Storage & CMS data at CC-IN2P3

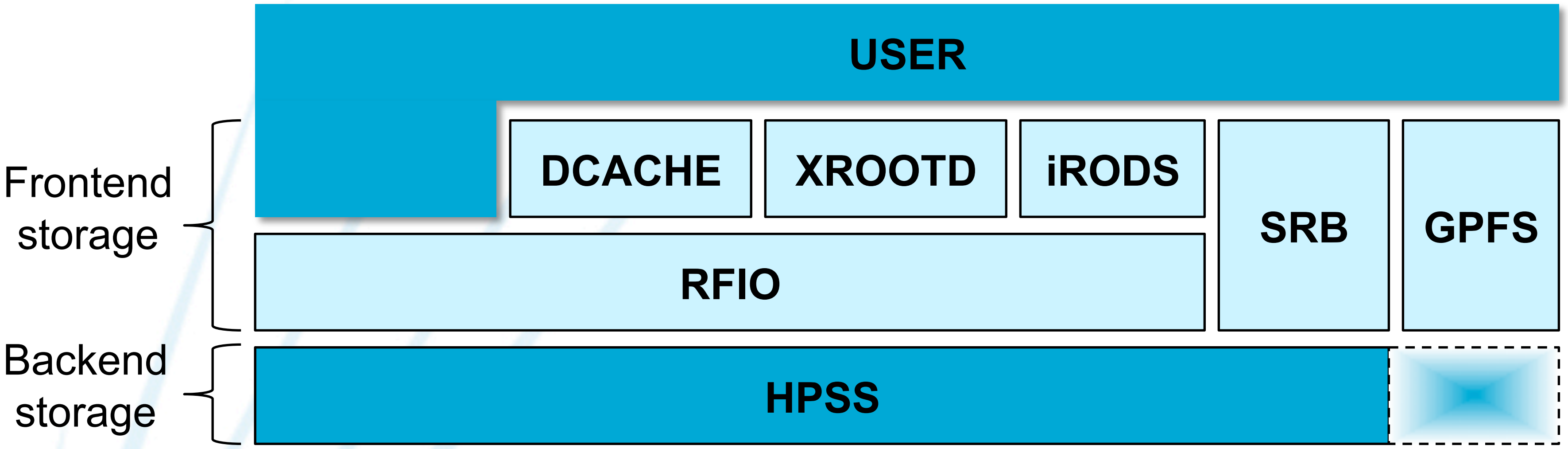
Benoit DELAUNAY

20091023

Storage at CC-IN2P3

- CC-IN2P3 provides computing and storage for the 4 LHC experiments and many others (astro particles...)
- A long history of service sharing between experiments
- Some dedicated resources for LHC experiments

Storage infrastructure



USER

Frontend storage

DCACHE

XROOTD

iRODS

SRB

GPFS

RFIO

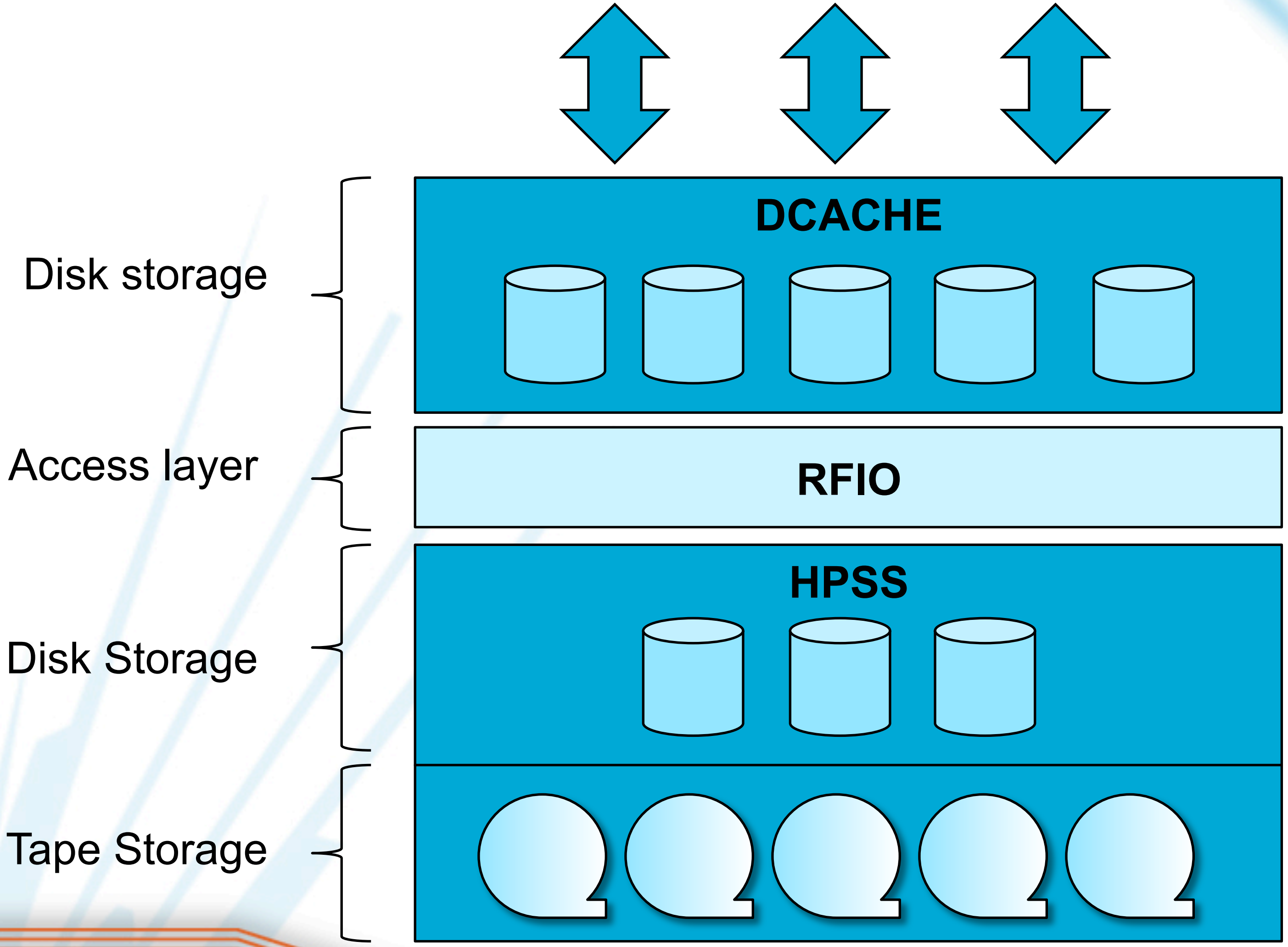
Backend storage

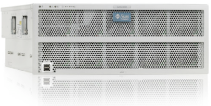
HPSS

Storage management

- 6 engineers involved in the management of the storage middleware
- 3 engineers for dCache
  - Lionel SCHWARZ, Jonathan SCHAEFFER, Yvan CALAS
- 3 engineers for HPSS
  - Pierre-Emmanuel BRINETTE, Andres GOMEZ, Benoit DELAUNAY

Storage for CMS





dCache hardware platform

4 Master Servers

- Scientific Linux 4
- 16GB memory
- 8 CPU cores

79 x4540 disk servers

- SUN Solaris 10
- 32TB disk storage
- 2 Gbps Nework Interface

Storage: ***2500TB***

Network: ***158Gbps***

HPSS hardware platform

1 Master Server

- IBM AIX 5.3
- 64GB memory
- 16 CPU cores

12 disk data movers

- RedHat Enterprise Linux 4
- 40TB disk storage
- 10Gbps Nework Interface

Storage: 480TB

Network: 120Gbps

27 tape data movers

- IBM AIX 5.3
- Mixed 2Gbps/10Gbps

Bandwidth: 70Gbps

3 libraries STK SL8500

- 10,000 slots each
- 13 x 9840 tape drives (20GB)
- 36 x T10KA tape drives (500GB)
- 32 x T10KB tape drives (1TB)

Max Capacity 30 PB

dCache and CMS

- dCache used as a end user storage system and also for data exchange between LCG sites
- Used for CMS Tier1 and Tier2 at CC-IN2P3
- Only one instance for the 4 LHC experiments
- Storage pools are dedicated to experiments

dCache changes

- Upgrade from dCache v1.9.0 to v1.9.4 (2009/09/22)

- Migrate from PNFS to CHIMERA (2009/09/28)
  - Complete shutdown during 3 days
  - First observations show that access to metadata has been improved

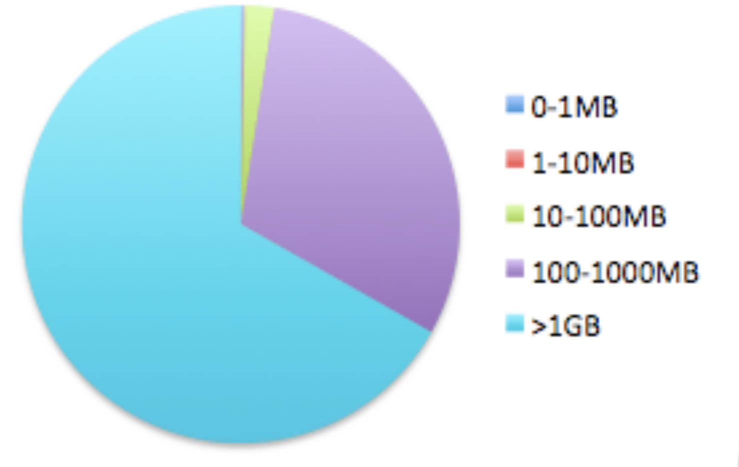
- Many storage servers engaged

dCache numbers for CMS data

- 39 buffer pools for a total of 661TB
  - T1 / Local read buffer : 360TB
  - T1 / Input buffer : 68TB
  - T1 / Output transfer buffer : 28TB
  - T2 / 205 TB

File size distribution

since june 2009 :



HPSS and CMS

- HPSS used as a backend storage for dCache

- System use not dedicated to LHC experiments, but dedicated storage resources
  - 1 logical instance (subsystem) of HPSS for CMS
  - Means dedicated disks and tapes resources
  - CMS data do not share HPSS disks and tapes with others

HPSS changes

- Major software upgrade to the version 6.2.2.2 in june 2009 (complete shutdown during 4 days)
- Hardware platform has been almost totally replaced
  - A new master server and new data movers
  - New tape drives STK T10KB (1TB/tape)
- HPSS is now more reliable, powerful and capacitive.

HPSS internals

- File size base storage policy

- 4 Class of Service

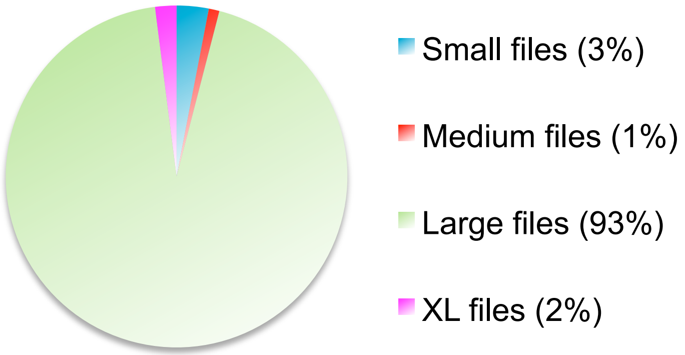
– COS10 : small files	0-64MB
– COS11 : medium files	64MB-512MB
– COS12 : large files	512MB-8GB
– COS14 : XL files	8GB-128GB

- Larger is the file, more powerful is HPSS !

HPSS numbers for CMS data

- 80 TB allocated disk
- 1.3 PB on tapes
- 986,000 files but, 523,000 never read (53%) !

File size distribution :



Data access improvement

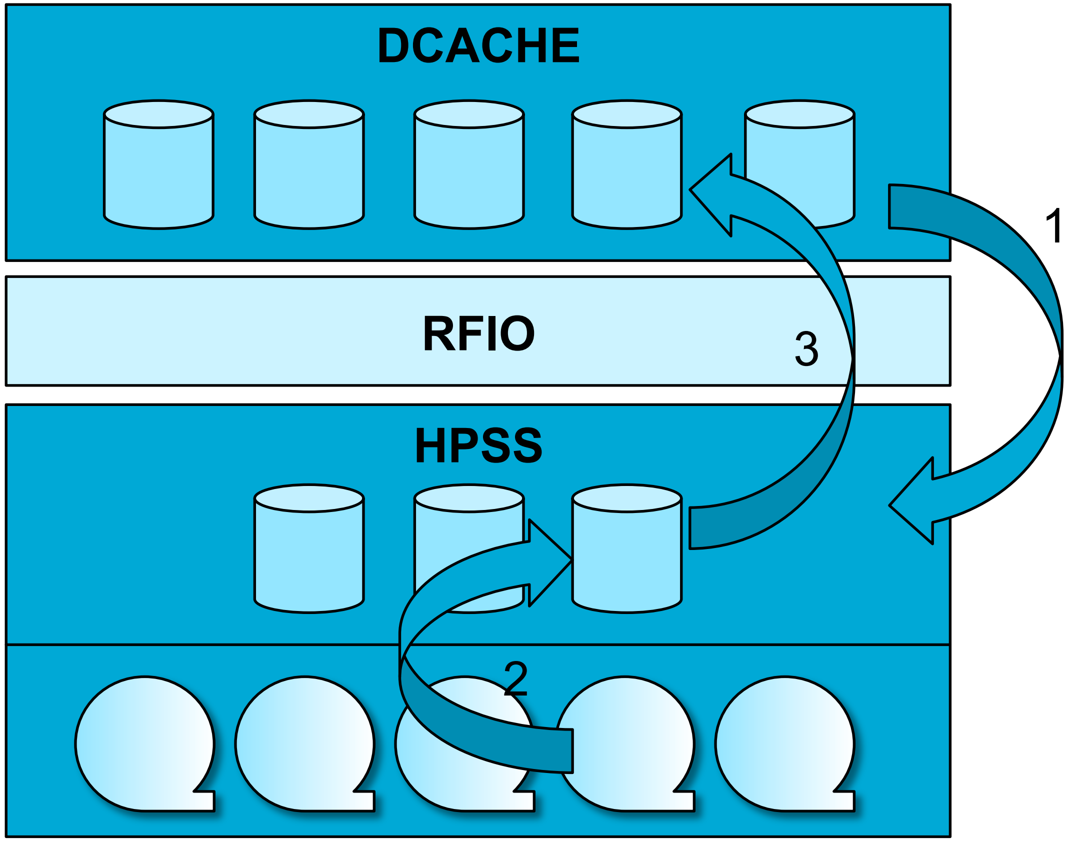
- dCache requests file staging one by one
- HPSS has a very basic behavior when reading files
  - Handles file read requests one by one in the order they were submitted (FIFO)
  - Could be very unefficient when the file lists is disordered and many files are stored on the same tapes
- A solution for that, submit ordered file list by tapes to HPSS using T-ReqS (Tape Request Scheduler)

dCache without T-ReqS

1. dCache asks for n files one by one to be read on m tapes (m<n)

2. HPSS stages files in disorder from tapes to disks (***n tape mounts!***)

3. dCache gets files from HPSS disks via RFIO one by one



dCache using T-ReqS

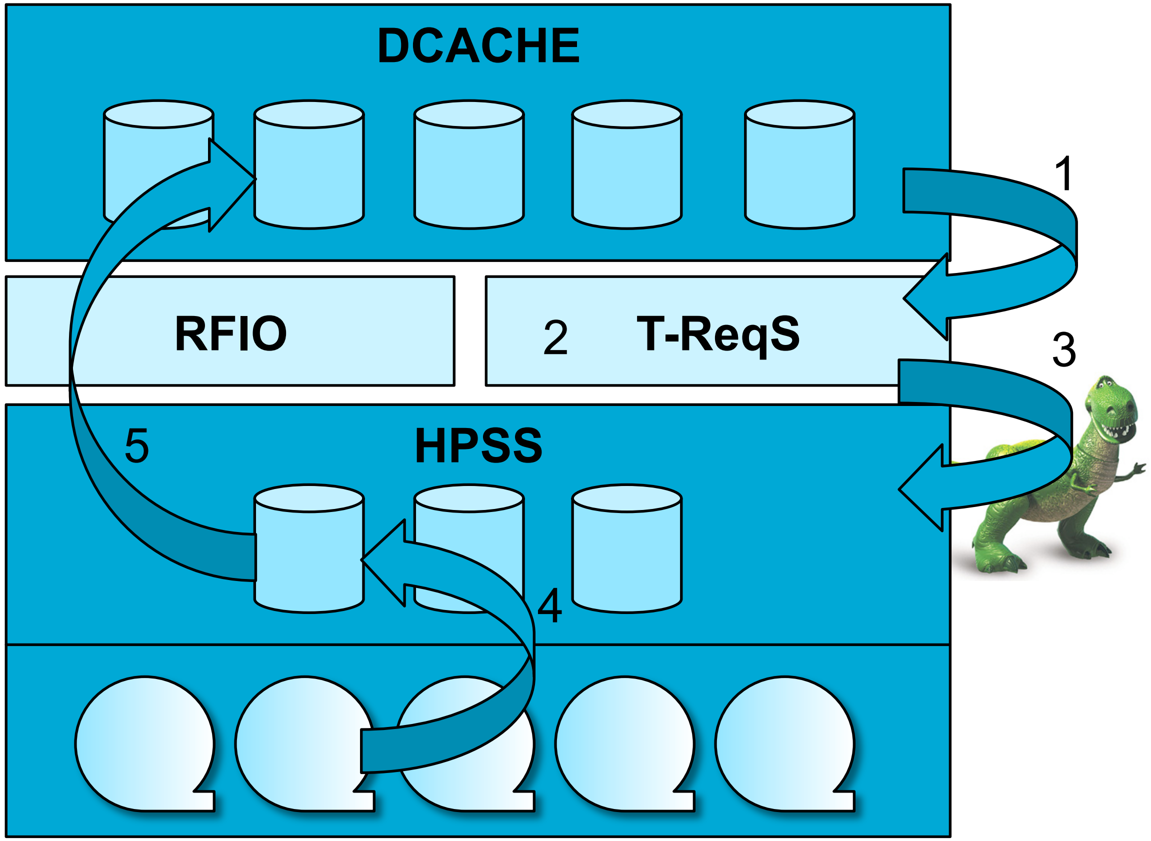
1. dCache asks for n files to be read on m tapes (m<n)

2. T-ReqS reorders file requests by tapes

3. T-ReqS requests the file staging to HPSS

4. HPSS stages files from tapes to disks (***m tape mounts!***)

5. dCache gets files from HPSS disks via RFIO



What’s next ?

- dCache upgrade to v1.9.5 (Golden release) on november 2009
- HPSS upgrade to v7.x on june 2010
- More disks and more tapes
  - A fourth SL8500 in january 2010 (10,000 tape slots + T10KB drives) => tape storage capacity extended to 40PB
  - Call of tender for new disk servers at the end of 2009

Conclusion

- A lot of work has been done this past year
  - On the hardware infrastructure
  - On the software (dCache and HPSS)

- We will be still quite busy the next year !

- We are confident in the ability of the storage system to cope with the LHC experiment requirements.

Thank you !