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

HPSS and Tape storage at IN2P3

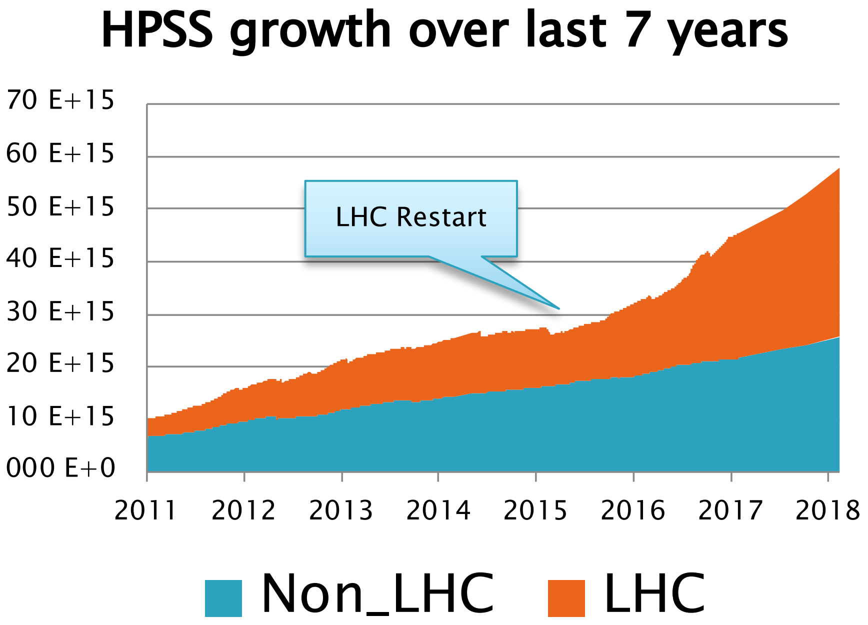
Pierre-Emmanuel Brinette, 2018-02-13
FJPPL 2018

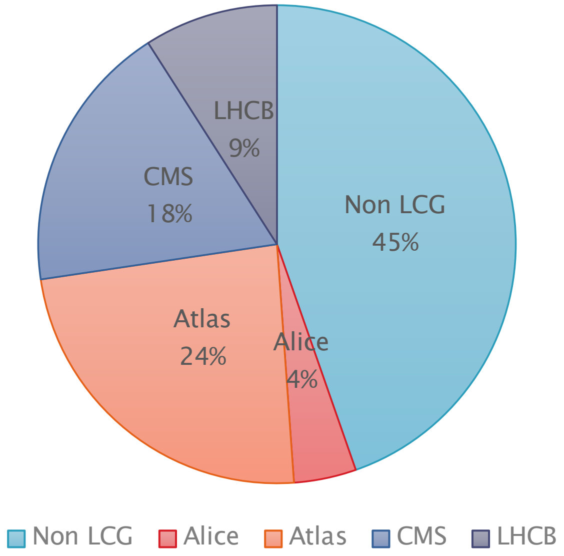
Overview

- HPSS and TREQS overview
- Tape infrastructure and evolution
- HPSS 7.5 Migration

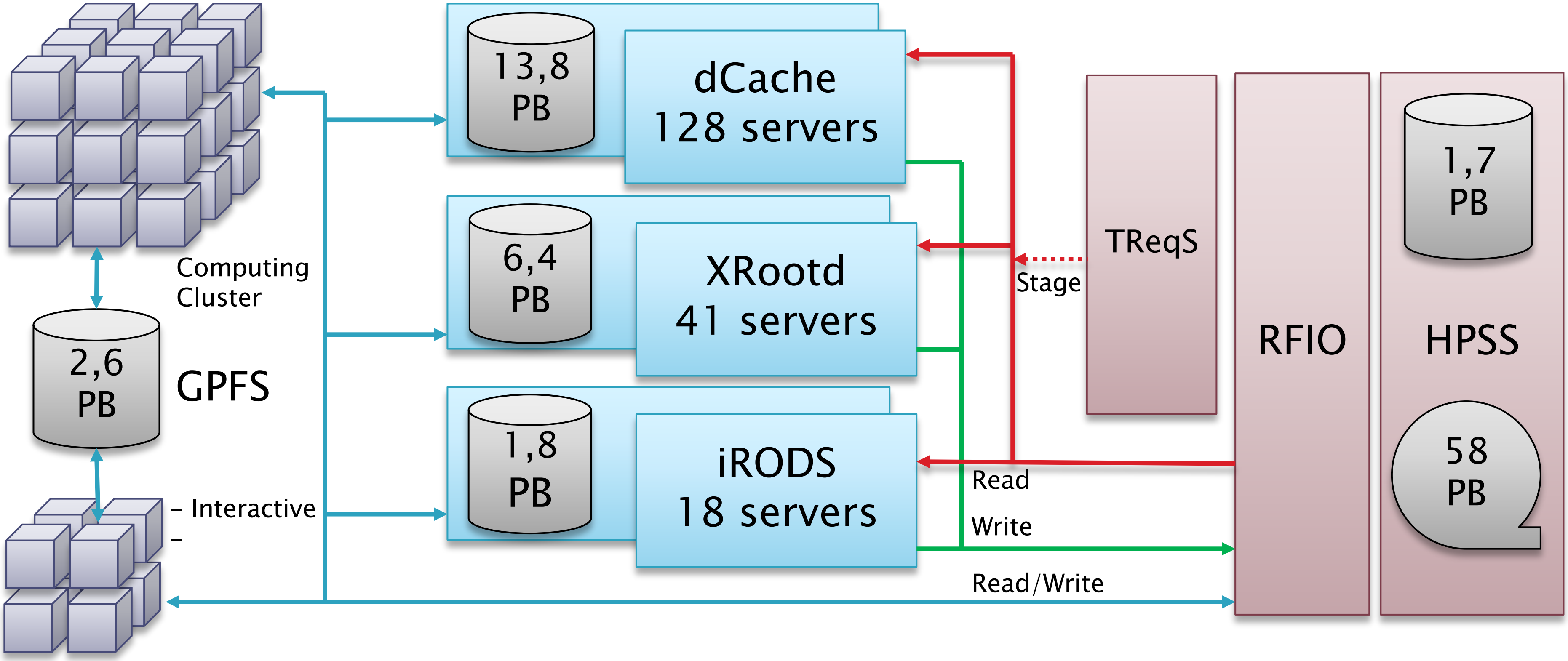
HPSS Overview

- HPSS is the main repository for scientific data
  - 80 different VO (groups) store data in HPSS
  - 55 % used for LHC data (Alice, Atlas, CMS, LHCb)
- Usage (Feb 2018)
  - 58 PB stored
  - 75 M of files
- Evolution over last year +11,7 PB (+26 %)
  - LCG : +8 PB (+34 %)
  - Non LCG : +3,7 PB (+ 17%)
- Forecast for 2018 : + 16 PB (~ 2000 tapes)





HPSS Overview



- HPSS v7.4.3p2
- HPSS Interface : RFIO with HPSS extensions
- 85 % of HPSS access are performed through storage middleware
  - **dCache** (LCG/egee),
  - **Xrootd** and **iRods**
- Still some direct access to HPSS but decreasing

- Disk cache renewed in 2017
  - + 8 new movers (DELL R730xd)
  - Total 13 movers (1,7 PB) @ 10Gbits
- Read operations from storage middleware are handled by TREQS 2

TREQS 2 Overview

- TREQS 2 is the IN2P3 tape scheduler for HPSS
  - Optimize read operations by sorting files by tapes and positions
  - Reduce the number of mounts / dismounts of the same tape.
  - Limit the number of drives used for staging
- Fully in production since June 2017
  - 4,5 M files / 8,5 PB proceed
- Features detailed at HUF 2017 [1]
- Product stable, no new development since the HUF.

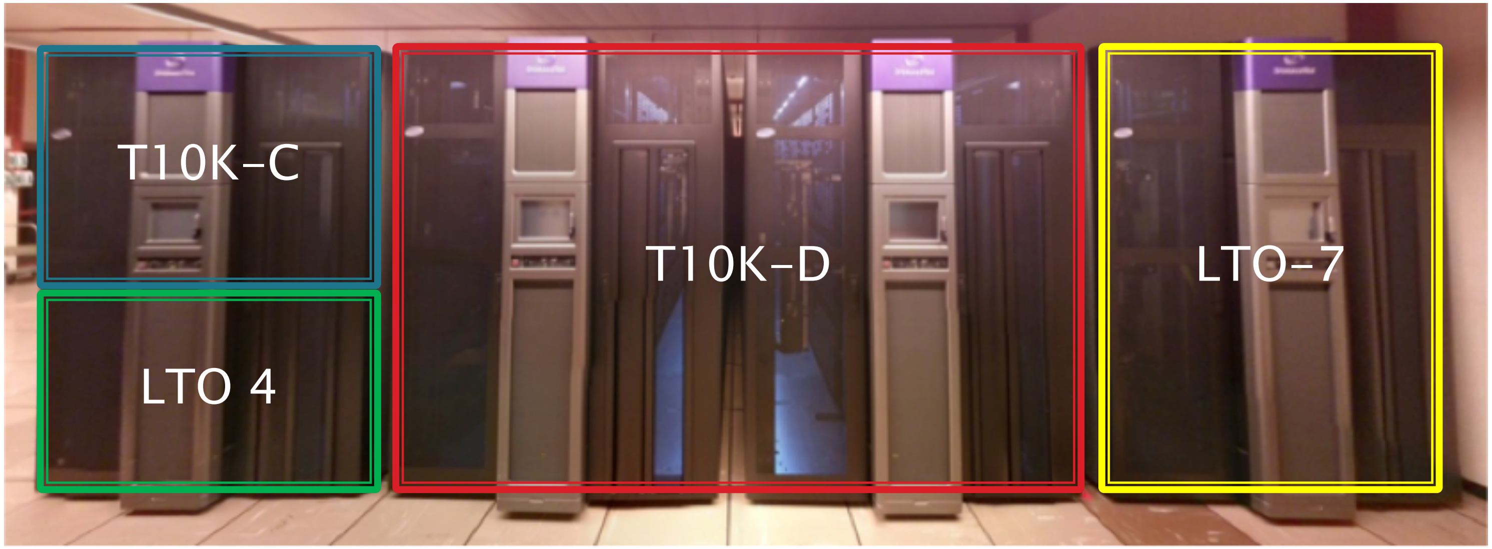
Tape infrastructure

- Tape Libraries
  - 4 Oracle SL8500 Libraries
  - Interconnected (with PTP)
  - Collocated with TSM (backup)
- 130 Tapes drives
  - T10K-B/C out of warranty used on tests system
  - LTO 4/6 used for TSM
- 50 Tapes drives in production for HPSS
  - 50 T10K-D (8,5 TB on T10K-T2)
  - +6 T10K-D (in Q1-2018)

- 22 000 Tapes

11500	T10000T2 (8,5 TB)
5 000	LTO 4
2 000	LTO 6
3 500	T10000T1 (to destroy)

- Daily tape mounts:
  - 2 000 average
  - > 6 000 peak
- HPSS Repacks
  - 23,000 T1 → T2 proceed in 2 years
  - 2,000 T10K-C → T10K-D in 2017



Tape infrastructure evolution

- Oracle stopped developing “Enterprise drives” (T10000)
  - T10000-E drives won’t be marketed
  - Need to move to a new technology

- 2 scenarios :
  - Move to IBM Entreprise class tapes drives (Jaguar)
  - Keep our libraries and use LTO 8 drives.

- IBM Enterprise tapes (Jaguar) :
  - Native capacity : 15 TB on a JD cartridge (TS1155)
  - Short media (”Sport” Tape) for storing small files.
  - Drive support latest’s advanced features
    - 64 landing zone allowing fast positioning
    - Tape Ordered Recall and End To End Data integrity
  - Drive is NOT supported on Oracle libraries → Need to purchase new libraries

- LTO 8
  - Native Capacity : 12 TB
  - Media cost 25% lower than Enterprise tape and may decrease quickly.
  - Use the same R/W head than Jaguar (TMR) head and BeFe media.
  - But Only 2 landing zones → Performance lower on random recall.
  - Advanced features not supported (TOR and E2EDI)

Tape infrastructure evolution

- Choice not evident
  - Reliability/performances of the LTO drives / media ?
    - LTO tapes can support our workload (6000 mount/day) ?
    - Today, we “break” about 10 drives T10K-D per month.
  - Service and support ?
    - Today, T10K-D drives are monitored by Oracle SDP2
    - Service Request opened automatically when a drive fail.
  - Our libraries getting old (10 years)
    - Maintenance cost will increase by 50 %
  - How long Oracle will continue in the tape business ?

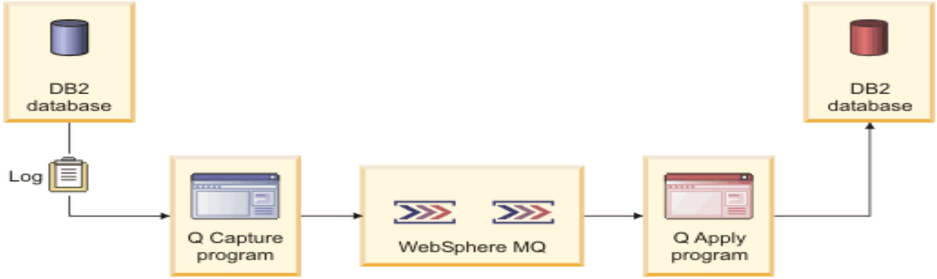
- Preliminary tests started on LTO-7
  - Tape filled with 2GB files
  - Good performances on LTO-7 at migration (writing)
    - Close to 300 MB/s
  - Read operations slower on LTO-7
    - Positioning slower on LTO-7 vs T10K-D (-10% to -30%)
    - But performance similar using Treqs (!)
  - Tests has to be made with small / medium files size (10 to 100 MB) and aggregates

- LTO 8 Tests planned in Q2-2018

Migration to HPSS 7.5

- HPSS 7.5.1 is the new major HPSS version
  - Features presented by J. Procknow at HUF 2017 [2]
  - Database partitioning
  - End To End Data Integrity
  - Tape Ordered Recall + ’Quaid’
  - Many changes in the metadata schema
    - Redesigned for improving NS performances (files creation / deletion)
    - SOID reduced from 32 bytes to 19 bytes

- Migration based on QREP
  - Designed to reduce downtime
  - Metadata converted while HPSS running



- Two scenarios :
  - In place metadata conversion (on the same machine)
  - Server to server conversion (data replicated and converted on a target server)

Migration to HPSS 7.5

- Started to migrate the test environment
  - HPSS 7.4.3p2 on Openstack VM (RHEL 6.9)
  - 3 subsystems and about 1.1 millions of files
  - Scenario 2 : Migration on a new machine (RHEL 7.4)
  - Documentation and tools provided by HPSS support
    - QREP and a set of python scripts
    - IBM Websphere + DB2 licence
- My feedbacks :
  - Some mistakes in the documentation
    - It's not clear which commands has to be run on the source or target server
    - Files and directories permissions has to be tuned
  - Many component need to be deployed on servers
    - Python 2.7.5 must be compiled for RHEL 6.9 servers
      - DB2 python module > 2.0.4 doesn't works
    - Websphere MQ use 10 GB is on the root filesystem
      - Need to create a dedicated partition
  - All the DB must be catalogued on both nodes
    - Both servers are able to access to source an target DB
    - But databases must be catalogued in different way depending the host
  - DB2 Instance need to be restarted anyway
      - To upgrade DB2 v10.5 fp8
      - To set Federated mode
  - Hard to troubleshot : Sometime no errors messages, but nothing happens

Migration to HPSS 7.5

My feedbacks (cont)

- Bug detected at “Verify” step
  - Problem due to default collating sequence of the DB that change the “ORDER BY” results
  - On source DB, default values is “SYSTEM_819” and on target DB, default value is “INDENTITY”
  - Problem quickly identified the HPSS support and a fix was delivered
- Some operations take lot time :
  - Ie : Initial load of the DB (“activate” step)
  - More 2h for 1,1 M files
  - → should take at least 24h on the production system
- Some commands are confusing :
  - ie : stop capture

    ```
    ./manage_qrep.py -c -s 1 -s 2 -s 3 --stop_capture
    ```

  - ie : restart replication after a reboot :

    ```
    ./manage_qrep.py -c -s 1 -s 2 -s 3 --stop capture --start_capture
    ```

Current status :

- Target databases synced with the sources databases
- Each changes on the source (while hpss running) is applied within ms on the target
- Next step : Stop the replication and switch HPSS to the target server

Schedule for the production :

- March 2018 : Setup QREP and start replication
- June 2018 : Migrate to HPSS v7.5.1p2

Thank you

[1] https://conference-indico.kek.jp/indico/event/28/session/10/contribution/25
[2] https://conference-indico.kek.jp/indico/event/28/session/6/contribution/9