as we exist.
  - Still not official policy but tend to go into this direction.
- What about data format migration:
  - Still up to the experiments.
  - With Arts & Humanities, more and more involvement on this.

Data access: virtualization

- Scientific collaborations spread world-wide:
  - Data can also be spread among different sites.
- Using heterogeneous:
  - storage technologies.
  - operating systems.
- Virtual organization needed:
  - Authentication and access rights to the data.
- Storage virtualization:
  - To be independent from technology and hardware evolution.
  - To be independent of local organisation of the files (servers, mount point etc…).

➔ Logical view of the data independent of the physical location.

Solutions ?

- Need for a « grid » middleware.
- SRB (Storage Resource Broker) is anwswering these needs and much more:
  - Developed by SDSC: start in 1998 (license General Atomics).
  - Developers in constant contact with the user community:
    - Functionnalities asked by the users.
  - Portable on many OS and platforms.
  - Support of a vast number of storage system, no limit.
  - Large user community.
- Competitors ?

Solutions ?

- Need for a « grid » middleware.
- SRB (Storage Resource Broker) is anwswering these needs:
  - Developed by SDSC: start in 1998.
  - Under the license of General Atomics.
  - Developers in constant contact with the user community:
    - Functionnalities asked by the users.
  - Portable on many OS and platforms.
  - Support of a vast number of storage system, no limit.
  - Large user community.

SRB in Lyon

HEP	***BaBar***	*SLAC « mirror » site*
	CMOS, Calice	*Data archival*
	Indra	*Data distribution and archival*
	Lattice QCD	*hundreds of TB / y*
Astroparticle	**Antares**	*Main center: ~200 TB / y*
	Auger, Virgo	*Main center: tens of TB / y*
	Edelweiss	*Main center: tens of TB / y*
	SN Factory	*Part of the online: ~GB / d*
Biomedical	***BioEmergence***	*European project ~ TB/y*
	Mammography	*Project with a computing lab*
	Neuroscience	*Lyon and Strasbourg hospital*

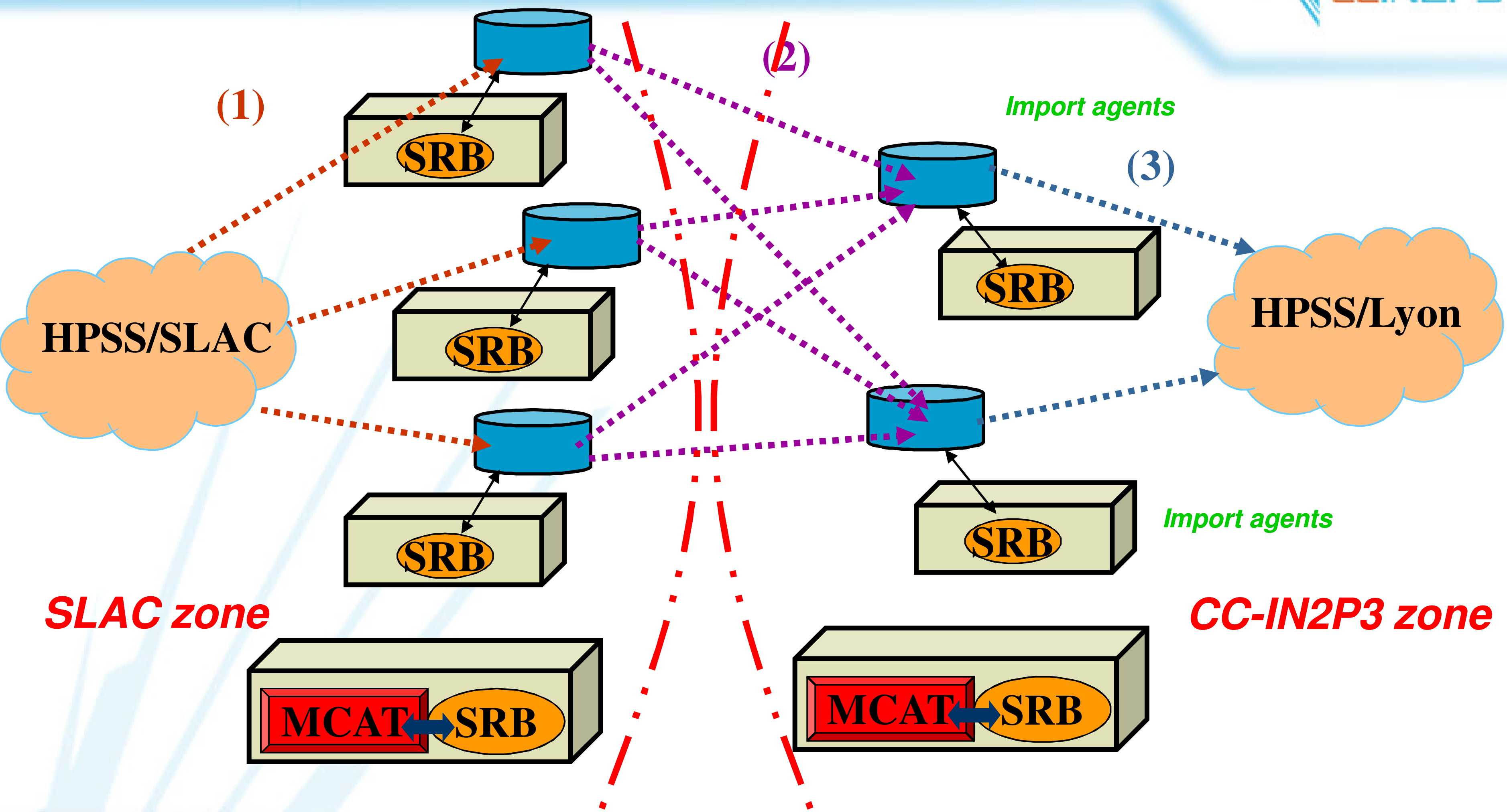
SRB in Lyon

- Being used since 2003.
- 15 servers (disks: 250 TB).
- Oracle 11g database cluster for the metacatalog.
- Interface with HPSS as the Mass Storage System back end, some SRB data backed up in Tivoli Storage Manager (TSM).
- Still very active and still growing:
  - Reaching 2 PBs of data in Dec. 2009.
  - Hundreds of thousands of connection per day.
  - Data stored on disks only and/or on tape.
  - Traffic can reach more than 10 TB / day, coming from everywhere in the world, from laptop to PC batch farms to SuperComputers.
  - Very different usage depending on the projects.

Example in HEP: BaBar

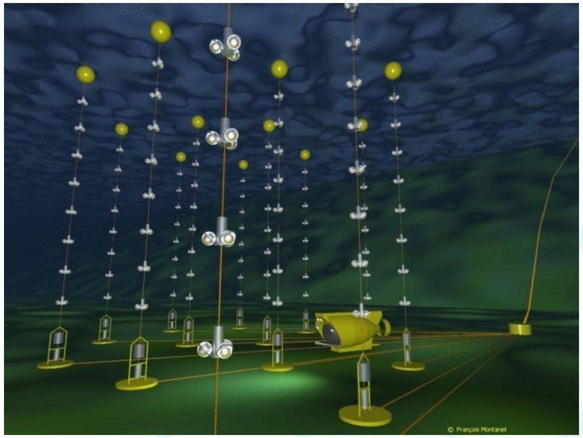
- Data import from SLAC to Lyon.
- SRB being used since 2004 in production.
- **Fully automated**:
  - New files created are registered in the SLAC catalog database.
  - Client application in Lyon: detection of files missing in the Lyon catalog database + transfer of these files.
  - Automated error recovery.
- Up to **5 TB / day** (max. rate observed).
- Usual rate: 2-3 TB / day (during production periods)
- 900 TB imported so far (since 2004), 2 million files.

Example in HEP: BaBar



Examples in astrophysics and astroparticles

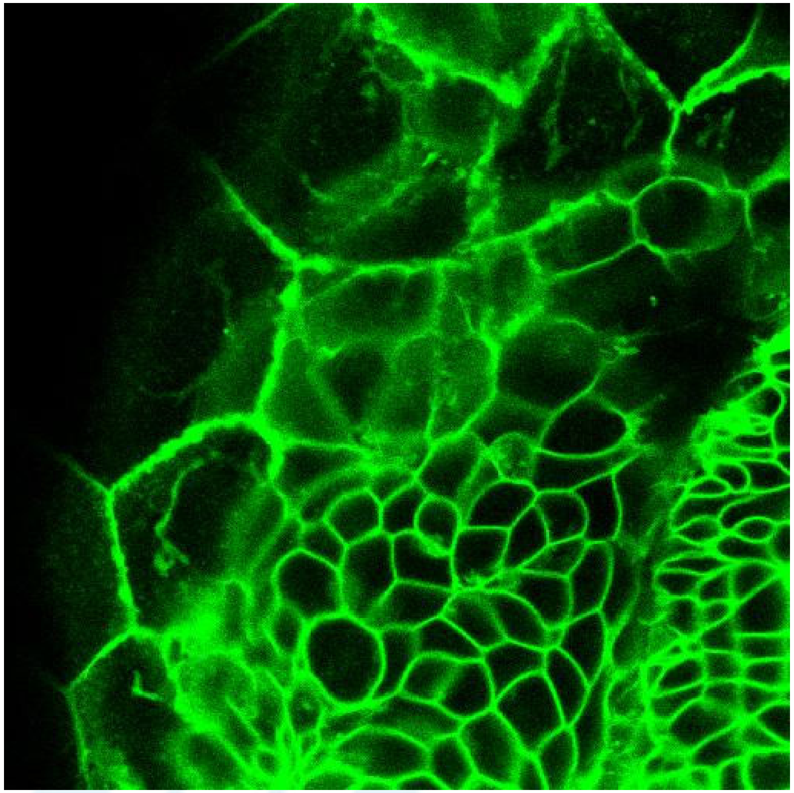
- Underwater: Antares



- in the pampa: Pierre Auger Observatory

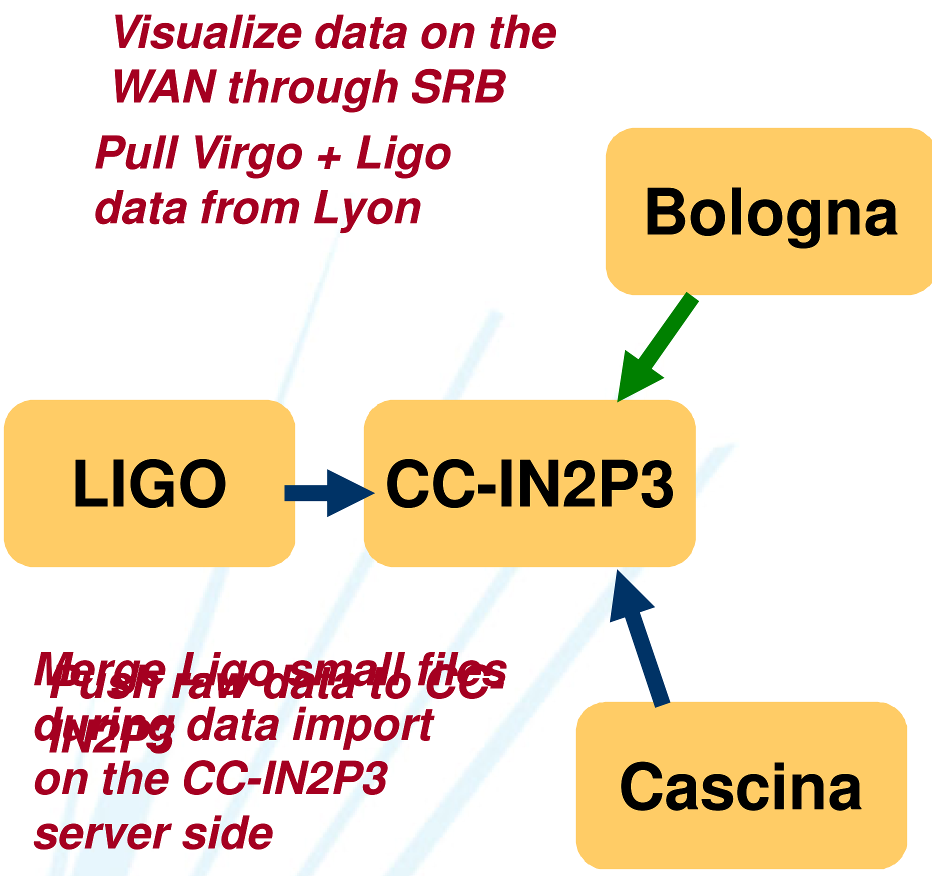
- At the top of the mountain: SuperNovae Factory in Hawaii

Example in biology: BioEmergence



- European projects involving 5 countries.
- Embryogenesis: zebra fish.
- 2 microscopes now (several in the future): amount of data could be huged (PB scale).
- Data pushed from the microscopes into the SRB.
- SRB integrated within their workflow.
- CC-IN2P3: core of the system.

Virgo: data sharing with Ligo



- Interferometer for gravitational waves detection (in production: 60 TB / y).
- Need for a reliable data distribution system.
- Distribute Ligo data (same experiment in the US) to the european sites: CC-IN2P3 and Bologna.
- Have been using bbftp so far.
- Test of EGEE tools not successful.
- SRB has replaced bbftp:
  - Bookkeeping system.
  - Interface with HPSS.
  - Ligo: interoperability.

SRB @ CC-IN2P3

- MCATs performance enhancement:
  - Reindexing made automatically on a weekly basis.
- Issues with Oracle performances in the past:
  - Some oddity in the way Oracle optimized requests.
  - Request analyzis done on all the MCATs on a daily basis.
- Database is one of the key component of the system.

  ➔Now OK: Oracle 11g servers dedicated to SRB.

  ➔Able to have ~ 0.1 s time response on SRB commands even in a millions of files catalog.

SRB in Lyon: future

- Still around for quite some time (2-3 years from now):
  - At least + 1 PB next year.
- Will start to migrate services to iRODS in 2010.
- No migration planned for experiments which have stopped data taking (BaBar, SNFactory, ...).

Assessment of the SRB usage

- Many functionalities used …
- … but not all of them ☹, for example:
  - Extensible MCAT.
- Some developpements were needed:
  - Server side (monitoring, compound resource management, …).
  - Client side (data management application for BaBar, neuroscience etc…).
- Documentation (FAQ):
  - People can be lost by the level of functionalities
- GUI applications (eg: inQ) are fancy but dangerous:
  - Too easy → can be used without being cautious.
- Also true for APIs, Scommands (shell commands)…

Assessment of the SRB usage

- Lack of control on the number of connections to the SRB system (but true for many computing software !):
  - Can be difficult to scale the system.
- Database has to be tuned properly:
  - Need for someone having DBA expertise.
- Sociological factors: fear to have data not under his control.
- Sometimes, lack of manpower on the experiment side in order to build customized client application.

Beyond SRB

- Storage virtualization not enough.
- For client applications relying on these middlewares:
  - No safeguard.
  - No guarantee of a strict application of the data preservation policy.
- Real need for a data distribution project to define a coherent and homogeneous policy for:
  - data management.
  - storage resource management.
- Crucial for massive archival projects (digital libraries …).
- No grid tool had these features until 2007.

Virtualization of the data management policy

- Typical pitfalls:
  - No respect of given pre-established rules.
  - Several data management applications may exist at the same moment.
  - Several versions of the same application can be used within a project at the same.

  ➔ *potential inconsistency.*
- Remove various constraints for various sites from the client applications.
- Solution:
  - Data management policy virtualization.
  - Policy expressed in terms of rules.

A few examples of rules

- Customized access rights to the system:
  - Disallow file removal from a particular directory even by the owner.
- Security and integrity check of the data:
  - Automatic checksum launched in the background.
  - On the fly anonymization of the files even if it has not been made by the client.
- Metadata registration:
  - Automated metadata registration associated to objects (inside or outside the iRODS database).
- Customized transfer parameters:
  - Number of streams, stream size, TCP window as a function of the client or server IP.
- … up to your needs …

iRODS

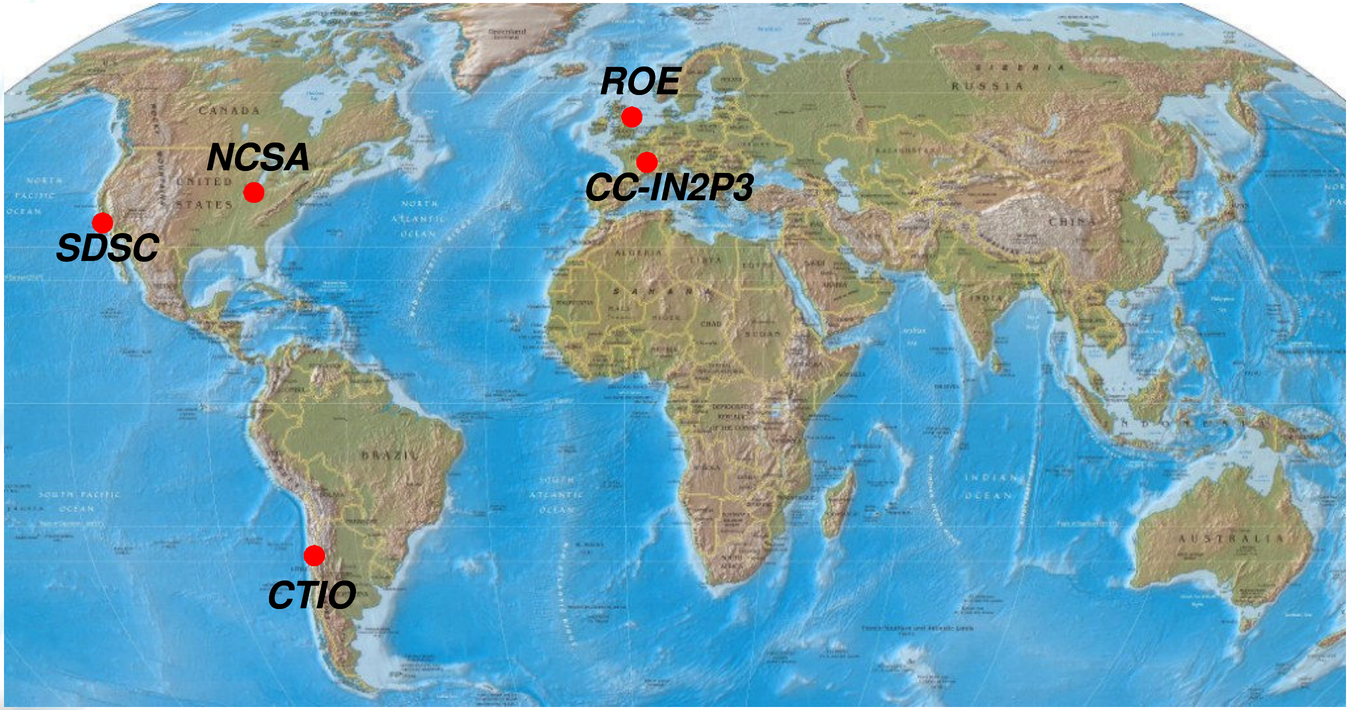
- **iR**ule **O**riented **D**ata **S**ystems.
- Project begun in January 2006, led by DICE team (USA).
- First version official in December (v 0.5).
- **Open source.**
- Financed by: NSF, NARA (National Archives and Records Administration).
- CC-IN2P3 (France), e-science (UK), ARCS (Australia): collaborators.

iRODS developments @ CC-IN2P3

- Tests scripts of the APIs through the shell commands.
- Stress tests.
- Micro-services:
  - Host based access control.
  - Tar and untar of files.
- Load balancing and monitoring system.
- Universal interface with any kind of Mass Storage System.

iRODS test beds

- With KEK (Japan): data transfer at high rate.
- LSST (telescope in Chile, 2015): data replication and workflow.



Production iRODS @ CC-IN2P3

- iRODS:
  - 6 servers with Oracle backend,180 TB.
  - Interfaced with our Mass Storage System (HPSS).
- Adonis (Arts & Humanities projects):
  - > 14 TB of data registered so far.
  - 2 millions of files.
  - Accessed from batch farm.
  - Micro-services needs to be used for one project (long term data preservation):
    - Data archived in CINES (Montpellier) and pushed to Lyon (tar files):
      - Automatically untar the files @ CC-IN2P3.
      - Automatically register the files in Fedora (external system).
      - Data integrity check also done (checksum).

Production iRODS @ CC-IN2P3: Adonis

- Adonis (federation of Arts & Humanities projects, ***Th. Kachelhoffer, P-Y Jallud***):
- > 14 TB of data registered so far.
- 2 millions of files:
  - Accessed from batch farms, laptop
- Micro-services needs to be used for one project (long term data preservation):
  - Data archived in CINES (Montpellier) and pushed to Lyon (tar files):
    - Automatically untar the files @ CC-IN2P3.
    - Automatically register the files in Fedora Commons (external system).
    - Data integrity check also done (checksum).

Production iRODS @ CC-IN2P3: Adonis

- Fedora Commons and iRODS fully interfaced:
  - Fedora storage is iRODS (using fuse).
- Web cluster will use fuse to connect to iRODS servers where data are stored:
  - Interesting for legacy web applications.
  - Easy to use iRODS for new projects (no need to use the PHP APIs, still need enhancement).
  - Able to upload large amount of data by other means than http + data management capabilities of iRODS can be used.
- Will ramp up to 100 TB of data during 2010 from various sources.

Production iRODS @ CC-IN2P3: Rhône-Alpes data grid

- Rhône-Alpes data grid (TIDRA: Y. Cardenas, P. Calvat) provide computing services for research labs.
- iRODS proposed for the data storage and management.
- Biomedical applications:
  - Human studies: anonymized files, msi for DICOM metadata extraction under development.
  - Soon, mice studies (brain MRI): push the data into iRODS and automatic extraction of metadata into iRODS metadata.
- Biology applications.
- Other applications coming soon.
- Very active, up to 60000 connections / day.

iRODS assessment

- Highly scalable for data management tasks.
- Many features and customization: very attractive to potential users.
- Already a large community interested by iRODS growing world-wide in various fields, for example:
  - Long term digital preservation.
  - Astrophysics.
  - Biology.
- Already stable and mature enough for production.
- DICE team very reactive in order to solve problem and open to include new features.
- Confident that iRODS is able to sustain 100 millions of files catalogs with our infrastructure.

iRODS future in Lyon

- Is replacing SRB:
  - Migration from SRB to iRODS.
  - New experiments: directly on iRODS.
- iRODS becoming one of the key services:
  - Plan to replace it for the light weight transfers (usually bbftp, scp …).
  - Proposed for new projects.
- Soon: LSST, DChooz (neutrino experiment) etc…

Data repositories: present and future

- Tools like SRB, iRODS … have changed the way we are dealing with data in data centres:
  - Files are not just 0s and 1s.
  - Participating much more deeply in data management policy.
- Metadata:
  - Getting richer and richer.
  - Could be a challenge on the database side.
- Data preservation:
  - Still a lot of thing to be done on this side.

Acknowledgement

- DICE research team.
- Pascal Calvat, Yonny Cardenas (CC-IN2P3), Jean Aoustet.
- Thomas Kachelhoffer (CC-IN2P3), Pierre-Yves Jallud (Adonis).
- Wilko Kroeger (SLAC – BaBar).
- Adil Hasan (University of Liverpool).