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

HPSS at IN2P3

Pierre-Emmanuel Brinette,
Bernard Chambon

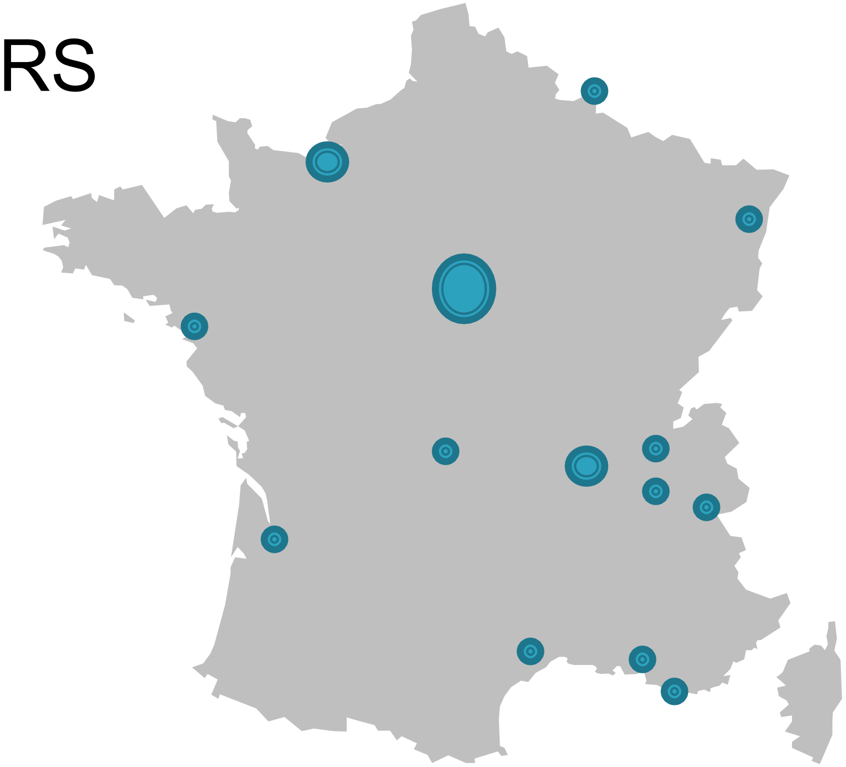
1ères rencontres HPSS France 15-16 mars 2017

Agenda

- IN2P3 in brief
- CC-IN2P3 in brief
- HPSS in detail
- TReqs2
- As a conclusion

IN2P3 in brief…

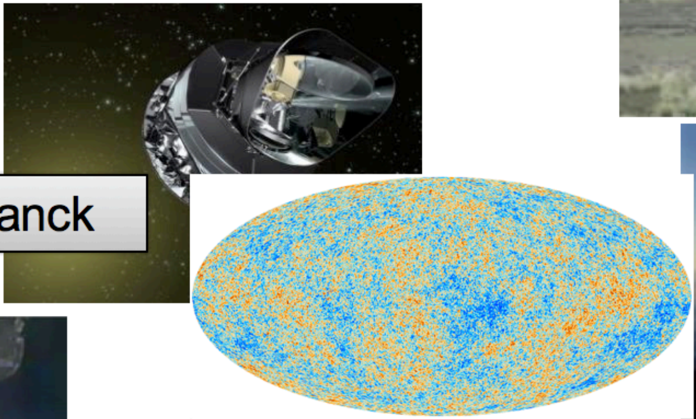
- National Research Institute for Nuclear Physics, Particle Physics and Astroparticle Physics
- One of the 10 Institutes of CNRS
- Composed of 25 laboratories
- Involve in 80 experiments
- Almost 5000 people
  - 1/3 researchers
  - 2/3 administrative, technical

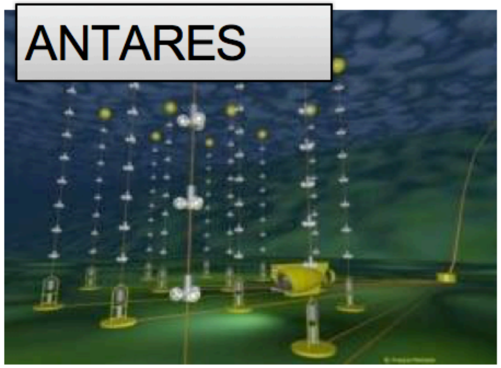


CC-IN2P3 in brief…

- Centre de Calcul de l'IN2P3 / CNRS
- Computing and data storage facilities for the IN2P3
  - Missions are to provide IT resources to the French High Energy Physics community
  - Also provide a common infrastructure for institutional services (collaborative, edms, development and project management tools…)
- People
  - 84 people (administrative, IT and facility management)
  - 74% are permanent positions, 26% are temporary
- Activities distributed across 10 teams, 7 for IT
- Provide resources to 80 experiments

Working for…





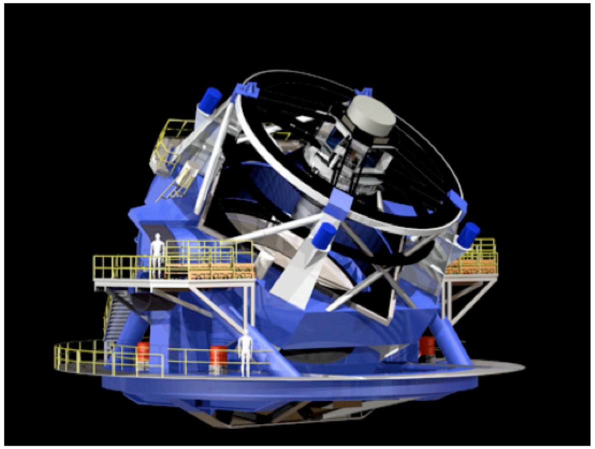
And prepare the future for…

LSST

Whole dataset available at CC-IN2P3

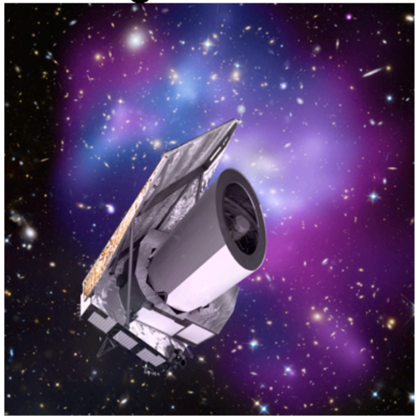
50% of the processing by CC-IN2P3

other 50% by NCSA



EUCLID

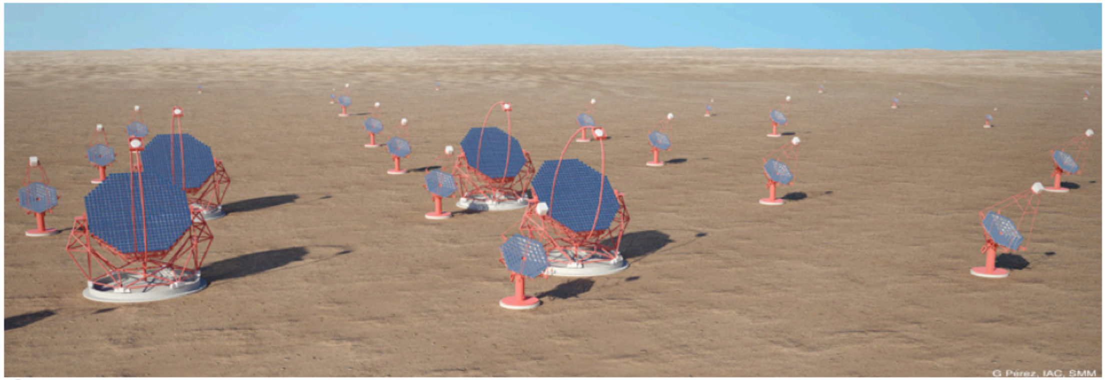
CC-IN2P3 is the French Data Center for processsing and data management



dark energy and dark matter

CTA

CC-IN2P3 should play a key role in the CTA data processing



G Pérez, IAC, SMM

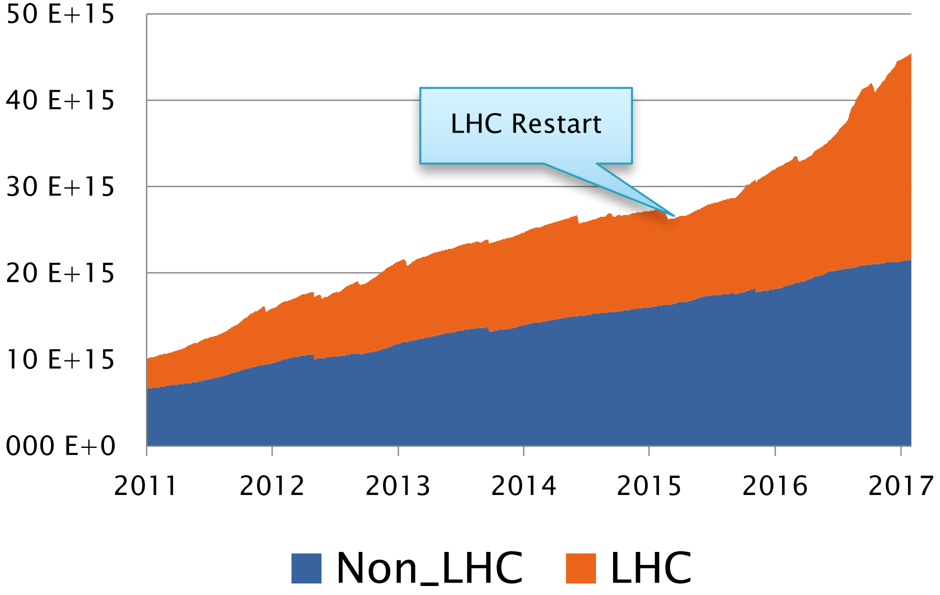
Gamma rays

HPSS

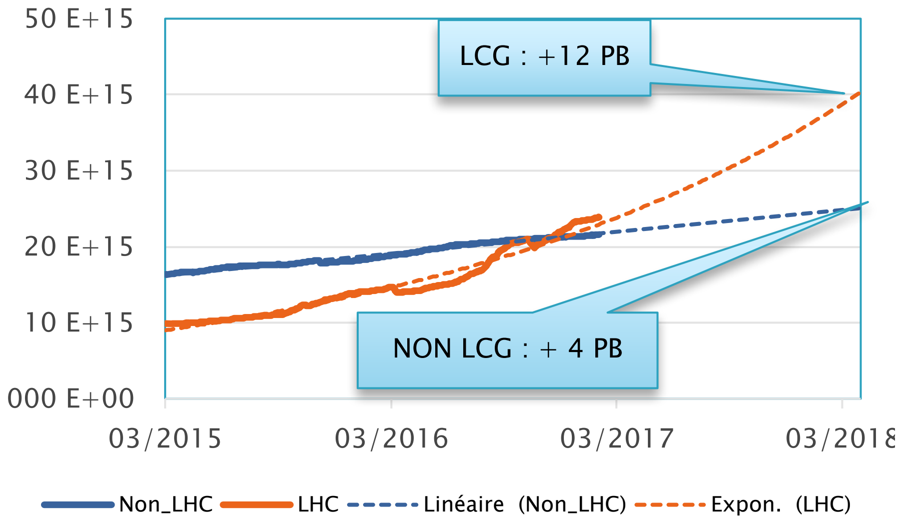
HPSS Overview

- IN2P3 is using HPSS since 1999 for BaBar experiment
- HPSS is the main repository for scientific data
  - 52 % used for LHC data (Alice, Atlas, CMS, LHCb)
- Usage (feb 2017)
  - 46 PB stored, single copy
  - 66 M of files
- Archive expected to grow up to 62 PB within next 12 months (+35 %)

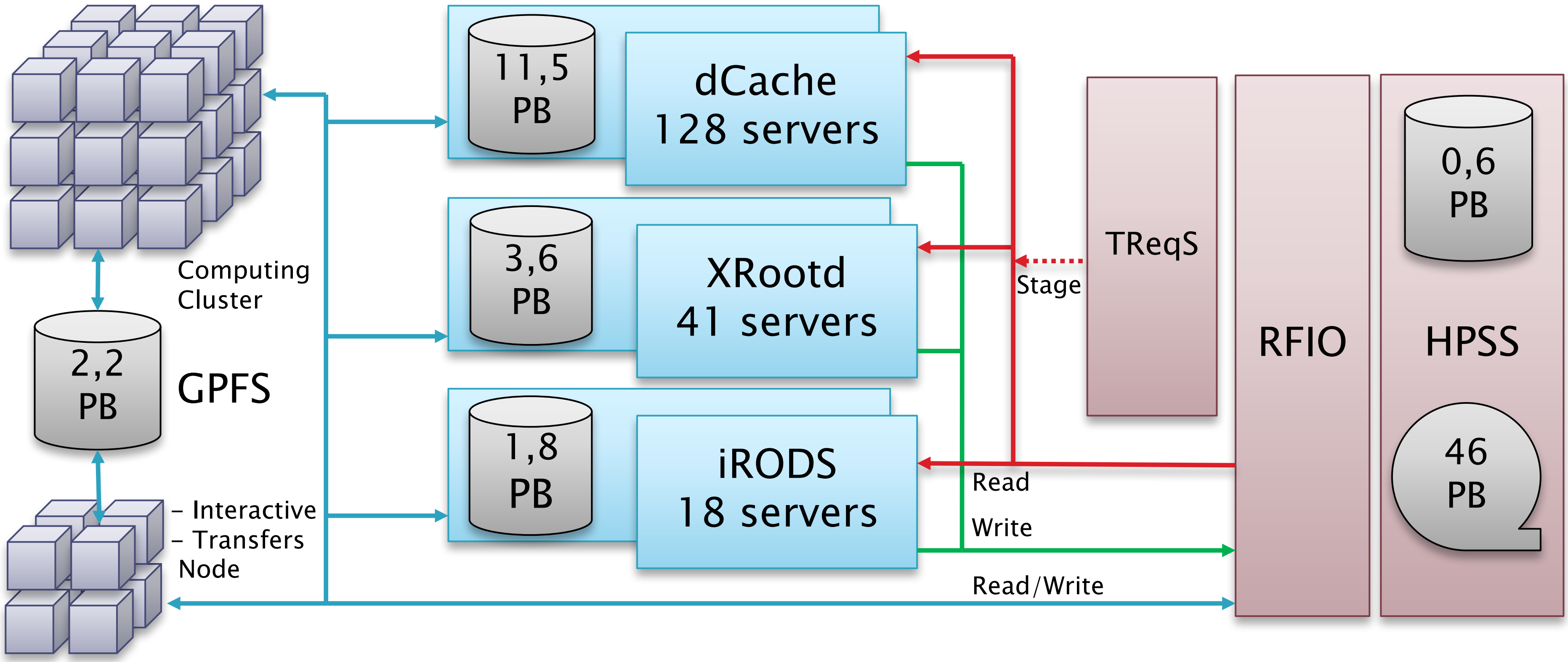
HPSS growth over last 6 years



HPSS growth forecast for next 12 month (LHC and non LHC)



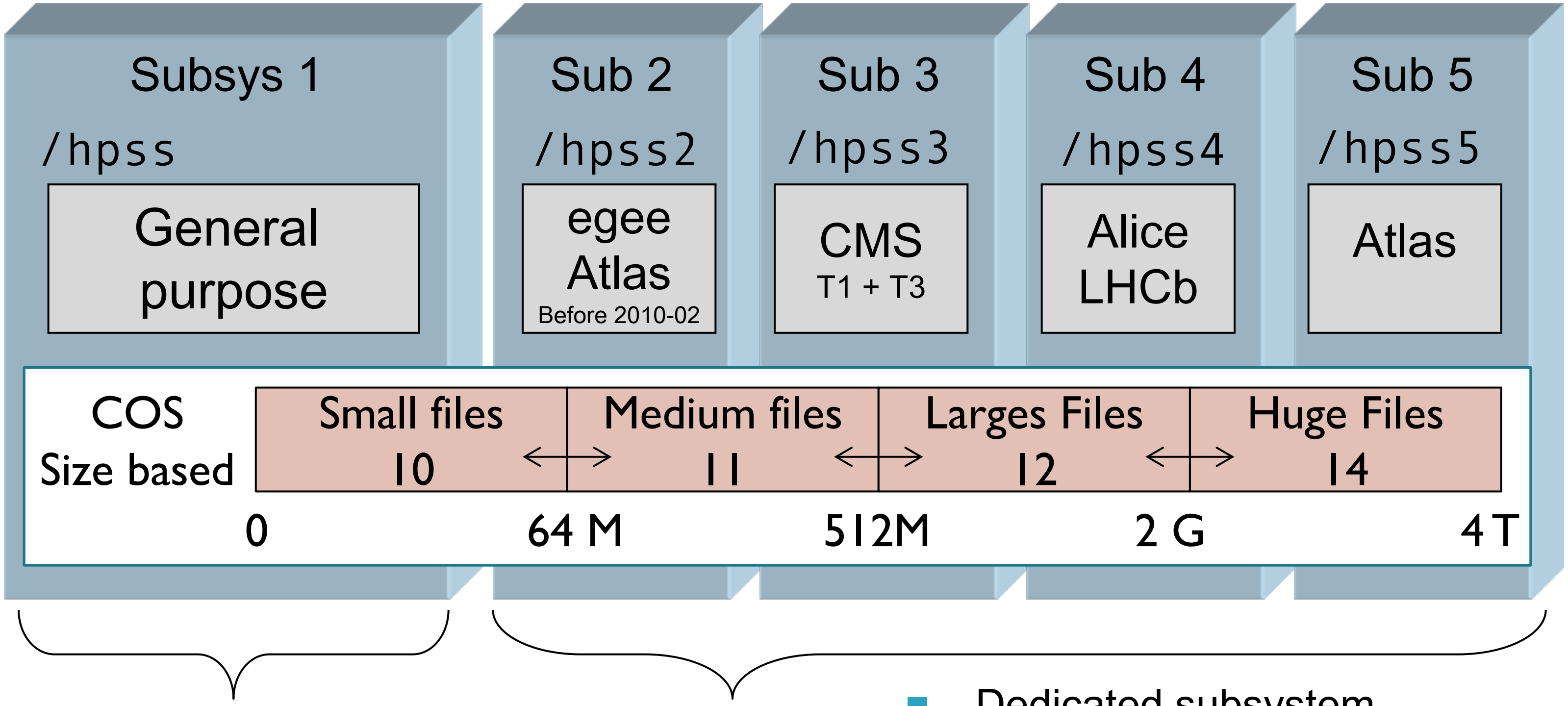
Storage Overview



- HPSS Interface : RFIO with HPSS extensions
- Historically, direct access from users/jobs to HPSS using RFIO
- Now, 85 % of access are performed through storage middleware
  - dCache (LCG/egee), Xrootd and iRods
  - Reduce stress on robotic due to the large disk cache
- Read operations from storage middleware handled by TReqS
  - Limit the numbers of drives used for large reading campaign
  - Optimize recall by sorting files on the same tape to speedup read

HPSS Storage policy

- 5 subsystems, 4 main COS Only (selected by size), tens of file families
- Different tape resources per COS (ie. Small files on “**T10K Sport**” tapes)



Subsys 1 /hpss General purpose

Sub 2 /hpss2 egee Atlas Before 2010-02

Sub 3 /hpss3 CMS T1 + T3

Sub 4 /hpss4 Alice LHCb

Sub 5 /hpss5 Atlas

COS Size based

Small files	Medium files	Larges Files	Huge Files
10	11	12	14

0 — 64 M — 512M — 2 G — 4 T

- Historical
  - 22 PB
  - ~2000 UID
  - 50 M files

- Newly created
  - 24 PB
  - 16 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 6 GB

Disk infrastructure

- DAS building block
  - 1 server + 1 disk tray
    - 1 server R730xd containing 12 disk drives
    - 1 SAS attached disk tray containing 12 disk drives
    - Hardware RAID => 120 TB
    - 10 Gbps network link

- Started with 2TB, 3TB, 4TB and now 8TB disk drives

- Scalability of this model proven by now
- Massively used for dcache/xrootd/irods
- Also used for HPSS Disk Mover (14 servers)

Tape infrastructure

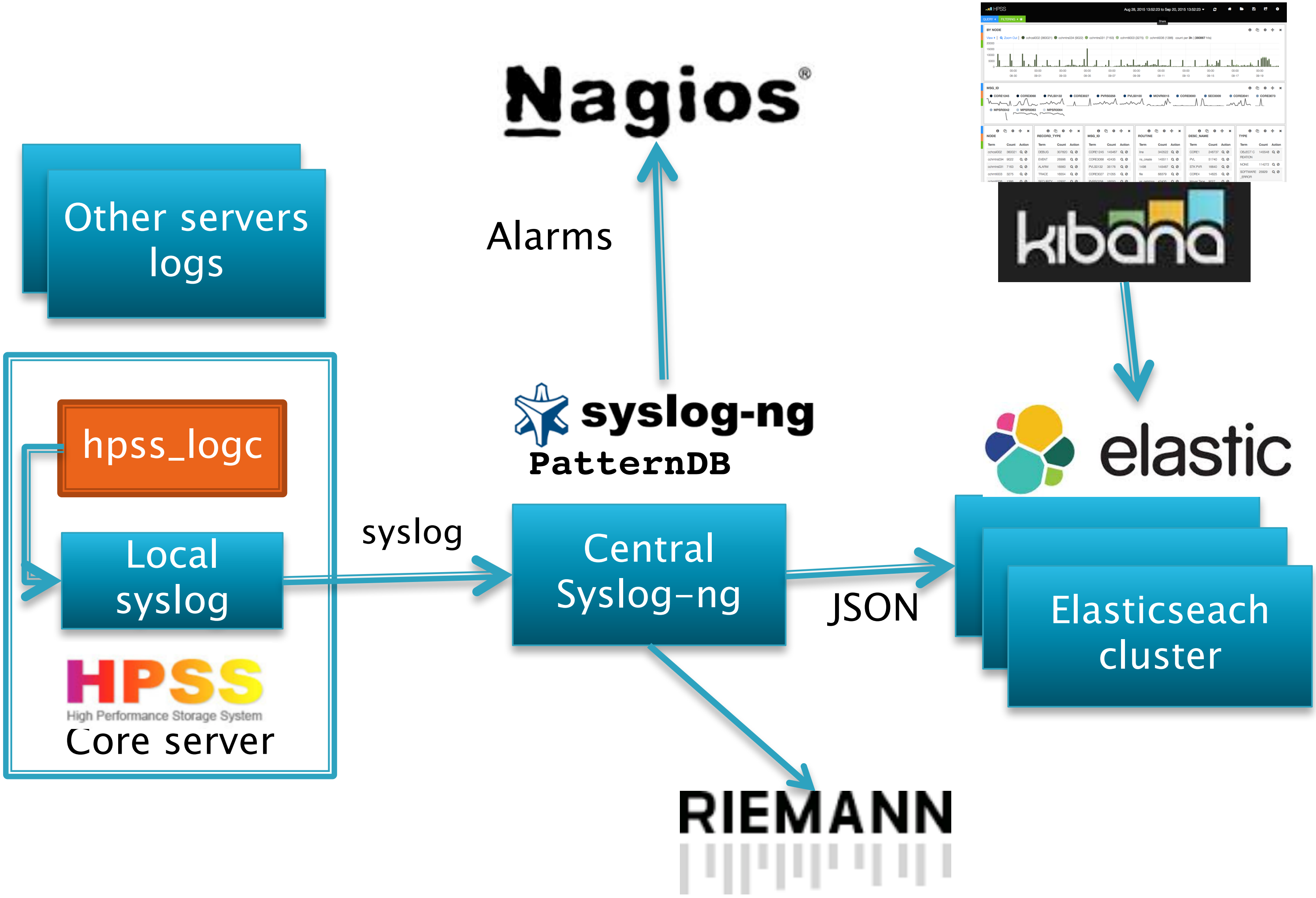
- Tape Libraries
  - 4 Oracle SL8500 Libraries
  - Interconnected (with PTP)
  - Shared with TSM (backup)
- 140 Tape drives
- 66 Tape drives for HPSS
  - 22 T10K-C (5,5 TB on T10K-T2)
  - 44 T10K-D (8,5 TB on T10K-T2)
  - +12 T10K-D (in 2017)

- 26 000 Tapes

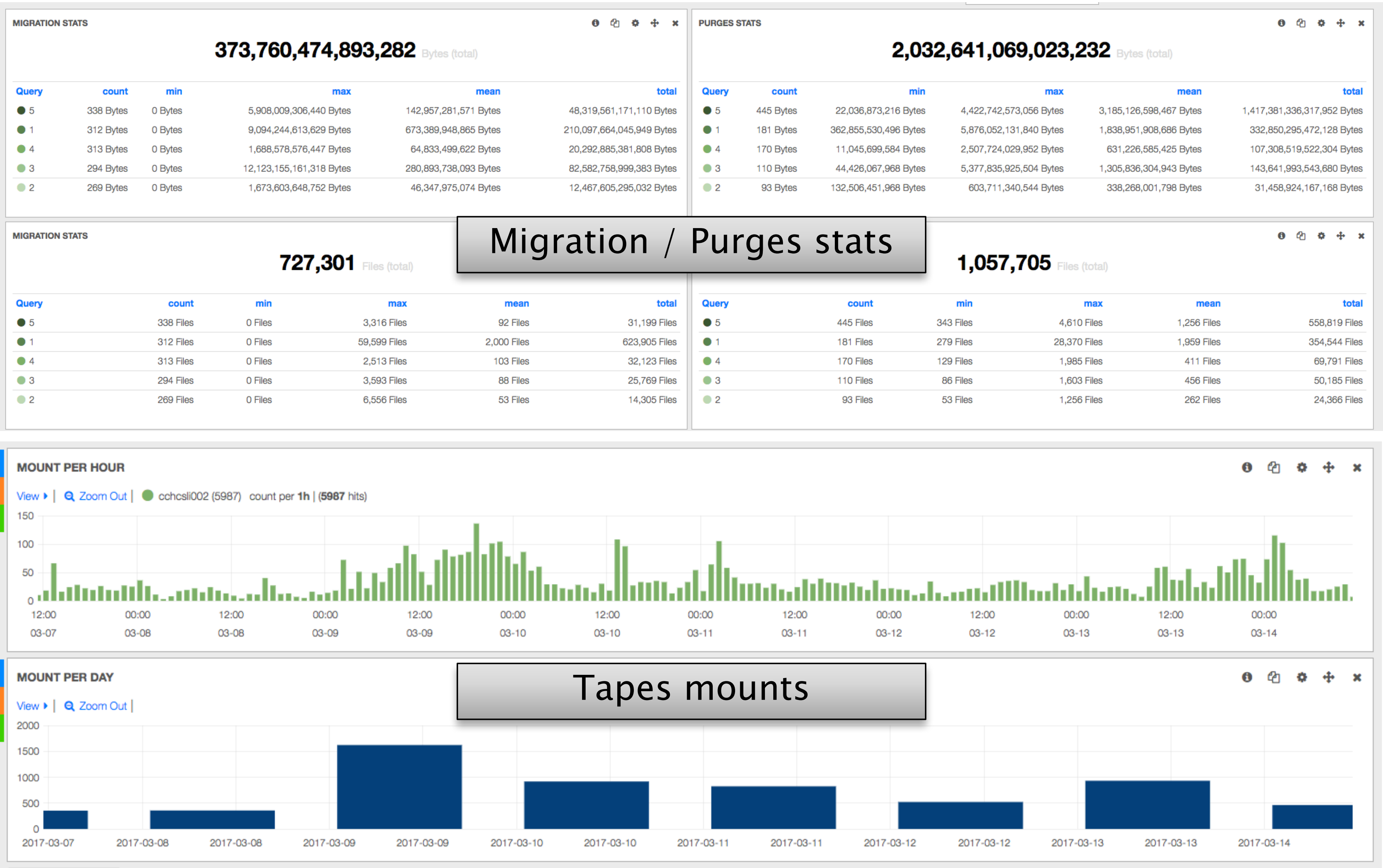
11 000	T10000T1 (to destroy)
8 000	T10000T2 (8,5 TB)
5 000	LTO 4
2 000	LTO 6

- Daily tape mounts:
  - 2 000 average (decreasing)
  - > 10 000 peak
- HPSS Repacks
  - 23,000 T1 → T2 proceed in 2 years
  - T10K-C to T10K-D in progress

Log based monitoring



Monitoring : Kibana Dashboard



Migration / Purges stats

MIGRATION STATS

373,760,474,893,282 Bytes (total)

Query	count	min	max	mean	total
● 5	338 Bytes	0 Bytes	5,908,009,306,440 Bytes	142,957,281,571 Bytes	48,319,561,171,110 Bytes
● 1	312 Bytes	0 Bytes	9,094,244,613,629 Bytes	673,389,948,865 Bytes	210,097,664,045,949 Bytes
● 4	313 Bytes	0 Bytes	1,688,578,576,447 Bytes	64,833,499,622 Bytes	20,292,885,381,808 Bytes
● 3	294 Bytes	0 Bytes	12,123,155,161,318 Bytes	280,893,738,093 Bytes	82,582,758,999,383 Bytes
● 2	269 Bytes	0 Bytes	1,673,603,648,752 Bytes	46,347,975,074 Bytes	12,467,605,295,032 Bytes

PURGES STATS

2,032,641,069,023,232 Bytes (total)

Query	count	min	max	mean	total
● 5	445 Bytes	22,036,873,216 Bytes	4,422,742,573,056 Bytes	3,185,126,598,467 Bytes	1,417,381,336,317,952 Bytes
● 1	181 Bytes	362,855,530,496 Bytes	5,876,052,131,840 Bytes	1,838,951,908,686 Bytes	332,850,295,472,128 Bytes
● 4	170 Bytes	11,045,699,584 Bytes	2,507,724,029,952 Bytes	631,226,585,425 Bytes	107,308,519,522,304 Bytes
● 3	110 Bytes	44,426,067,968 Bytes	5,377,835,925,504 Bytes	1,305,836,304,943 Bytes	143,641,993,543,680 Bytes
● 2	93 Bytes	132,506,451,968 Bytes	603,711,340,544 Bytes	338,268,001,798 Bytes	31,458,924,167,168 Bytes

MIGRATION STATS

727,301 Files (total)

Query	count	min	max	mean	total
● 5	338 Files	0 Files	3,316 Files	92 Files	31,199 Files
● 1	312 Files	0 Files	59,599 Files	2,000 Files	623,905 Files
● 4	313 Files	0 Files	2,513 Files	103 Files	32,123 Files
● 3	294 Files	0 Files	3,593 Files	88 Files	25,769 Files
● 2	269 Files	0 Files	6,556 Files	53 Files	14,305 Files

1,057,705 Files (total)

Query	count	min	max	mean	total
● 5	445 Files	343 Files	4,610 Files	1,256 Files	558,819 Files
● 1	181 Files	279 Files	28,370 Files	1,959 Files	354,544 Files
● 4	170 Files	129 Files	1,985 Files	411 Files	69,791 Files
● 3	110 Files	86 Files	1,603 Files	456 Files	50,185 Files
● 2	93 Files	53 Files	1,256 Files	262 Files	24,366 Files

MOUNT PER HOUR

View ▸ | Zoom Out | ● cchcsli002 (5987) count per **1h** | (**5987** hits)

MOUNT PER DAY

View ▸ | Zoom Out |

Tapes mounts

Issues

- Failures when writing files
  - Creation and transfer run fine
  - Error appears at close()
  - Only affect transfers that use more than 1 SS
  - Appears after HPSS migration (7.4.1.2 → 7.4.2.1)

- Error rate is ~ 0,1 %
- Non critical as client includes retries

Treqs 2

TREQS 2

- TReqS : Tape Request Scheduler
  - It’s a software companion to HPSS, that re-organize HPSS staging requests
  - Increase staging throughput, by re-ordering files to be staged from same tape, according to (logical) File Position on Tape (FPOT)
  - Control number of allocated drives for staging
- Positionning
  - Between storage middleware and HPSS (current clients of TReqS are dCache, XRootD)
  - For HPSS staging only (tape → disk)
- History
  - Running old implementation developed 7 years ago, but not fully reliable nor maintainable
  - New implementation, started from scratch at fall 2015 (TReqS-2)

TREQS 2

- Business point of view
  - Aggregate requests over time per tape, sorting files according to FPOT : → queue
  - Limit number of simultaneous running queues, per tape model
    - (ie: 10 drives allocated for T10K-D)
  - Provide role management (user’s role = ADMIN, USER or NONE)
  - Provide control (on/off) on tape, on tape-model, on HPSS access, on queues processing, on submission of client requests
  - Provide cancelation of client requests
  - Provide persistence for requests (useful for server stop & start)
  - Provide archiving for ended requests (built-in CSV archiver)
- Implementation point of view
  - REST API, JSON format, HTTPS support
  - JSW : Java Service Wrapper, to run application as a UNIX service (stop | start | status)
  - Out-of-the-box monitoring web pages

TREQS 2

- Client/Server model
- Server
  - Written in Java (18,000 lines of code) and C++ (500 lines of code)
  - Using JMS for internal exchanges,
  - H2 DB for persistence,
  - HPSS API via JNI
  - Providing a REST API with JSON over HTTPS
- Client
  - Written in Python (2,000 lines of code), using REST API
  - Authentication is based on login/password
- Project access
  - Code under LGPL-V3 licence, access to granted user on https://gitlab.in2p3.fr
  - Build procedure available in ADMIN-GUIDE
  - Docs : README, ADMIN-GUIDE, CHANGELOG
- Close to be generally available in production
  - Used mostly for dcache/atlas
  - During last atlas reprocessing: 302,000 files, 720 TB staged in 7 days

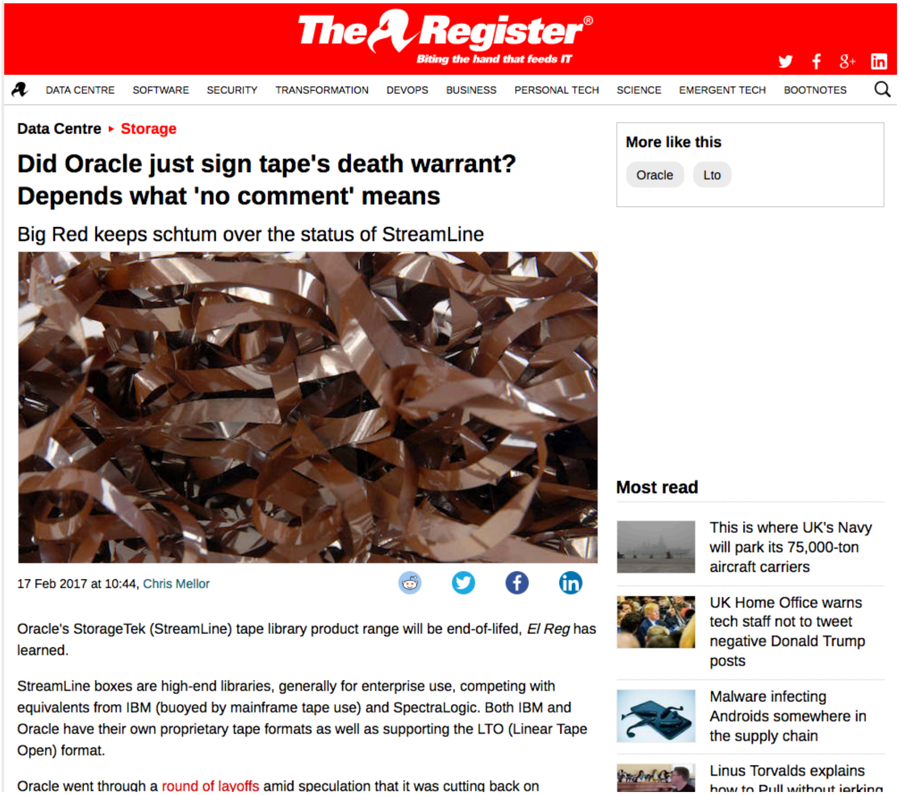
As a conclusion

Plans for 2017

- HPSS
  - Migration from HPSS 7.4.2.1 to version HPSS 7.4.3.2 next week
  - Explore RAIT (HPSS 7.4.3.2)
    - Currently setup on test system (RAIT 2+1) with T10K-B
    - Solution for long term archiving ?
  - Explore HPSS 7.5.x new features
      - Tape Ordered Recall (TOR) to increase TReqS performances
- TAPE
  - Retire T10K-C drives before end of year
    - 1500 Tapes to repack as background activity
  - T10000-E :
    - Will support 12,5 TB on a T10000-T2
    - Not before 2018

Future of entreprise tapes ?

- Rumor about Oracle entreprises tapes drives
  - T10K-E should be marketed in early 2017
  - But won't be released ….



The Register®
Biting the hand that feeds IT

Data Centre ▸ Storage

Did Oracle just sign tape's death warrant? Depends what 'no comment' means

Big Red keeps schtum over the status of StreamLine

17 Feb 2017 at 10:44, Chris Mellor

Oracle's StorageTek (StreamLine) tape library product range will be end-of-lifed, *El Reg* has learned.

StreamLine boxes are high-end libraries, generally for enterprise use, competing with equivalents from IBM (buoyed by mainframe tape use) and SpectraLogic. Both IBM and Oracle have their own proprietary tape formats as well as supporting the LTO (Linear Tape Open) format.

Oracle went through a round of layoffs amid speculation that it was cutting back on

More like this

Oracle Lto

Most read

This is where UK's Navy will park its 75,000-ton aircraft carriers

UK Home Office warns tech staff not to tweet negative Donald Trump posts

Malware infecting Androids somewhere in the supply chain

Linus Torvalds explains how to Pull without jerking

https://www.theregister.co.uk/2017/02/17/oracle_streamline_tape_library_future/

Backup Slide : HPSS Overview

Merci