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

Journées LCG France
Stockage WLCG - HPSS

Pierre-Emmanuel BRINETTE

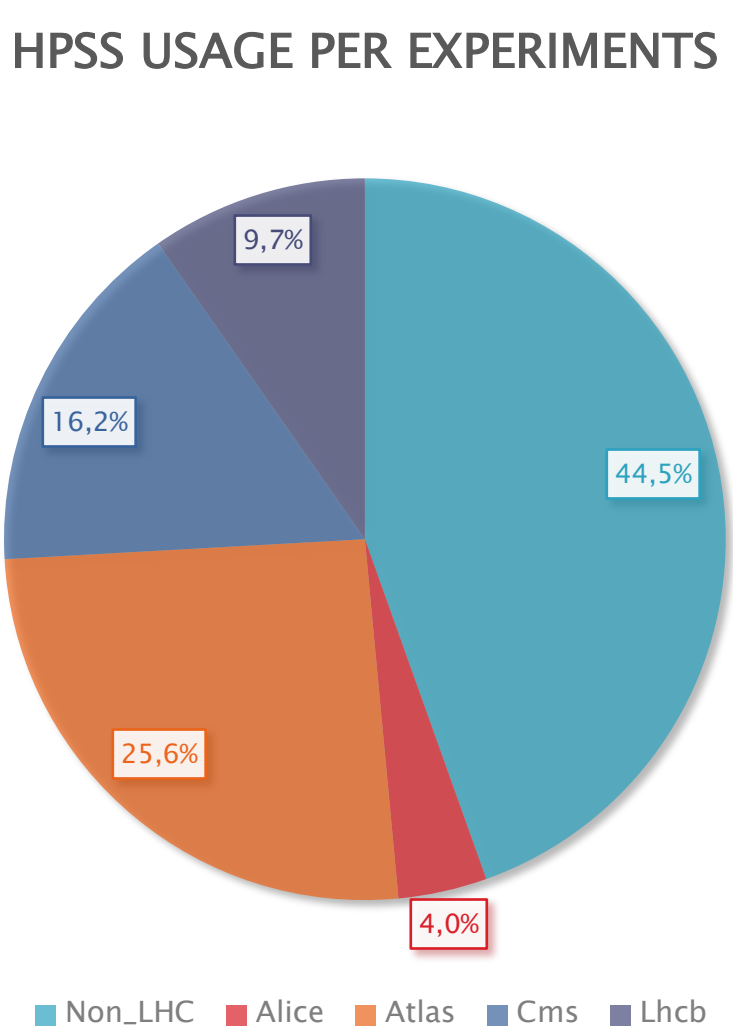
21 juin 2018

CCIN2P3

cnrs

in2p3

Utilisation HPSS

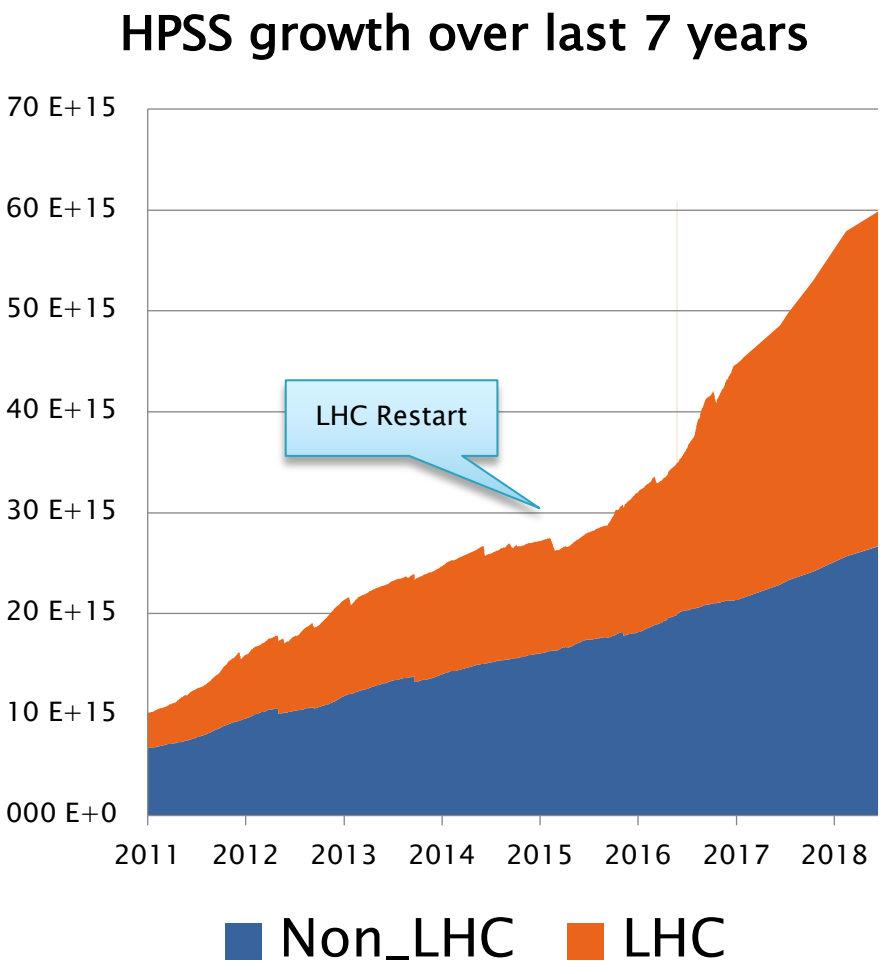


NB : CMS a récemment effacé ~1,5 Po

- Données stockées dans HPSS au 13/06/2018
- Total : 59,8 Po
- 76 M fichiers.

- LCG : 33,1 Po
- 55,5% de la volumétrie
  - Alice : 2,3 Po (4 %)
  - Atlas : 15,3 Po (25,6%)
  - CMS : 9,6 Po (16,2 %)
  - LHCB : 5,8 Po (9,7%)

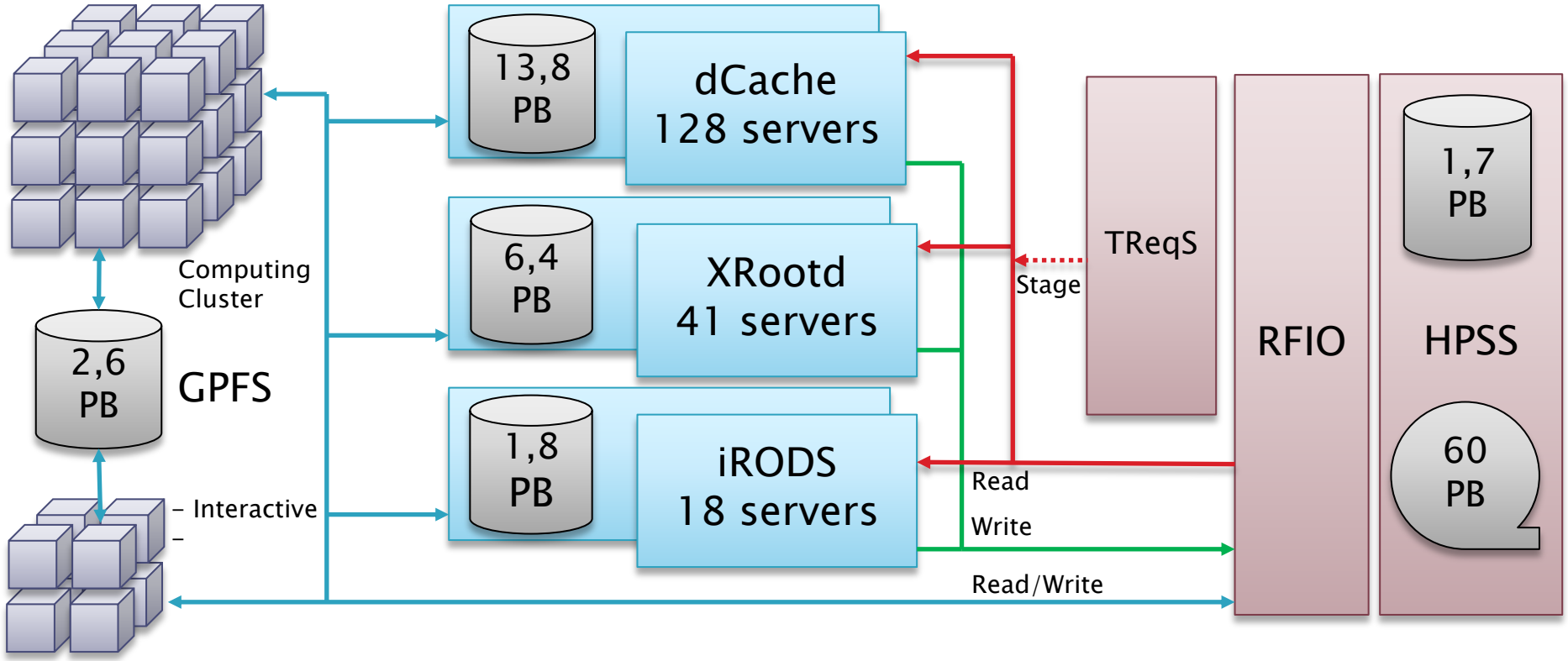
Évolution de la capacité



NB : CMS a récemment effacé ~1,5 Po

- Evolution sur 12 mois
- Total : + 11,3 Po (+ 23 %)
  - Non LCG : +3,7 Po (+17 %)
- LCG : + 7,5 Po (+ 29 %)
  - Alice : +0,4 Po (+ 19 %)
  - Atlas : + 4 Po (+ 35 %)
  - CMS : + 1,2 Po (+ 15 %)
  - LHCB : + 1,9 Po (+ 48 %)
- Prévisions :
  - 80 Po fin 2019
  - ~100 Po fin 2020

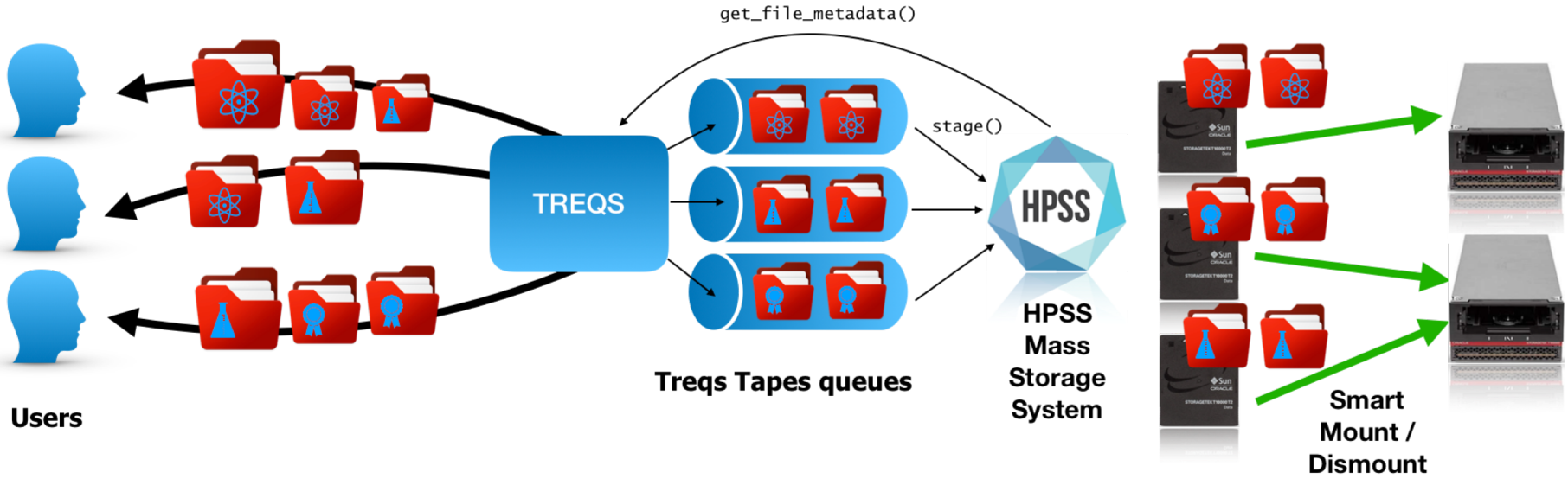
Infrastructure de stockage et HPSS



- 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
  - Total 12 movers (1,7 PB) @ 10Gbits
- Read operations from storage middleware are handled by TREQS 2

TREQS 2

- Outil d'optimisation des « staging » (relecture de puis une bande)
- Principe :
  - Agréger et ordonner les requêtes de lecture en fonction des bandes
  - Limiter le nombre de montage / démontage d'un même bande
- L'ancienne version (jTreqs) avait attend ses limites
- Nouvelle version développée par Bernard Chambon et Lionel Schwarz
- Pleinement en production depuis juin 2017
  - 9,2 M de fichiers / 16,3 Po traités
- Gère toutes les relectures pour LCG



Infrastructure Robotique

- 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
- 56 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

Infrastructure Robotique

- Technologie de bandes utilisé pour HPSS :
  - Oracle Enterprise T10K-T2 : Capacité 8,5 To
  - 9000 bandes utilisées
  - 3000 bandes neuves en stock (~24 Po)
    - Achat des 1500 bandes en juin 2018
- Toute les données HPSS sont stockées sur T10K-D
  - Migration T10K-C → T10K-D terminée (aout 2017).
  - Bande migrée partiellement réutilisée
    - Potentiellement +1000 bandes reconditionnable
      - Sous réserve ! Ces bandes ont déjà bien vécu !
- Infrastructure utilisée pour HPSS
  - 56 lecteurs T10K-D répartis dans 2 robots (sur 4)
  - 12 serveurs disques R730xd
    - Bande passante théorique maximum : 12 Go/s
  - 9 serveurs bandes

Changement de technologie de bande magnétique

- Oracle arrête le développement des ses lecteurs Entreprise.
  - Pas de T10K-E
- 3 scénarios :
  - Conserver les librairies et migrer sur LTO-8 (Cartouche 12 To)
  - Changer de librairie et passer sur la technologie Entrerprise IBM (Jaguar 15/20 To).
  - Continuer avec T10K-D et attendre la sortie du LTO-9 en 2020/2021 (cartouche de 18 à 24 To)
- Choix non arrêté à ce jour.
  - Amortissement du parc de lecteurs T10K-D
  - Lecteurs et bandes LTO réputés moins fiables que les technologies Entreprise (T10K/Jaguar)
  - Buffer plus petit (1Go), pas de fonctionnalités avancées RAO
  - Dépendra du mode de financement : massif ou lissé
- Tests LTO-8 en cours.
  - Prêt d'un lecteur par Oracle

Tests en cours sur l'infra de préproduction HPSS

Test	T10K-D	LTO-8	Commentaires
Ecriture de gros fichiers > 2Go	~ 180 Mo/s +/- 20	183 Mo/s	Résultats similaire Paramètres HPSS à vérifier
Lecture séquentielle de la bande (gros fichiers)	252 Mo/s	323 Mo/s	LTO : 30% + rapide
Lecture de 100 gros fichiers en ordre aléatoire	(pas d'échantillons)	52 Mo/s	Résultat à affiner, il faudrait plus d'échantillons
Ecriture de fichiers « moyens » (100 Mo)	38 Mo/s	22 Mo/s	LTO : 38% + lent
Ecriture d'agrégats de 1Go (10*100 Mo)	145 Mo/s	123 Mo/s	LTO : 17% + lent
Lecture de 100 fichiers de 100 Mo en ordre inverse	14 Mo/s	11,2 Mo/s	LTO : 20% + lent

Résultats préliminaires. Les valeurs sont susceptibles d'évoluer

Perte de données sur bande magnétique

- Nombreuses erreurs de relectures
  - Quelque dizaines de fichiers de certaines bande sont illisibles
  - Fichiers écrits entre mi 2016 et début 2017 (?)
  - Corruption silencieuse à l'écriture due à 1 ou plusieurs lecteurs
  - Lecteur(s) impacté(s) non identifié
- Les erreurs apparaissent lors de la relecture des données
  - Erreurs détecté plusieurs mois après l'écriture des données
  - Impossible de connaître a priori les fichiers corrompus.
  - L'état des dégâts sera connu lorsque toutes bandes T10K-D seront repackées (2020 ?)
- Parc de lecteur assainie suite à une mise à jour de FW en mai 2017
- Etat actuel :
  - 50 bandes identifiées
  - Plusieurs centaines de fichiers, toute VO confondu.
- Bandes envoyés chez Oracle pour analyse/restauration
  - Faible probabilité de récupérer les fichiers.

WLCG Archival Storage WG

- Groupe de travail initié par le CERN
  - Vladimir Bahyl et Oliver Keeble
  - Wiki : https://twiki.cern.ch/twiki/bin/view/HEPTape/WebHome
- Objectifs [3] :
  - Partager les connaissances entre experts exploitant des systèmes de stockage robotisé pour WLCG.
  - Définir comment monitorer l'utilisation de tels systèmes.
  - Définir les meilleures pratiques pour optimiser l'utilisation de ces système par les expériences.
- Réalisations :
  - Sondage sur les systèmes mis en place sur les différents site. [1]
  - Mise en place d'une plateforme de monitoring commune
    - Grafana du CERN : [2]
- Réflexions initiées autour du « Data carousel »
  - Anticiper l'utilisation intensive des bandes comme support de stockage principal.
    - Une petite portion des données serait accessible sur disque, le reste serait « stagé » périodiquement.
  - Rôle du WG:
    - Décrire ***comment*** utiliser efficacement des systèmes de stockage bande
    - Les expériences doivent décrire ce qu'***elles souhaitent faire***.
  - Voir les présentations du WG à Hepix [3] et au GDB [4]

WLCG Archival Storage WG

1. https://twiki.cern.ch/twiki/bin/view/HEPTape/Survey
2. https://monit-grafana.cern.ch/d/000000675/_user-dichrist-tape-reporting?orgId=6
3. https://indico.cern.ch/event/676324/contributions/2967985/
4. https://indico.cern.ch/event/651354/contributions/3019331/attachments/1666669/2672252/WLCG-GDB-DataCarousel.pdf