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

HPSS et Robotique au CC-IN2P3

Pierre-Emmanuel Brinette - 15 avril 2019

3ème rencontres HPSS France

CCIN2P3

cnrs

in2p3

Plan

- Statut HPSS au CC-IN2P3
- Evolution des besoins de stockage
- Evolution de la robotique

Statut HPSS

Quelques nombres

- HPSS v 7.5
  - hpss-7.5.1.2-20190116.u9
- Utilisation (Avril 2019)
  - Total 72 Po
  - LHC : 44 Po (61 %)
- Evolution sur 1 an
  - + 12 Po (+ 20%)
  - 80 M fichiers (+ 6 %)
  - Effacement de 4 Po de données de l'expérience d0
- Prévision 2019/2020
  - + 16 Po
- Voir ma présentation au HPSS FR 2018 :
  - https://indico.in2p3.fr/event/17222/contributions/61421/

Environnement HPSS

- Pas d'évolution récente de l'infrastructure
- Core serveurs :
  - 2 x DELL R720 + MD3220
  - RHEL 7
  - Mise en service 2013
  - Fin de garantie mai 2020
- Mover disque :
  - 12 x DELL R730xd + MD 1200
  - 10 Gbits
  - CentOS 7
  - Total : 1,7 Po
- Mover Tape :
  - 9x DELL R720/R640
  - 10 Gbits
  - SL 6 / CentOS 7
  - 6 drives T10K-D / mover
- TREQS
  - Tape REQuest Scheduler
- Optimise l'accès aux données sur bande
  - Mise en queue des requêtes de lecture
  - Relecture « ordonnée » des bandes
  - Limite le nombre de relecture //
- Version 1.2.2 stable !
- Version 1.3 en développement
  - Récupération de métadonnées à des fin de statistique
    - Date de création du fichier
    - Date du dernier accès
    - Etc …
  - Fonctionnalité de « bypass »

Migration HPSS 7.5

- Migration laborieuse
  - Mécanismes de réplication complexes à mettre en œuvre
  - Complexité augmentée due à nos 5 sous-systèmes
  - Nombreux tests réalisés sur l'environnement de préprod.
  - Voir ma présentation au HPSS FR 2018 :
    - https://indico.in2p3.fr/event/17222/contributions/61421/

- Migration réalisée à la 3ème tentative
  - 1ère tentative annulée par manque de temps (03/2018)
  - 2ème tentative annulée suite au remplissage de l'espace disque (06/2018)
    - Les logs de la base de donnée n'étaient pas nettoyés
    - Impossible de reprendre la réplication pour 1 des BDD
    - Nécessaire de détruire/récréer l'environnement de réplication
  - 3ème tentative a aboutie (09/2019).
    - Processus de réplication régulièrement contrôlé avant la migration

- Installation / Mise à jour par RPM
  - RPM téléchargés sur le Wiki HPSS
  - Déployé dans un repos PULP
  - Y compris sur le core serveur !

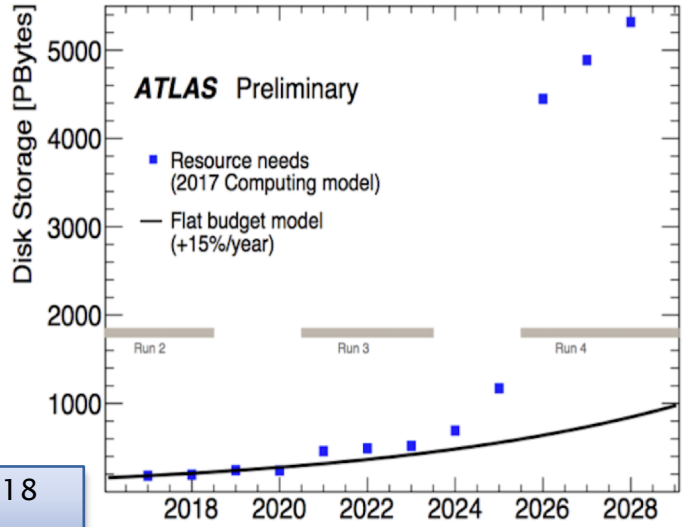
Evolution des besoins de stockage

Prévisions de stockage à l'horizon 2025

ATLAS perspective on the data storage challenge of HL-LHC:

- 'Opportunistic storage' basically doesn't exist
- Format size reduction and data compression are both long-term goals, require significant efforts from the software and distributed computing teams
- Tape storage is 3~5 times cheaper than disk storage, increasing tape usage is a natural way to cut into the gap of storage shortage for HL-LHC

Xin Zhao « Update on Atlas Data Carousel R&D » – Hepix Fall 2018
https://indico.cern.ch/event/730908/contributions/3153161/

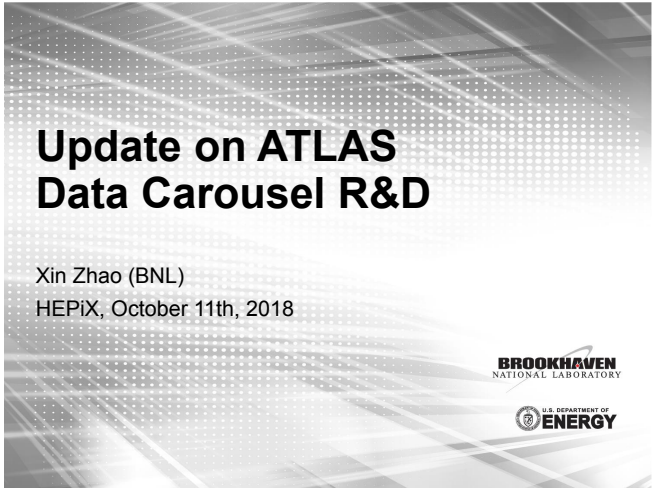


- 2025 : HL–LHC (High Lumiunosity LHC), Mise à jour majeure des détecteurs
- Les besoins de stockages des expériences LHC va exploser
  - Le budget reste constant
- Besoin de solution de stockage bon marché
  - Piste : Utilisation intensive des bandes magnétiques.

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



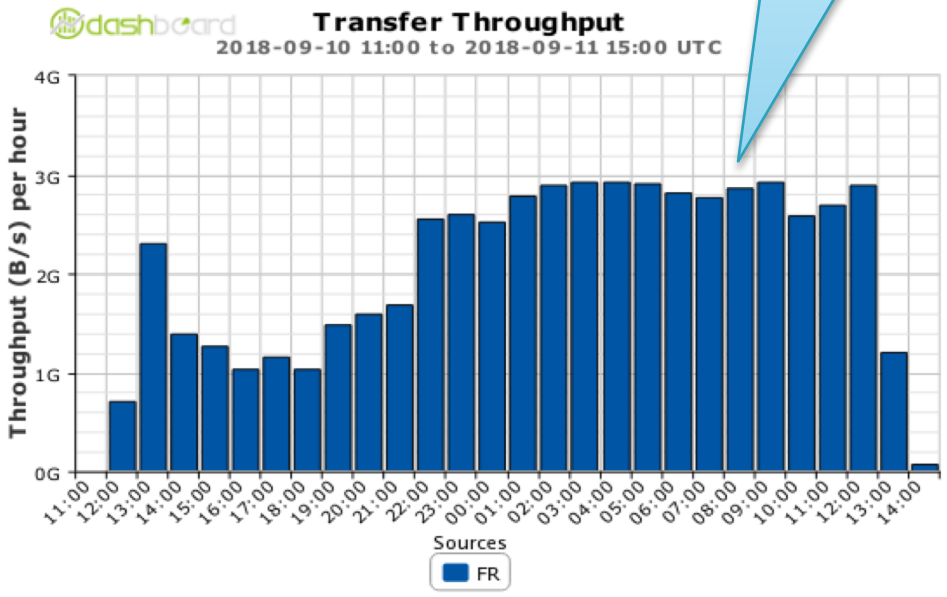
- Jeux de test :
  - Données de production : AOD, utilisés pour des jobs de « derivation »
  - Conditions d'utilisations réalistes :
    - Rucio → FTS → SE
    - Staging et copie des fichiers depuis les pool DATATAPE → DATADISK

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

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

Meilleurs résultats des T1
3 GB/s



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

Pistes d'amélioration

- Problématique : Réduire le taux de remontage
- 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 »
  - → Mais comment gérer les évolution de capacité des médias (repack)
- Exploiter les fonctionnalités RAO des lecteurs enterprise
  - Recommanded Access Order : Chemin optimal d'accès au fichiers sur bande
  - Nécessite de modifier treqs.

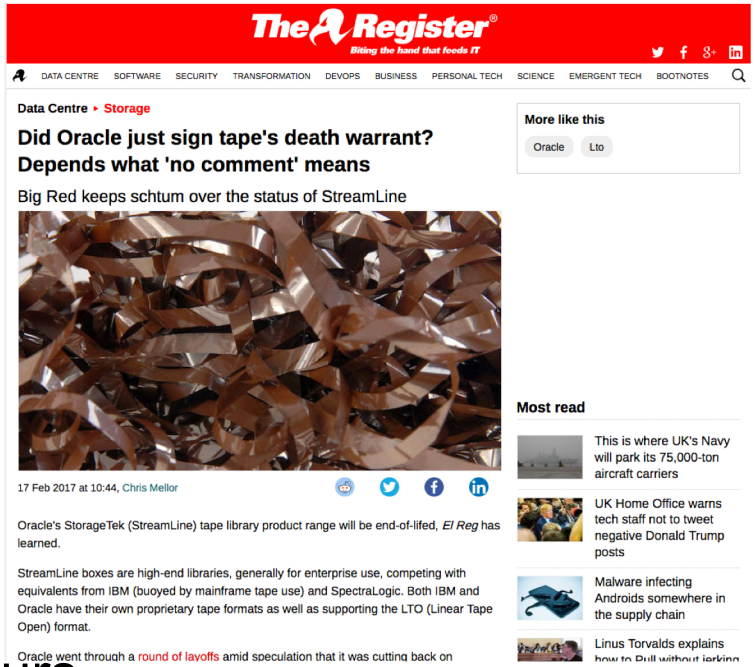
Evolution de la robotique

Habillage des librairies

Pour à peine le prix de 10 cartouches !

Rappel des faits

- HP arrête la fabrication de ses lecteurs LTO
  - Mi 2016
- Oracle arrête le développement des lecteurs Enterprise T10000
  - Non annonce début 2017



The Register®
Biting the hand that feeds IT

DATA CENTRE SOFTWARE SECURITY TRANSFORMATION DEVOPS BUSINESS PERSONAL TECH SCIENCE EMERGENT TECH BOOTNOTES

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

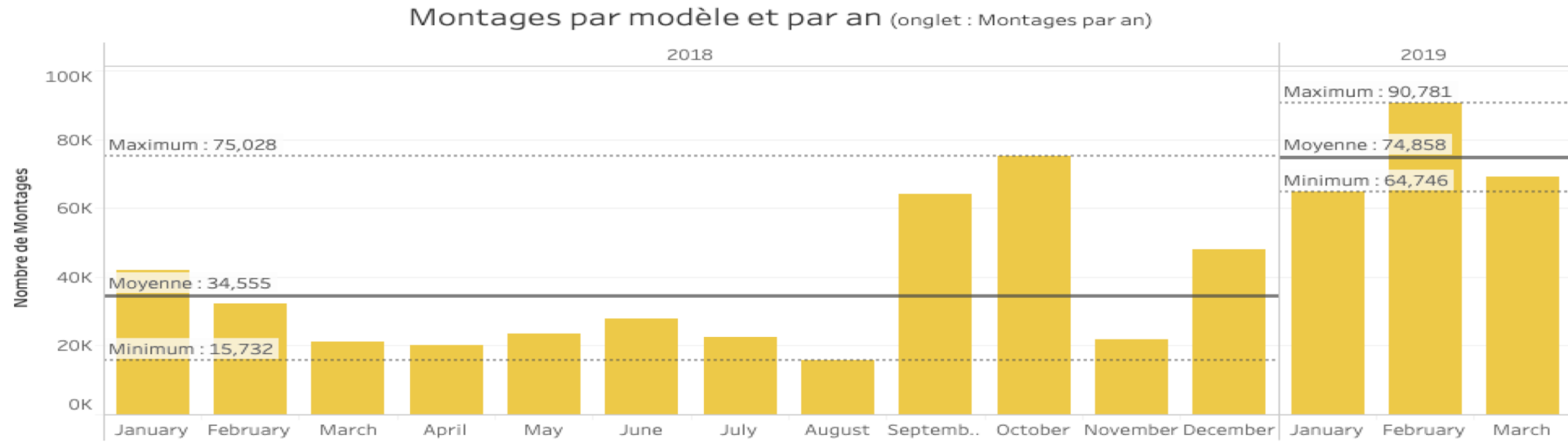
Linus Torvalds explains

- IBM est le seul fabricant de lecteurs
  - LTO et Enterprise
- 2 fabricants de cartouches
  - Fujifilm et Sony
  - Procès concernant les brevets entrainant une pénurie de LTO-8
    - → A ce jour (mars 2019) le contentieux n'est toujours pas réglé

Au CC-IN2P3

- HPSS utilise la technologie Oracle T10000 depuis 2007
  - 4 librairies Oracle SL8500 / capacité 40000 slots
  - 56 lecteurs Enterprise T10000D
  - 8,5 To / cartouche
  - 13000 T10K-T2 bandes pour les robot
  - 20000 emplacements **libres**
- Technologies LTO utilisées jusqu'à présent pour le backup uniquement
- (LTO4 / LTO7)
- **Impossible** d'utiliser les technos IBM Enterprise dans les librairies Oracle.

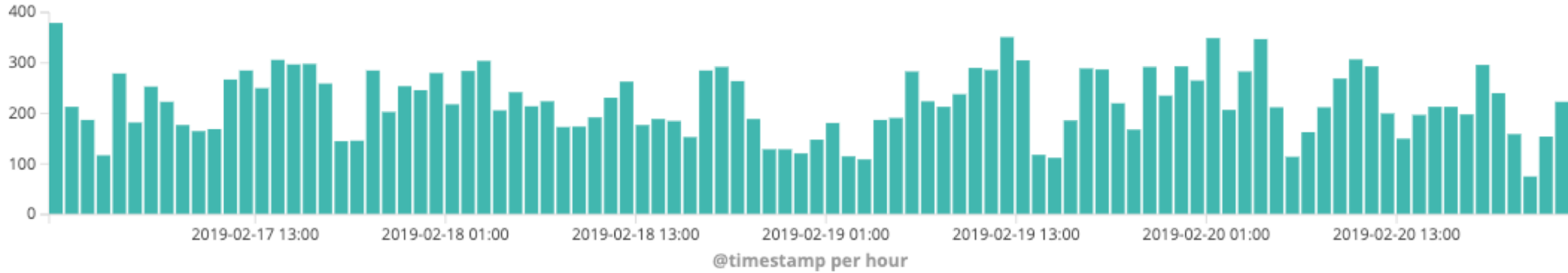
Statistiques de montage (T10K)



16-21 fev 2019

Max : 380 m/h

Max 5500 / j : (222 /h avg)



LTO ou Entreprise ?

- Le CC **doit** faire évoluer son infrastructure bande suite au retrait d'Oracle du marché des bandes Entreprise
- Quel stratégie choisir ?
  - Lecteurs et bandes LTO :
    - Médias bon marché mais performances « a priori » plus faible qu'en version « Enterprise »
  - Lecteurs et bandes « Enterprise » IBM Jaguar
    - Technologie performante, mais « a priori » plus onéreuse
- Beaucoup « d'aprioris »

Avantage Entreprise vs LTO

- Capacité native de la cartouche plus élevé
  - (Cartouche JE : 20 To vs LTO8 : 12 To)
- Média physiquement plus robustes
  - Conçu pour une utilisation intensive (Moins vrai depuis le LTO8)
- Meilleurs performances :
  - Débit des lecteurs plus élevé
  - Optimisation de l'écriture des petits fichiers (Fast Sync)
  - Optimisation de la relecture de fichiers (**RAO**)
  - Précisions de positionnement plus élevée (1/64 vs ½ LZ)
  - Buffer plus important (2 GB vs 1GB)
  - → **Moins de lecteurs nécessaire pour la même charge de travail**
- Réutilisation possible des médias sur 2 générations de lecteurs
  - Réécriture des bandes à plus haute densité → réduction cout €/To
  - Investissement plus pérenne
- Support de multiples interfaces (FC, Ethernet, RoCE etc)

Mais :
- lecteurs et bandes + chers
- Cout/To ~ x2

Tests LTO 8

Test	T10K-D (non RAO)	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	69 Mo/s	52 Mo/s	LTO : 25% + lent
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

- LTO sont ~25% plus lents à la relecture
  - Pas de RAO
  - 2 landing zone → Positionnement plus lent
- Pour le moment TREQS ne permet pas d'exploiter les fonctionnalités RAO

- RAO : Recommanded access ordering
  - Disponible depuis HPSS 7.5.1.2
  - Utilise des fonctionnalités du lecteurs pour déterminer l'ordre optimal lors de la lecture de plusieurs fichiers sur un même bandes
  - Fonctionnalité supporté par les lecteurs « Enterprise » uniquement

Lecture séquentielle :
326 s

Lecture optimisée RAO :
126 s
Gain : 151 %

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

Tests RAO sur T10K-D / LTO8

- Relecture de fichiers de test avec et sans RAO
  - Echantillons :
    - Fichiers de 2200 MB
    - Bande T10K-D contenant 3646 fichiers
    - Bande LTO-8 contenant 5205 fichiers
    - Echantillons tirés aléatoirement
    - Staging avec hpss_cache (non ordonnée et ordonné) et quaid (RAO)

Test effectué	Echantillon	Durée	Débit
Relecture non ordonnée	25 fichiers	19m41s	41 Mo/s
Relecture ordonnée par position logique (offset)	25 fichiers	19m05s	48 Mo/s
Relecture ordonnée RAO	25 fichiers	8m0s	114 Mo/s
Relecture non ordonnée LTO8	25 fichiers	23m26s	39 Mo/s
Relecture ordonnée LTO8	25 fichiers	24m9s	38 Mo/s
Relecture ordonnée LTO8 Quaid	25 fichiers	25m5	37 Mo/s
Relecture ordonnée par position logique (offset)	50 fichiers	34m58s	58 Mo /s
Relecture ordonnée RAO	50 fichiers	13m10s	139 Mo /s

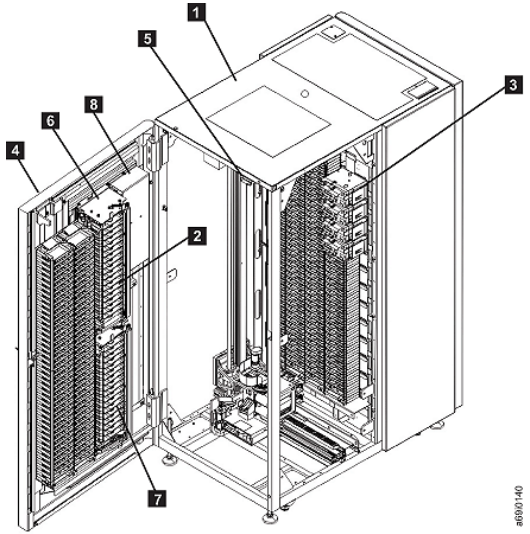
Projet évolution robotique

- Plan initial : Configuration « démonstrateur » proposant un mix technologique LTO/Enterprise
  - 6 lecteurs de chaque technologies
  - Capacité utile : ~ 2,5 Po en LTO-8
  - Capacité utile : ~ 2,5 Po en Jaguar D/E
- Suite au derniers tests réalisés, la technologie LTO ne semblent pas répondre à nos besoins
  - Les performances offertes seraient moins bonnes que celle délivrées aujourd'hui
- Nouvelle configuration « démonstrateur »
  - 12 lecteurs TS1160
  - 20 Po de capacité
  - Evolutive
- Intégrée à la production
- But du test :
  1. Tester l'exploitation d'une nouvelle technologie robotique dans notre contexte HPSS
  2. Tester le « service » commercial et support proposé par le revendeur de la solution
  3. Cout du support et des extensions

Proposition : IBM

- IBM TS4500 :
  - 3 frames
    - 1 L25 : Lecteurs Jaguar + 550 slots
    - 1 D25 : Extension Cartouche Jaguar
    - 1 D55 : Lecteurs LTO + 770 slots
  - 2 accesseurs (2 robots)
  - 6 lecteurs de chaque technologie
    - 6 LTO-8
    - 6 TS1160

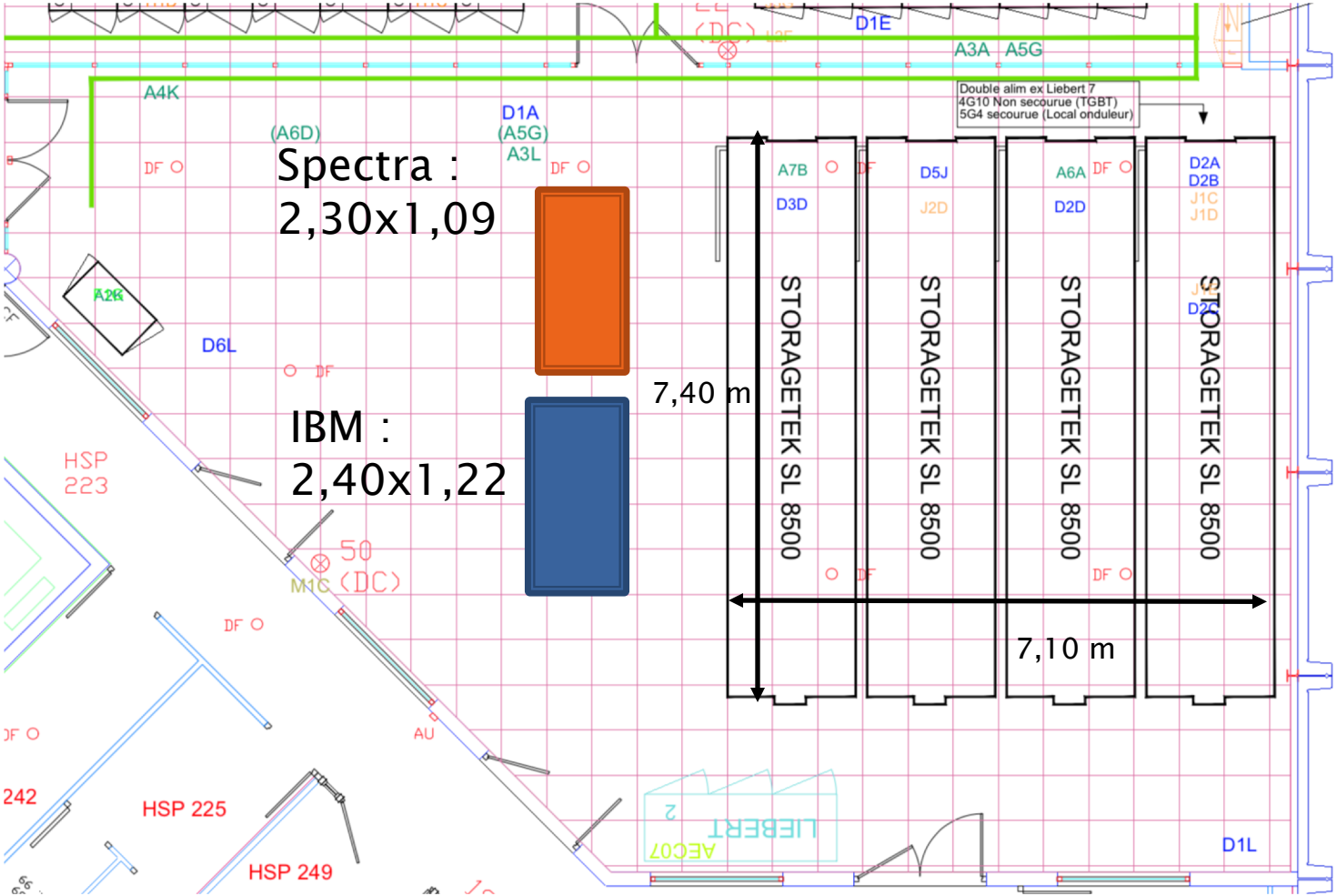
- Solution technique :
  - Frame dédié par type média et lecteurs
    - On ne peut pas mélanger les technos au sein d'une frame
  - Extension de la librairie par ajout de « frame » dans la longueur
  - Il faut prévoir des mix entre frame lecteurs + cartouche et cartouche seules
  - Il existe des frames « haute densité » permettant d'empiler des cartouches en profondeur



Proposition : Spectralogic

- Spectralogic
  - Société US peu présente en Europe
- Spectra Tfinity
  - 3 frames
  - 2 bras
  - 6 lecteurs de chaque technologies
- Solution technique :
  - Utilise des « terapack » pour le stockage des média
  - Densité :
    - 9 bandes Enterprise
    - 10 bandes LTO
  - Mix de technologies possibles au sein des mêmes frames
  - Extension de capacité par frame :
    - 1300 LTO-8 / 990 Jaguar
  - Interrogations sur le cout des média/terapack
- TAOS ™ :
  - Time-based Access Order System
  - Simule les fonctions RAO au niveau library
  - https://edge.spectralogic.com/index.cfm?&fuseaction=home.displayFile&DocID=5035

Implantation des librairies



Que font nos confrères (mars 2019)

- RAL : Spectralogic Tfinity + TS1160 / LTO8
- CNAF : IBM TS4500 + Mix LTO 8 /TS1160 (a l'étude)
- CERN : IBM TS4500 + LTO8
- BNL : POC Evaluation librairie IBM/Spectra + TS1150
- JLAB : TS4500 + LTO8
- NERSC : TS4500 + TS1155
- DESY : SL8500 + LTO8

Planning ?

- Rédaction / publication appel d'offre (fin avril 2019)
- Choix de la solution et du fournisseur (fin mai 2019)
- Installation (été 2019)
- Intégration HPSS (septembre 2019)
  - Acquisition des serveur Tape
- Démarrage des tests (Octobre 2019)
- Bascule d'utilisateurs pilotes (novembre 2019)
  - VO non LHC utilisant Xrootd : ie KM3NET ?
  - VO LHC : Atlas (Dataset « Tape Caroussel »), Alice ?
- Objectif 2020 :
  - Retour d'expérience, choix de la solution définitive (Lecteurs / Robot)
  - Rédaction d'un appel d'offre pour un libraire de capacité N x100 Po

Le futur

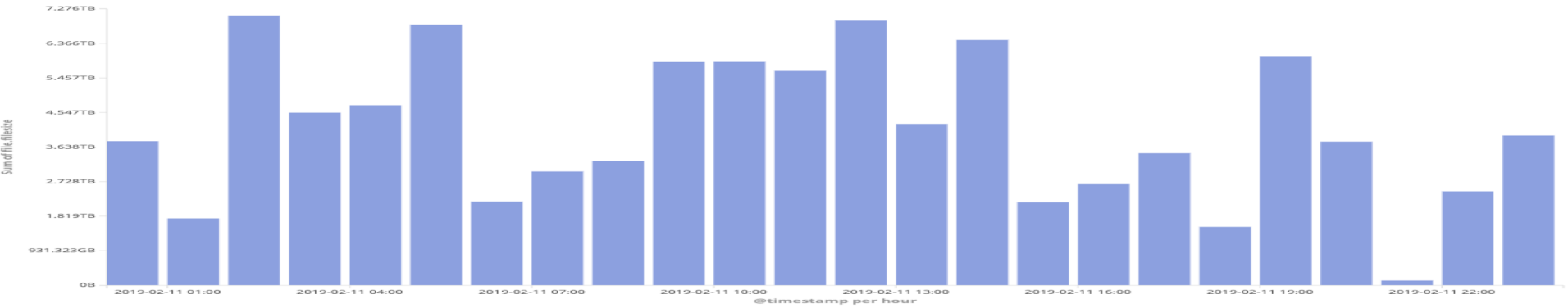
- Nouvelles fonctionnalités de HPSS 7.5.3
  - RAO
  - Log des accès fichiers via syslog (depuis 7.5.2)
  - Limite le nombres de drives utilisés pour la relecture.
    - Equivalent aux fonctionnalités offertes par TREQS
- TS1160 + Jaguar E + HPSS 7.5.3
  - Solution adaptés pour les performances en relecture.
- LTO 9 :
  - Sortie fin 2020 / Début 2021
- TODO :
  - Poc Lustre HSM (ca fait 5 ans que je dis ça …)

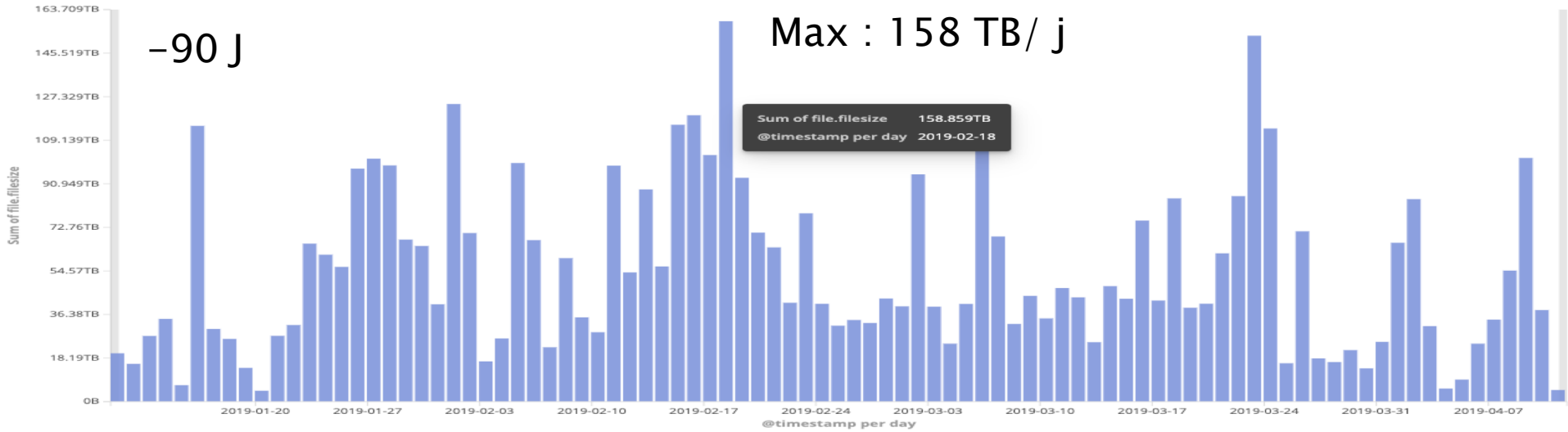
Backup slide

Stats treqs

11 fev 2019

Max : 7,1 TB/h (2 Go/s) **(36 lecteurs)** → 55 Mo/s/lecteur (avg)
98,5 TB/j (4,5 TB/h, 1,14 Go/s)



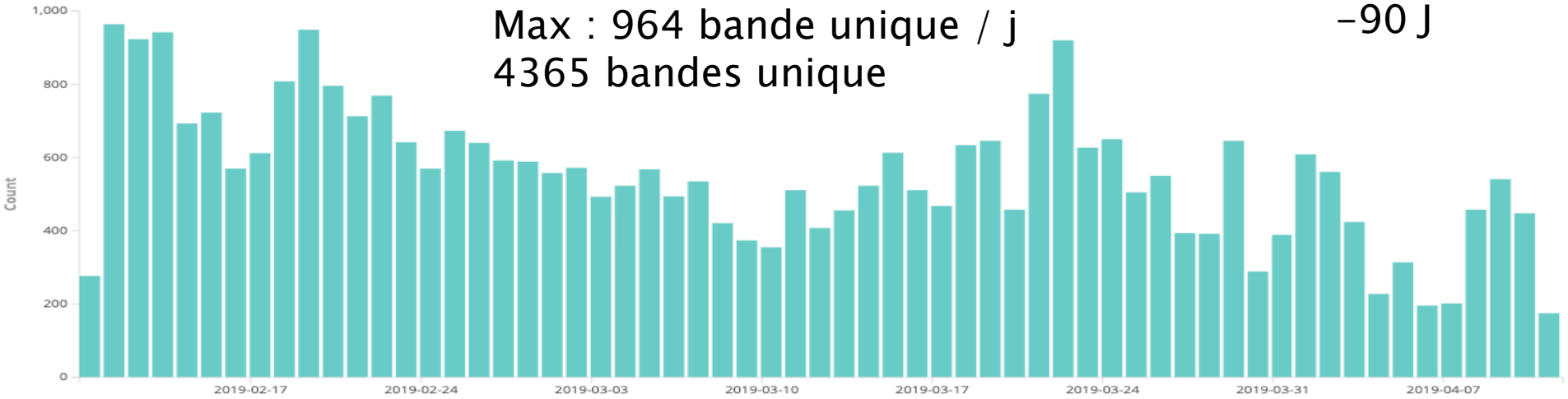


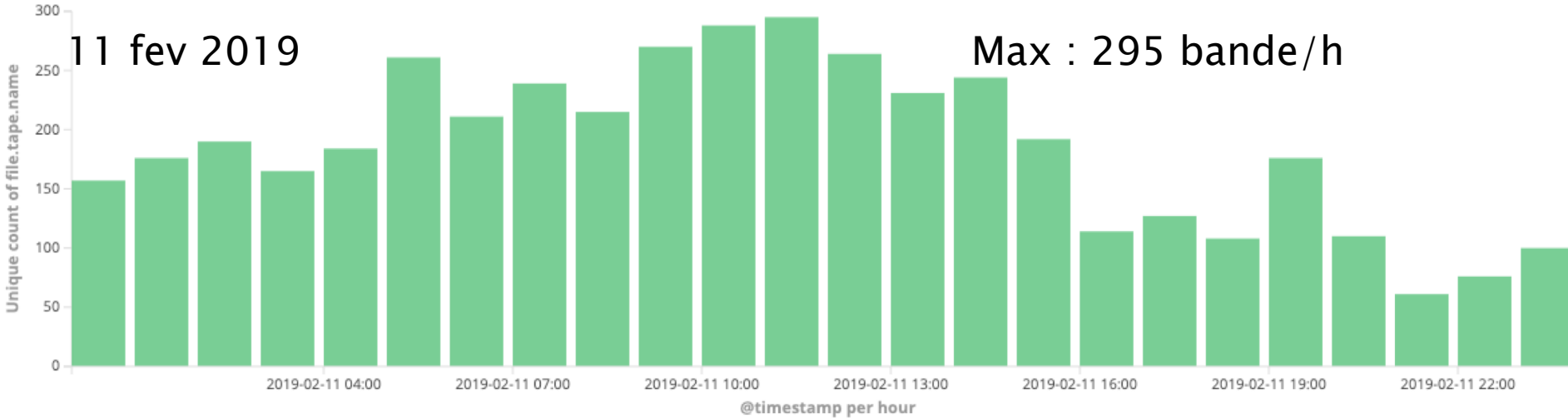
-90 J

Max : 158 TB/ j

Sum of file.filesize 158.859TB
@timestamp per day 2019-02-18

Stats Treqs


Max : 964 bande unique / j
4365 bandes unique
-90 J
Count
1,000
800
600
400
200
0
2019-02-17
2019-02-24
2019-03-03
2019-03-10
2019-03-17
2019-03-24
2019-03-31
2019-04-07


11 fev 2019
Max : 295 bande/h
Unique count of file.tape.name
300
250
200
150
100
50
0
2019-02-11 04:00
2019-02-11 07:00
2019-02-11 10:00
2019-02-11 13:00
2019-02-11 16:00
2019-02-11 19:00
2019-02-11 22:00
@timestamp per hour