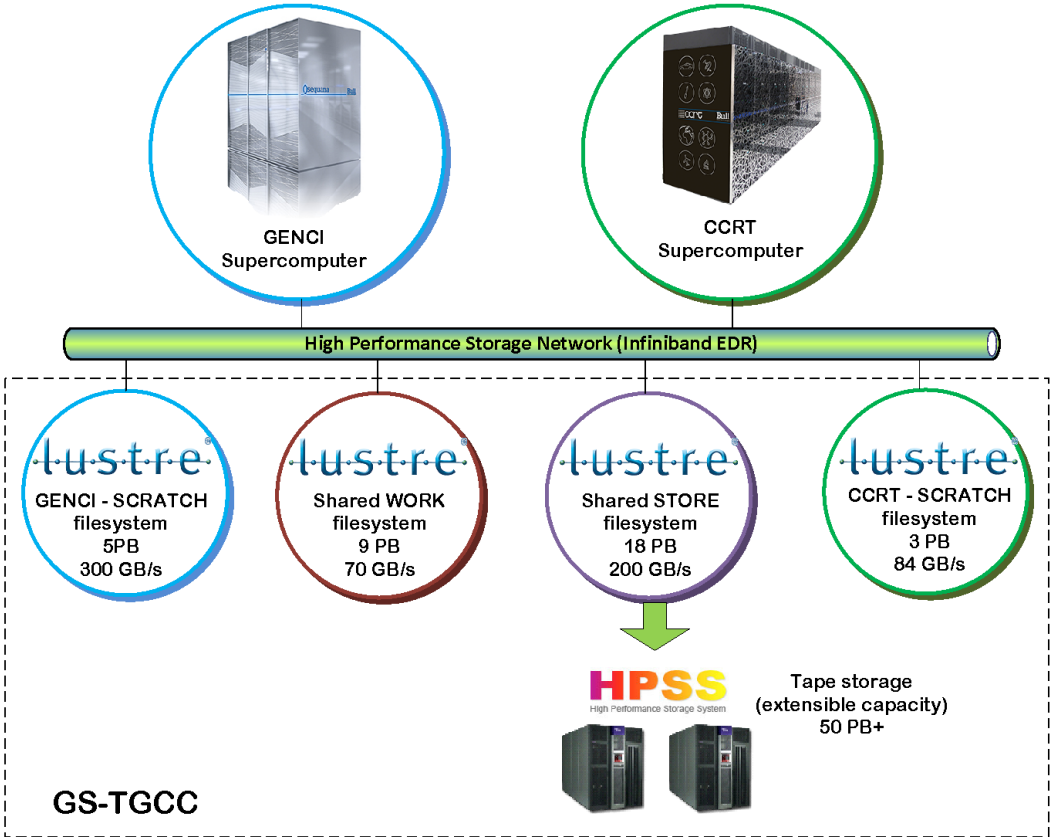
DE LA RECHERCHE À L'INDUSTRIE

HPSS au CEA

Rencontres HPSS France – Bruyères le Châtel – 15 & 16 avril 2019

Thomas Leibovici <thomas.leibovici@cea.fr>

ARCHITECTURE DU STOCKAGE (TGCC)



2 systèmes HPSS de production :

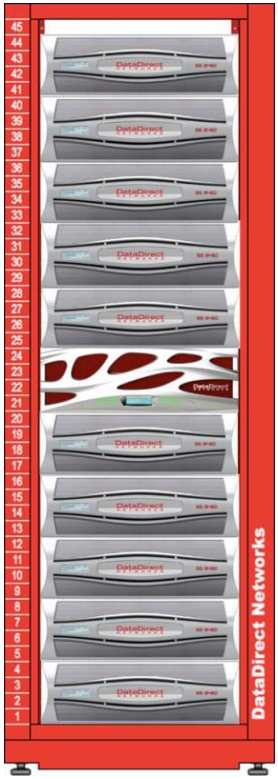
- TERA :
  - Volume stocké : 116 Po (+37 Po/an)
- TGCC :
  - Volume stocké : 52 Po (+13 Po/an)

- HPSS version 7.4.3p2u2
  - Ajoute notamment le support des LTO8

HPSS : niveau disque

- DDN SFA 14KX-e
  - Movers disques « embarqués »
  - Les movers sont des VMs embarquées dans les contrôleurs
  - Installation libre du contenu des VMs
    - Dans notre cas : Linux Ocean (basé sur CentOS 7.x)
- 1 SFA = 2 controleurs :
  - Bi-sockets Intel Broadwell
  - 256 GB RAM
  - 2 Mellanox Infiniband Connectx4 EDR dual ports
- x610 Hitachi Hard Drives 8 TB (enclosures 8462)

=> 5Po et 8 HPSS movers dans 1 SFA (1 rack)



HPSS **TERA** :

- Robots :
  - 3 x SL8500 (full extension)
- Bandes : **T10K-D**
  - 43 lecteurs
  - Repacks des anciennes technologies terminé
- Quelques statistiques (par jour) :

Par jour	Moyenne	Max
Volume écrit	145 To	457 To
Volume lu	3 To	29 To
Fichiers écrits	28k	127k
Fichier lus	500	16k
Montages	290	920

Problème rencontré avec les T10K-D

- Symptôme : sur certaines bandes, le repack révèle un fichier qui n'est pas lisible
- Actions : màj FW drives, Driver Linux, FW carte FC...
- Résultat :
  - On n'arrive pas mieux à relire les fichiers en question
  - Plus d'apparition de nouvelles bandes non lisibles?
- **Difficulté à déterminer les lecteurs problématiques**
- Problème similaire sur LTO6 il y a 2 ans :
  - L'outil ITDT avait permis d'assainir le parc de lecteurs
  - Pas d'équivalent sur T10K ?
  - Câblage Ethernet en cours pour utiliser MDVOP

HPSS **TGCC** :

- Robots :
  - 3 x SL8500 (full extension)
- Production principalement sur **LTO8**
  - 40 lecteurs
  - Petits/moyens fichiers : bande + mirror
  - Gros Fichiers : Storage Class en RAIT (2+1)
  - En cours : migration LTO6->LTO8 (voir slides suivants)
- Communauté génomique sur LTO5
  - Idem : petit/moyen = mirror ; gros = RAIT 2+1
  - A venir : migration vers LTO6 (voire LTO8?)

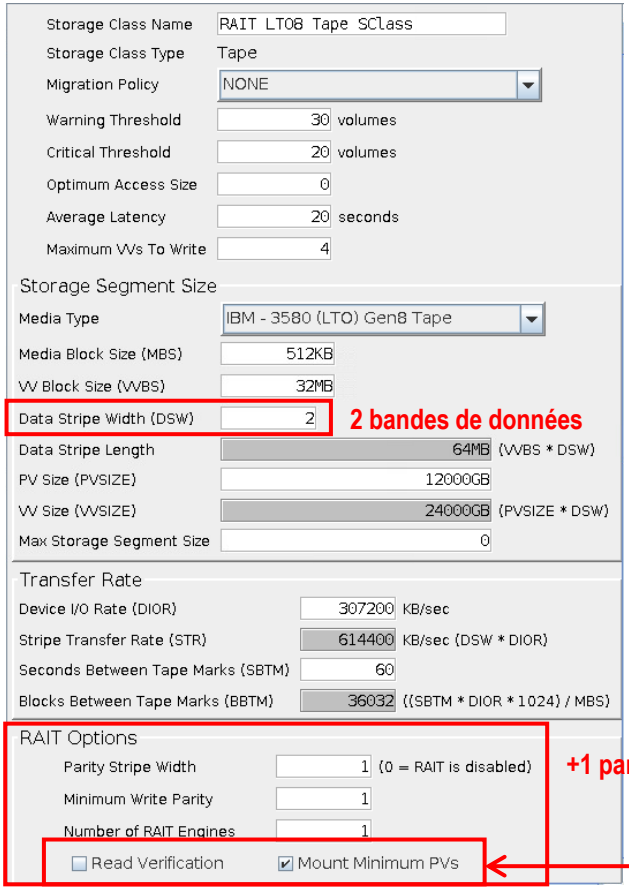
STATISTIQUES TGCC

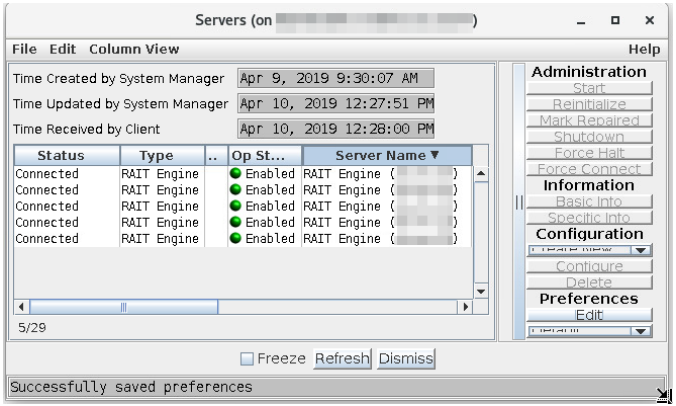
- Statistiques d'accès à HPSS TGCC :

Par jour	Moyenne	Max
Volume écrit	33 To	50 To
Volume lu	5 To	35 To
Fichiers écrits	13k	68k
Fichier lus	650	8k
Montages	315	650

- Statistiques de pannes :
  - Remplacement d'environ 1 lecteur par mois

RAIT PAR L'EXEMPLE





Serveurs « RAIT engines »
(ici, répartis sur les movers)

- Réparer une bande endommagée = repack du VV
- Optimisations des robots :
  - SL8500 : 1 VV = 3 bandes dans 3 robots
  - Spectra : VV groupés dans les « Terapack »

Seulement 2 montages nécessaires pour la lecture

MIGRATION DE LTO6 VERS LTO8

- Au départ : 1 SC LTO6 simple
- Objectif : 1 SC LTO8 en RAIT 2+1
- Problème :
  - La procédure « Storage technology replacement »
    ne s'applique pas en cas de changement de *stripe width*
- Solution mise en œuvre :
  - Une longue campagne de *changecos* qui durera plusieurs mois/années
  - stage & changecos par volume
- Toutefois…
  - Les développeurs ont indiqué que la possibilité de faire une migration de technologie en changeant le stripe width demanderait peu de changement dans le code.

  => Anticiper un CR si vous avez un tel projet…

Problème d’approvisionnement en LTO8…

Acquisition pour évaluation

(système de pré-production)

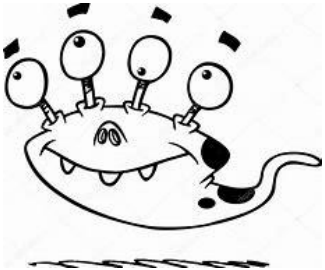
- 1 robotique SpectraLogic T-Finity avec :
  - 2 lecteurs TS1155 (FC)
  - 2 lecteurs TS1155e (RoCE)
  - +10 lecteurs LTO6 existants
- Plan de test:
  - Avec HPSS 7.5.3
    - Optimisations spécifiques pour SpectraLogic
  - Test drives RoCE nécessite :
    - RHEL récent (7.6)
    - Switch Ethernet supportant la fonctionnalité DCB*
      (*Data Center Bridging)

Projets en cours / à venir :

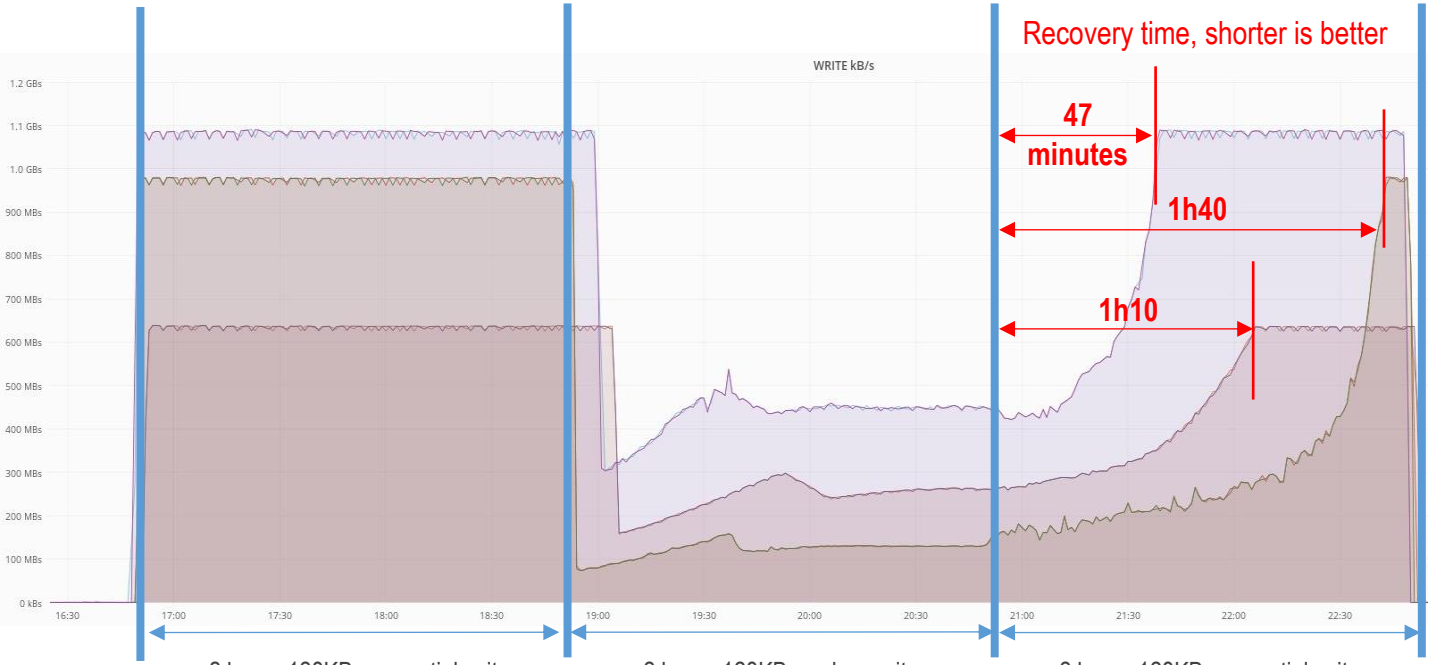
- T1K-F : système de fichiers Lustre full-flash
  - Débit : 1To/sec
  - Capacité : 2 Po
  - Mise en service : Q2 2019
- FENIX: plateforme de services pour HBP
  - Interactive computing
  - Cluster OpenStack
  - Stockage Lustre flash
  - Object store (interface OpenStack Swift)
  - Mise en service : Q1 2020
- Stockage pour EXA-1 (successeur de TERA 1000)

Une créature complexe… très différente des disques durs

- DWPD (Drive Writes per Day)
  - Indique la durée de vie du media (cycles de réécriture)
  - Mais aussi : influe sur la performance *intra-day*
- Un SSD, ce n'est pas seulement du hardware
  - Mécanismes de garbage-collection complexes
  - Maitrise inégale selon les constructeurs
- « Apprentissage » du pattern I/O
  - S'adapte au cours du temps (voir slide suivant)



COMPORTEMENT D'UN DISQUE SSD



3DWPD Hitachi SS530

1DWPD Samsung PM1643

3DWPD Samsung PM1643 Low power

MERCI DE VOTRE ATTENTION

Questions ?