Tests Data Carousel Atlas au CC
LCG France & DOMA-FR

Pierre-Emmanuel Brinette
9 Novembre 2018

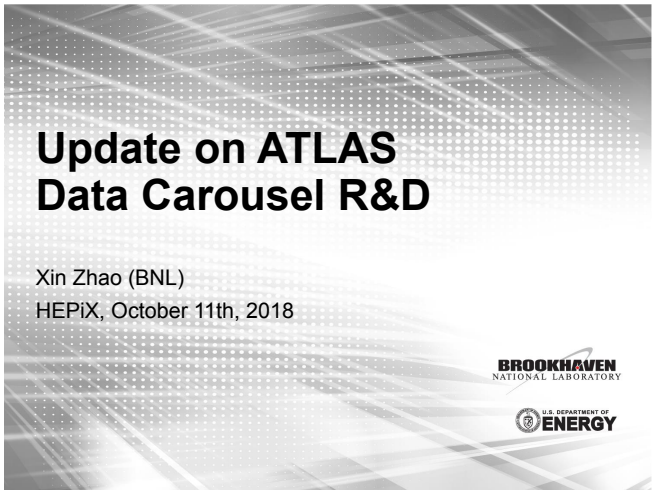
CCIN2P3

cnrs

Data / Tape carousel

- Concept de Data / Tape Carousel
  - Projet initié par Atlas pour évaluer l'utilisation des bandes comme support des données.
    - Faire face à l'explosion de données attendue (HL-LHC)
  - Seule une petite portion des données est présente sur disque
  - Relecture en continue des données des Bandes → Disque

- Tests préliminaires effectués durant l'été sur presque tous les T1
  - Résultats présentés à Hepix par Xin Zhao



https://indico.cern.ch/event/730908/contributions/3153161/

Objectifs du test ATLAS

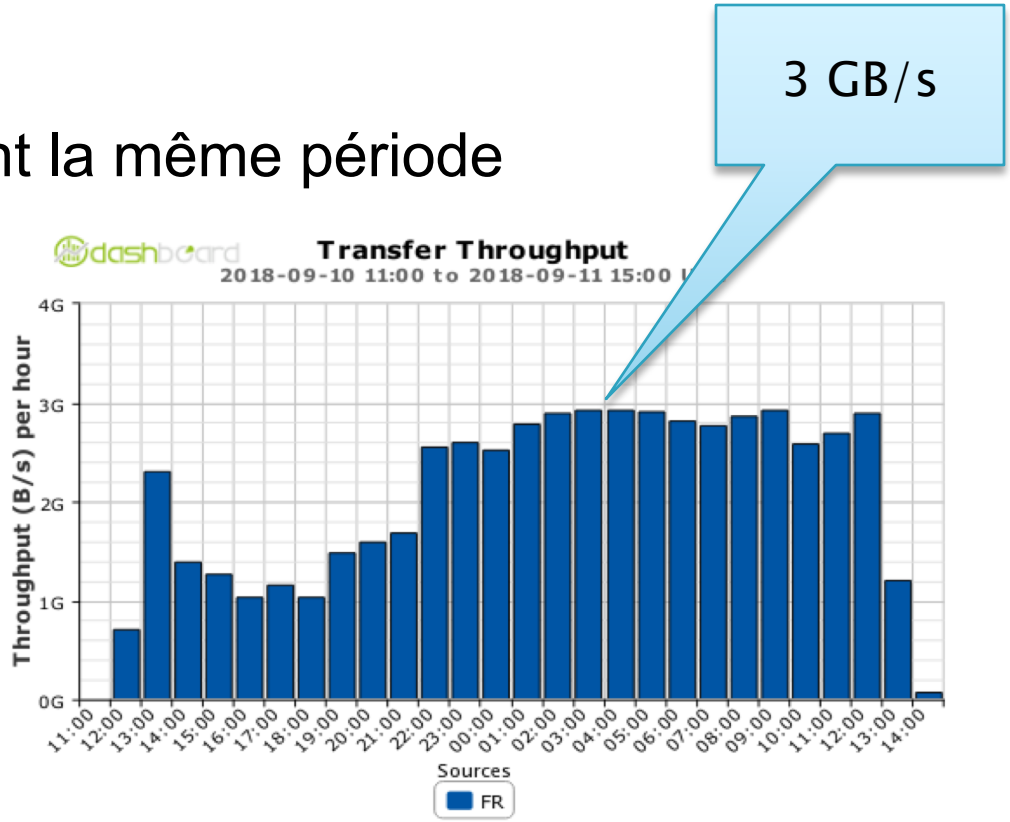
- Etudier la faisabilité d'exécuter les différents traitements ATLAS (« workload ») depuis les stockages bandes

- But du test :
  - Etablir une mesure de **référence** des performances atteignables sur les services bandes des différents T1
  - Pour les sites :
    - Test de charge
    - Identifier les points de blocages
  - Possibilité de **rejouer** le même test à la demande.

- Jeux de test :
  - Données de production : AOD, utilisés pour des jobs de « derivation »
  - Conditions d'utilisations réalistes :
    - Rucio → FTS → SE
    - Staging et copie des fichiers depuis les pool DATATAPE → DATADISK

Résultats

- 204 TB stagés dans dcache en 25h30
  - ~ 84000 fichiers
- Performances
  - Débits moyen: 2,2 GB/s
  - Pics soutenus: 3 GB/S
- Activité totale HPSS durant la même période (atlas + autres VO)
  - 231 TB
  - 106,000 fichiers
- Tous les staging ont été traités par TREQS



Preliminary Results

- Throughput

Site	Tape Drives used	Average Tape (re)mounts	Average Tape throughput	Stable Rucio throughput	Test Average throughput
[1]BNL	31 LTO6/7 drives	2.6 times	1~2.5GB/s	866MB/s	545MB/s (47TB/day)
FZK	8 T10KC/D drives	>20 times	~400MB/s	300MB/s	286MB/s (25TB/day)
INFN	2 T10KD drives	Majority tapes mounted once	277MB/s	300MB/s	255MB/s (22TB/day)
PIC	5~6 T10KD drives	Some outliers (>40 times)	500MB/s	[2]380MB/s	400MB/s (35TB/day)
[1]TRIUMF	11 LTO7 drives	Very low (near 0) remounts	1.1GB/s	1GB/s	700MB/s (60TB/day)
CCIN2P3	[3]36 T10KD drives	~5.33 times	2.2GB/s	3GB/s	2.1GB/s (180TB/day)
SARA-NIKHEF	10 T10KD drives	2.6~4.8 times	500~700MB/s	640MB/s	630MB/s (54TB/day)
[4]RAL	10 T10KD drives	n/a	1.6GB/s	2GB/s	1.6GB/s (138TB/day)
[5]NDGF	10 IBM Jaguar/LTO-5/6 drives, from 4 sites	~3 times	200~800MB/s	500MB/s	300MB/s (26TB/day)

[1] dedicated to ATLAS
[2] with 5 drives, later increased to 6 drives
[3] 36 is the max number of drives, shared with other VOs who were not using them during the test
[4] 8 drives dedicated to this test. Will have 22 shared with other VOs in production.
[5] federated T1, 4 physical sites have tapes

Meilleurs résultats des T1 !

Mais :
- 36 drives utilisés
- Taux remontage : 5,33 x / bande

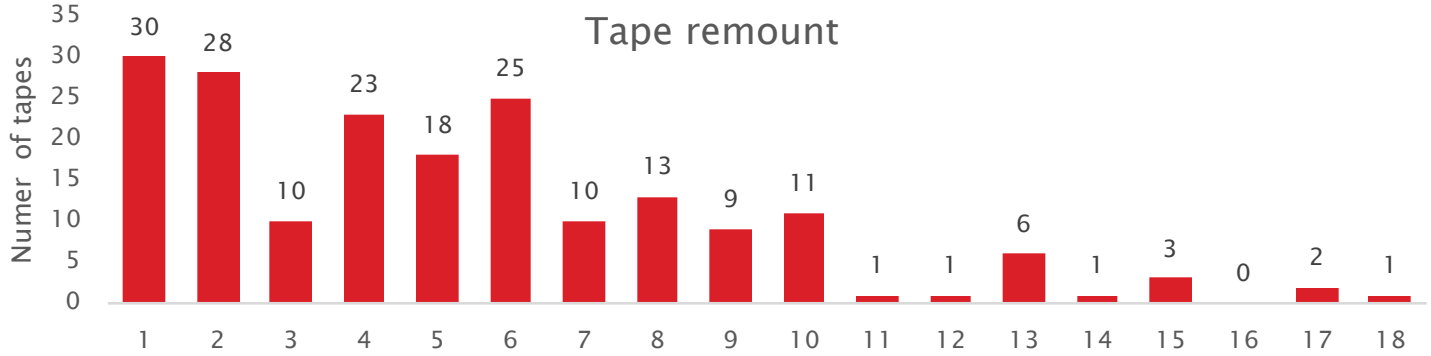
U.S. DEPARTMENT OF ENERGY

BROOKHAVEN NATIONAL LABORATORY

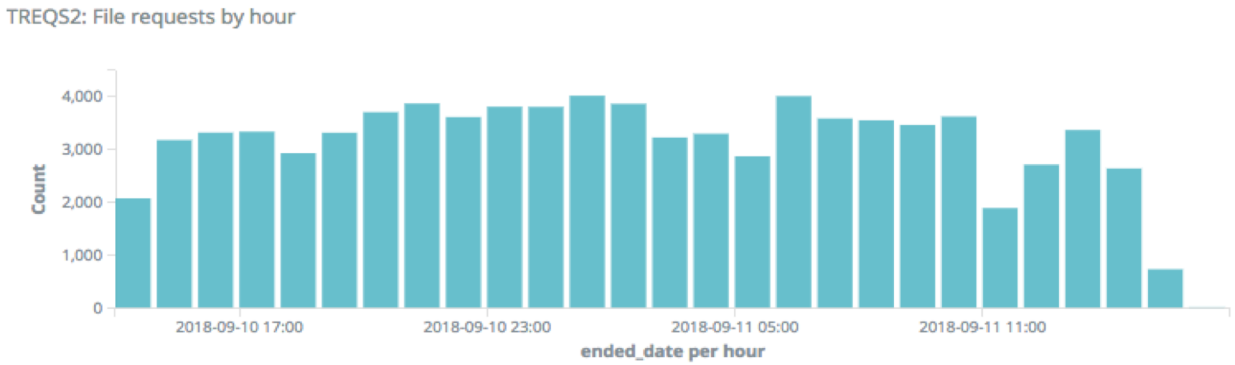
- 192 unique tapes for the 84,000 files

# Files per tape	% total tape	# Tape
1	7%	13
2-9	14%	27
10-99	17%	33
100-500	34%	66
500-1000	15%	28
1000-2000	10%	19
2000-3289	3%	6

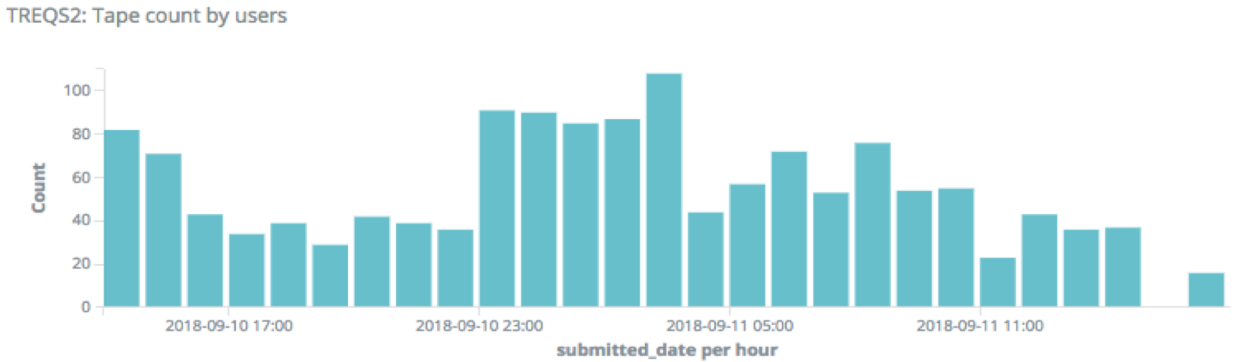
- Total : 1025 mounts / 25 h : AVG : 5,33



Dans le détail



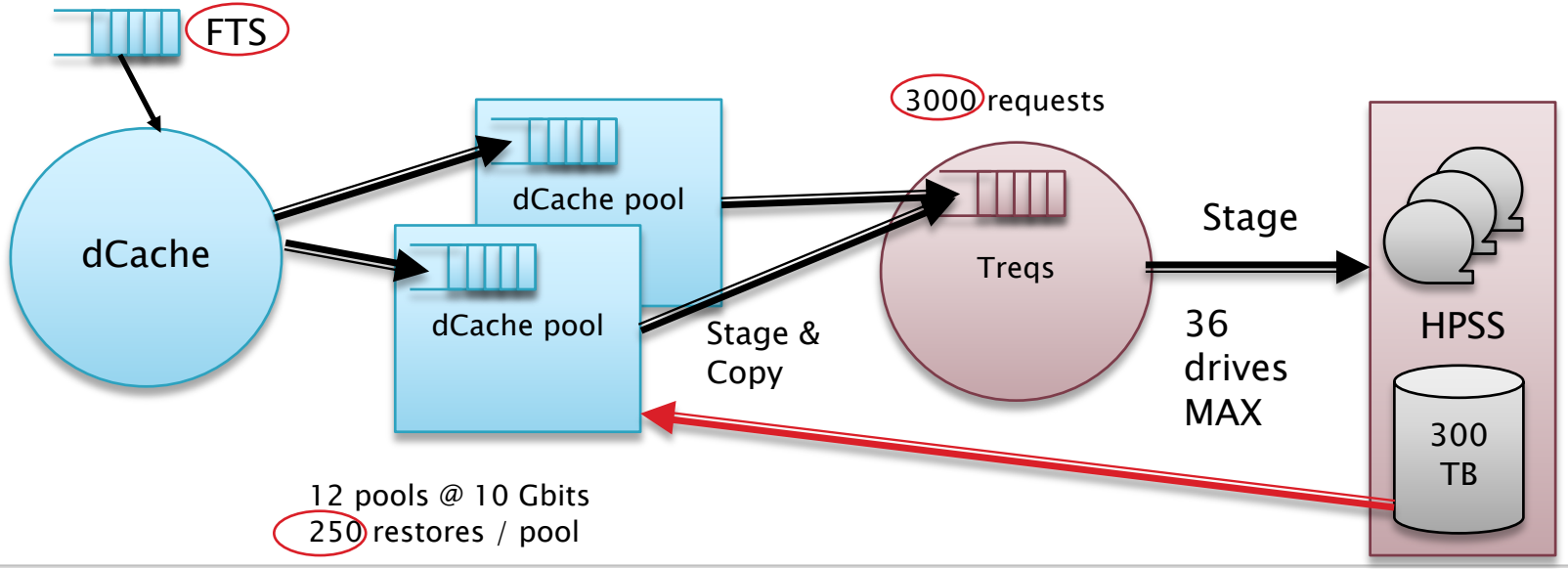
- Fiché stagés / h
  - 0,9-1 Hz
  - Stable



- Nombre de bandes montées /h
  - ~ 40 bandes /h
  - Pic 100 / h

Pistes d'amélioration

- Configuration :
  - Rucio sent ~5000 files / h to FTS
  - Treqs queue handle ~ 3000 requests at time
  - So tape containing lot of files may be mounted multiple time within 26 hours
- Improvements:
  - Sent more requests to FTS at time (ie: 10 k/h)
  - Increase # of restore / pool to increase (ie: 500/h)



Pistes d'amélioration

- Augmenter le temps d'intégration dans Treqs
  - De 2mn à 10 mn ?
- Organiser les données sur bande au moment de l'écriture
  - Regrouper les données d'un même dataset sur les même bandes
  - « Tape Familly »
- Exploiter les fonctionnalités RAO des lecteurs
  - Recommanded Acess Order : Chemin optimal d'accès au fichiers sur bande
  - Nécessite de modifier treqs.

LTO ou Entreprise ?

- Le CC **doit** faire évoluer son infrastructure bande suite au retrait de Oracle du marché des bandes Entreprise
- LTO-8 : Média bon marché (< 10 € / To)
- IBM Jaguar : Média plus capacitif (20 To), lecteurs plus performants pour la relecture (RAO)
- Projet démonstrateur (POC):
  - Petite librairie équipée de drives LTO et Enterprise
  - Rejouer les tests Atlas sur les différents média

https://indico.in2p3.fr/event/17206/contributions/64147/