DE LA RECHERCHE À L'INDUSTRIE

www.cea.fr

HPSS@CEA

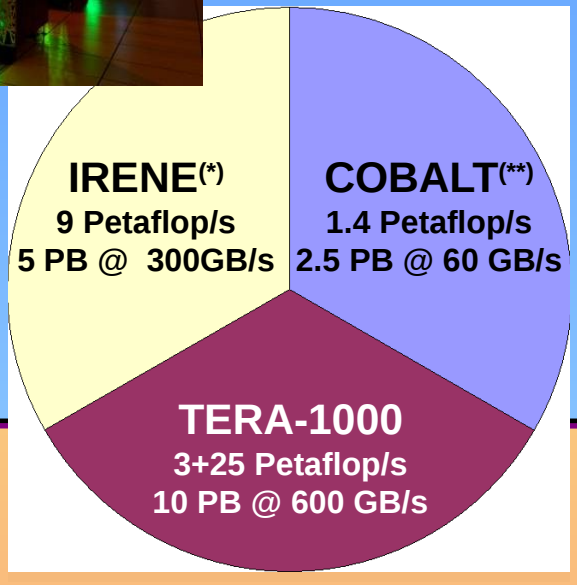
Rencontres HPSS France
Villeurbanne, 5-6 avril 2018

Thomas Leibovici
thomas.leibovici@cea.fr

CEA/DAM/DIF

CEA Computing Complex

Public research
French (20%) Europe (80%)
(Tier-1, Tier-0)



CEA and Industrial partners
(CCRT)

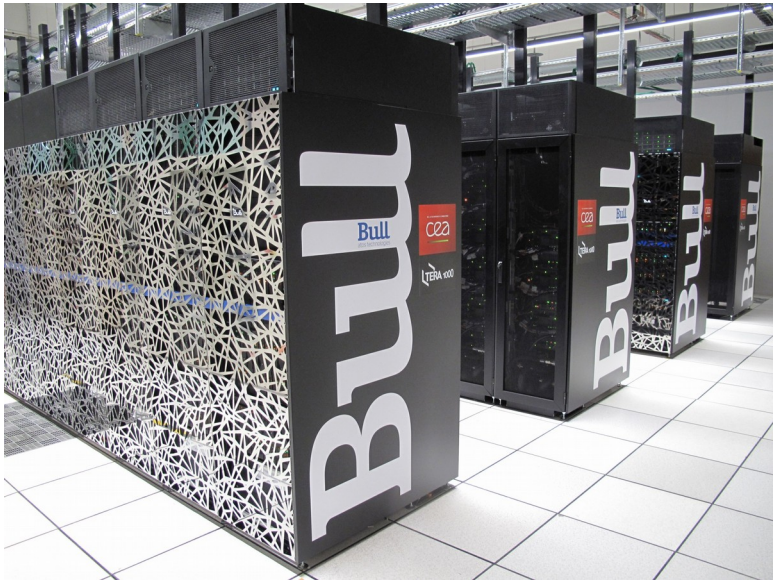
TGCC building

TERA building

CEA/DAM Classified
computations (TERA)

(*) Irene supercomputer, made available by GENCI, is the French contribution to PRACE Infrastructure. Irene is operated by CEA in its TGCC facility.
(**) Co-funded by CEA and industrial partners

TERA 1000 supercomputers



TERA 1000-1

- 2196 nodes : 2*HSW 16 cores @ 2.3GHz
  - +100 nodes with 2*NVidia K80
- 128 GB memory/node
- 2.99 Pflops peak
- Interconnect Infiniband FRD
- Power: 1MW

TERA 1000-2

- 8004 nodes : 2*KNL 68 cores @ 1.4 GHz
- 192 GB memory/node + 16GB MCDRAM
- 25 Pflops peak
- Interconnect BXI 1.2
- Power: 4MW

ARM experimentation

ARM cluster installed Q3 2018

- Cavium ThunderX2 ARM processors
  - 159 nodes x 2*THX2 30 cores @ 2.2GHz
- 256 GB memory/node
- Interconnect Infiniband EDR
- Installation: September 2018

Calculateurs du TGCC et CCRT

TGCC : Irène (Curie 2)

- SKL system:
  - 1656 x 2*SKL 24 cores @ 2.7 GHz
  - Infiniband EDR
  - 6.68 PFlops
- KNL system:
  - 828 x KNL 68 cores @ 1.4 GHz
  - BXI v1.2
  - 2.36 Pflops

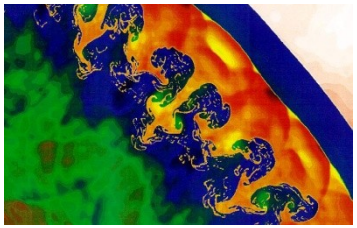
CCRT : Cobalt

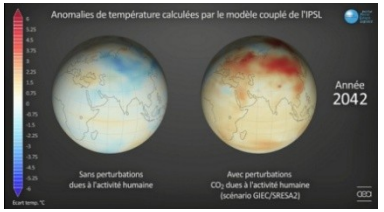
- 1296 nodes x 2*BRW 14 cores @ 2.4 GHz
- Infiniband EDR
- 1.36 Pflops
- Planned extension: SKL, 0.5 Pflops

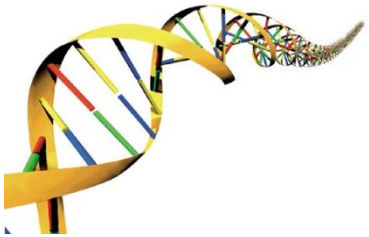
Systèmes HPSS au CEA

Systèmes HPSS :

- Centre de calcul **TERA :**
  - Depuis 1998
  - Volume stocké : 79Po (+17 Po/1 an)
- Centre de calcul **TGCC :**
  - Depuis 2010
  - Volume stocké : 39 Po (+12Po/1 an)
- Stockage **France Génomique :**
  - Depuis 2013
  - Fusionné dans le système TGCC en 2017
- Système de tests (TERA+)







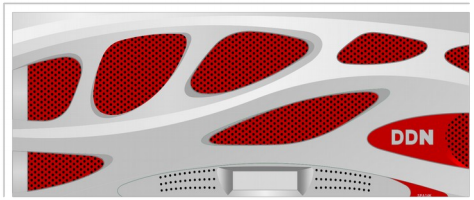
Versions des logiciels

Versions :

- **HPSS 7.4.3p2e3**
- **DB2 v10.5 fp5**
- Linux:
  - Core server: **CentOS 6.9**
  - Movers et clients: **CentOS 7.4**
- **MOFED 4.2**

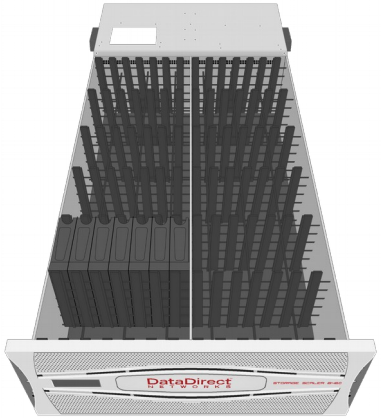
Matériel : disques

- **Movers disques HPSS « embedded »**
  - DDN SFA 14KX-e
  - Movers are VMs running in the controllers



- x2 singlets with this configuration :
  - Bi-sockets Intel Broadwell
  - 256 GB RAM
  - 2 Mellanox Infiniband Connectx4 EDR dual ports
  - 2 Ethernet Gigabit ports
  - x4 Internal SSD Toshiba 480 GB
  - x5 SSD Toshiba 480 GB

→ 8 HPSS movers in one SFA

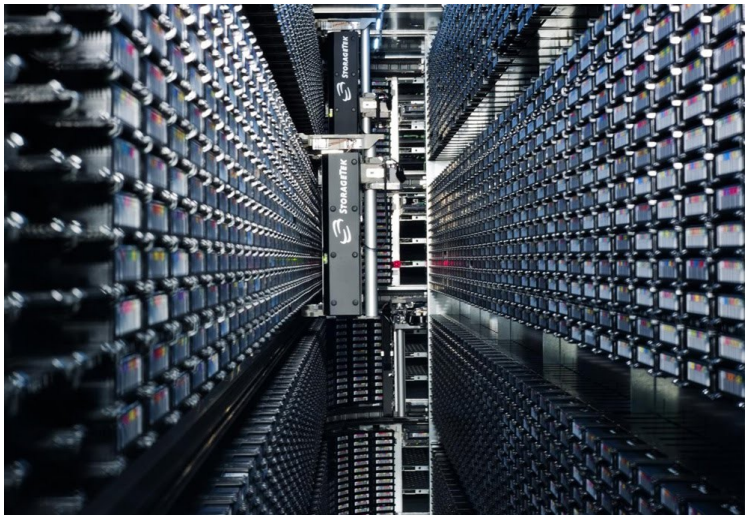


- x410 Hitachi Hard Drives 8 TB (enclosures 8462)

  => 2.5 PB of disks usable

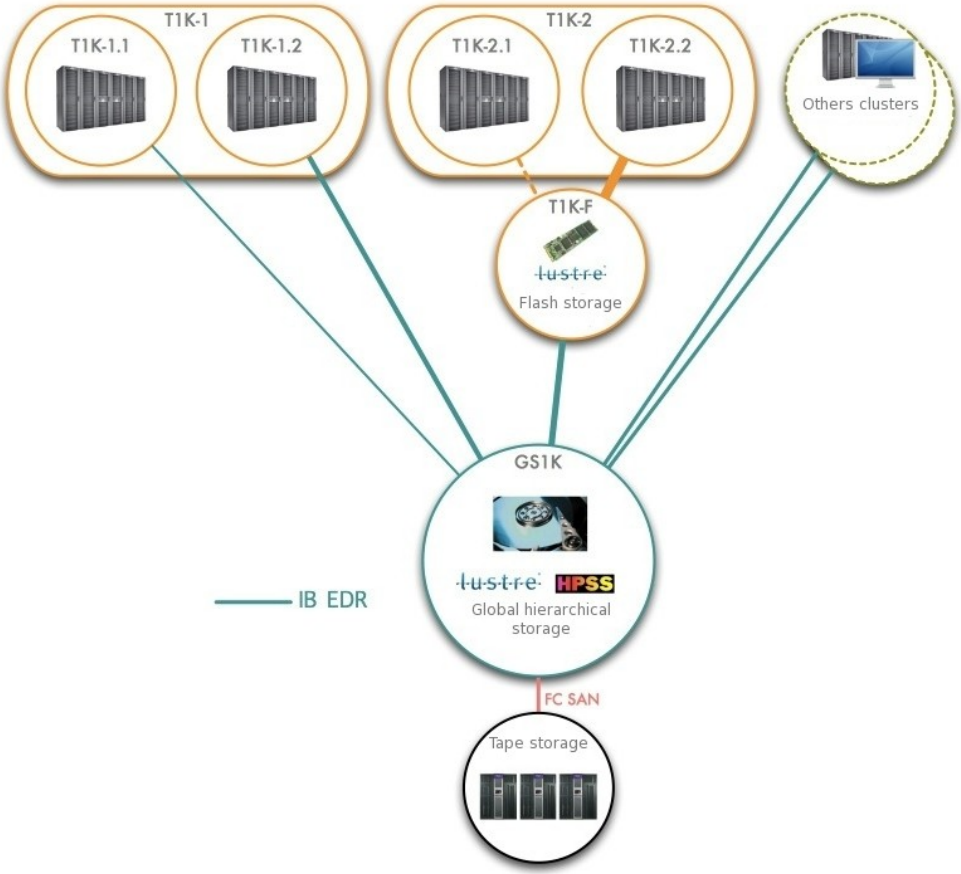
Matériel : bandes et robotiques

- **Robotiques : STK SL8500**
  - TERA x3, TGCC x3
- TERA :
  - Production sur **T10K-D**
  - Réforme des LTO5 (repacks finis)
    - 14 bandes difficiles à relire / 25.000
- TGCC :
  - Production sur **LTO6**
    et **LTO5 (en RAIT 2+1)**
  - Passage à **LTO8** cette année (mi-2018)



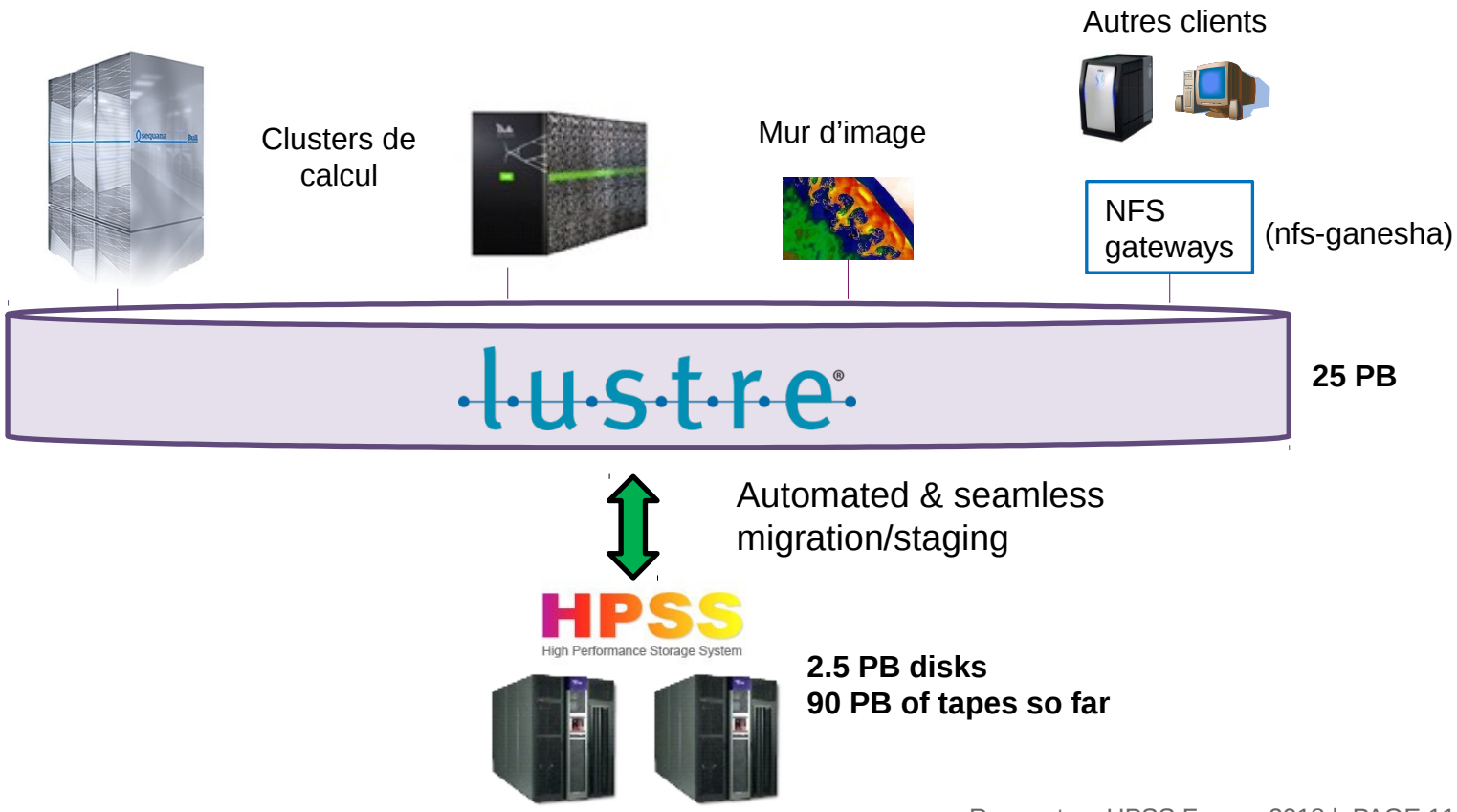
Integration au centre de calcul

Architecture du centre de calcul TERA1000



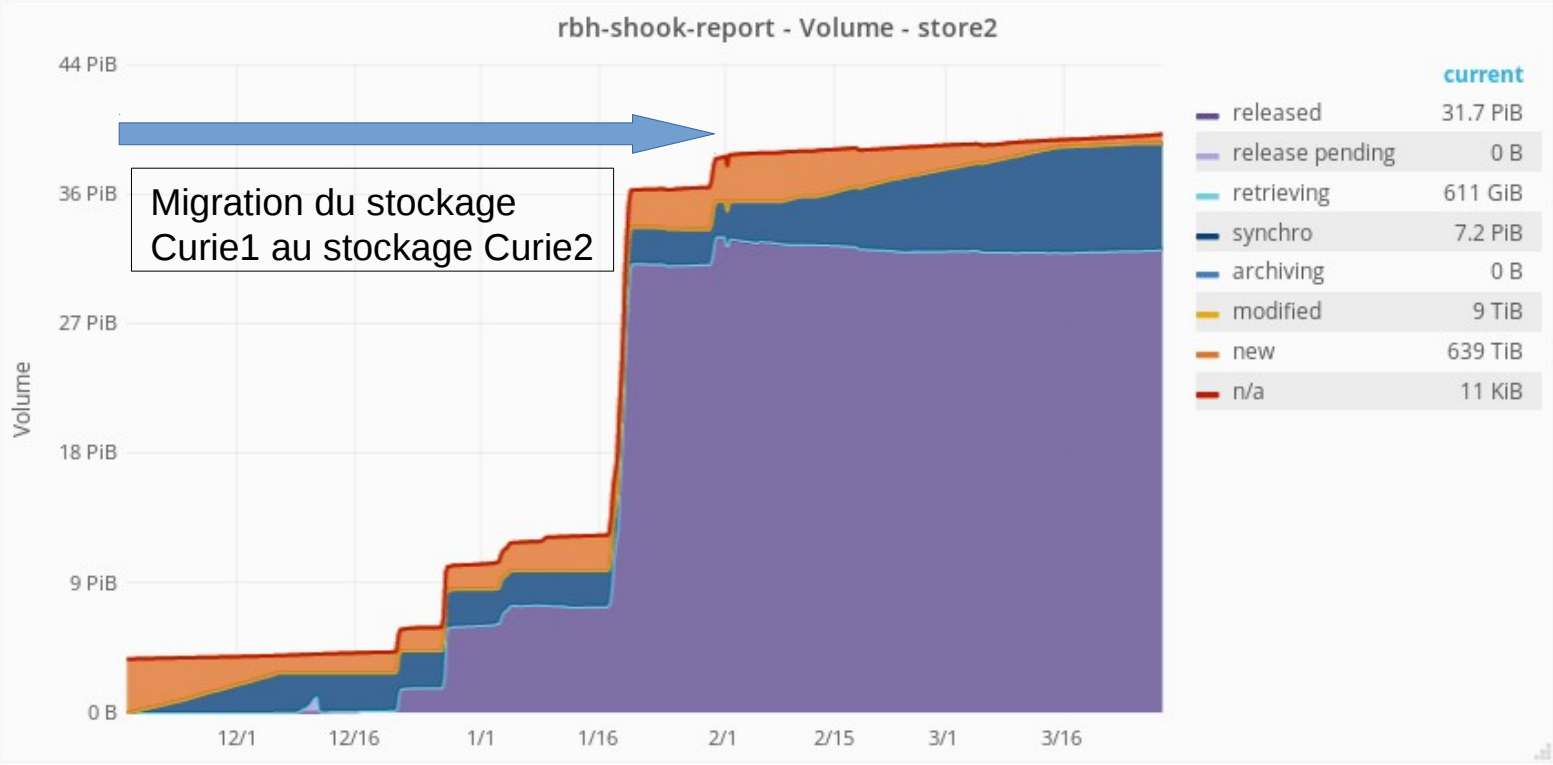
Interface utilisateur

- Seule interface utilisateur : système de fichiers Lustre global (accès POSIX)
- Pas d'interface directe entre utilisateurs et HPSS
- HPSS est un backend Lustre/HSM



Localisation des données

Statut HSM des fichiers au TGCC - Mars 2018
(Système de stockage Curie 2)



Restrospective 2017 (1/3)

Mise en production du RAIT

- En configuration 2+1
- Si bande endommagée => repack du volume recontruit la bande manquante
- Les performances sont au niveau attendu (débit x2)
- Rait Engines répartis sur les movers tape
- Option “check-on-read” désactivée
  - Nécessite moins de dérouleur pour les lectures
  - Mais... pour un triplet de bandes [A,B,C], parfois monte A+B, parfois A+C, ou B+C...

Fusion du système France Génomique avec le HPSS TGCC

- Beaucoup de données encore dans le niveau Lustre
- Il a suffit de pousser ces données dans le HPSS TGCC

Restrospective 2017 (2/3)

Passage à HPSS 7.4.3p3e3

- Beaucoup de bandes passées en EOM dès les premiers fichiers écrits
- Dû à une taille de block d'accès bande trop grande pour le driver (e.g. 256K ou 512K)
- Spécifique au driver « lpfc »
- Tuning nécessaire du module **lpfc** : lpfc_sg_seg_cnt=256
- **Reconstruction de l'initrd nécessaire pour prise en compte**
- « dracut -f » a résolu le problème

Plan de disaster recovery et mise en oeuvre

- Mise au point d'un plan de reprise en cas de perte complète du core server
- Test complet effectué : récupération des sauvegardes, redéploiement d'un core server, restauration de la base DB2, démarrage du core server
- REX :
  - Restauration des sauvegardes : ouvertures de ports nécessaires
  - Difficultés à changer le nom du core server (pour éviter d'impacter le système en production !)
  - Plus long à préparer qu'à faire : cela vaut le coup de s'y préparer !

Restrospective 2017 (3/3)

Nouveau stockage TGCC (stockage Curie 2)

- Nouveau cluster de stockage (renouvellement complet du matériel)
- Métadonnées DB2 :
  - Sur 2 baies Netapp E5600 avec SSD TOSHIBA 400Go (idem TERA)
- Disques :
  - Utilisation de movers embedded dans DDN (idem TERA)
- Bandes :
  - Initialement prévu : T10K-E... Abondonné par Oracle
  - Finalement : passage à LTO8 dans les prochains mois

Evolutions à venir (2018)

Test de HPSS 7.5.1

- Dans les prochains mois
- Tests fonctionnels
- Validation de l'opération de migration 7.4.3->7.5.1 (avec Qrep)

Mise en production HPSS 7.5.1

- Dans le courant de l'année (2e semestre)
- Nous permettra de nous débarrasser de notre dernier serveur en EL6 : le core server HPSS
- Tirer parti des nouvelles fonctionnalités
  (notamment groupement des staging par bande)

Nouveaux projets stockage

ICEI / Fenix : https://fenix-ri.eu/

- Sous-projet du projet européen “Human Brain Project”
- Vise à fournir une e-infrastruture de recherche fédérée entre **BSC**, **CEA**, **CINECA**, **CSCS**, **JSC**
- To be hosted at CEA:
  - OpenStack cluster with interactive computing nodes
  - Flash storage for active data
  - Archive data (likely an object store) accessed through OpenStack swift
  - R&D: Swift over Lustre

=> Appel d'offre en septembre 2018

Stockage pour EXA-1

- Veille techno et R&D sur les technologies flash et leur intégration
- Stockage objet
- I/O proxies

DE LA RECHERCHE À L'INDUSTRIE

cea

Questions ?

Commissariat à l'énergie atomique et aux énergies alternatives
Centre DAM Ile-de-France | 91297 Arpajon Cedex, France
T. +33 (0)1 69 26 40 00

CEA
DAM
DIF

Etablissement public à caractère industriel et commercial | RCS Paris B 775 685 019