Centre de Calcul de l'Institut National de Physique Nucléaire et de Physique des Particules

Data storage services at CC-IN2P3

Jean-Yves Nief

CCIN2P3

cnrs

Irfu

Agenda

- Hardware:
  - Storage on disk.
  - Storage on tape.
- Software:
  - Storage services.
- Pitfalls and lessons learned.
- Storage needs in the near future.
- Prospects.

Storage at CC-IN2P3: disk

Hardware

Direct Attached Storage servers (DAS):

- Dell servers (R720xd + MD1200)
- **~ 240** servers
- Capacity: **15 PBs**

Disk attached via SAS:

- Dell servers (R620 + MD3260)
- Capacity: **1.9 PBs**

Storage Area Network disk arrays (SAN):

- IBM V7000 and DCS3700, Hitachi HUS 130.
- Capacity: **240 TBs**

Software

Parallel File System: GPFS (**1.9 PBs**)

File servers: xrootd, dCache (**14 PBs**)

- Used for High Energy Physics (LHC etc…)

Mass Storage System: HPSS (**600 TBs**)

- Used as a disk cache in front of the tapes.

Middlewares: SRM, iRODS (**840 TBs**)

Databases: mySQL, PostGres, Oracle **(57 TBs)**

Storage at CC-IN2P3: tapes

Hardware

4 Oracle/STK SL8500 librairies:

- **40,000** slots (T10K and LTO4)
- Max capacity: **320 PBs** (with T10KD tapes)
- **111** tape drives

1 IBM TS3500 library:

- **3500** slots (LTO6)

Software

Mass Storage System: HPSS

- **25 PBs**
- Max traffic (from HPSS): **100 TBs / day**
- Interfaced with our disk services

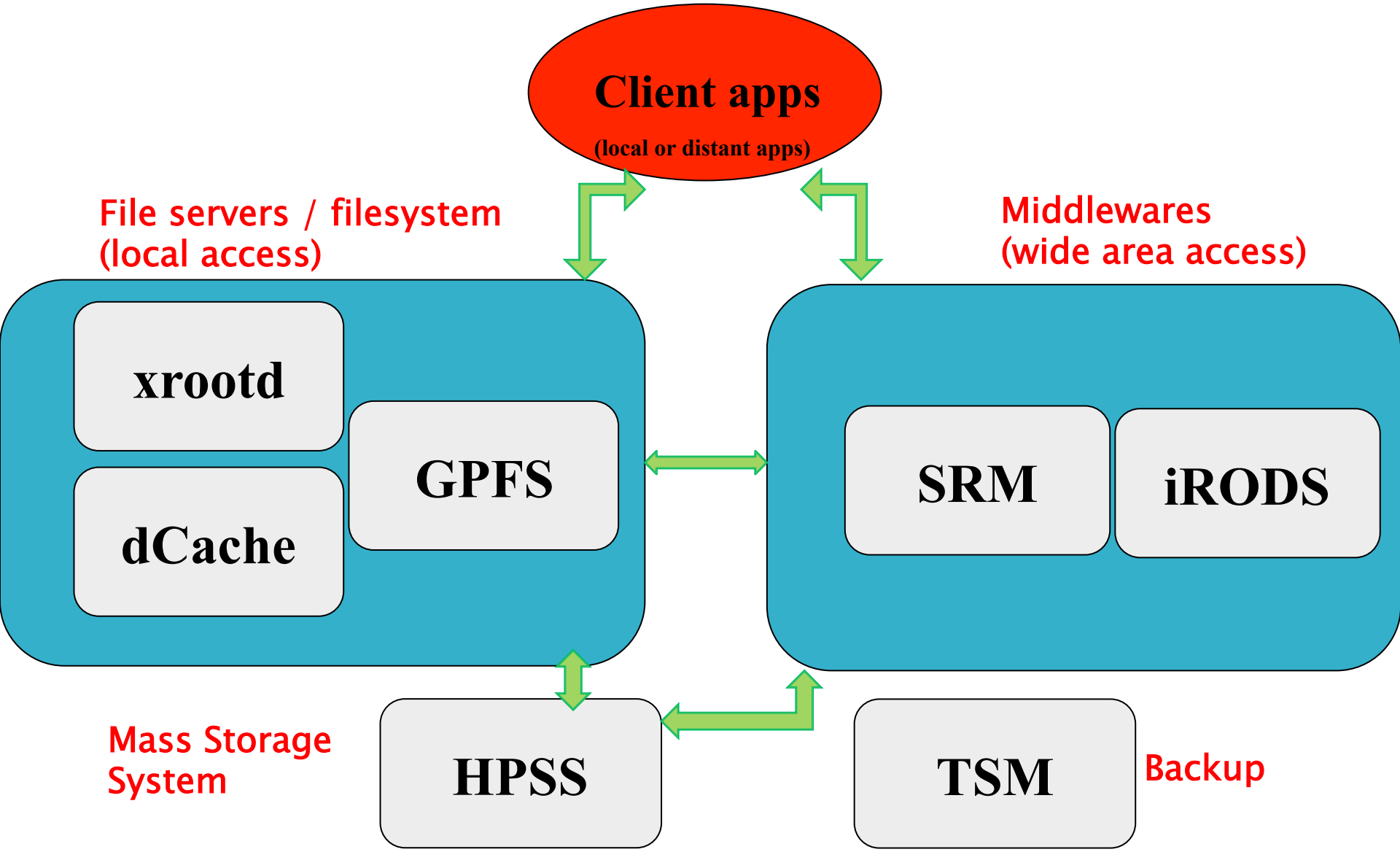
Backup service: TSM (**1 PB**)

Storage hardware as of today

- 4 Oracle/STK SL8500 libraries (4 x 10000 slots) for HPSS and TSM.
- 1 IBM TS3500 library (3500 slots) for TSM.
- Tape media:

Media	Capacity	Max Bandwidth	Number of drives	Service
T10K-B	1TB	120 MB/s	59	HPSS
T10K-C	5 TB	240 MB/s	21	HPSS
T10K-D	8.5 TB	252 MB/s	31	HPSS
LTO4	800 GB	120 MB/s	20	TSM
LTO6	2.4 TB	160 MB/s	6	TSM

How experiments are dealing with the storage ?



Local access: distributed filesystem

- Purpose:
  - For heavy I/O operations by many batch jobs in //.
  - Not a permanent repository.

- 60 servers, 1.9 PBs of disk space.
- 1000 GPFS clients installed on our

- Used by:
  - up to 25% of the batch farm.
  - 84 groups.

- Max activity observed: 500 TBs read in 5 days (Planck)

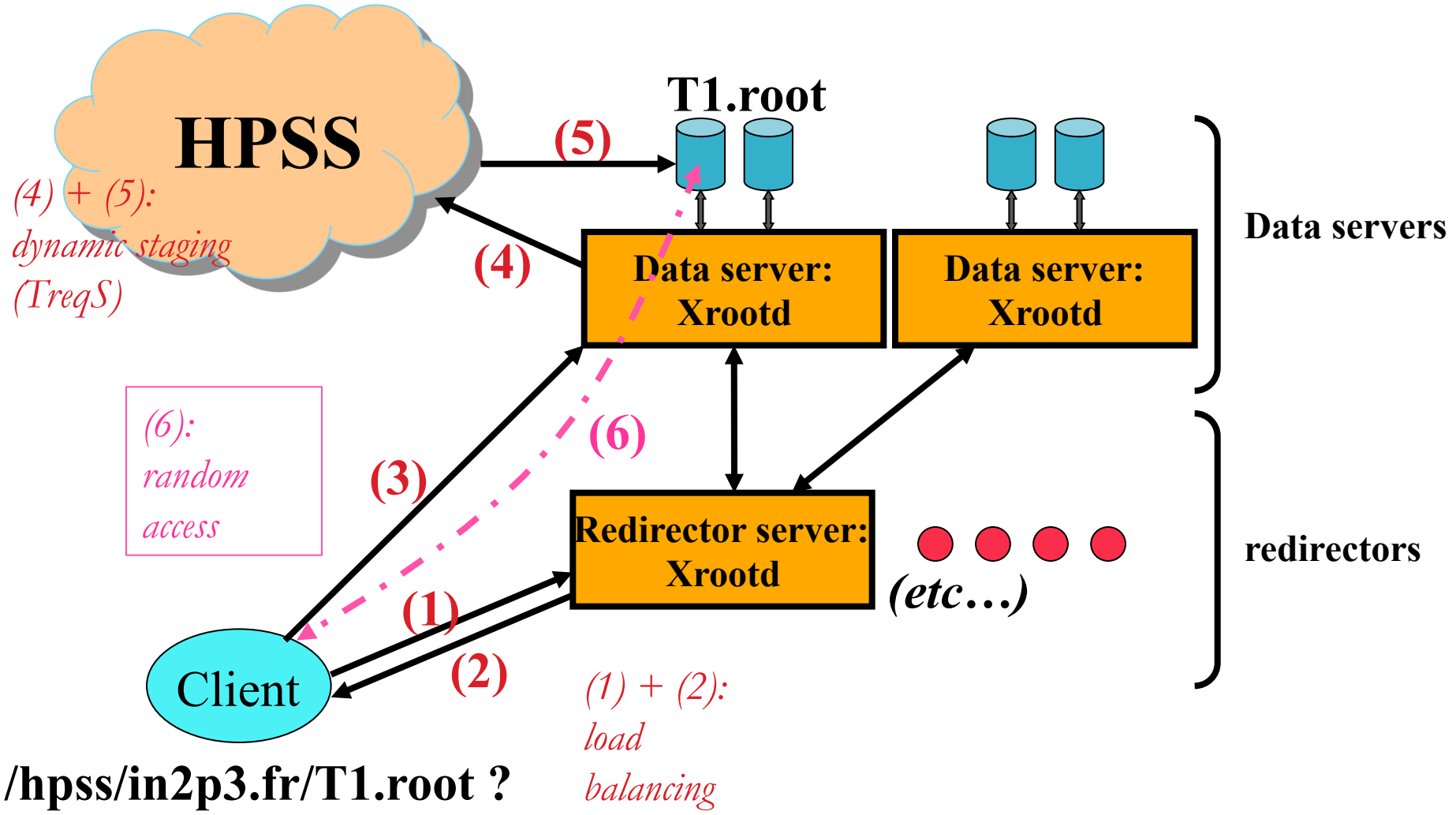
- To do:
  - Migration of metadata storage to flash technology (NetApp EF560).
  - Reduction of the number of space (35 right now) down to 4.

Local access: dCache-xrootd

- Purpose:
  - Data access by batch jobs (local + grid).
  - Pattern: remote access (random I/O, sequential).
  - Interface with HPSS:
    - Using Treqs (Tape request scheduler).
  - dCache service part of the Atlas and CMS federations:
    - xrootd proxy servers (gateway to our local dCache service).
  - dCache head nodes db: use of SSD.
- To do:
  - xrootd: lots of tape staging (eg: 20k requests per day) ➔ « automatic » prestaging for some large « productions » needed.

service	# servers	capacity	experiments	# clients in //	access protocol
dCache	165	8.5 PiB	LCG + « EGEE » VOs (17)	15000	dcap, xrootd, WebDAV, gsiftp
xrootd	44	2.5 PiB	Alice + HEP + astroparticle (13)	8000	xrootd, *http*

Local access (xrootd/dCache)



Mass Storage service: HPSS

- Purpose:
  - Interface to our back end storage: tapes (main storage).
- 29 servers (disk + tape movers).
- Massive activity:
  - 27 PiB on tapes (rank #12 in HPSS world), 597 TiB of disk space (Dell R510, R720 servers).
  - **Up to 100 TiB in 22h between HPSS and other storage services.**

- New drive technology every **30 months**.
- Life cycle of a tape generation **8 years**:
  - Ramping up in 4 years.
  - 4 more years in production.

- T10KB decommissionning:
  - 20000 tapes to be repacked until end 2017 ⇔ 5 PiB / year.
  - Represents:
    - ***1000*** T10K-T2 ➔ T10K-C
    - Or ***750*** T10K-T2 ➔ T10K-D
- Tape lifetime: avg 7 years (max 10 years)**.**

- Redefine the class of services.

- Refactoring of Treqs (Tape request scheduler) by our dev team.

Local access: pitfalls and lessons learned

- Dynamic staging may result in:
  - Massive staging requests.
  - Long delays to access the files.
  - Tapes capacity is getting bigger!

  ➔Increased job inefficiency.

- To leverage these drawbacks:
  - Big files on tapes (several GBs).
  - Reordering of file requests per tape id (reduce # of mounts/ dismounts): up to 10000 per day.
  - Prestaging.
  - Avoid file scattering on a large amount of tapes (tape families ?).

- Refactoring of Treqs (Tape request scheduler) by our dev team.

Middleware: SRM, iRODS

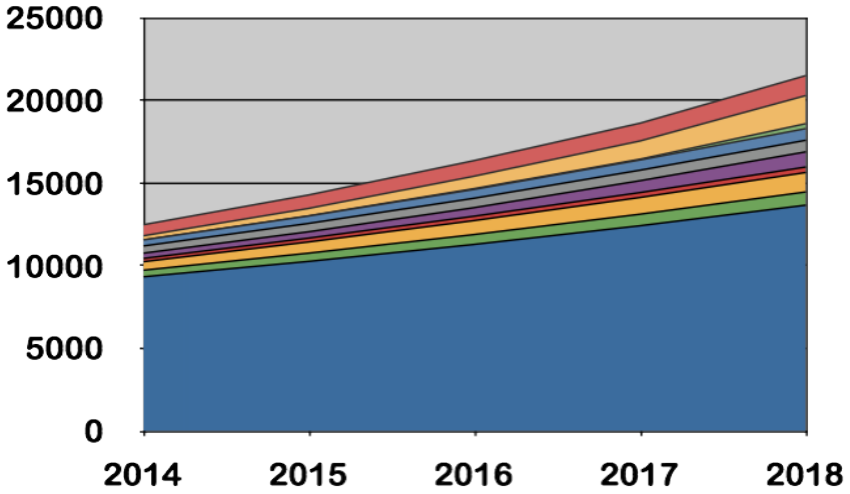
- Purpose:
  - Data management (storage virtualization, policies).
  - Data sharing, access, archival etc… in a distributed and heterogeneous environment.
- SRM:
  - Used by Atlas, CMS, LHCb + EGEE/EGI virtual organizations.
- iRODS:
  - Managing > 9 PBs of data (HEP, astro, biology, Arts & Humanities): includes data stored on tapes.
- Web interfaces, HTTP/REST, webdav.

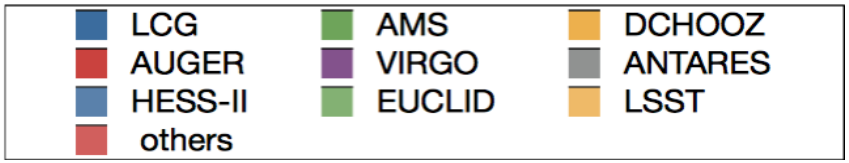
Focus: LSST storage usage

- GPFS:
  - 130 TB (50% being used at the moment)
- HPSS:
  - 78 TB
- iRODS:
  - 90 TB

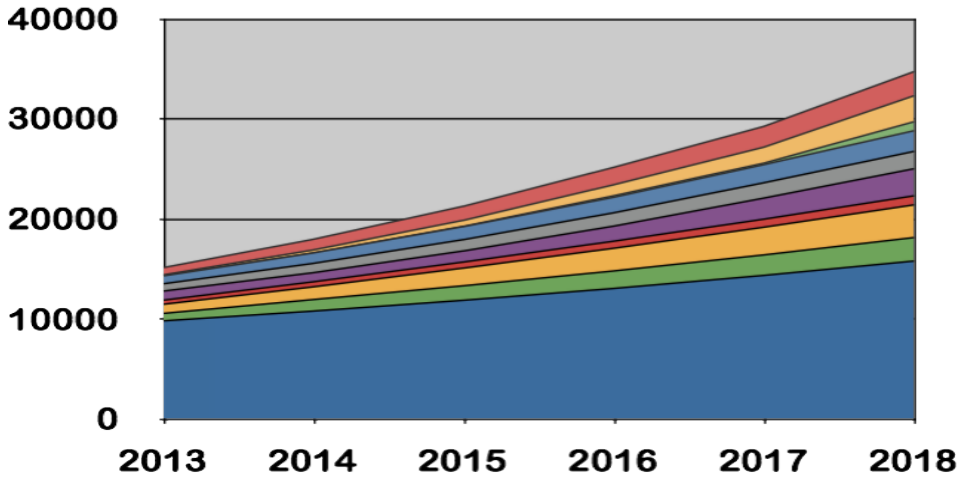
Future storage needs (Hard Sci Fi)

Cumulated Disk (TB)





Cumulated Tape (TB)



Hypothesis:

- **LCG: +10% / year**
- **HEP like experiment: ratio disk/ tape = 30%**

CTA not included

Credit: Rachid Lemrani

Qserv

- Used to debug Qserv.

- 1 interactive node:
  - 1 virtual machine *ccqservbuild.in2p3.fr*.
  - Mainly used for:
    - the compilation of the Qserv software.
    - the deployement of Qserv on the cluster.
  - Accessible from SLAC.

- 50 Qserv nodes:
  - Dell / CC-IN2P3 partnership.
  - *ccqserv{100..149}.in2p3.fr*.
  - Dell servers R620 and R630
    - Intel Xeon, 16GB RAM
    - 7 TB of usable storage
  - run the Qserv software (1 master, 49 slaves).
  - private subnetwork (nodes accessible from *ccqservbuild* only).

(Credit: Yvan Calas)

Storage prospects

- Tape will still have a major role:
  - Balance between nearline and pure archive storage ?
- Hard drives still preferred to SSD:
  - prices will converge in the next couple of years.
  - SSD usage limited at the moment for key apps (GPFS metadata, HPSS and dCache databases etc…).
- Heavy I/O still and even heavier:
  - Lots of « random » I/O: still strong needs for seek, read, write.
  - Simple put/get interaction not sufficient.
- Big metadata challenge (storage system metadata).
- Data life cycle improvements:
  - > 100 groups, 5000 users (~ 2000 active users).
  - Data Management Plan.
  - archival service beyond simple bit preservation ?
  - OAIS, dublin core metadata.
- Provide interfaces for Open data ?

Storage prospects

- Assessment of object storage:
  - Swift: testing it at the moment.
  - Swift extra value wrt what we already have: to be proven.
  - Ceph (but it can be more than an object storage): evaluation starting in the next couple of months.

- Hadoop, MapReduce type of systems: no plans yet as not much needs expressed by our user community.

- Metadata: will increase massively with increased storage.
  - Limit the amount of files => increased file size.