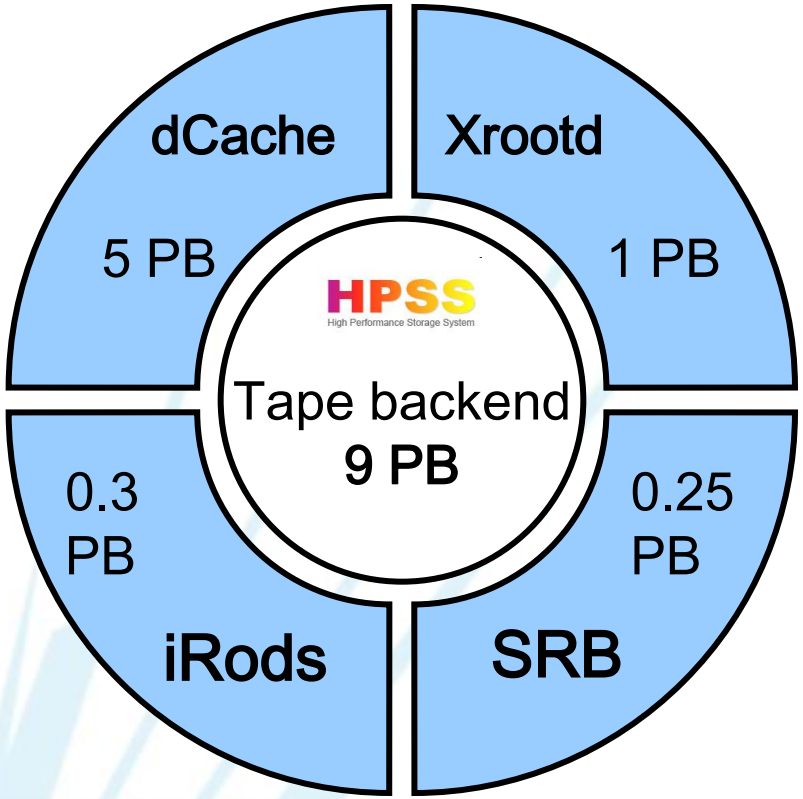
Storage @ CC-IN2P3

Pierre-Emmanuel Brinette
IN2P3-CC Storage Team

HEPiX Fall 2010
2 nov. 2010

Storage services overview

Mass Storage Systems



Semi permanent Storage

GPFS

$HOME

AFS

BACKUP

TSM

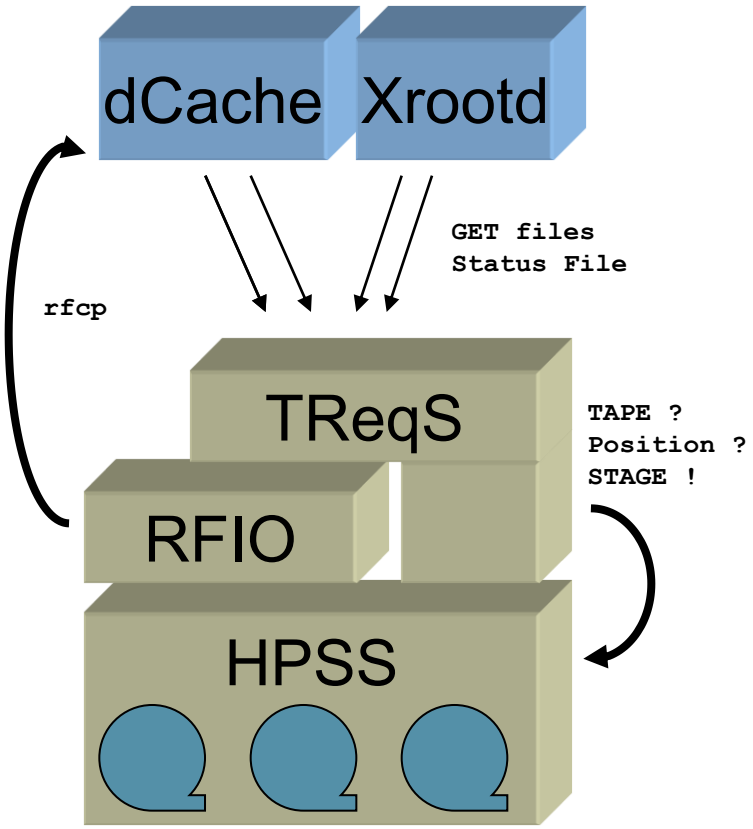
90 Groups (VO)
(HEP,astro,bio,H&SS)
3000+ Users

The Storage team

- Design and manage the storage services of IN2P3-CC
- 8 peoples for services management
  - 2 FTE dCache, FTS (+1 new)
  - 2,5 FTE HPSS
  - 0,75 FTE GPFS
  - 1,5 FTE TSM
  - 1 FTE SRB, xRootd, iRods
- Storages infrastructure deployed by the system team (1,5 FTE)
- AFS services are managed by the system team (1,5 FTE)
- Robotics libraries are managed by operation team (2 FTE)

Tape Backend : HPSS

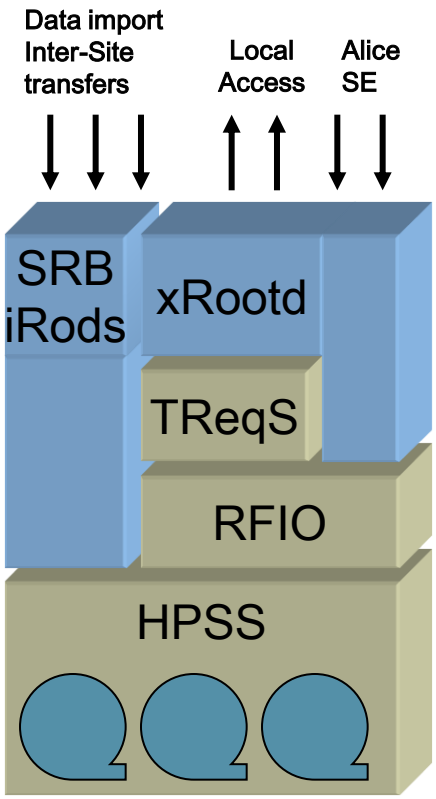
- HPSS v6.2.3 (HSM)
- Repository for experimental Data
  - 9.5 PB (30 % LHC)
  - + 4PB / Year
- Access with RFIO
  - Fork of CASTOR 1.3.5 (2002)
  - Simple and lightweight client (rfcp, rfdir, …)
  - Good performances (directs transfers from disks servers to clients)
- Staging from dCache/xRootd is controlled by TReqS [1]
  - Optimizing read operations on tape (sort requests by Tape ID & Position)
  - Efficient control of tape drive allocation
  - Better performances (reduce mount/dismount)
- Migration to HPSS v7.3 Soon
  - Small file aggregation



[1] http://indico.cern.ch/contributionDisplay.py?contribId=45&sessionId=10&confId=61917

SRB / Xrootd/ iRods

- SRB :
  - Data management middleware
  - Used by HEP / astro / bio experiments
  - Connected to HPSS for tape backend
  - Now used for inter-site transfers and data management
- Xrootd :
  - Intensive I/O server
  - Mainly for local reads operations from computing farm
  - Act as disk cache for files in HPSS
  - Main SE for Alice, R/W permitted
- SRB → iRods Migration
  - Metadata Migration only, the data remains
  - Clients non compatible
  - Users Applications need to be rewritten for iRods
  - Migration planning : 2 years



dCache

- 2 instances
  - LCG : T1 for ATLAS & LHCB, T1/T2 for CMS
  - Egee (10+ VO)
- Access via dcap, gsidcap, gFTP, xrootd
- Servers Pool shared between LHC VO
  - Side effect on other VO when a VO overload pools
- Troubles :
  - Transfer problems (export T1 → T2) still indeterminate (Solaris machines ? , network configuration ? , last version of dCache (v1.9.5-22) ?)
  - Scalability : Too many simultaneous SRM requests (misbehaviors of jobs from computing farms)
- Future
  - Some discussion to create dedicate dCache instance for each LHC VO
  - Some discussion to dedicate servers pool to a single VO

Semi Permanent storage

- GPFS since 2006 in replacement of NFS
- Used for medium term disk storage
  - Automatic Cleanup policies
- IBM support contract renewed for next 3 years for CC + several IN2P3 laboratories
- HSM interface ?
  - Not yet, to many small files
- Why not Lustre ?
  - Data placement policies (different QoS on same namespace)
  - Reliable transparent online data migration :
    - Decommissioned servers
    - Filesystem size changes
  - Oracle…

OpenAFS

- /afs/in2p3.fr
- Used for $HOME et $GROUP, experimental data & VO software toolkit
  - ATLAS Software release issues fixed
  - 3 RO servers sustain 6K jobs
- Migrate File Server from SAN to DAS
  - Migration from 25 V24x (Sparc) to 34 Fire X42xx (x86) 2TB SAS
  - Solaris 10 + ZFS
- Client performance issues
  - Particularly on new hardware (DELL C6100, 24 Core HT)
  - Linux kernel tuning
  - OpenAFS linux client on SL5 (cache size increase, # daemon, …)

BACKUP

- IBM TSM :
  - User data (AFS),
    still using the old TSM/AFS client
  - Experimental data (astro)
  - 19 IN2P3 laboratories all over France over WAN 1/10Gb links
  - 1 billion files / 700 TiB / 1500 LTO 4
- Future
  - Migrate to v6.x (end 2010)
  - Metadata on DB2 DB
  - 2 → 4 servers
  - TSM/AFS client :
    - Internal tools based on Anders Magnusson client.
  - Waiting for LTO 6

Hardware : Disk

High End

2 DS8300 =~ 256 TiB
SAN / FC Disk 12 k
Direct FC attachment to server

Usage :
Intensive IO/s , reliability
AFS, Metadata, Web cluster, TSM, …

High Capacity

7 DDN DCS9550 =~ 1.2 PiB
SAN / SATA 7.2 k
Direct FC attachment to server
Usage :
High throughput storage
GPFS , HPSS

LEGO® Bricks

250 Thumper/Thor =~ 8 PiB
DAS / SATA 7.2 k
Standalone server, 2/4 Gb/s
Solaris 10 / ZFS
Usage :
Distributed storage software
dCache, xRootd, SRB, iRods

Hardware : Disk

AFS

34 SUN (Oracle) Fire X4240
16 SAS 10K disk / server

Oracle Cluster

Pillar Data System
Axiom 500

OPERA experimental data
40 TiB → 120 TiB

Mixed SATA/FC Disk
SAN attached

Hardware : Robotics & Tape

Powder Horn
STK 9840
STK 9940
Retired

4 SL8500 / 10000 Slots
Redundant bots
multihost

94 T10K-A/B drives
16 LTO 4
Theoretical capacity : 40 PB

- 3 type of drives
  - T10K-A (120/500 GB) and T10K-B (1TB) for HPSS
  - LTO 4 for TSM
- SAN : Brocade Director 48000
- Monitoring using StorSentry [2]
  - Preventive recommendations ("*Change drive*", "*Copy & replace tape*", …)
  - Good results, simplify robotics operations.
- Short-Term plan
  - LTO 4 for HPSS (archiving of BaBar data from SLAC)
- Mid-Term plan
  - T10K-C for HPSS, remove T10K-A Sport

[2] http://indico.cern.ch/contributionDisplay.py?contribId=23&confId=61917

Hardware : Next Purchase

- Financial Crisis !
  - → Budget reduction in the next years (2012 …)
- Sun X45xx (Thor) sold by Oracle
  - → Prices increases (>> 1 € / GiB)

DELL PowerEdge R510
12 * 2TB (3”5) 7.2k SAS
+
DELL PowerVault MD1200
12 * 2TB (3”5) 7.2k SAS
Or
DELL PowerVault MD1220
24 * 600 GB (2”5) 10k SAS

0.27 € / GiB
54 TiB usable
(2 MD1200)
4*1 Gbits
5 year support

Benchmarks results :

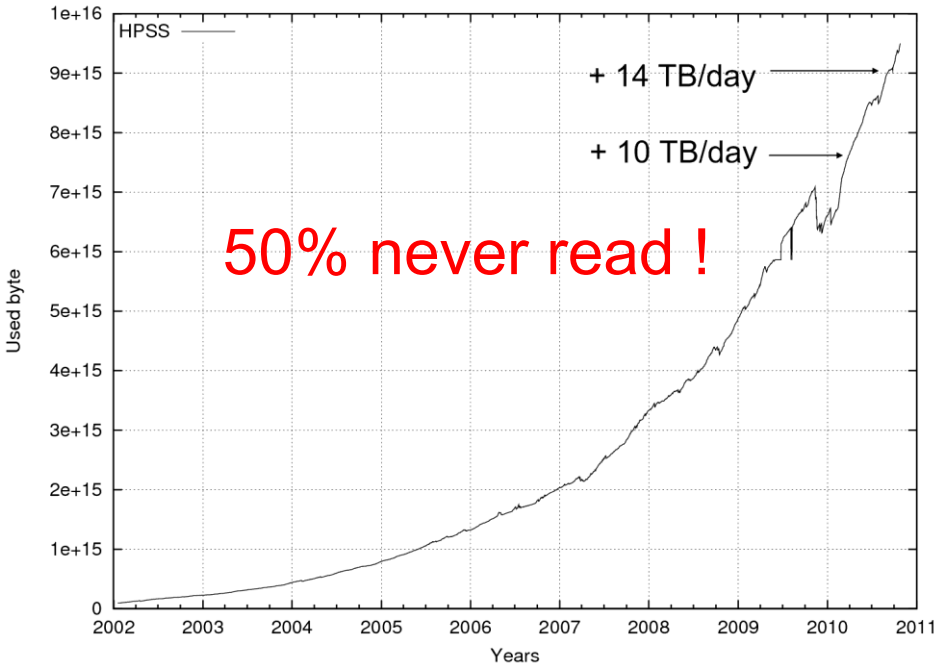
- **MD1200 : OK for dCache but not for xRootd**
- **MD1220 : To be tested**

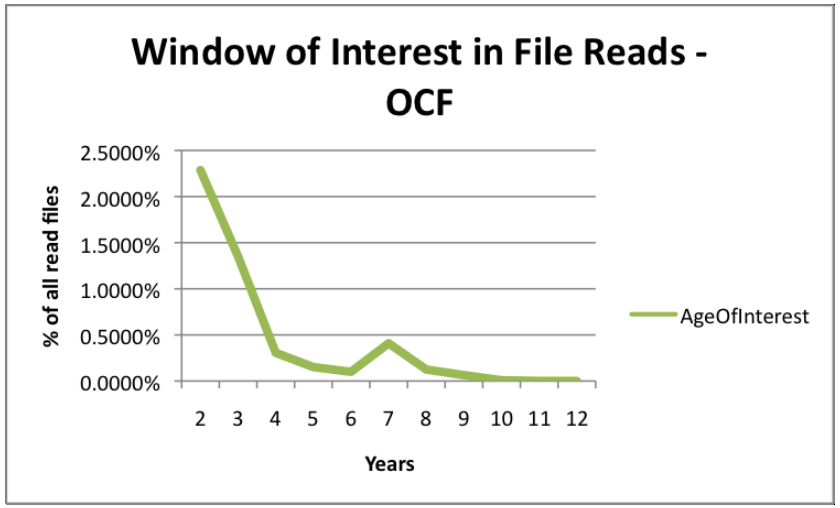
Disk Benchmark

- XIO : Simple and lightweight disks benchmark
  - Multithread / Multiplatform (Linux, AIX, SunOS)
  - Works on FS & RAW devices
  - Able to define different simultaneous I/O workload
  - Applications profiles :
    - dCache : 128 Thread, R&W 1MiB sequential blocks, 60 % read operations
    - Xrootd : 128 Thread, write 1MiB sequential block, read 16 KiB random blocks, 95% read operation
  - Output statistics in CSV
  - Used for hardware purchase/tests to evaluate performances for the last 5 years

- mfiles/cfiles : Simple filesystems benchmark
  - Filesystems benchmark/disk stress tools
  - Small Configuration file for creating large filesystem (M files)
  - Read & control files (compute checksum) → CPU & Disk burning
  - Used to control the real power usage

Data lifecycle





Todd Herr / LLNL
HPSS User Forum 2010

- Reasons:
  - Personal production, archiving, log files stored with data files, …
  - Tape backend for storage middleware (dCache, …)
  - The storage is infinite from user the user point of view !
- Idea :
  - Involve the experiments (customer) representative to manage the users data
  - Define a data lifecycle with experiments

Summary

- dCache
  - LCG : v1.9.5r22 (LCG / v1.9.0r9 (EGEE)
  - +300 pools on 140 servers (x4540 Thor on Solaris10/ZFS)
  - 4,8 PiB
  - 16 M files
- SRB
  - V 3.5
  - 250 TiB disk + 3 TiB on HPSS
  - 8 Thor 32 TB 10/ZFS + MCAT on oracle
- Xrootd :
  - (march 2010 Releasse)
  - 30 Thor, 32/64 TO
  - 1.2 PiB
- iRods
  - V2.4.1
  - 9 Thor 32 TB
- HPSS
  - V 6.2.3 on AIX p550
  - 9.5 PB / 26 M files
  - + 2000 different users
  - 60 drive T10K-B / 34 drives T10K-A on 39 AIX (p505/p510/p520) Tape servers
  - 12 IBM x3650 attached to 4 DDN Disk (4*160 TiB) / 2 FC 4Gb / 10 Gb Eth
- AFS
  - V1.4.12.1
  - 71 TB
  - 50 servers
- GPFS
  - V 3.2.1r22
  - ~ 900 TiB, 170 M files , 60 filesystemes
  - 750 TiB on DCS9550, 16 I/O node IBM x3650
  - 100 TiB on DS8300 (Metadata & small FS)
- TSM
  - 2 servers (AIX 5) w/ 4 TiB on DS8300
  - 950 M files,
  - 16 LTO 4 drives
  - 1500 LTO 4
  - 3 TB day