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

HPSS at IN2P3

Pierre-Emmanuel Brinette,
2019-06-24

CCIN2P3

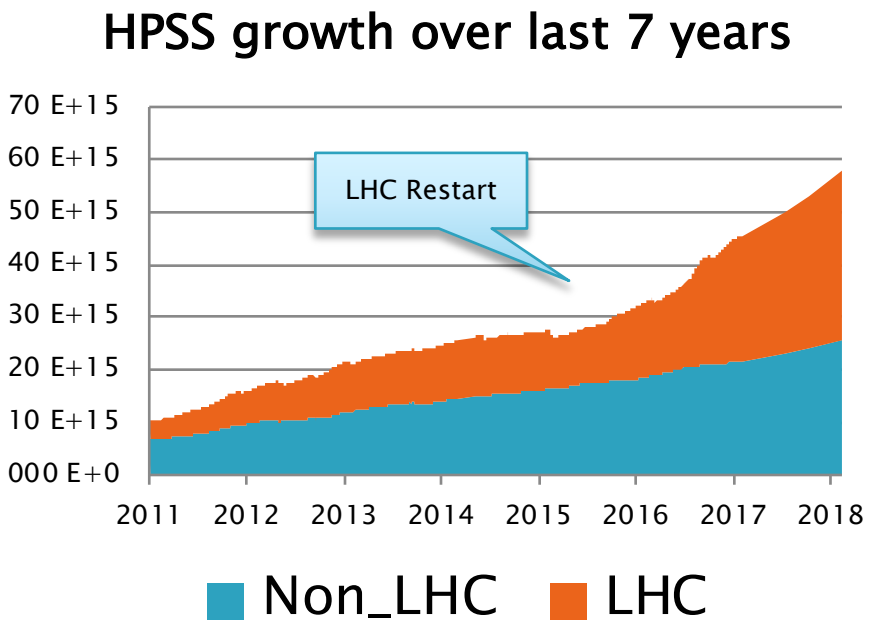
cnrs

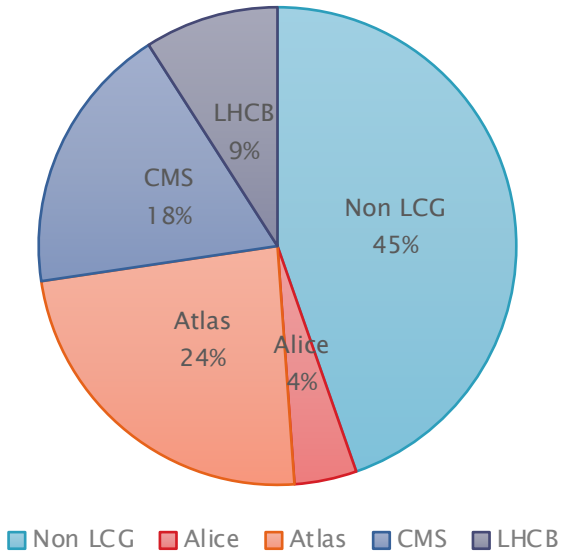
Overview

- HPSS and TREQS overview
- HPSS Monitoring
- Tape infrastructure and evolution
- Future

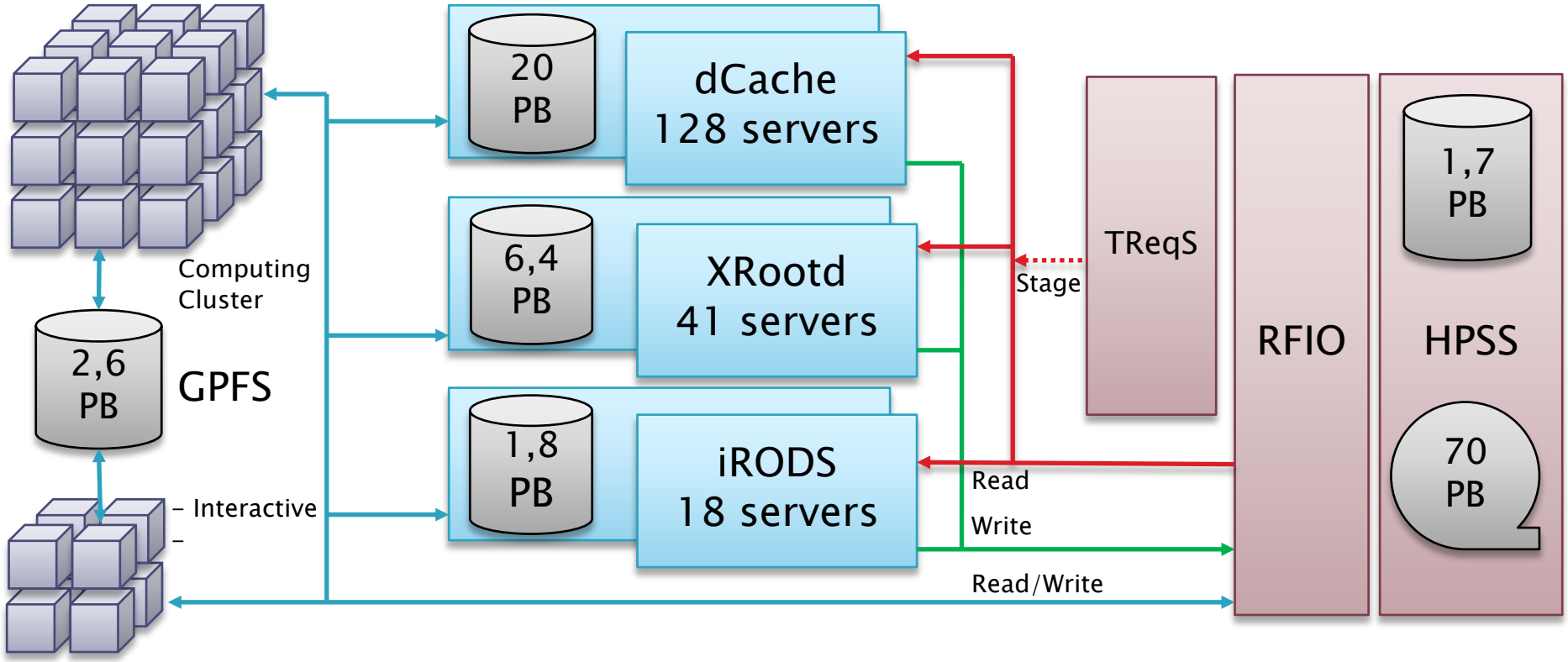
HPSS Overview

- 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)
- In production since 2001
- Usage (Jun 2019)
  - 70 PB stored
  - 78 M of files
- Evolution over last year +11,7 PB (+26 %)
  - LCG : +8 PB (+34 %)
  - Non LCG : +3,7 PB (+ 17%)
- Forecast for 2019 : + 16 PB (~ 2000 tapes)





HPSS Overview

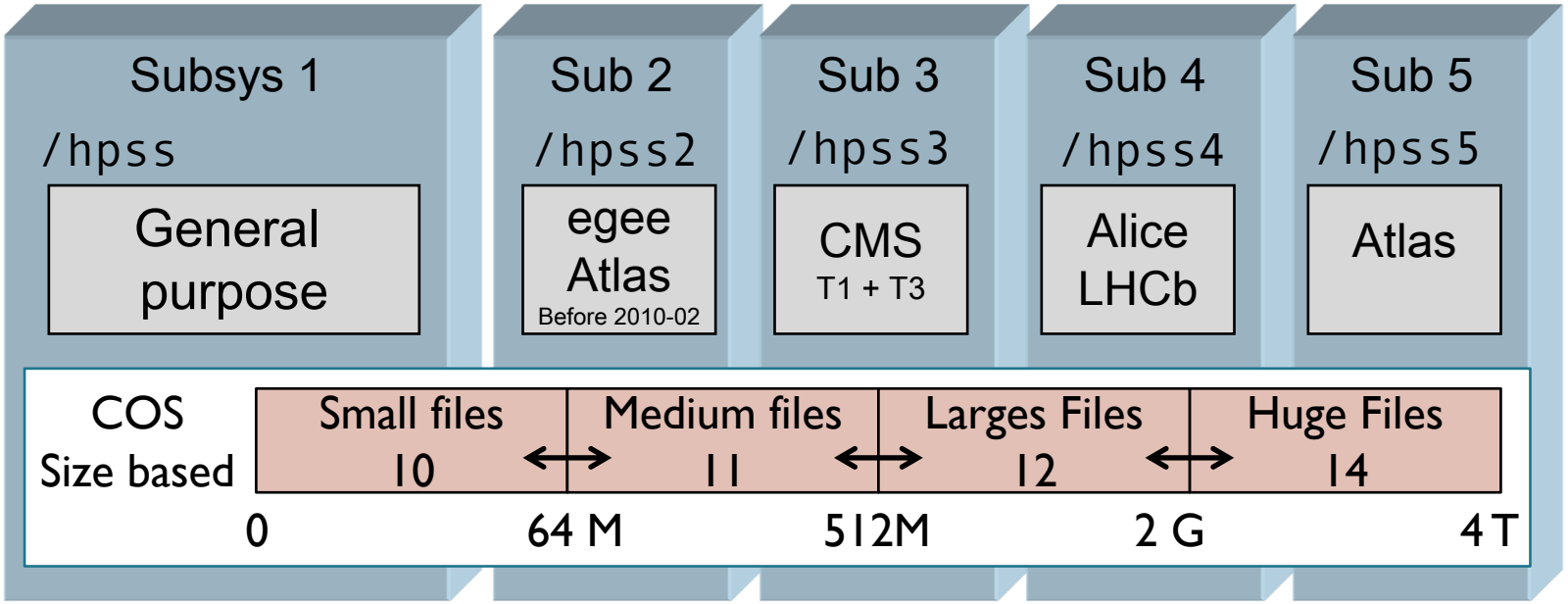


- HPSS hpss-7.5.1.2-20190116.u9
- 85 % of HPSS access are performed through storage middleware
  - **dCache** (LCG/egee),
  - **Xrootd** and **iRods**
- Still some direct access to HPSS but decreasing

- HPSS Interface :
  - RFIO with HPSS extensions
  - Read operations from storage middleware are handled by TREQS 2

HPSS Storage policy

- 5 subsystems, 4 COS Only (selected by size), 10th file families
- Different tape resources per COS (ie. Small files on “Sport” tapes)



- Historical
  - 24 PB
  - ~500 UID
  - 47 M files
- Newly created
  - 46 PB
  - 31 M files
  - Mainly used for LHC Data
- Dedicated subsystem
  - Allow to dedicate DISK resources for specific set of users when using **automatic COS selection**
  - Specific database for a set users → faster query
    - Subsys 1 : 40 GB
    - Subsys [2-5] : 1.5 to 10 GB

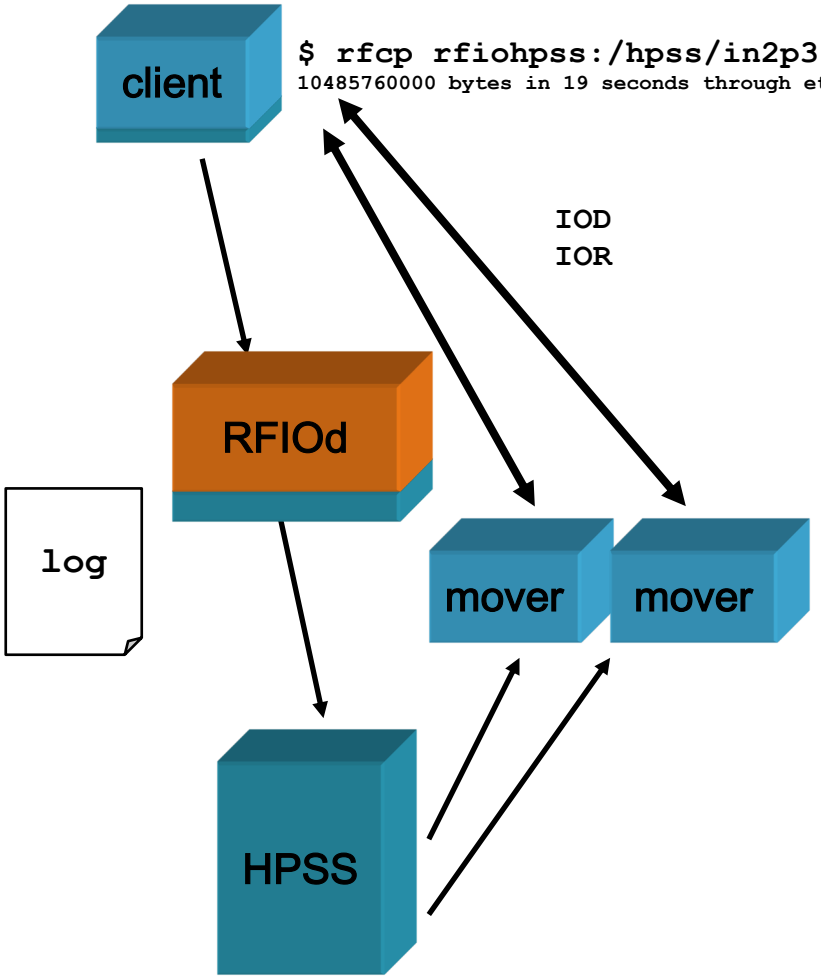
- Commodity hardware

- Core servers :
  - 2 x DELL R720 + MD3220
  - « Manual Failover »
  - RHEL 7
  - In production since 2013
  - EOL 2020

- Disk movers :
  - 12 x DELL R730xd + MD 1200
  - Hardware RAID 6
  - 10 Gbits
  - CentOS 7
  - Total : 1,7 Po

- Tape Movers :
  - 9x DELL R720/R640
  - 10 Gbits
  - SL 6 / CentOS 7
  - 6 drives T10K-D / mover

RFIO : IN2P3 HPSS Copytool



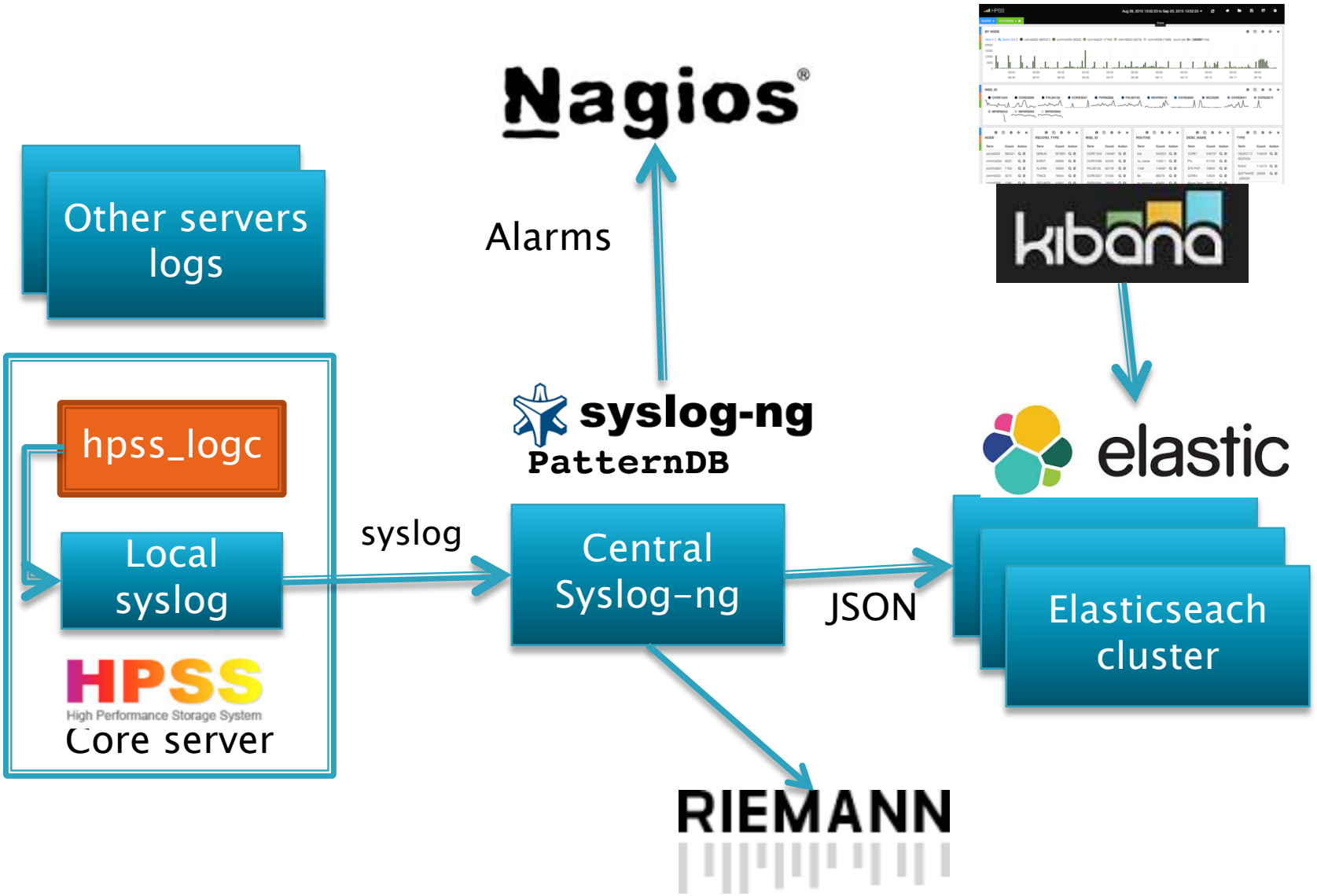
- Designed by CERN and maintained by CC-IN2P3
  - Provide Unix like command to end users : rfcp , rfdir, rfrm, …
  - rfcp client use HPSS API
    - `hpss_readlist`
    - `hpss_writelist`
- Benefits
  - Good performances, direct transfers from movers to clients
  - hpss libraries statically linked on the client
  - Access trough a control server
    - Limit simultaneous Cx
    - Access Logging
- Not compatible anymore with package dpm of EGI Grid MW

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
  - 8 M files / 14,5 PB proceed
    - 2 M files on cache
- Features detailed at HUF 2017 [1]
- Product stable, no new development since the HUF.
- Code available for the HPSS community
  - https://gitlab.in2p3.fr/cc-in2p3-dev/treqs2
  - License : GPLv3
  - Account opened on request
- See Bernard Presentation

HPSS Monitoring

Log based monitoring



Treqs Kibana based dashboard :

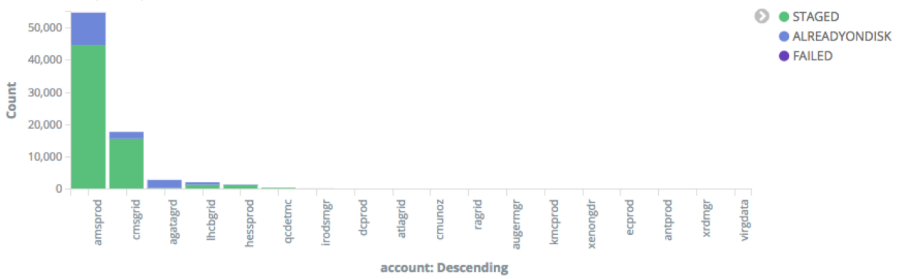
81,035
Requests

109.947TB
Total Size

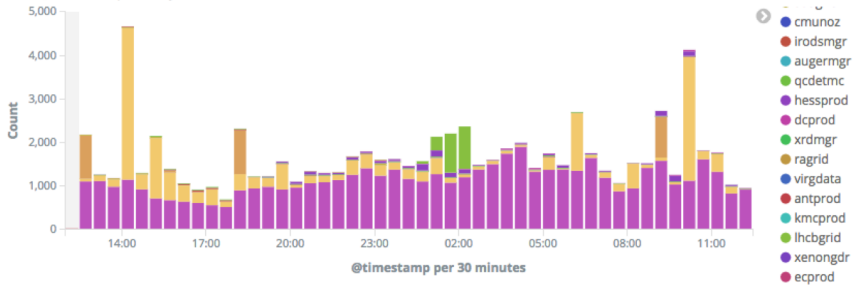
Status	Count	File size
STAGED	64,885	91.115TB
ALREADYONDISK	16,108	18.79TB
FAILED	42	43.245GB

Export: Raw Formatted

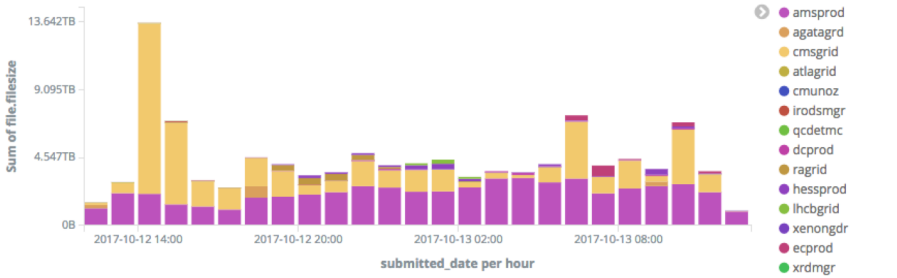
TREQS2 : Requetes par utilisateurs



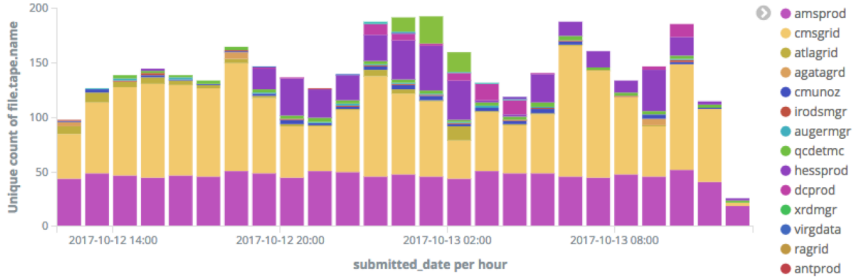
TREQS2: File requests by hour



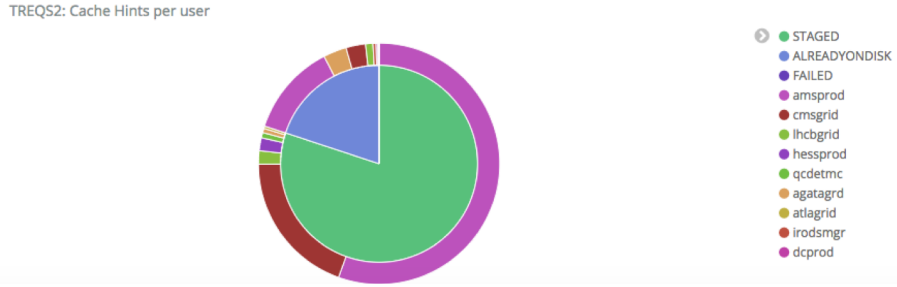
TREQS2: Stage rate by users



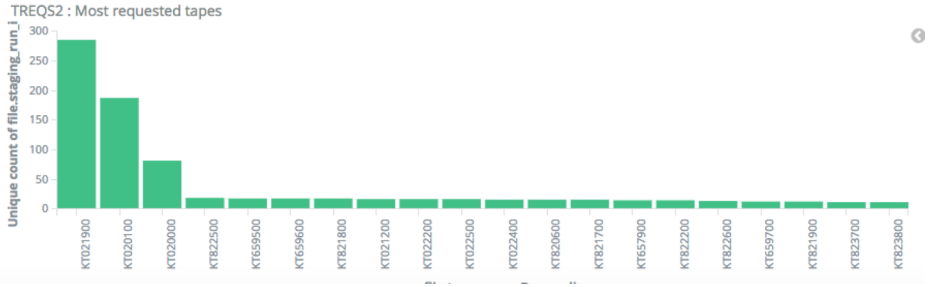
TREQS2: Tape count by users



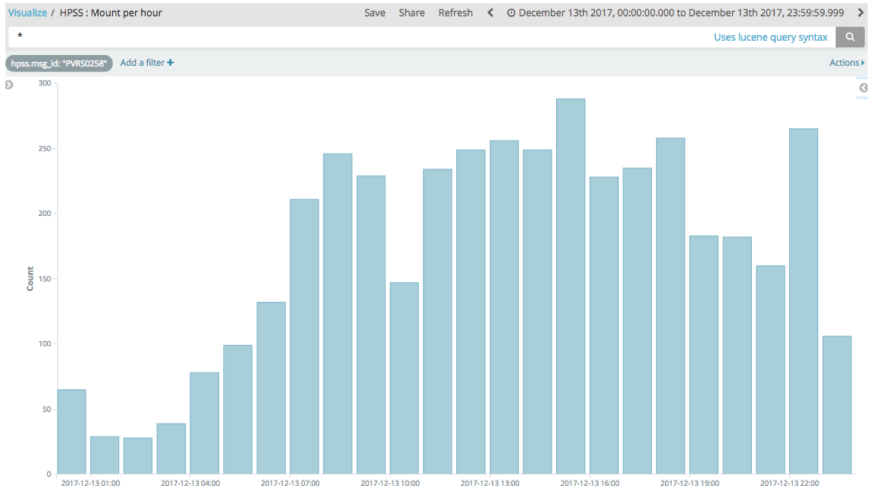
TREQS2: Cache Hints per user



TREQS2 : Most requested tapes



HPSS Kibana Dashboard



Visualize / HPSS : Migrations / purges stats — Save Share Refresh — December 13th 2017, 00:00:00.000 to December 13th 2017, 23:59:59.999

Uses lucene query syntax

hpss.sclass_id: "61" hpss.sclass_id: "1" Add a filter + Actions

Subsys	SC	Migrated bytes	Migrated Files	Purged bytes	Purged Files
1	10	1.727GB	256	0B	0
1	11	125.176GB	509	0B	0
1	12	367.804GB	347	0B	0
1	14	15.897TB	1,224	64.634TB	5,260
2	10	21.608GB	2,755	153.75GB	10,285
2	11	102.348GB	870	250.344GB	1,766
2	12	302.106GB	267	0B	0
2	14	2.621TB	810	0B	0
3	10	0B	0	0B	0
3	11	0B	0	0B	0
3	12	0B	0	0B	0
3	14	786.308GB	198	17.675TB	4,762
4	10	9.294GB	386	10.403GB	227
4	11	18.218GB	101	82.109GB	464
4	12	673.765GB	905	0B	0
4	14	4.369TB	1,152	29.136TB	8,266
5	10	12.325GB	1,280	0B	0
5	11	100.044GB	1,045	0B	0
5	12	1,007.754GB	1,257	0B	0
5	14	43.466TB	5,833	56.903TB	22,356
		69.8TB	**19,195**	**168.833TB**	**53,386**

- HPSS Dashboard :
  - Mount per hour
  - Migration/purge stats

Tape Libaries evolution

Tape infrastructure

- Tape Libraries
  - 4 Oracle SL8500 Libraries
  - Interconnected (with PTP)
  - Collocated with TSM (backup)
- 116 Tapes drives
  - 22 T10K-C (EOL)
  - 20 LTO 4 (TSM)
  - 17 LTO 7 (TSM)
  - 56 T10K-D (HPSS)
  - 1 LTO8 (test)

- ACSLS 8.5 on RHEL 7
- 20 000 Tapes
  - 13000 T10000T2 (8,5 TB)
  - 5 000 LTO 4
  - 2 000 LTO 6
- Daily tape mounts:
  - 2 000 average
  - > 6 000 peak (300/h)

Tape infrastructure evolution

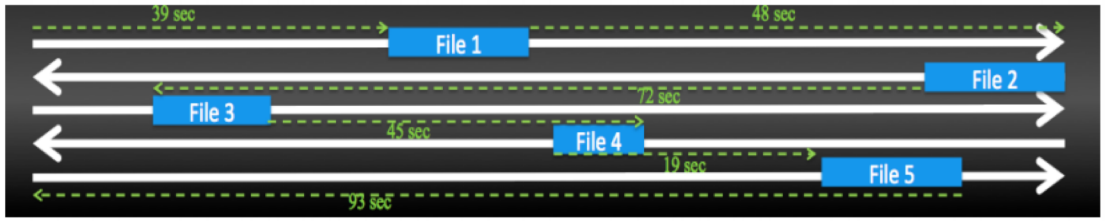
- Oracle stopped developing “Enterprise drives” (T10000)
  - T10000-E drives won’t be marketed
  - Need to move to a new technology

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

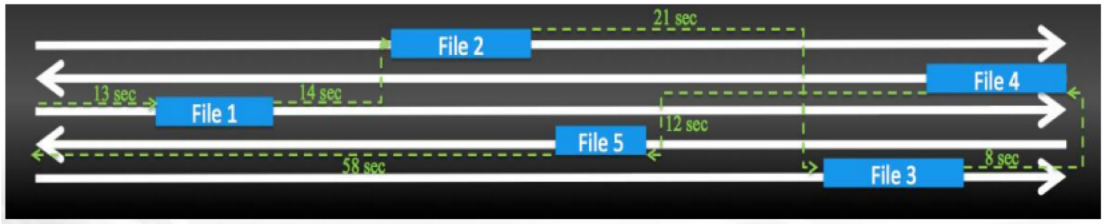
- IBM Enterprise tapes (Jaguar) :
  - Native capacity :
    - 15 TB on a JD cartridge (TS1155)
    - 20 TB on a JE cartridge (TS1160)
  - Short media (”Sport” Tape) for storing small files.
  - Drive support latest’s advanced features
    - 64 landing zone allowing fast positioning
    - Tape Ordered 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)

- RAO : Recommanded access ordering
  - Drive feature to find the better path to recall a bunch of files.
  - Fully available since HPSS 7.5.1.2
  - Features only supported on « Enterprise » tapes drives



Seqquential Read :
326 s



RAO Optimized Read :
126 s
Gain : 151 %

« Performance Evaluation for Tape Storage Data Recall with TS1150 Drive »
Guangwei Che – BNL – HUF 2018

RAO Tests : T10K-D vs LTO8

- Tests with HPSS 7.5

- Recall of tests files with and without RAO
  - Sample :
    - Tests files : 2200 MB
    - 1 T10K-D filled 3646 files
    - 1 LTO-8 filled 5205 files
    - Samples drawn randomly
    - Staging with hpss_cache (ordered and non ordered) and quaid (RAO)

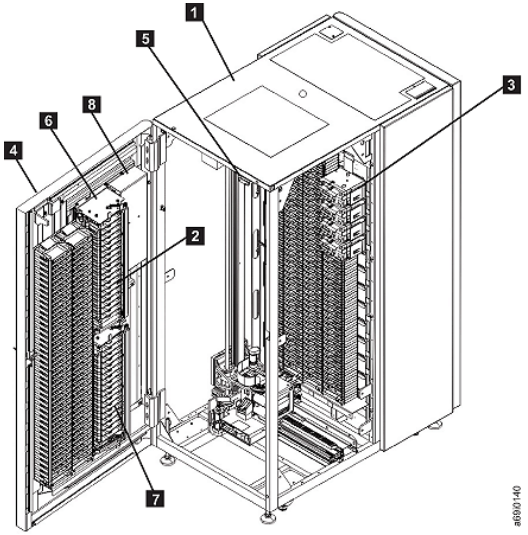
Test	Sample	Duration	Rate
T10K-D : Unordered read	25 files	19m41s	41 MB/s
T10K-D : Offset ordered read	25 files	19m05s	48 MB/s
	50 files	34m58s	58 MB /s
T10K-D : RAO ordered read	25 files	8m0s	114 MB/s (gain : 137%)
	50 files	13m10s	139 MB /s (gain: 139%)
LTO8 : Unordered read	25 files	23m26s	39 MB/s
LTO8 : Offset ordered read	25 files	24m9s	38 MB/s
LTO8 : Quaid ordered read	25 files	25m5	37 MB/s

- For CC-IN2P3 Enterprise drive seems to be the right direction
  - Today LTO 8 performances are lower than T10K-D

- Acquisition soon of a « small » libraries
  - 12 TS1160 drives
  - 20 PB minimum
  - Proposal for tender in progress
  - Put in production in HPSS in Q4 2019

- IBM TS4500 :
  - 3 frames enterprise (L25/S25)
  - 2 accessors
  - 12 TS1160

- Technical solution:
  - Frame dedicated for technology
  - Not possible to mix LTO / Enterprise within the same library
  - “Easily” upgradable
    - “Just add frames”
    - Only few hours downtime



Spectralogic Proposal

- Spectra Tfinity
  - 3 frames
  - 2 accessors
  - 12 TS1160
- Technical solution :
  - « terapack » for media storage
  - Density / terapack :
    - 9 Enterprise tapes
    - 10 LTO tapes
  - Technology mix possible within the same frames
  - Costs of media terapack ?
- Support ?
  - Only few installation in EU
- TAOS ™ :
  - Time-based Access Order System
  - Emulate RAO over LTO at library level
  - https://edge.spectralogic.com/index.cfm?&fuseaction=home.displayFile&DocID=5035

Future

Future

- Short term:
  - Migrate to hpss 7.5.3
    - New feature : drive limitation for recall
- Propose an alternative to RFIO
  - RFIO use deprecated readlist/writelist API
  - May be remove from HPSS API in HPSS 8
- Alternate tools
  - HSI / HTAR ?
  - Xrootd ?
- Expose HPSS filesystem to end users?
  - GHI ?
    - Expensive !
  - VFS over NFS
    - Problem with COS selection when creating file
    - Performances
  - Lustre HSM ?

Thank you